Tags:

Use comma for seperation between tags

<ID>	REF 0.0 - title of the comment
<TYPE>	$$$BMR $$$Power $$$Presentation $$$Annotation $$$Network $$$Hierarchy $$$CellLine $$$Stemness $$$Validation $$$NoveltyPos $$$NoveltyNeg $$$Minor $$$Validation $$$Other
<ASSIGN>	@@@XYZ
<PLAN>	&&&AgreeFix - agree and fix &&&DisagreeFix - disagree but we fix, obsequious, and we're safe &&&OOS - out of scope &&&Defer - help me &&&MORE : Go above and beyond the scope of the question and indicates more analyses to be done
<STATUS>	%%%TBC: To Be Continued %%%50DONE: response done (MS+figure to be updated) %%%75DONE: response+calc+figure done (MS to be updated) %%%100DONE: all done. MS+figure+response done %%%CalcDONE: calculation done

Formatted Table

PLEASE NOTE $$$ @@@ &&& %%% are reserved as shown above.
PLEASE USE ### only for all other tags.

Usage example:

<ID>REF 0.0 - Overall comments on the paper
<TYPE>$$$BMR
<ASSIGN>@@@MG,@@@JZ,@@@DL,@@@JL,@@@WM,@@@PDM,@@@Peng,@@@TG,@@@XK,@@@STL,@@@MTG
<PLAN>&&&AgreeFix

<STATUS>%%%TBC

Format:

Referee Comment: Courier New, 10pt
Author Response: Helvetica Neue, 12pt
Excerpt From Revised Manuscript: Times New Roman, 10pt

Referee expertise:

Referee #1: cancer genetics, mutational processes
Referee #2: statistical genetics
Referee #3: human genetics
Referee #4: gene expression
Referee #5: cancer genomics

Cover Letter

Deleted:

Dear Orli,

We are enclosing our revised version of the ENCODEC manuscript. As you can see, we have attempted to completely and definitively address all of the referees' concerns. In the attached sheets which have a point by point response.

Deleted: referee's

We corresponded a bit about this manuscript before so I will be brief here and simply say that we consider this paper to be an integral part of the ENCODE package and the main analysis group to do large-scale integration across various types of assays and the only group that provides a network perspective on the annotations. We think cancer is a great application for this. But this, as we have mentioned before this is not a cancer genomics paper.

Deleted: as

In the revision version, we have summarized our efforts to highlight the application and integration of ENCODE data on cancer, which includes

- Effect of various genomic features on structures variations in strictly matched cell types
- Another CRISPR validation of the SVs effects on extended gene annotations
- A targeted validation on the effect of key regulators to well-known oncogenes expressions
- Analysis of numerous cancer-associated TF effects on overall gene expression patterns
- Normal-Tumor-Stem comparisons from both transcription and regulatory network aspects

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

We realize that this response is quite long. To make it easier for you and the referees we have made each response to each referee completely self-contained (at the risk of repeating some text between referees. Thus each referee just needs to go sequentially through his or her comments. We hope you like the manuscript and we look forward to hearing from you.

Yours sincerely,
marK

Editor:

<ID>REF 0.1 - Overall comments on the paper

<TYPE>$$$Presentation
<ASSIGN>@@@MG
<PLAN>
<STATUS>%%%65

Referee Comment	The referees have raised a range of technical concerns on the analyses, including for the background mutation rate, the need to include statistical significance to support many of the claims, and the limitations of this data including cell lines used.
Author Response	We have tried to revise our manuscript to completely and definitively address all of the referee's comments. We felt many of them are good suggestions, so we expanded upon them extensively while keeping the focus of our manuscript. In particular, we have expanded the manuscript to address suggestions related to - Highlight the overall value of this resource to cancer genomics - Extend analysis of genes' effects on somatic and germline SNVs or SVs - Normal-tumor-stem comparisons from network and expression profiles - Discuss SUB1 as an example to highlight the cancer network biology - SVs' effects on networks and extended genes - CRISPR-based validations on SV effects *Regarding the misunderstanding on the BMR section* One misunderstanding we wish to clarify is that the main goal of the BMR section is to demonstrate how the richness of ENCODE data can improve BMR estimation, and not so much to discover novel drivers genes. Hence, we feel that detailed cancer driver comparisons are outside the scope of our manuscript. Another point we want to emphasize is the necessity of including many features due to the heterogeneous nature of tumor data, which **was also accurately pointed out by referee 4**. Usually, there are numerous non-cancerous cells, such as immune, fibroblasts, and blood cells, within and around the tumor cells, which may play important roles in cancer \cite{xxx}. We have shown that ENCODE dramatically increases the available genomic data by more than a factor of 10 compared to the current methods (2069 vs 169). We want to further point out that the majority of such data

Deleted: . ... [1]

Deleted: TBC

Formatted Table

	are actually from real tissues (1339 out of 2069). We have shown that the inclusion of more data noticeably improves BMR estimation.

<ID>REF0.2 – Regarding context with prior studies

<TYPE>$$$Presentation
<ASSIGN>@@@MG,@@@JZ
<PLAN>
<STATUS>

Referee Comment	The referees also find that the current manuscript provides limited context with prior studies using similar approaches for use of prior ENCODE and Epigenome Roadmap datasets in cancer genomics. They detail the need for clearer presentation in context of prior studies as well comparisons to demonstrate advance.
Author Response	We thank the referees for this comment, and we have tried to provide better context with prior work in our revised manuscript. We note that we have cited many of these works in our initial submission. Some papers came out well before we submitted our paper in Aug 2017. Martincorena et al 2017, was published in Nov 2017 (this was work from the lab of Peter Campbell, and we excluded him due to a conflict of interest in our initial submission). We want to further point that the main focus of this work from Dr. Peter Campbell's lab was not at all on BMR estimation, but rather selection patterns in coding regions in cancer (abstract below). BMR estimation and noncoding regions are not even mentioned in the abstract or the main manuscript associated with that work. As suggested, we now cite this paper in our revised manuscript, and we make it clear how our paper is different from this one. However, we feel that it may not be entirely reasonable to carry out detailed comparisons with that work. In fact, after our submission, several new studies were released that linked the noncoding genomes to cancer, such as Zhang et al 2018. We strongly believe that our ENCODEC resource would benefit such analyses, so we have updated our reference list in this revised version.

Formatted Table

	"Universal Patterns of Selection in Cancer and Somatic Tissues: Cancer develops as a result of somatic mutation and clonal selection, but quantitative measures of selection in cancer evolution are lacking. We adapted methods from molecular evolution and applied them to 7,664 tumors across 29 cancer types. Unlike species evolution, positive selection outweighs negative selection during cancer development. On average, <1 coding base substitution/tumor is lost through negative selection, with purifying selection almost absent outside homozygous loss of essential genes. This allows exome-wide enumeration of all driver coding mutations, including outside known cancer genes. On average, tumors carry 4 coding substitutions under positive selection, ranging from <1/tumor in thyroid and testicular cancers to >10/tumor in endometrial and colorectal cancers. Half of driver substitutions occur in yet-to-be-discovered cancer genes. With increasing mutation burden, numbers of driver mutations increase, but not linearly. We systematically catalog cancer genes and show that genes vary extensively in what proportion of mutations are drivers versus passengers.

<ID>REF0.3 – Regarding the advance to the ENCODE paper

<TYPE>$$$Presentation
<ASSIGN>@@@MG,@@@JZ
<PLAN>&&&DisagreeFix
<STATUS>

Referee Comment	The referees also recommended that the current manuscript does not represent a distinct advance to the main ENCODE manuscript, as it does not report separate new datasets, methods, or clear novel findings. Some referees also recommended that this may be more suitable as Perspective in a specialized journal that further highlights the use on the current ENCODE datasets for cancer genomic studies.
Author Response	We thank the referees for pointing out potential sources of confusion about whether this is a novel biology paper or a resource paper, as well as for raising their questions regarding the relationship between our paper and the whole ENCODE package. In our revised version, we have tried to make these points more explicit. ***Regarding the objectives of our paper and how to relate it to the whole package:*** • this paper should be considered as a "*resource*" paper, not a novel biology paper • this work is the main integrative paper that provides deep annotation for several cell types, while the main encyclopedia paper

Formatted Table

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: be

is focused on broad and universal annotations (for all cell types) based on 4 assays

- this is the only paper in ENCODE that provides comprehensive networks from ENCODE3 and this is the only paper that incorporate novel data types from the ENCODE functional charaterization center

Regarding data in this paper

- our paper is the only one that incorporates multiple novel assays in ENCODE3, such as STARR-Seq, Hi-C, TF knockouts
- it is the only one with unique validations that have been carried out with various techniques, such as luciferase assays, CRISPR engineering, and knockout experiments
- ENCODE 3 "data" are not explicitly tied to any paper. Unlike previous rollouts, ENCODE 3 does not associate particular data sets with specific papers (as codified in an agreement with NHGRI.)

Regarding the new methods in this paper

As summarized below, we have many under-appreciated methods for integrating multiple assays for deep annotations. We have tried to make these more clear in our revised version:

- Multiple methods regarding enhancer predictions
    - CRISPER: Pattern recognition-based enhancer prediction that integrate more than 10 histone modification marks
    - ESCAPE: Enhancer predictors based on STARR-Seq methods
    - CARE: Compact and AccuRate Enhancer prediction by integrating STARR-Seq and genomic features
- A method for enhancer-gene linkage predictions: JEME+Hi-C
- A gene community-based method to analyze network rewiring
- A integrative new method to prioritize regulators based on burdening, rewiring and expression regulations
- A new pipeline for variant prioritization

Deleted: [JZ2MG: do you say >=20 assays?]

Deleted: (JZ2MG: can we say we are

Deleted: representing

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Deleted: centerl? Or some PI from there? Is this confidential to Orli, can the reviewers see it?)

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Formatted: Outline numbered + Level: 2 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 1"

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: (JZ2MG: dangerous here. Think about it more carefully)

Referee #1 (Remarks to the Author):

<ID>REF1.0 – Preamble

<TYPE>$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%65DONE

Overall the reviewer mentioned that this is an interesting resource, but was of the opinion that the novelty of our paper is lacking. We first want to thank the referee for his/her acknowledgement of the potential popularity of our resource for cancer genomics. In our revised version, we have tried to address the reviewer's comments by better clarifying the value of the resources in this paper through improved communication of our main results and validations. Specifically, we would like to emphasize two points.

1. The novel results and resources in this paperin the context of the ENCODE package

We have tried to make it more clear that the objectives of our work include providing deep and accurate annotations focusing on several data-rich cell types. The breadth and accuracy of our annotations are not possible in the main encyclopedia paper (because of limited data), which aims to provide universal annotations for all cell types based on just 4 assays.

We also try to emphasize that the new ENCODE3 release (used in this paper) can greatly benefit cancer research because this new release is vastly more expansive than those in previous works. This ENCODE3 release includes

- 2,017 histone ChIP-seq data (1,339 from tissues/primary cells; compare to 169 in Marticorena et. al. 2017)
- 51 replication timing Repli-chip and Repli-seq data (compared to 16 in Polak et. al. 2015)
- 1,863 TF ChIP-Seq from 143 cell types (compare to 958 in ENCODE2)
- 103 tumor-normal matched TF ChIP-seq data (common TF antibodies between K562 and GM12878 shown; compare to 42 in ENCODE2)
- CRISPR and RNAi-based 661 TF/RBP knockdown data (compare to none in ENCODE2)
- Numerous novel assays, including whole genome STARR-seq, Hi-C, ChIA-PET, and eCLIP

Deleted: noted

Deleted: our main goal and clearly organizing our analysis to illustrate

Deleted: [JZ2DL: please fill
Formatted: Font:Bold, Not Highlight
Deleted: xxx, only focus
Deleted: this paper and its distinct role in
Formatted: Font:Bold, Not Highlight
Deleted: data we used] . ... [2]
Deleted: whole
Deleted: 2017
Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"
Deleted: Seq
Deleted: 1339
Deleted: in contrast
Deleted:
Deleted: 52
Deleted: sets from xx tissues (
Deleted: with
Deleted: Xxx
Deleted: xxx
Deleted: vs. xx
Deleted: Xxx
Deleted: Seq for xxx cancer types (vs. xxx for only
Deleted: Xxx
Deleted: knockdows
Deleted: xxx in xxx cell types (vs. xx
Deleted: A number of
Deleted: such
Deleted: Seq

We have tried to make it more clear that we have developed many new methods in this paper to deeply annotate several cancer-associated cell types from multiple aspects, including

- Multiple-level compact and accurate enhancer predictions
- Integrative gene-enhancer linkages
- Extended gene definitions that incorporate numerous types of regulatory elements in a gene-centric way
- Universal and tissue-specific regulatory networks built using ChIP-seq and eCLIP data for 1,863 TFs and 112 RBPs
- Matched TF regulatory profiles and their rewiring status
- Normal-tumor-stem distance quantifications based on expression and network profiles

We have also tried to illustrate the utility and value of this resource to prioritize both key regulators and genomic variations (SNVs and SVs). We further validated our results using various techniques, such as luciferase assays, CRISPR, and knockdowns. Collectively, we believe that all of these illustrate the value of our resource to cancer genomics.

2. Regarding the BMR section

With respect to the BMR estimation part in particular, the reviewer noted that there have been several prior publications focusing on applications such as cancer driver detection. We thank the referee for pointing out this body of related work.

Recent interest by the cancer genomics community suggests that there is value in identifying methods to improve BMR estimation. As suggested, we have tried to provide better context for previous work in our revised manuscript. These references are summarized in Table R1.

Second, we would also like to emphasize that the main goal of our paper is not to present novel methods of driver discovery, but rather to illustrate that the richness of the ENCODE data can be leveraged to noticeably improve the accuracy of BMR estimation. Hence, we feel it is slightly outside the scope for our ENCODE resource paper to make detailed comparisons with driver gene discovery. In the revised version, we have clearly highlighted the value of ENCODE data in our updated Fig. 1.

Third, we want to point out that the BMR application is just **one out of many** potential ENCODE data applications. We have also provided results and validations of our resource related regulator/SNV/SV prioritization, network rewiring, and stemness measurement that are of value in cancer genomics (and other disease contexts).

Table R1. status of the related references

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: Seq

Deleted: xxx

Deleted: xxx

Deleted:)

Deleted: assay, CRISP

Deleted: the

Deleted: had

Deleted: many existing

Deleted: We thank

Deleted: referee for pointing out a body of related work.

Deleted: of

Deleted: (see Table R1 below). We would also like to point out that some

Deleted: were either published after our initial submission (such as Marticorena et al. 2017) or with a different focus (i.e., other than BMR estimation; see Table R1).

Deleted: We

Deleted: 2

Deleted: Even for Figure 2, we

Deleted: include SV

Deleted: GWAS germline

Deleted: analyses. There are many other ENCODE applications, such as regulatory activity,

Deleted: , which are also key to interpreting and prioritizing variants effects

Deleted: .

Formatted: Highlight

Reference	Initial	Revised	Main point	Comments
Lawrence et al, 2013	Cited	Cited	Introduce replication timing and gene expression as covariates for BMR correction	Replication timing in one cell type
Weinhold et al, 2014	Cited	Cited	One of the first WGS driver detection over large scale cohorts.	Local and global binomial model
Araya et al, 2015	No	Cited	Sub-gene resolution burden analysis on regulatory elements	Fixed annotation on all cancer types
Polak et al (2015)	Cited	cited	Use epigenetic features to predict cell of origin from mutation patterns	Use SVM for cell of origin prediction, not specifically for BMR
Martincorena et al (2017)	No (out after our submission)	Cited	Use 169 epigenetic features to predict gene level BMR	No replication timing data is used
Imielinski (2017)	No	Yes	Use ENCODE A549 Histone and DHS signal for BMR correction	Limited data type used from ENCODE
Tomokova et al. (2017)	No	Yes	8 features (5 from ENCODE) for BMR prediction and mutation/indel hotspot discovery	Expand covariate options from ENCODE data
huster-Böckler and Lehner (2012)	Yes	Yes	Relationship of genomic features with somatic and germline mutation profiles	NOT specifically for BMR
Frigola et al. (2017)	No	Yes	Reduced mutation rate in exons due to differential mismatch repair	NOT specifically for BMR
Sabarinathan et al. (2016)	No	Yes	Nucleotide excision repair is impaired by binding of transcription factors to DNA	NOT specifically for BMR
Morganella et al. (2016)	No	Yes	Different mutation exhibit distinct relationships with genomic features	NOT specifically for BMR
Supek and Lehner (2015)	No	Yes	Differential DNA mismatch repair underlies mutation rate variation across the human genome.	NOT specifically for BMR

<ID>REF1.1 – Positive comments on the resource releases

<TYPE>$$$NoveltyPos
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	This manuscript describes how the ENCODE project data could be utilized to derive insights for cancer genome analysis. It has

Comment [1]: PDM has commented on this table before:

Requires modification (or transition to short narrative format).

-- Reference formatting is non-uniform.
-- 'Yes' and 'cited' are interchangeable and should be unified.
-- It's unclear what content is in the 'Comments' section.
- Most importantly, there is significant risk of offending reviewers associated with these papers with dismissive or inaccurate comments or summaries.

Reference	Initial	Revised
Lawrence et al, 2013	Cited	Cited
Weinhold et al, 2014	Cited	Cited
Araya et al, 2015	No	Cited
Polak et al (2015)	Cited	cited
Martincorena et al (2017)	No (out after our submission)	Cited
Imielinski (2017)	No	Yes
Tomokova et al. (2017)	No	Yes
huster-Böckler and Lehner (2012)	Yes	Yes
Frigola et al. (2017)	No	Yes
Sabarinathan et al. (2016)	No	Yes
Morganella et al. (2016)	No	Yes
Supek and Lehner (2015)	No	Yes

Deleted:
Formatted Table

	several examples to illustrate this point, e.g., how to better estimate background mutation rate in a cancer genome, how to modify gene annotation for finding mutation-enriched regions (e.g., by bundling enhancer regions to target genes using Hi-C/ChIA-PET), and describing the changes in regulatory networks in cancer. Obviously, the ENCODE project involves a great deal of planning and a lot of experimental work by many groups, and the overall aim of re-highlighting the ENCODE as a resource to cancer research seems worthwhile in general, perhaps even in a high-profile journal.
Author Response	We thank the referee for this positive feedback.

<ID>REF1.2 – BMR: comparison with existing literature

<TYPE>$$$BMR,$$$Text
<ASSIGN>@@@JZ,@@@WM,@@@PDM
<PLAN>&&&OOS
<STATUS>%%%95DONE

Referee Comment	Just to take the first application as an example, the problem of estimating background somatic mutation rate accurately in order to better identify cancer drivers has been studied extensively in the literature. One paper, "Mutational heterogeneity in cancer and the search for new cancer-associated genes" (Nature 2013), is cited in the current manuscript, but there are many others. For instance, Weinhold et al, 2014 (Genome-wide analysis of noncoding regulatory mutations in cancer, Nat Genetics), Araya et al, 2015 (Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations, Nat Genetics), and similar non-coding mutation identification papers all include steps to account for epigenetic features in their background rate calculation.
Author Response	We thank the referee for pointing out these works. Modelling background mutation rate has been an important topic of inquiry, as even modest improvements can be of great benefit to the ascertainment of driver mutations in cancer. As suggested, we have cited all the references mentioned above, and we have tried to provide better context of previous work in the revised manuscript.

Formatted Table

	In our revised manuscript, we have explicitly clarified how the new ENCODE data can be useful for BMR estimation. Our contribution is to provide data in a ready-to-use format that is considerably more expansive than those in previous works (2069 features vs. 169 in Matincorina et al 2017). We have shown that this scale of data can benefit previous models to better characterize BMR.
Excerpt 1.2-A (in main text)	Wait for main text

Deleted: We note that, in fact, we did notice previous efforts for driver detection, and we have cited parts of these references (such as Weinhold et al, 2014). In the revised version, we have tried to make it more clear that we are not claiming to have developed a new model for BMR estimation for driver detection, or presenting a new discovery that "matched" features are better correlated with BMR. Instead, we

Deleted:

Deleted: -- our work includes data on

Formatted: Font:Helvetica Neue

Deleted: histone modification and 52 replication time.

Deleted: larger

Deleted: many models described in

Deleted: works

Deleted: From ... [3]

<ID>REF1.3 – BMR: Match

<TYPE>$$$BMR,$$$Text
<ASSIGN>@@@JZ,@@@WM
<PLAN>&&&DisagreeFix
<STATUS>%%%50DONE

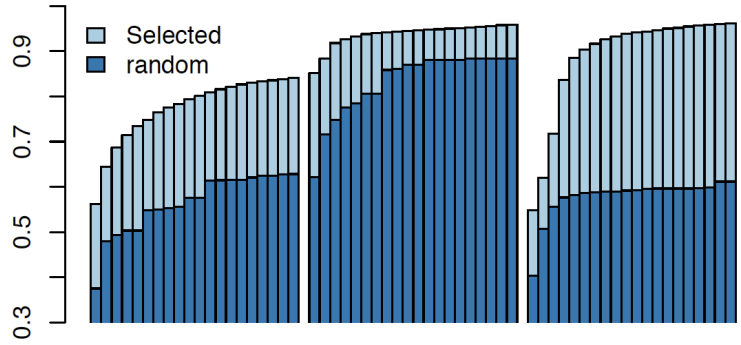
Referee Comment	Most large-scale cancer genome sequencing papers also have models at various levels sophistication, most of them including the issue of proper tissue-type matching. "matched" cell lines are better than unmatched or addition of more epigenetic features results in some improvement is almost trivial at this point. Which marks contribute to this is also not new.
Author Response	We thank the referee for this comment, and we have tried to better clarify our main goal in our revised manuscript. We made it very clear that we are not claiming to have proposed the use of negative binomial regression with epigenetic features on BMR estimation. Instead, our key points are that: • The ENCODE3 rollout dramatically expands the number of high quality genomic data available for this type of regression by more than an order of magnitude (**2069** compared to 169 in Matincorina et al 2017), many of which are from real tissue samples or primary cells.

Formatted Table

Deleted: point is

Deleted: the

Formatted: Don't add space between paragraphs of the same style, Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

	There is significant technical challenge in processing this scale of data to create a ready-to-use resource that may be applied to BMR estimation. • This expanded data provides a significantly larger pool to find the best match for a given cancer type • More data are useful due to tumor heterogeneity. While it is valuable to match cancer to its cell of origin, tumors are highly heterogeneous (as clearly pointed out by referee 4 also), so a combination of different data sets provide the best overall fit to mutation rates. We have shown this in the updated version of Figure 2 (see Excerpt 1.3-A and 1.3-B). [[WUM+PDM2all: WUM suggested that a clear demonstration of the value of increasing numbers of features, would be to compare the accuracy of using only the 169 features used by Inigo et al. to the accuracy we achieve with all 2069 features. PDM agrees that this would be useful and clear.]]
Excerpt 1.3-A (main text and figure)	The 2,017 uniformly processed histone modification and 51 replication timing data may serve as a resource to significantly improve BMR estimation accuracy. We also showed that BMR estimation can be improved dramatically by selecting appropriate combination of multiple features from ENCODE. 
Excerpt 1.3-B (cross validation in	To avoid overfitting problem, we performed 5 fold cross validation using the selected model for each cancer type and listed the performance as below.

Deleted:

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: is useful from two aspects: ... [4]

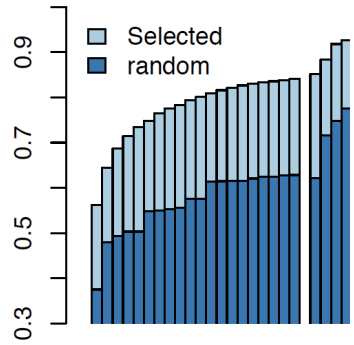
Deleted: is

Deleted: excerpt below).

Deleted: 2017…,017 uniformly processed histone modification and 52 ... [6]

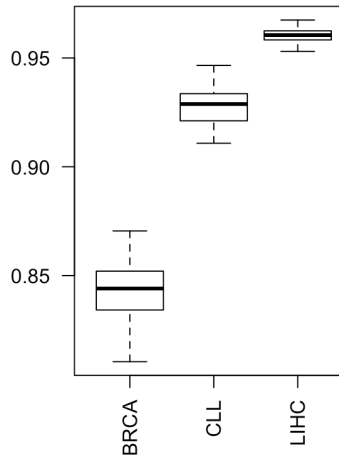
Deleted: From ... [5]

Formatted: Justified



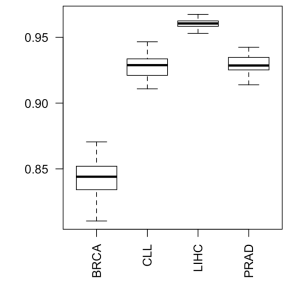
Deleted:

Moved down [1]: To avoid overfitting problem, we performed 5 fold cross validation using the selected model for each cancer type and listed the performance as below.



Deleted:

Moved (insertion) [1]

supplement)	

<ID>REF1.4 – BMR: cell of origin features vs. many features

<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ,@@@JL
<PLAN>&&&DisagreeFix,&&&More
<STATUS>%%%70DONE

Referee Comment	Importantly, Polak et al, 2015 (Cell-of-origin chromatin organization shapes the mutational landscape of cancer, Nature) in fact show that cell-of-origin chromatin features are much stronger determinants of cancer mutations profiles than chromatin feature of matched cancer cell lines, and that cell type origin can be predicted from the mutational profile.
Author Response	We thank the referee for raising this point about features from cells-of-origin, and we have expanded upon the relevant discussion in our revised manuscript. In summary, we have made the following changes. 1. We have added more to the discussion section that accurate cell-of-origin definitions are challenging. Distinct subtypes of tumor cells may derive from different 'cells of origin' \cite{21248838}. (see Excerpt 1.4-A) 2. **In contrast to the results of Polak et al., we suggest that linear combinations of cancer cell lines may provide a basis for a more accurate determination of cancer mutation profiles than either cell-of-origin, or a single matched cancer cell line. [[Consistent with the stemness discussion etc., would need to flesh out argument or provide suggestive evidence.]]**

Formatted Table

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

Deleted: discussions

Deleted: within an organ

Deleted: excerpt

Deleted: our goal is to better predict BMR, instead of finding the cell-of-origin. A good combination of multiple features can provide better fits overall (details in Excerpt 1.3 above).

PEAK FITS B/C HETRO

Excerpt 1.4-A (added to disc. sect)	Recently work has pointed out the effect from cell-of-origin on tumor from multiple aspects, such as mutational process and tumor classifications. However, to accurately define tumor cell-of-origin is sometimes challenging. For example, even different subtypes of tumor from the same organ may originate from different cell types. The richness of ENCODE data provides us a larger pool to find the best representative cell of origin.

Deleted: Newly added to the discussion section: ... [7]

Deleted: From

Deleted: Revised manuscript

<ID>REF1.5 – BMR: Tissues vs. Cell lines

<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ,@@@JL
<PLAN>&&&DisagreeFix,&&&More
<STATUS>%%%70DONE

Referee Comment	Stepping back, it is not obvious to me that using the ENCODE cell lines, despite the availability of more epigenetic data, is the best approach to calculating the background rate in the first place—they briefly mention that using cell lines (rather than tissues) can be problematic, but do not explore this further. If this were a regular research paper, the authors would have to shown how the proposed approach is different and how it is better than methods already available.
Author Response	We thank the referee for raising this question about cell line data usage in our paper, and we feel this is a good opportunity to clarify that ENCODE is not just about cell lines. In our revised manuscript, we have extensively discussed the use of different types of data from multiple aspects in both the main manuscript and the supplements. (not double counting roadmap) JZ2DL: pls double check the roadmap data *Regarding the cell line data in the BMR part* • Certain data types, like TF ChIP-seq, are only predominantly available in cell lines (Excerpt 1.5-C). **Although whole tissue data could theoretically provide a closer match, this data is not obtainable due to current technical limitations. Cell line data reflects the current best possible data for these data types.** We added a table to clarify that the features extracted from ENCODE data are not just from cell lines. The majority are from tissues or primary cells (Excerpt 1.5-A). •

Formatted Table

Formatted: Font:12 pt

Deleted: as if clarifying

Formatted: Font:12 pt

Deleted: is a great suggestion.

Formatted: Font:12 pt

Formatted: Font:12 pt

Formatted: Font:12 pt

Formatted: Font:12 pt, Highlight

Deleted:

Comment [2]: The number (tissue/primary cell) includes roadmap data, but they are small number compared to whole ENCODE3

Formatted: Font:12 pt

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Comment [3]: May be the most important point, and should be placed early on.

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: we used in is

Formatted: Font:12 pt

Deleted: excerpt

Formatted: Font:12 pt

Formatted: Font:12 pt

Comment [4]: 'comparable, at least in some cases' does not sound like a strong justification for the use of cell line data. Would suggest deleting.

Deleted: <#>We added figures (in the supplement) to demonstrate how cell line data can show comparable performance (excerpt 2). ... [8]

Regarding the robustness of using cell line inference on real patient data

- added a whole new external validation section to compare with our conclusions drawn from cell lines (Excerpt 1.5-E). **Cells + tissues only from cells side by side comparisons**

Subset data

Excerpt 1.5-A (in Supp.)

In total, we have used 2,017 histone ChIP-seq and 51 replication timing Repli-chip and Repli-seq features to predict BMR. We did a PCA of the signals these features and selected the best combination of 20 PCs for BMR prediction. It is worth pointing out that the majority of our data is from real tissue or primary cells. A summary of cell types of these features were given below. **[[WUM's comment: Could we show a back-of-the-envelope power analysis that shows the improved capability of identifying a rare driver variant based on marginal improvements in BMR.]]**

Table S1. Summary of ENCODE histone ChIP-seq data **[[WUM suggests and PDM agrees that this data may be more clearly presented as a pie chart]]**

Cell Type	# histone marks
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

Table S2. Summary of ENCODE3 Replication timing data

Cell Type	Repli-seq	Repli-chip
cell line	101	10
in vitro differentiated cells	0	35
primary cell	12	5
stem cell	6	11
induced pluripotent stem cell line	0	2

Table S3. Summary of 51 replication timing features from Repli-chip and Repli-chip

Comment [5]: This is interesting, but perhaps not relevant to the reviewers comments. Would suggests removing.

Moved down [2]: . ... [9]

Deleted: *global comparison of cell lines and tissu* ... [10]

Formatted: Font:Arial, 12 pt, Not Italic, No underline

Formatted: Font:12 pt

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Comment [6]: This analysis seems valid - but could perhaps use slightly more description of the relevance of the results. e.g., to prioritize variants/regulators for follow-up?

Deleted: excerpt 5).

Comment [7]: This analysis seems valid - but could perhaps use slightly more description of the relevance of the results. e.g., to prioritize variants/regulators for follow-up?

Formatted: Font:12 pt

Deleted: there are 2017

Deleted: 52 Replication

Deleted: From . ... [11]

Deleted: .

Formatted: Font:11 pt, Bold

Moved (insertion) [3]

Formatted: Font:11 pt

Deleted: .

Formatted: Font:Bold

Cell Type	# hist mark
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

Deleted: ... [12]

Comment [8]: PDM + WUM believe this table is included to suggest that this data is available through ENCODE3 and not elsewhere, and that we are the suppliers of this data. If so, this point could be made more clear.

Formatted: Font:10 pt

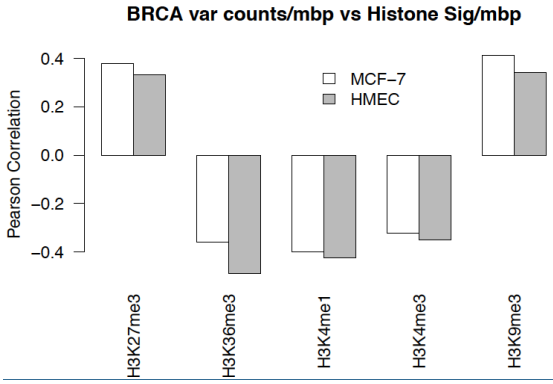
Formatted: Font:10 pt

Deleted: data]

Cell State	Repli-chip/Repli-seq
Pluripotent	8
DE	3
Liver/Pancreas	6
Neural crest/Early mesoderm	7
Late mesoderm	6
NPC	2
Myeloid/Erythroid	5
Lymphoid	5
Cancer	9

Excerpt 1.5-B (Supp. - mutation rate vs. cell line & tissue)

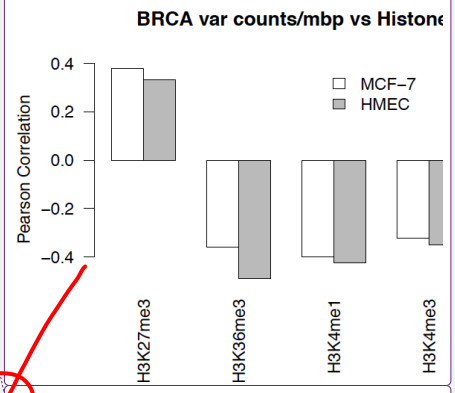
We calculated the pearson correlation of the breast cancer mutations count per Mbp vs. various histone modification features in tissue and cell line. Cell line data provides comparable (and sometimes better) correlation with mutation counts.



Excerpt 1.5-C (added to disc. sect.)

Some features, like TF binding events, have been shown to affect somatic mutation rates but the majority of such data are mainly available in cell lines. Hence, we systematically investigated the RNA-seq and TF ChIP-Seq data and found that many of the cancer transcriptome/TF binding landscape are quite similar to each other, as compared to the initial of primary cells. This has also been mentioned by previous reports, such as Lotem et al. 2005 and Hoadley et al. 2014. The fact that cancer cells lose diversity and showed a distinct pattern from the primary cells highlights the values of cell line data.

Deleted: Excerpt 2 From
Regarding the comparison of mutation rate vs features in tissue/cell lines:



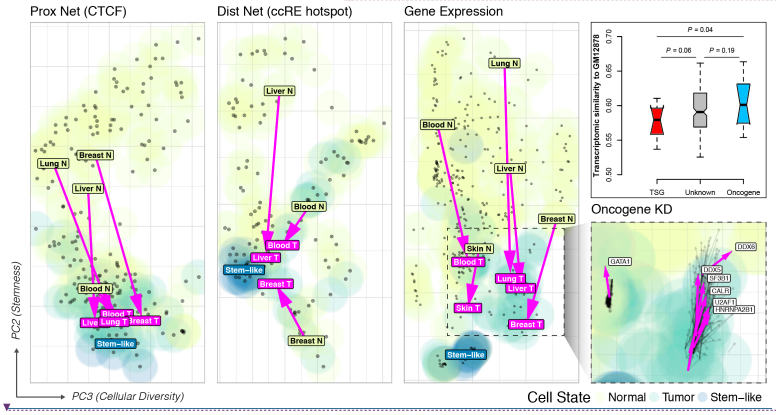
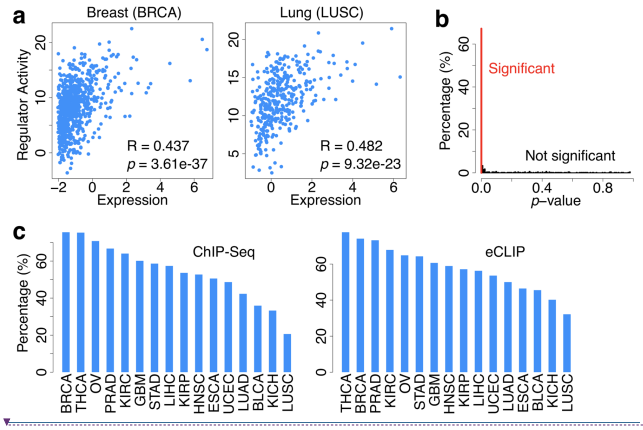
Deleted: Regarding the comparison of mutation rate vs features in tissue/cell lines: ... [13]

Comment [9]: WUM + PDM suggest that the more favorable comparison is tissue vs. tissue + cell line.

Formatted: Font:Arial, 11 pt

Deleted:

Deleted: 3 From ... [14]

Excerpt 1.5-D (rewiring in main figure)	We performed RCA/PCA analysis on RNA-Seq, shRNA RNA-Seq, and TF ChIP-seq data and found that cancer cells demonstrate a consistent pattern to be more similar to stem cells, as compared to their primary cells of origin. **Relevance?** 
Excerpt 1.5-E (validation of cell line data)	We predicted the regulatory activities of transcription factor (TF) MYC using a ChIP-Seq profile in MCF-7 cells. We found that the MYC regulatory activity is highly correlated with the MYC expression across TCGA breast tumors. For most TFs, their regulatory activities predicted using ENCODE ChIP-Seq profile in cell lines are significantly correlated with their expression levels across breast tumors. Moreover, using the same MCF-7 ChIP-Seq profile, the MYC regulatory activity predicted for lung tumors is also significantly correlated with MYC expression level in TCGA lung cancer. These results indicate that the ChIP-Seq profiles from a particular cell line can capture regulatory targets in human tumors from diverse cancer types. To select ChIP-Seq or eCLIP profiles that are representative of the regulatory targets in human cancers, we only reported the results of TFs or RBPs whose regulatory activities are significantly correlated with their gene expression level in each TCGA cohort. 

Deleted: 4 From ... [15]

Formatted: Justified

Comment [10]: This is an interesting analysis, but not sure it applies to the reviewer comments. Would suggest mentioning elsewhere in the response.

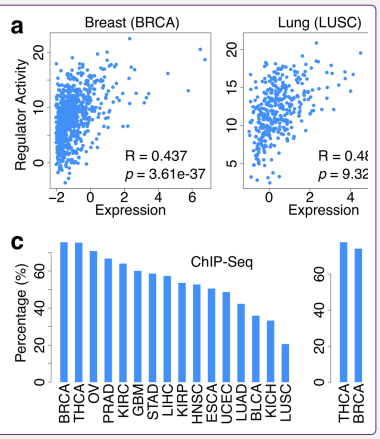
However, perhaps the analysis is impressive enough that we want to mention it even when it is not directly relevant? Not sure.

Formatted: Font:Bold

Deleted:

Deleted: Regarding the validation of cell line conclusions on real patient data: ... [17]

Deleted: From Regarding the validation of cell line conclusions on real patient data: ... [16]



Deleted:

	Supplementary Figure X. The clinical relevance of ENCODE cell line data in human primary tumors. **(a)** The correlation between *MYC* expression level and regulatory activity across tumors. The MYC regulatory activity in each tumor was predicted using the ChIP-Seq profile in MCF-7 cell line. The Pearson correlation between MYC gene expression level and regulatory activity were computed across tumors in each cancer type. The statistical significance of Pearson correlation was tested by the two-sided student t-test. BRCA: breast invasive carcinoma. LUSC: lung squamous carcinoma. **(b)** The distribution of correlation *p*-values in TCGA breast cancer. For each TF, we tested the statistical significance of Pearson correlation between TF expression levels and regulatory activities predicted across tumors through two-sides student t tests as panel a. For TCGA breast cancer cohort, most *p*-values are very significant with a few non-significant values. The fraction of regulators with statistically significant correlations in different cancer types for ChIP-Seq and eCLIP networks. In each TCGA cancer type, we computed the correlations between regulator expression levels and regulatory activities across tumors for all regulators (TFs, or RBPs). We selected regulators with statistically significant correlations through two-sided student t test (FDR < 0.05).

<ID>REF1.6 – Difference between ENCODEC and Prev. prioritization methods

<TYPE>$$$BMR,$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&DisagreeFix
<STATUS>%%%90DONE

Referee Comment	That ENCODE data helps in prioritization of non-coding variants has been well demonstrated already (including by some of the authors on this paper), and so the value of the described analysis less clear.
Author Response	**The prioritization of non-coding variants is a major frontier in genomics and cancer genomics, and these prior publications suggest the importance of this topic** We have tried to clarify that the uniqueness of our method lies in that fact that • It not only prioritizes variants, but also regulators, which is not included in the other papers. We have highlighted this in revised Fig. 3 (Excerpt 1.6-A) and performed targeted validations on key

Formatted Table
Formatted: Font:Bold
Deleted: referee pointed out that we and others have tried to prioritize
Formatted: Font:Bold
Deleted: elements before. This
Formatted: Font:Bold
Deleted: definitely true
Formatted: Font:Bold
Deleted: we have tried to make it more clear in our revision that we are not claiming to be among
Formatted: Font:Bold
Deleted: first to attempt
Formatted: Font:Bold
Deleted: .
Formatted: Font:Bold
Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

	regulators (Excerpt 1.6-B). [[WUM+PDM2all: Is this related to the **prioritization of regulators in MCF-7 (REF 1.5)?]]** • For variant prioritization, we added discussions to emphasize the integration of various novel assays in a tissue-specific manner, which was not possible in previous works (Excerpt 1.6-C). The fact that we coupled this with successful validation demonstrates the considerably greater value of the integrated ENCODE data. [PDM+WUM2all: This analysis could use more specifics on what was done, and for what reason -- Excerpt 1.6C itself is about the same length as this summary point, and provides more specifics.]]
Excerpt 1.6-A (TF regulation in main fig.)	New legend of figure 3. Figure to put here Ask Feng's group to write up here! [JZ2MG: wait]
Excerpt 1.6-B (regulator validation in supp.)	**[[PDM+WUM2all: The following text could use more explanation as to why this analysis is relevant. It currently reads like an excerpt from a methods section, and the figure has no accompanying caption.]]**To detect predicted common target gene of MYC and SUB1, shRNA plasmids containing 4 targets sites of each gene were used to transfected to HepG2 cell using LipofectamineTM 3000 following the manufacturer's instructions (Invitrogen) (target sites for each gene are listed in Sup table 1). Briefly, 0.12 M HepG2 cells were seeded in each well of one 24-well plates 24 hours before transfection. 500 ng plasmids containing either single shRNA or 4 shRNA plasmids as pool were mixed with 0.75 uL LipofectamineTM 3000 in Opti-MEM I medium (Invitrogen) and loaded to HepG2 cells in each well. Blank plasmids without shRNA target sequence was used as control. To improve transfection efficiency, 2 ug/mL puromycin was used to select successful transfected cells. 72 hours after transfection, total RNA was extracted using RNeasy Mini Kit (Qiagen) and followed by cDNA generation using SuperScript III (Invitrogen). Knockdown efficiency and target gene expression level were quantified and compared to BACTIN by qPCR using KAPA SYBR® FAST qPCR Master Mix (2X) Kit (Sigma). The qPCR primers were listed in Sup table 2.

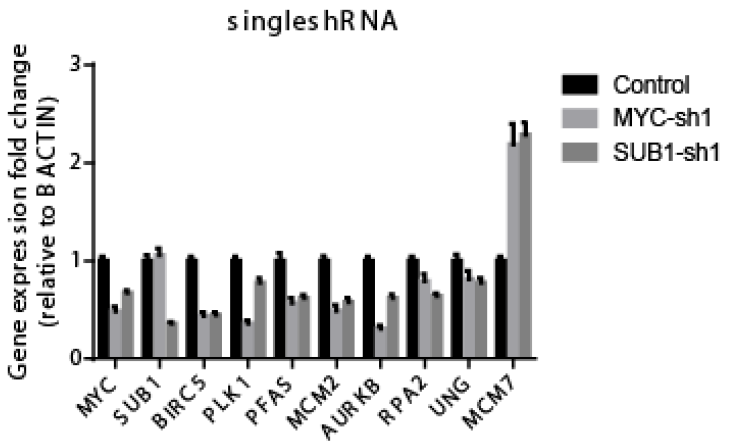
Deleted: 2).

Deleted: 3

Deleted: From ... [18]

Deleted: Feng's validation to come here

Deleted: 2 from Revised figure and supplement

	
Excerpt 1.6-C (added in disc. sect.)	In particular, our prioritization framework takes into account the STARR-seq data, the connections from Hi-C, the better background mutation rates, and the network rewiring data, which is only possible in the context of the highly integrated and their data available on certain cell lines.

Deleted: 3 From ... [19]

Referee #2 (Remarks to the Author):

<ID>REF2.0 – Preamble

<TYPE>$$$Text
<ASSIGN>@@@MG,@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

Let's focus more on the statistical genetics.

Deleted: ... [20]

Reviewer 2 raised questions about our statistical choices. These were very helpful questions, which we took as an opportunity to think carefully about our model choices and to highlightlt . We want to make reviewer 2 happy, but it's not the point of this paper. The point is about the encode data, which we desceibe below. Robust .Just to put this in perspective, ...

We greatly appreciate the referee's feedback, especially the positive comments regarding the overall value of our resource, the extended gene, and the network rewirings.

As suggested, we have tried to address the reviewer's comments, and we further extend and reorganize our analyses to illustrate the value of the resources in this paper.

Specifically, in our revised version, we have tried to provide deep and accurate annotation focusing on several data-rich cell types. We developed new methods to deeply annotate several cancer-associated cell types, which include:

- Multiple-level compact and accurate enhancer predictions
- Integrative gene-enhancer linkages
- Extended gene definitions that incorporate numerous types of regulatory elements in a gene-centric way
- Universal and tissue-specific regulatory networks built using ChIP-seq and eCLIP data for 1,863 TFs and 112 RBPs
- Matched TF regulatory profiles and their rewiring status
- Normal-tumor-stem distance quantifications based on expression and network profiles

We emphasize that this paper is unique in highlighting a number of ENCODE assays (e.g., replication timing, TF/RBP knockdowns, STARR-seq, ChIA-PET, and Hi-C), its deep, integrative annotations combining a wide variety of assays in specific cell types, and its analysis of networks. Note also that while we do NOT feel this is a cancer genomics paper, we do feel that cancer is the best application to illustrate certain key aspects of ENCODE data and analysis - particularly deep annotations and network changes.

Deleted: make it more clear that this is the main integrative paper in ENCODE3 to

Deleted: Such breadth and accuracy of our annotation is not possible in the main encyclopedia paper, which aims to provide universal annotations for all cell types based on 4 assays (due to limited data in other cell types).

Deleted: multiple

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: integrative

Deleted: extended

Deleted: universal

Deleted: network

Deleted: on

Deleted: Seq

Deleted: xxx

Deleted: xxx

Deleted: matched

Deleted: normal

<ID>REF2.1 – Comment on utility of the resource

<TYPE>$$$NoveltyPos
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	However, there is a possibility that the resource would be very popular among cancer genomics researchers. Also, results on extended genes and rewiring are of interest.
Author Response	We thank the referee for the positive comment.

Formatted Table

<ID>REF2.2 – Comparison of negative binomial to other methods

<TYPE>$$$BMR,$$$Text,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&OOS
<STATUS>%%%85DONE

Referee Comment	1) The negative binomial regression (Gamma-Poisson mixture model) was introduced in Nik-Zainal et al. Nature 2016 and Marticorena et al., Cell 2017. Why was not this available method applied, and what is the benefit for the procedure used by the authors?
Author Response	We thank the referee for pointing out the previous efforts on cancer driver detection by negative binomial regression. We certainly agree with the reviewer that negative binomial regression is a standard technique to handle overdispersion in count data. A number of earlier works (such as Imielinski et al 2016) also used negative binomial regression. In our revised manuscript, we have cited those works and tried to provide a better context of related work. We also try to make it more clear that we are not claiming to provide a novel negative binomial regression-based driver detection method, but rather to use this as a showcase for the value of ENCODE data. We did, in fact, use very similar methods to Marticorena et al. these are well established stat methods and there's lots of R packages for this.

<ID>REF2.3 – Questions about the Goodness of fit of the Gamma-Poisson Model

<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix,&&&OOS
<STATUS>%%%100DONE

Referee Comment	Also, does Gamma-Poisson model fits data for most cancers well or is it just an approximation? One can use non-conjugate priors but this is probably beyond the scope of this work.
Author Response	We thank the referee for mentioning the goodness-of-fit of the Gamma-Poisson model. As suggested, we now provide more figures in our supplement to investigate this.

Formatted Table

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

Formatted: Font:Arial

Deleted: There are three reasons to explain why we

Deleted: not directly applied available

Deleted: :
Excerpt From
Table S1. Summary of ENCODE3 histone ChIP-Seq data
Histone ChIP-seq ... [21]

Deleted: paper came out in Nov 2017, which was almost three months after our initial submission,

Deleted: it is more about positive selection in coding regions than BMR estimation.
Excerpt From
Table S1. Summary of ENCODE3 histone ChIP-Seq data
Histone ChIP-seq ... [22]

Deleted: the Marticorena et al paper is not on BMR estimation or mutational burden. For the part mentioned about BMR, BMR estimation or mutational burden are ONLY applied

Deleted: the coding regions, and no source code or software package is available for the whole genome.
Excerpt From
Table S1. Summary of ENCODE3 histone ChIP-Seq data
Histone ChIP-seq ... [23]

Moved up [3]: Table S1. Summary of ENCODE3 histone

Moved down [11]: ... [29]

Moved down [12]: ... [30]

Deleted: Excerpt From
Table S1. Summary of ENCODE3 histone ChIP-Seq data
Histone ChIP-seq ... [24]

Deleted: Excerpt From
Table S1. Summary of ENCODE3 histone ChIP-Seq ... [28]

Moved up [3]: Table S1. Summary of ENCODE3 histone

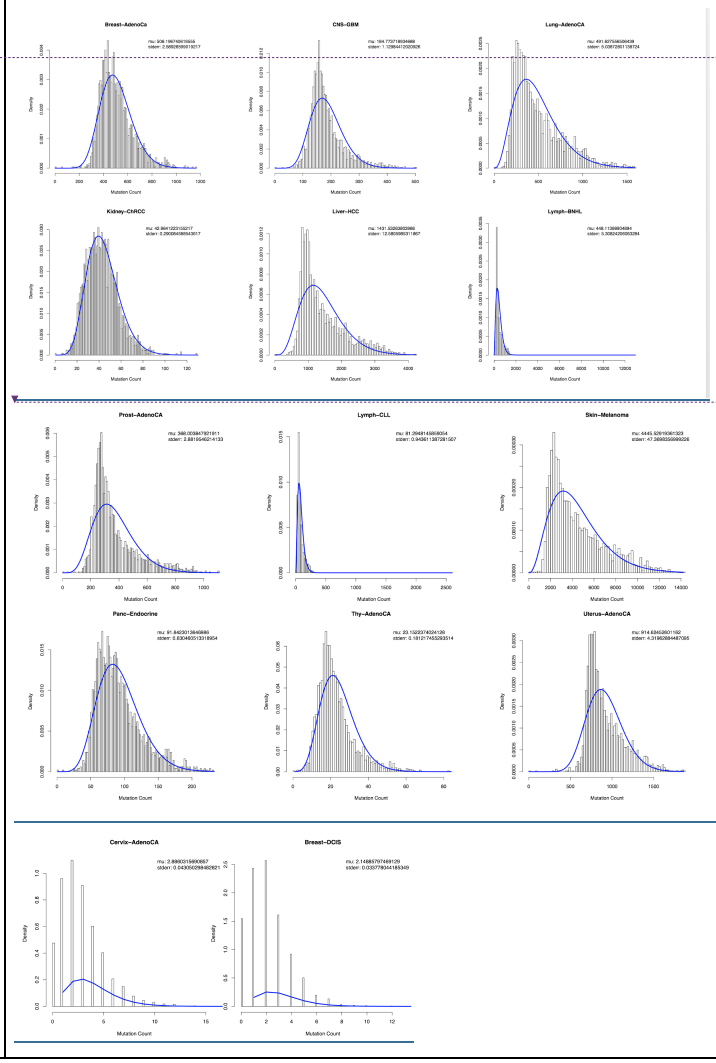
Moved down [11]: ... [25]

Moved down [12]: ... [26]

Deleted: ... [27]

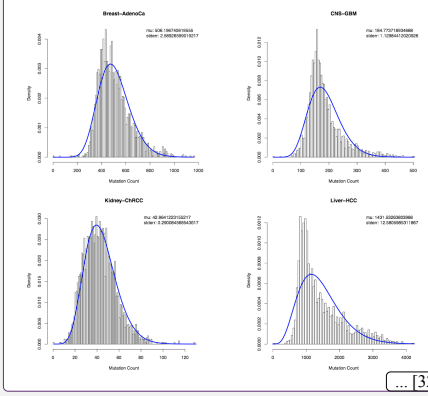
Deleted: ... [31]

Formatted Table

	For most cancer types, fitting a Gamma-Poisson is pretty good (as seen in the figures below). However, we agree that it is interesting to investigate other non-conjugate priors. As the referee mentioned, this is out of scope, but we have noted this in the text.
Excerpt 2.3-A (added in Supp.)	

Deleted: ... [33]

Deleted: From



... [32]

<ID>REF2.4 – Was the Poisson Model used for low mutation cancers

<TYPE>$$$BMR,$$$Text,$$$Cale
<ASSIGN>@@@JZ,@@@JL
<PLAN>&&&AgreeFix
<STATUS>%%%80DONE

Referee Comment	2) It seems that the Poisson model was not rejected for cancers with very low mutation counts (liquid tumors). Is this a power issue rather than the property of the mutation process?
Author Response	We thank the reviewer for mentioning this, and we feel this is a good point. We think higher mutation rate is often associated with overdispersion, but the rejection of a poisson model is not just due to limited power. We carried out further analyses in our revised manuscript. • We added a new plot to show the average mutation rate vs. the overdispersion parameter (Excerpt 2.4-A). • We added a new supplementary figure of the QQ-plot using Poisson and NBR, and we found that they provide similar results. We need to check two key aspects, enough covariate correction and separating the kmers, before considering overdispersion (Excerpt 2.4-B). • Other papers only based on poisson regression with good covariates, and kmer separation works well (https://www.biorxiv.org/content/early/2017/12/19/236802). In summary, it is simpler to avoid introducing additional parameters. However, we think it is better to check how heterogeneous the count data can be, even after correcting for the effects of enough covariate.
Excerpt 2.4-A (added in Supp.)	We plotted the overall mutation count under different 3mer context vs. the estimated overdispersion parameter (using the AER package) in R in the following figure. On one side, it is obvious that for those 3mers with more variants, there is a tendency to introduce overdispersion and accept the Gamma-Poisson model.

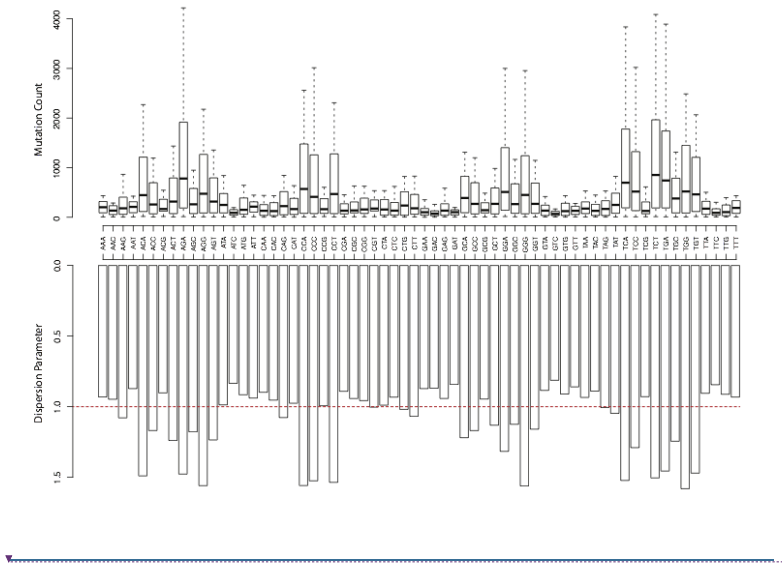
Formatted Table

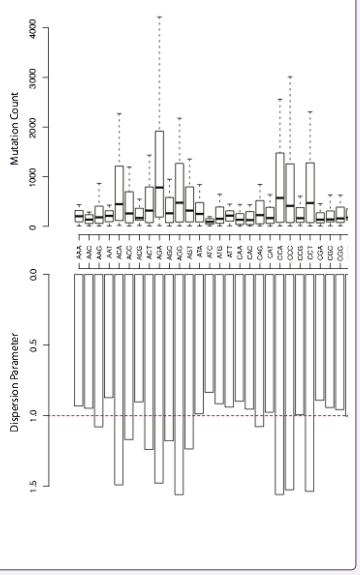
Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: . (details please see excerpt 1)

Deleted: .

Deleted: 1 From ... [34]

	
Excerpt 2 4-B (QQ-plot of Poisson vs. NBR)	We have used both Poisson and Negative Binomial distribution on the mutation burden calculation with exactly the same covariate set. QQ-plot of p values on breast cancer CDS region were given below. QQ-plots from these distributions look similar. Similar to the conclusion by wadi et al, it is necessary to first check whether covariate effects have been corrected and local kmer context has been calibrated and then test the level of overdispersion.

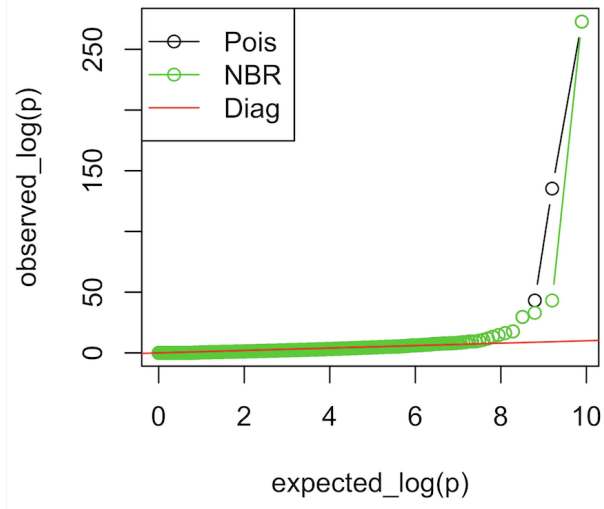


Deleted:

Deleted: poisson

Deleted: negative binomial

Deleted: From . ... [35]



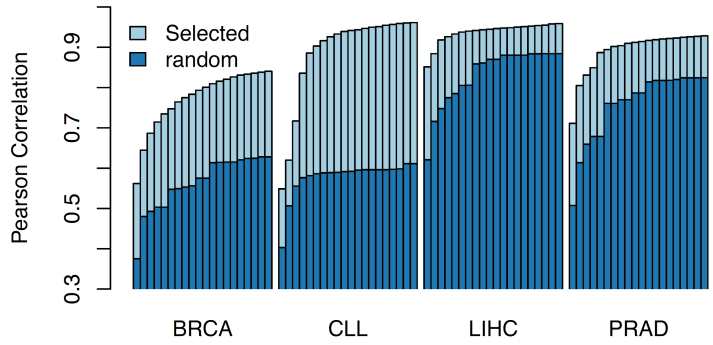
Deleted: JZ2JZ: remember to put the figure in.

<ID>REF2.5 – BMR: use of principal components

<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE,%%%CalcDONE
Add the cross validation in this response section

Referee Comment	3) The approach with principal components used for the BMR estimation does not seem to work well. Starting with the second PC most components have roughly the same prediction power. One possibility is that higher principle components do not capture the additional signal and reflect noise in the data, and the correlation with mutation rate is due to an overfit of the NB regression (it is unclear whether it was analyzed with cross-validation). Another possibility is that the signal is spread over many components. In the latter case, this is not an optimal method choice.

Formatted Table

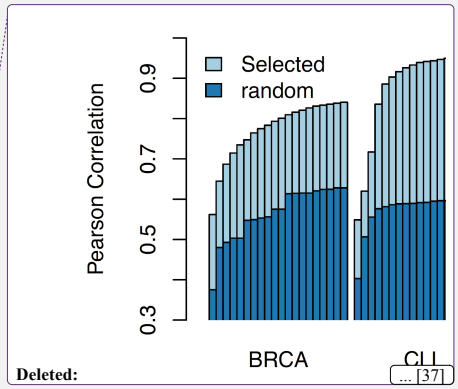
Author Response	We thank the referee for pointing out the limited contribution from the higher-order principal components. In the revised version, we have tried to better illustrate our main point: the wealth of the ENCODE data for BMR estimation. In summary, we have • revised figure 2 by directly using a combination of features via forward selection (Excerpt 2.5-A), and we have moved the PCA part into the supplement. • added a supplementary figure of cross validations (Excerpt 2.5-B)
Excerpt 2.5-A (modified main fig.)	At 1mb bin resolution, we compared the performance of models using random features vs. computationally selecting best features sequential (forward selection). It has shown that by adding features appropriately from ENCODE3, we can noticeably improve the performance of BMR accuracy. 
Excerpt 2.5-B (added in Supp.)	To avoid overfitting problem, we performed 5 fold cross validation using the selected model for each cancer type and listed the performance as below.

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: details in excerpt 1

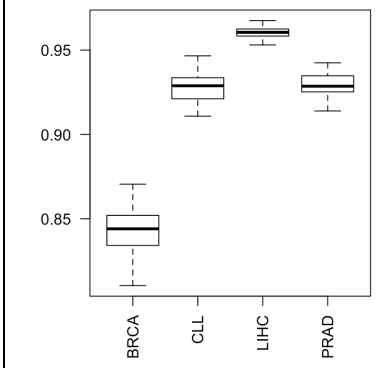
Deleted: details in excerpt 1)

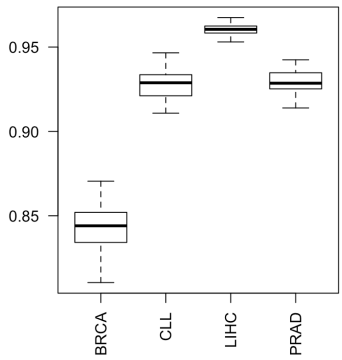
Deleted: 1 From ... [36]



Deleted: ... [37]

Deleted: From ... [38]





Deleted:

<ID>REF2.6 – Comments on the power analysis and compact annotations

<TYPE>$$$Power,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%80DONE
[JZ2JZ: more equations to come]

Referee Comment	4) I do not agree with the power analysis presented to support the idea of compact annotations. I understand that this is a toy analysis neglecting specific properties of mutation rate known for regulatory regions and also sequence context dependence of mutation rate. The larger issue is that the analysis assumes that ALL functional sites are within the compact annotation. In that case, power indeed would decrease with length. However, in case some of the functional sites are outside the compact annotation power would not decrease and is even likely to increase with the inclusion of additional sequence. Is there a justification for all functional sites to reside within compact annotations? Can this issue be explored? Some statistical tests incorporate weighting schemes.
Author Response	We thank the referee for this feedback, and we certainly agree with the referee. As suggested, we have largely expanded our somatic burden

Formatted Table

	power calculations under various assumptions. In summary, we have now included: • an entirely new section on power analysis and the effect of test region functional site ratios (Excerpt 2.6-A) • more discussion (in the main text) about the pros and cons of merging test regions (Excerpt 2.6-B) • real examples in supplement (Excerpt 2.6-C) • a new section of quality metrics of the compact annotations to capture functional sites and remove noise(Excerpt 2.6-D)
Excerpt 2.6-A (in Suppl.)	Suppose that we define the following parameters. l_i^n : noise region length for region i l_i^r : noise region length for region i μ_i : BMR in region i λ_i : effect size in risk region i $\rho_i = \frac{l_i^r}{l_i^r + l_i^n}$ Then under the null hypothesis, the probability to observe at least one mutation per patient is $p_0 = 1-(1-\mu_i)^{\frac{l_i^r+l_i^n}{C}}$ Under the alternative hypothesis, $p_1 = 1-(1-\mu_i)^{l_i^n}(1-\lambda_i\mu_i)^{l_i^r}$ We did a simulation by starting from a very noisy test region with pretty low true risk loci percentage. We have showed that by trimming the nosie loci, statistical power can be increased. But after we have removed the noise and start to trim the true functional loci, the statistical power drops quickly.

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: see supplement and excerpt 1 below

Deleted: see in excerpt

Deleted: see in excerpt 3

Deleted: catpure

Deleted: rm

Deleted: see in excerpt 4

Deleted: 1 From ... [39]

Deleted: l_i^n : noise region length for region i l_i^r : noise region length for region i μ_i : BMR in region i λ_i : effect size in risk region i $\rho_i = \frac{l_i^r}{l_i^r + l_i^n}$

Deleted: hypotheis

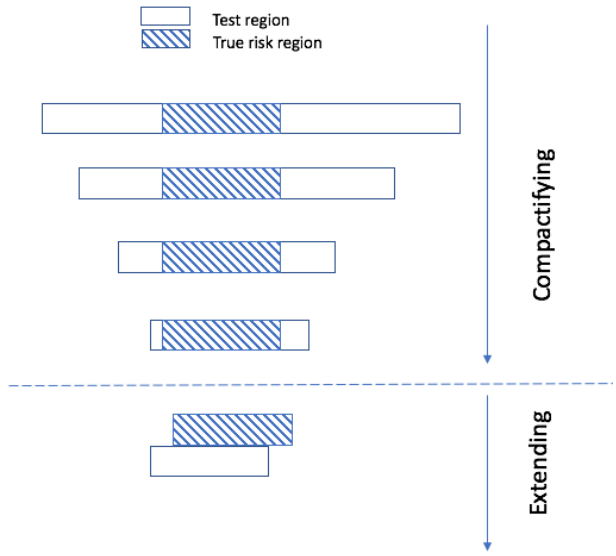
Deleted: proability

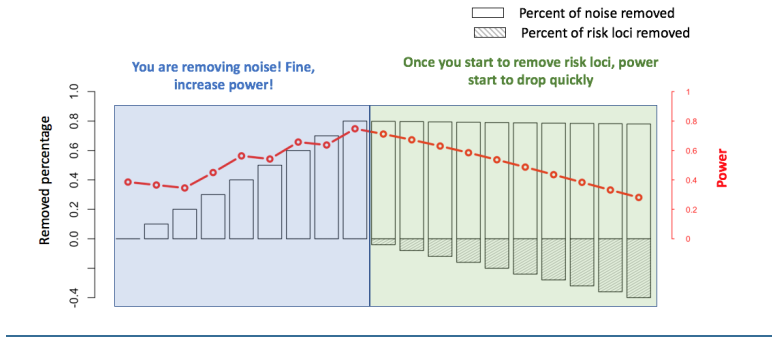
Deleted: $p_0 = 1-(1-\mu_i)^{\frac{l_i^r+l_i^n}{C}}$

Deleted: hypotheis

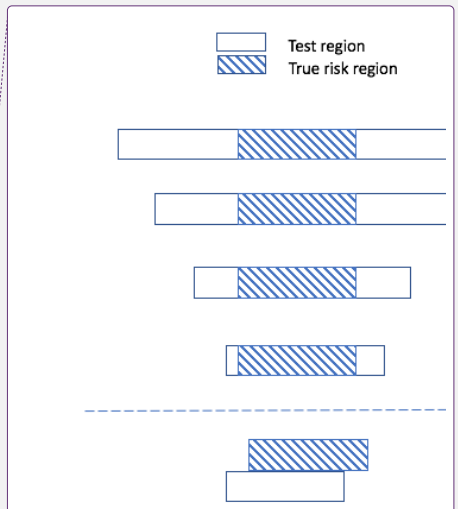
Deleted: $p_1 = 1-(1-\mu_i)^{l_i^n}(1-\lambda_i\mu_i)^{l_i^r}$

Deleted: quicktly



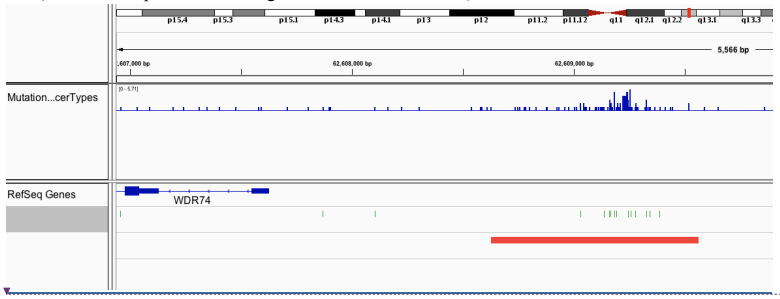


Excerpt 2 (added in main text)	In summary, our claim is that first we provide compact annotations to pick up functional nucleotides and remove noisy ones through the guidance of many functional characterization assays. Then we hope to join the distributed functional sites together to increase statistical power.



Deleted: ... [40]

Deleted: From ... [41]

Excerpt 2.6-C (in Suppl.)	We provided two examples to explain the motivation of our compact and extended gene annotations and why we feel our assumptions for the power analysis is reasonable. 1) Enhancers: Traditionally, enhancers were called as a 1kb peak regions, which admittedly introduced a lot of obviously nonfunctional sites. We believe we can get functional region more accurately by trimming the enhancers down using the exact shapes of many histone marks and further integration with STARR-seq and Hi-C data. 2) TFBS hotspots around the promoter region of WDR74. Instead of testing the conventional up to 2.5K promoter region, we can trim the test set to a core set of the promoter region where many TFs bind, which perfectly correlates with the mutation hotspots (red block) for this well-known driver site (blue line for pan-cancer and green line for liver cancer). 
Excerpt 2.6-D (in Suppl.)	• *Regarding the qualities of enhancers* As for the enhancer part, with the ensemble method, for example, we can get more accurate annotation and pin-point to sequences where transcription factors would actually bind to. To estimate the false positive rate would not be very practical at this stage as there is no gold-standard experiment that could assert an predicted enhancer is definitely negative. Here we took the FANTOM enhancer data set and assess the overlap percentage of our enhancer annotation in each ensemble step. We showed that each ensemble step indeed increases the percentage of overlap between our annotation and the FANTOM enhancer set. The overlap percentage for our annotation is much higher than that of the Roadmap annotation, and is also higher than the main encyclopedia enhancer annotation annotation (ccRE).

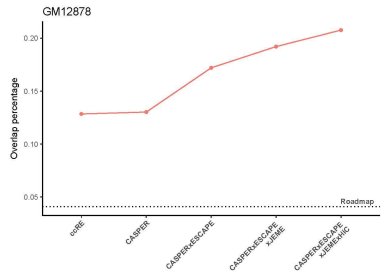
Deleted: 3 From ... [42]

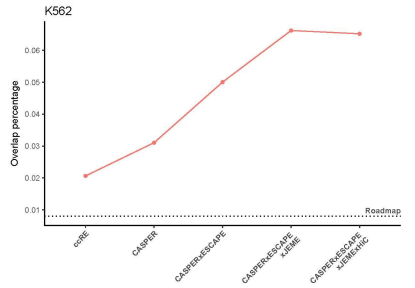
Deleted: assupmptions

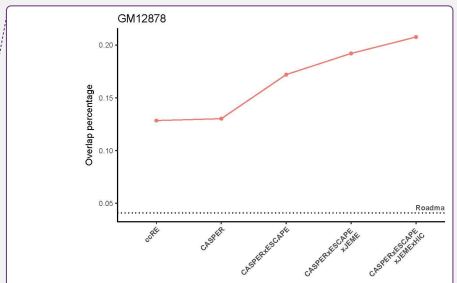
Deleted:

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

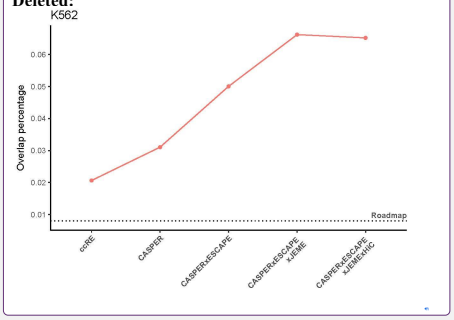
Deleted: 4 From ... [43]







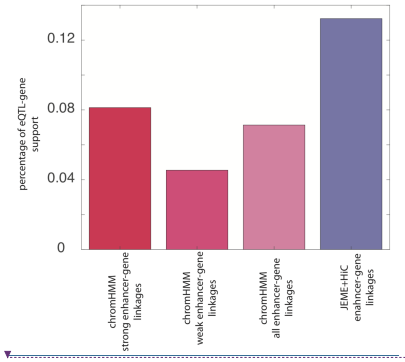
Deleted:

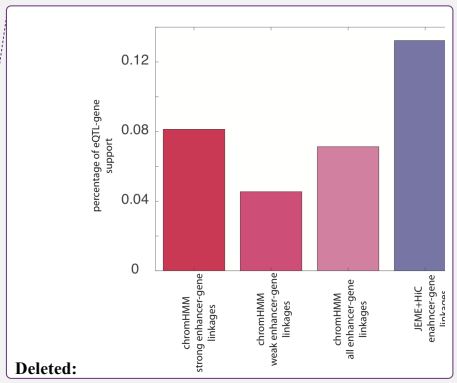


- *Regarding the quality of enhancer-gene linkages:*

To show how our JEME+Hi-C approach captures enhancer-gene linkages compared to existing linkages, we used published chromHMM derived enhancer-gene linkages (cite chromhmm) as the comparison dataset and GTEx whole blood eQTLs as the benchmark. We found the linkages, which the enhancer has an eQTL that changes the expression of the target gene significantly. After finding all the eQTL supported linkages for chromHMM and JEME+Hi-C, we calculated the fraction of enhancer-gene linkages that has eQTL support for various types of linkages in chromHMM and in JEME+Hi-C. As can be seen in figure below, JEME+Hi-C has higher fraction overlapped with eQTL-gene linkages.

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"



Deleted:

<ID>REF2.7 – Value of the extended gene

<TYPE>$$$NoveltyPos
<ASSIGN>
<PLAN>&&&AgreeFix,&&&MORE
<STATUS>%%%75DONE

Referee Comment	6) The idea of extended genes and the use of multiple information sources to construct them is a strength of the paper. It would be great to see a formal analysis about how extended genes increase power of cancer driver discovery.
Author Response	We thank the reviewer for the positive remarks of the extended gene. As suggested, we further highlighted this part in our revised manuscript. We also tried to make it more clear that our goal here is to illustrate how the extended gene concept can be used in cancer. We have also re-organized all our related analysis to better illusrate the value of our extended gene resource, which includes • GWAS germline variant enrichment analysis across different annotations in the main figure (Excerpt 2.7-A) • A new figure panel to stratify patient expression levels based on the mutation status from various annotations. We found that extended genes perfromed better than others (Excerpt 2.7-B)

Formatted Table

Deleted: in the orginal supplement to the main text

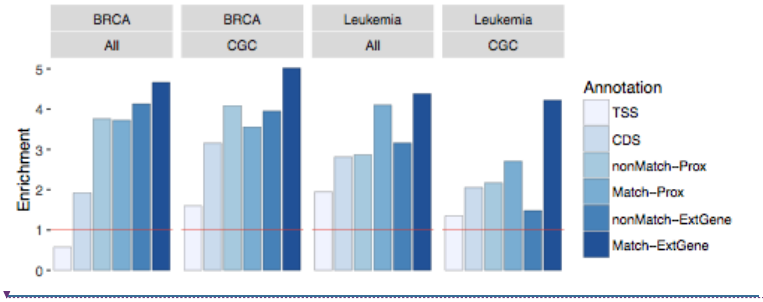
Deleted:

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: see in excerpt 1

Deleted: stratift

Deleted: see in excerpt

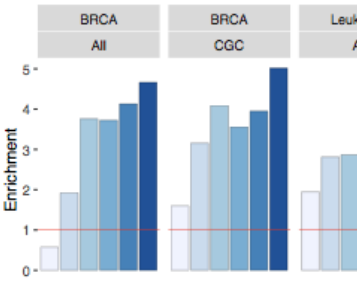
	• A new figure in the supplement to show variant effect in extended gene regions on regulator activities (Excerpt 2.7-C) • A CRISPR based validation of onco-gene activation based on extended genes (Excerpt 2.7-D)
Excerpt 2.7-A (main Manuscript)	We extracted all the breast cancer GWAS variants from GWAS Catalogue and only kept those with European ancestry. Then we extracted all the LD SNPs within 500kb of the GWAS SNP (r2>0.8) to calculate variant enrichment in different annotations sites. The R package VSE was used (https://cran.r-project.org/web/packages/VSE/vignettes/my-vignette.html). We found that extended gene regions showed significantly larger variant enrichment than the CDS regions and TSS regions. 
Excerpt 2.7-B (main Manuscript)	For a given gene, we tried to separate patients into groups with or without mutations under certain annotations, such as CDS, UTR, TF/RBP binding sites, enhancers, and our extended gene. We then tried to test difference of gene expressions (FPKM) from these two groups based on two-sided Wilcoxon. We found that our extended gene annotation provides better expression separation between these two groups. As an ex to illustrate the value of the extended gene for expression analyis, we show a well-known splicing factor SRSF2, which has been recently reported to drive liver cancer development \cite{28082404}, gives the strongest p-value for stratifying expression out of all genes in liver cancer.

SHORTEN

Deleted: see in excerpt 3

Deleted: see in excerpt 4

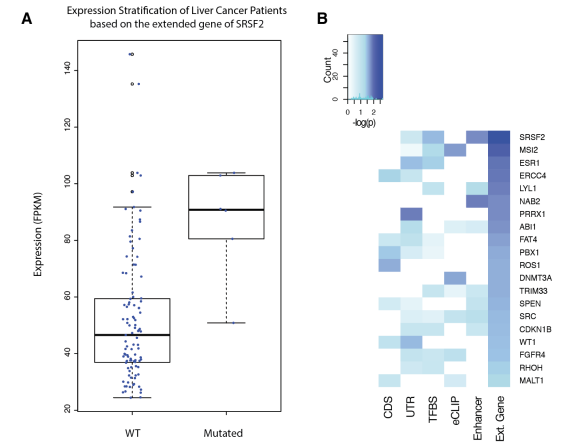
Deleted: 1 From ... [44]



Deleted:

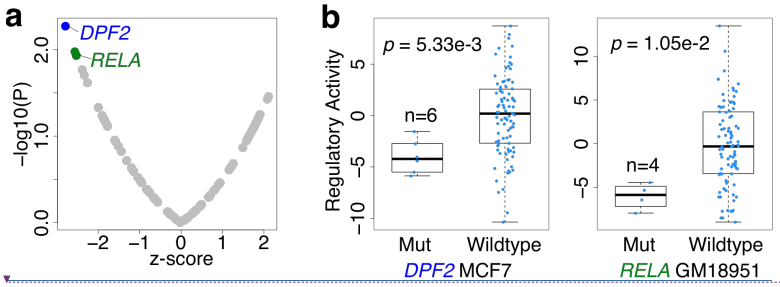
Deleted: From ... [45]

Deleted: Specifically, we found

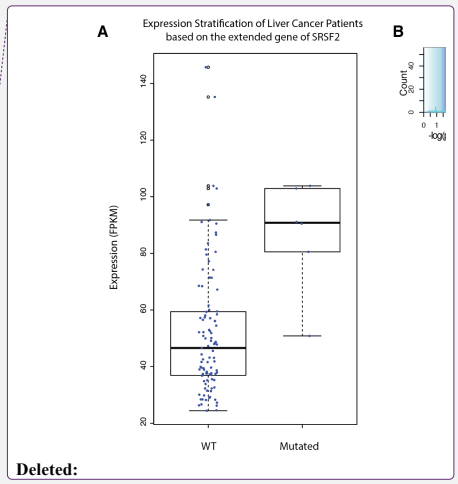


Excerpt [2.7-C (in Suppl.)]

We analyzed the association between TF mutations in extended gene region and TF regulatory activity in three cancer types (breast, liver, and leukemia). Between each pairs of mutation type (e.g., ENH1, TF, eCLIP, UTR) and cancer type, we tested the association between mutation status and TF regulatory activity by two-sided rank-sum test and converted the *p*-values into FDRs by Benjamini-Hochberg procedure. Only the combination between liver cancer and ENH1 mutation has statistically significant results (FDR < 0.25, panel a). A mutation in the enhancer region of DPF2 or RELA indicates a lower TF regulatory activity (panel b). These results indicate that mutations in enhancers may cause TF loss-of-function in certain cancer types.

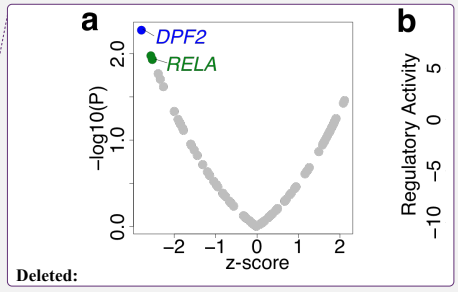


Supplementary Figure X. Mutations in level one enhancers affects the activity of nearby TFs. (a) The association between TF regulatory activity and mutation in enhancer regions. For each cancer type, the association between TF regulatory activity computed using ChIP-seq data and mutation status of nearby enhancer region was tested by two-sided rank-sum test. Only liver cancer has significant associations (FDR < 0.25) for TF DPF2 and RELA, and the results for liver cancer are shown with volcano plot. X-axis represents the z-score of rank-sum test and Y-axis represents the negative log p-values. (b) The regulatory activities of significant TFs in panel a in tumors with mutated or wild-type TF genes. The comparison between two groups was done by two-sided rank-sum test.



Deleted:

Deleted: 3 From ... [46]

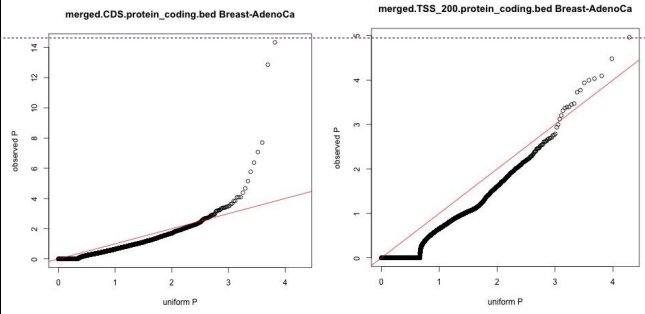


Deleted:

Excerpt 2.7-D (main manuscript)	Ask Feng's group for text and wait for figure to come in

Deleted: 4 From ... [47]

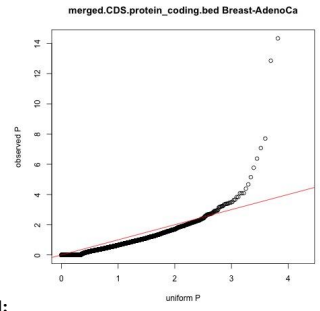
<ID>REF2.8 – Q-Q plots
<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&Defer
<STATUS>%%%90DONE

Referee Comment	Some of the QQ-plots in supplementary figures look problematic. Also, for some tumors with low count statistics QQ-plots are expected to always be deflated, so the interpretation of QQ-plots may be non-trivial.
Author Response	We thank the referees for this comment. We have updated the QQ-plots in our revised manuscript. It is actually due to a minor issue when we are using R for P value calculation. For negative binomial (or Poisson), the test on the right tail should be P(X>=x_obs). However, in R pnbinom(x, size, prob, mu, lower.tail = F, log.p = FALSE) actually calculated the P(X>x_obs), which will introduce a slight p value inflation in our orginal submission. We have corrected this and provided the updated QQ-plot as below.
Excerpt 2.8-A	

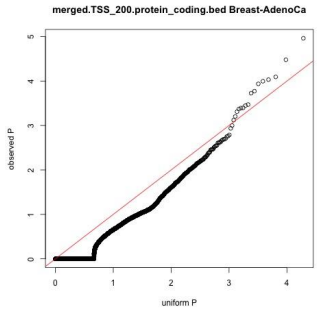
Formatted Table

Deleted: 5)

Deleted: and they look fine



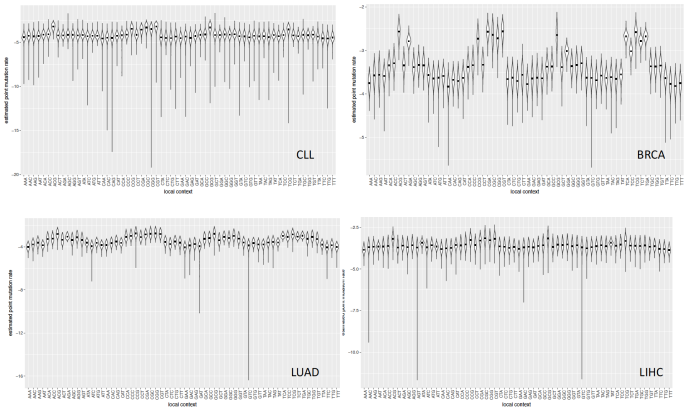
Deleted:



Deleted: From ... [48]

<ID>REF2.9 – BMR effect on local tri-nucleotide context

<TYPE>$$$BMR,$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%90DONE

Referee Comment	However, it is unclear whether the analysis takes into account complexities of the mutation model in regulatory regions. The influence of tri- or even penta-nucleotide context can be significant.
Author Response	We thank the referee for pointing out this. We have considered the influence of tri-nucleotide effect in our original submission. As suggested, we have tried made it more clear in our revised manuscript that the influence of local text is significant.
Excerpt 2.9-A (main text)	We feel local context and covariate correction are two main factors to confound somatic burden analysis. In our BMR model, we performed separate trainings for all 3mers and allow then two chage differently with various genomic features.
Excerpt 2.9-B (org. Suppl.)	Consistent with previous literature, we observed large mutational heterogeneity over the genome for all 3-mers in all cancer types. As seen in Figure S 2-2 , the mutation rate changes significantly over different regions of the genome. (large region of each violin bar) and over different local contexts. Figure S 2-2 (TL, ⅈ) Violin plot of estimated BMR over local context and genomic locations 

Formatted Table

Formatted: Font:12 pt

Deleted: the

Formatted: Font:12 pt

Moved (insertion) [4]

Deleted: *The newly added sentence in the main text:*

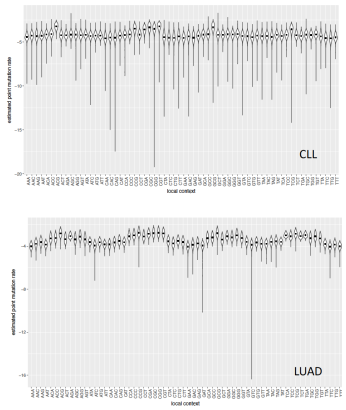
Moved up [4]: We feel local context and covariate correction are two main factors to confound somatic burden analysis. In our BMR model, we performed separate trainings for all 3mers and allow then two chage differently with various genomic features.

Deleted: ... [49]

Formatted Table

Deleted: From main text and *The newly added sentence in the main text:*

Figure S 2-2 (TL, ⅈ) Violin plot of estimate



Deleted:

<ID>REF2.10 – Confounding factors

<TYPE>$$$Other
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%85DONE

Referee Comment	Next, TF binding and nucleosome occupancy is known to interfere with the activity of DNA repair system.
Author Response	We thank the referee to bring out this important point. Actually many of the current background mutation rate estimation method assumes a constant rate in a fairly large region, such as a within a gene (including the long introns in between) or up to Mbp fixed bins. In such large scale, it is difficult to small scale features such as TF binding, nucleosome occupancy, histone modification (which changes sharply in less kbps). Hopefully, with accumulating cancer patient data in the future could help to build up site specific background models to investigate more about such effects. We added this point in our discussion section.
Excerpt 2.10-A (main text)	Hower, most of the current BMR models are focused on larger scale mutation rate variations by integrating many features at 50 kb to 1 Mb resolution while ignoring small scale perturbations introduced by TF binding and nucleosome occupancy. Improvement of such finer scale features in the future could further improve BMR estimation.

Formatted Table

Deleted: From . ... [50]

<ID>REF2.11 – minor: comment on burden test

<TYPE>$$$Minor,$$$Presentation,$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

Referee Comment	1) I would not use the term "burden test". This usage is slightly confusing because this term is commonly used in human genetics where it refers to a case-control test.
Author Response	We thank the referee to point out his confusion about the term "burden test". This is where some of the confusions of this paper come from. Originally we intended to use this term because we want to emphasize that our

Formatted Table

	resource is not just for somatic variant analysis such as cancer driver detection. We have other applications such as case-control GWAS variant interpretation. We have re-organized our analysis to better convery our idea. Please check details to the response in REF 2.7 above.

<ID>REF2.12 – Minor: comment on terminology

<TYPE>$$$Minor,$$$Presentation,$$$Text
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

Referee Comment	2) Similarly, it is unclear what is meant by "deleterious SNVs" as the term is commonly used in human genetics in reference to germline variants under negative selection.
Author Response	We thank the referee to point out this. "Deleterious SNVs" in our manuscript means somatic mutations that disrupts gene regulations. To avoid potential confusion, we changed it in our revised manuscript.

Formatted Table

Referee #3 (Remarks to the Author):

<ID>REF3.0 – Preamble

<TYPE>$$$Text
<ASSIGN>@@@MG,@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

In relation to the supplement, the referee points out that it is sometimes hard to see full documentation of our methods in the main part and one has to look at the extensive supplements. We are well aware of this fact. The very large scale of supplement is quite typical for large genomic paper, such as the previous roll outs of the ENCODE publications \cite{encodenet and the main encode paper}.

The whole ENCODE publication committee, in fact, has been actively discussing with Nature Publishing and other companions journals about the supplement with regard to the main text. We have attempted to put important things in the supplement and to structure it very carefully.

Based on suggestions from Nature and the editor, we are prepared to work very hard to make the structure of the supplement understandable. As suggested, we have tried to revise it to make it clearer and also to move more method descriptions into the main text, though we think given the current main text limitations of a typical Nature paper and the scale of data and analytical results in this paper, it is almost impossible to put everything into the main text. We are preparing to work constructively with the referees and the others to make this clear.

<ID>REF3.1 – Presentation of the paper

<TYPE>$$$Presentation
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%25

Referee Comment	It is difficult to understand the significant novel findings in this paper (compared to the main ENCODE paper). Perhaps, some of this is due to the data not being presented in a concise and clear manner. For example, I wonder whether the authors can add more details and straightforward directions when citing supplementary

Deleted: commitee

Deleted: We admit that maybe this construction is not that intuitive. We

Deleted: TBC

Formatted Table

	information. In the current main manuscript, the authors cited all supplementary information as (see suppl.). It might be hard for the reader to check where the authors refer to in the supplementary information. I think more direction, such as sup Fig1, sup Table 1, or section 7.2S etc, would be very helpful.
Author Response	We thank the referee to raise this comment about our supplementary file. Our original thinking was some of the contents are distributed in multiple sections. For example, each step in the final prioritization scheme is corresponding to a separate part in the supplements. As suggested, we have added the specific sections in our revised manuscript to make it easier to check the technical details.

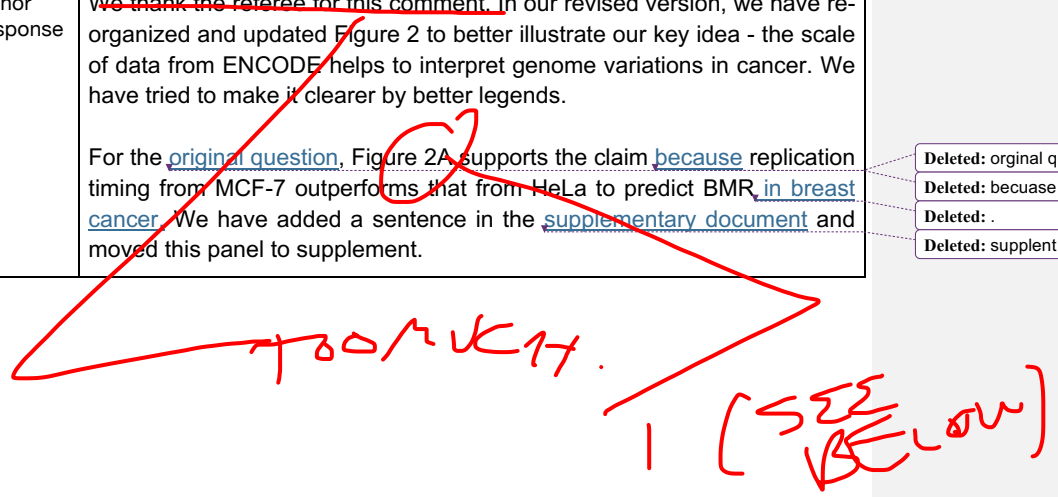
Deleted: orginal
Deleted: are
Deleted: section

<ID>REF3.2 – Benefits of using multiple cancer types in BMR

<TYPE>$$$BMR
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	In the second paragraph of page 3, it says 'using matched replication timing data in multiple cancer types significantly outperforms an approach in a which one restricts the analysis to replication timing data from the unmatched HeLa-S3 cell line.' This statement is confusing and does Figure 2A or 2B supported it?
Author Response	We thank the referee for this comment. In our revised version, we have re-organized and updated Figure 2 to better illustrate our key idea - the scale of data from ENCODE helps to interpret genome variations in cancer. We have tried to make it clearer by better legends. For the original question, Figure 2A supports the claim because replication timing from MCF-7 outperforms that from HeLa to predict BMR in breast cancer. We have added a sentence in the supplementary document and moved this panel to supplement.

Formatted Table
Deleted: orginal quetion
Deleted: becuase
Deleted: .
Deleted: supplent



Excerpt 3.2-A	Wait for new figure 1

Deleted: From ... [51]

<ID>REF3.3 – Presentation of the data figure

<TYPE>$$$Presentation
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	In Figure 1, "top tier" should point to cell types that is mentioned in the content. However, we also see SNV, SV, Mutation, etc.
Author Response	We thank the referee for this comment. In fact, by integrating many assays such as whole genome sequencing and Jrys, we called the SNV and SVs for several top tier cell lines, and release them together with our resource (see excerpt 2). In the revised figure 1, we have made it clearer that our resource include these SVs and SNVs.
Excerpt 3.3-A (main Fig)	Wait for updated Fig 1
Excerpt 3.3-B (suppl.)	JZ2DL: could you pls make a table from Feng's data and deposit it to our resource?

Formatted Table
Deleted: , xxx,
Deleted: xxx
Deleted: serveral
Deleted: ... [52]
Deleted: From ... [53]
Formatted: Highlight
Deleted: From ... [54]

<ID>REF3.4 – Regarding enhancer detection algorithm

<TYPE>$$$Presentation

<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	What is a single shape algorithm? The authors point to Supplementary data, but there is no definition there either. Do the authors mean the complete graphs or connected components?
Author Response	We thank the referee for the comment. It is based on a method pattern recognition method to identify the double peaks. We have updated the supplementary and provided more detailed indexing in the main text.
Excerpt 3.4-A	JZ2MTG: may need something more about CASPER. Please add here

Formatted Table

Deleted: CRASPER

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:Times New Roman, 10 pt

Deleted: From ... [55]

<ID>REF3.5 – Regression coefficients of BMR

<TYPE>$$$BMR
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	For Figure 2B, what does 'regression coefficients of remaining features' mean? Does that means beta_0 or the remaining regression noise? From Figure 2B, the coefficient to regression is rounded to -0.001 and 0.001. How should we understand these values? **If the coefficients are for the main features, we would be expecting higher coefficients, wouldn't we? In this case, does it means the lower the better?**
Author Response	To better illustrate the value of ENCODE data and our extended gene annotation, we reorganized our analysis to provide a new figure and moved this to the suppl. We have also fixed the text to describe our method and specifically answer the referee's questions (details in the excerpt below).

Formatted Table

Formatted: Font:Bold

Excerpt 3.5-A(Suppl.)	Our model incorporated many genomics features. Here features only means functional genomics data, such as H3K27ac and DHS. The absolute value of regression coefficient is closely related to how we normalized the data. For the genomic features, we calculated the average signal per 1mbs and transformed it into Z scores. It is worth mentioning that we also had an offset parameter, which means we are trying to estimate the point mutation rate (~10E-6 in some cases), so 0.001 is not a small value. Regarding the interpretation of the regression coefficient, the larger absolute value means better BMR estimation.

Formatted: Font:Times New Roman
Deleted: one set of
Formatted: Font:Times New Roman
Deleted: From [56]
Deleted: [57]
Formatted: Font:Times New Roman
Deleted: with
Formatted: Font:Times New Roman

<ID>REF3.6 – definition fo the extended gene

<TYPE>$$$Annotation
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	For Figure 2C, more explanation is needed on how to form an extended gene.
Author Response	We thank the referee for this comment and we have added a paragraph in the supplement to better describe how we generated the extended genes. (Excerpt 3.6-A)
Excerpt 3.6-A	There are four important basic elements in our extended gene definition: CDS, TFBS, RBP binding sites, and enhancers. For each gene, we extracted all the TFBS within 2.5kb of the tss sites of the protein_coding transcript, all the eCLIP binding sites of the whole transcript (and upstream 200bp and downstream 1500bp), all the linked enhancers, and then merged these annotations together to form the extended gene.

Formatted Table
Deleted: see excerpt below
Deleted: definitoin
Deleted: From [58]

<ID>REF3.7 – Validations

Deleted: validations

<TYPE>$$$Annotation
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Deleted: <TYPE>$$$Annotation
For the Figure 2D and its description on the third paragraph of page 4 (as well as Figure 3A), did the authors validate all the genes systematically? [59]

Referee Comment	For the Figure 2D and its description on the third paragraph of page 4 (as well as Figure 3A), **did the authors validate all the genes systematically?** Is there any validation rate showing the precision rate of the method?

Formatted Table
Formatted: Font:Bold

Author Response

We thank the referee for raising the question of validations.

For Figure 2D, it is about the somatically burded genes. We fully agree with the referee that it is useful to compare our BMR to established benchmarks. We are aware of community efforts and are very involved with the PCAWG effort to do whole genome cancer analysis. One of our authors is the co-leader of the non-coding annotation group. PCAWG, which is a hybrid of TCGA and ICGC, has not developed any explicit BMR benchmark. Instead, we have provide literature support for our discovered genes and added them into a supplementary table (Excerpt 3.7-A).

Please note that we do have explicit validation for the prioritized SNVs and SVs in the paper. For instance, Figure 2C shows a validation of extended gene that initiate oncogene transcription (Excerpt 3.7-B). For Fig. 3A, We have used TF/RBP knockdown experiments to validate several key regulators, such as MYC and SUB1. We have also used external data to validate our conclusion. These analysis were added into our revised supplements (Excerpt 3.7-C).

Regarding the validation rate, we have prioritized SNVs at the end of our manuscript, 6 out of 8 SNVs were shown to affect gene expressions (Excerpt 3.7-C).

Excerpt 3.7-A (for Fig. 2D in Suppl.)

We have listed the literature supporting our discovered genes with higher than expected mutations.

BRCA

Gene	Cancer Type	Literature Support (PMID)	Known Cancer Gene (CGC)
CBFB	Breast	22722202, 16959974, 20668451	YES TSG
HIST1H2BF	Breast	26113056	
HIST1H2AD			
HINT3			
HIST1H3D	Breast	26113056	
PIK3CA	Breast	26028978, 29636477, 25176561, 27358378	YES Oncogene
TP53	Breast	11879567, 12619115, 8013000	YES TSG/Oncogene

LIHC

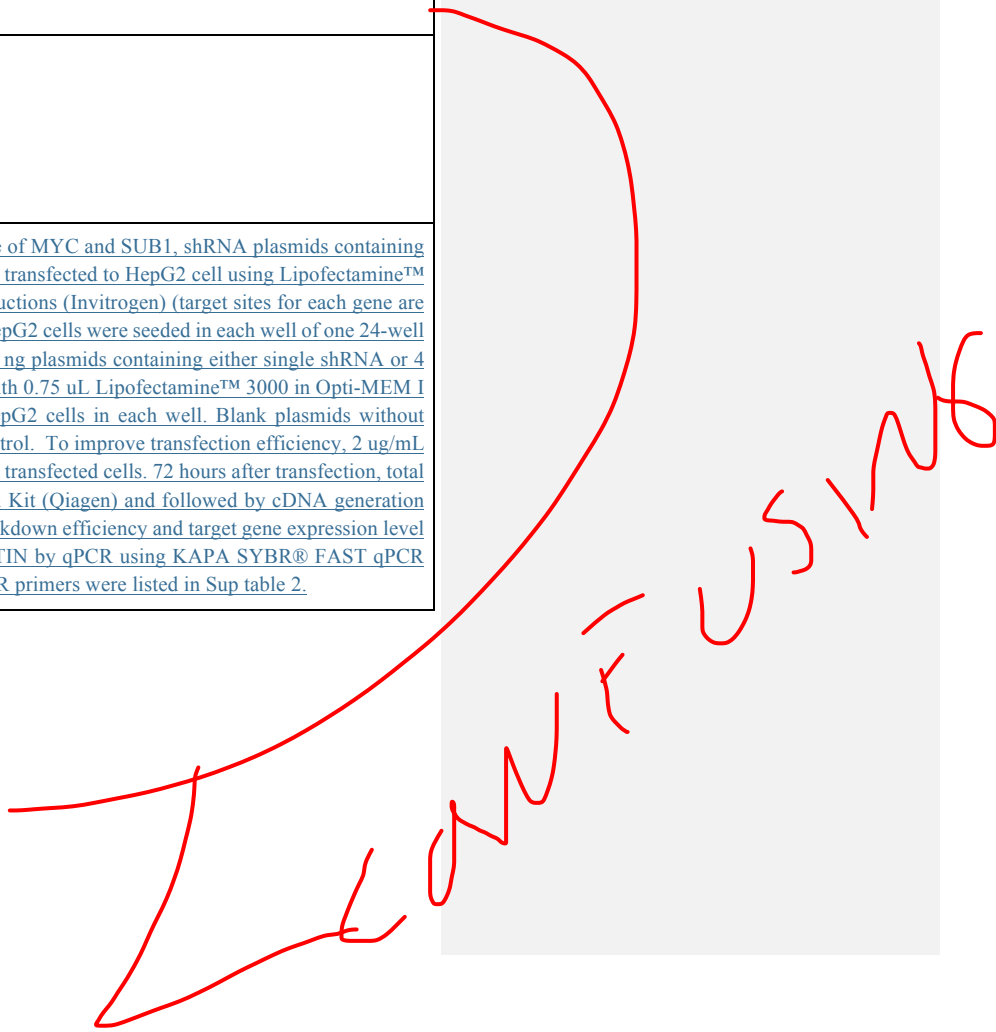
Formatted: Font:12 pt
Deleted: this issue of quality metrics of our annotations, such as
Formatted: Font:12 pt
Deleted: enhancers.
Formatted: Font:12 pt
Deleted: important
Deleted: provide such information. We have struggled hard
Formatted: Font:12 pt
Deleted: explain the much greater accuracy
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: annotations than previous effort, such as the chromHMM based enhancers purely from computation and imputed network based on DHS only.
Formatted: Font:12 pt
Moved (insertion) [2]
Formatted: Font:Arial, 12 pt, Not Italic, No underline
Deleted: As suggested, we have added a whole section in our revised our manuscript to discuss the qualityies of annotations, including:
Deleted: From ... [60]

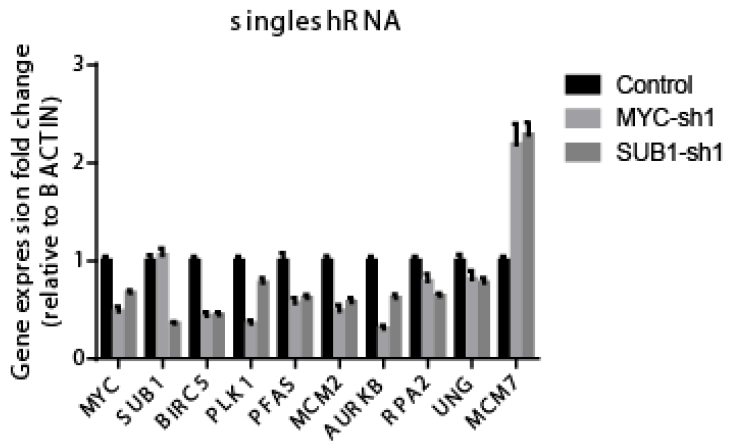
Gene	Cancer Type	Literature Support (PMID)	Known Cancer Gene (CGC)
TERT	Liver	26336998, 25267585, 28947783	YES
KRTAP5-11			
NFE2L2	Liver	22459801	YES
SETDB1	Liver	26471002, 26481868, 27334461	
ARID2	Liver	21822264, 26169693, 22095441	YES TSG
DUSP22			
IFI44L	Liver	27254796	
PHLDB2	Liver	22681909	
AL590714.1			
APOB	Liver	23723369	
APOA2			
PLCXD2			
ZNF595			
ALB	Liver	24663086	
CTNNB1	Liver	26715116	YES Oncogene
TP53	Liver	17401425	YES TSG/Oncogene

CLL

Gene	Cancer Type	Literature Support (PMID)	Known Cancer Gene (CGC)
NXF1	CLL	27060156	
ATM	CLL	26113859, 22952040	YES TSG
SYVN1			
WDR74			
LTB	CLL	12801841	
SF3B1	CLL	25371178	YES

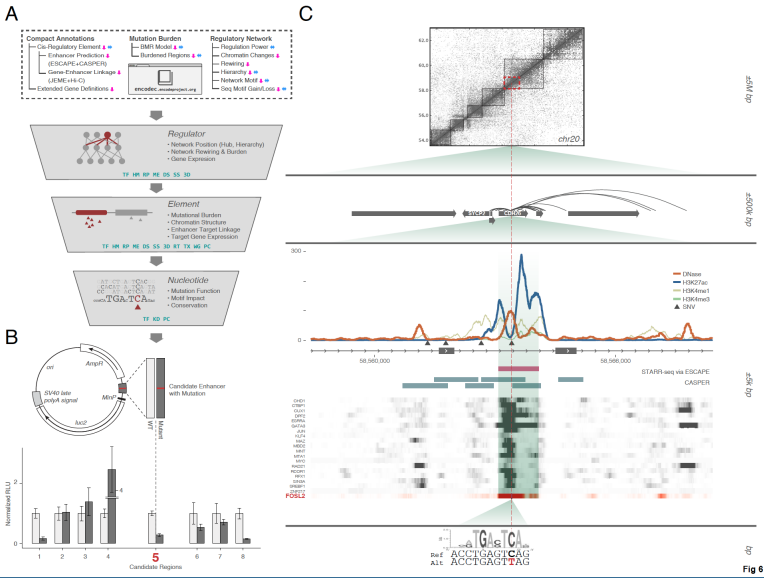
	BTG2			
	RPL11	CLL	12200376	
	BCL7A	CLL	23043359	YES Oncogene
	CXCR4	CLL	24855209, 20501831	YES Oncogene
	BACH2			
	BCL2	CLL	27069256	YES Oncogene
	TP53	CLL	27742075	YES TSG/Oncogene
	BCL6	CLL	19367498	YES Oncogene
Excerpt 3.7-B (for Fig2. C in main text)	Add Feng's text to b			
Excerpt 3.7-C (for Fig3 in main text)	To detect predicted common target gene of MYC and SUB1, shRNA plasmids containing 4 targets sites of each gene were used to transfected to HepG2 cell using Lipofectamine™ 3000 following the manufacturer's instructions (Invitrogen) (target sites for each gene are listed in Sup table 1). Briefly, 0.12 M HepG2 cells were seeded in each well of one 24-well plates 24 hours before transfection. 500 ng plasmids containing either single shRNA or 4 shRNA plasmids as pool were mixed with 0.75 uL Lipofectamine™ 3000 in Opti-MEM I medium (Invitrogen) and loaded to HepG2 cells in each well. Blank plasmids without shRNA target sequence was used as control. To improve transfection efficiency, 2 ug/mL puromycin was used to select successful transfected cells. 72 hours after transfection, total RNA was extracted using RNeasy Mini Kit (Qiagen) and followed by cDNA generation using SuperScript III (Invitrogen). Knockdown efficiency and target gene expression level were quantified and compared to BACTIN by qPCR using KAPA SYBR® FAST qPCR Master Mix (2X) Kit (Sigma). The qPCR primers were listed in Sup table 2.			





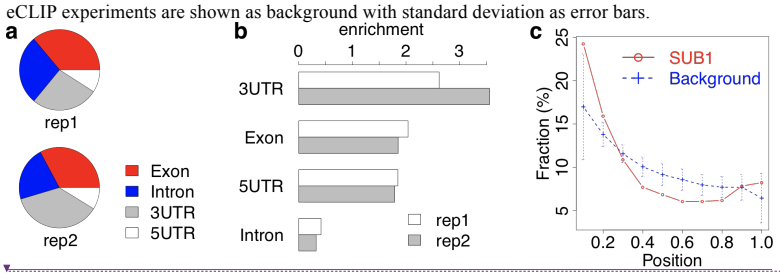
Excerpt 3.7-D (for SNV)

Figure 6.

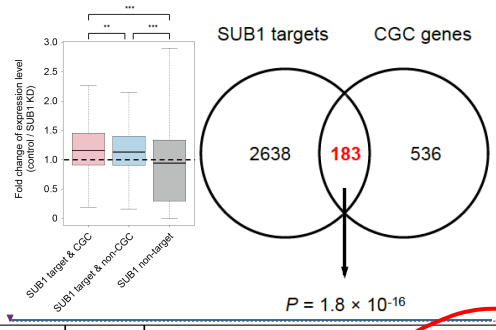


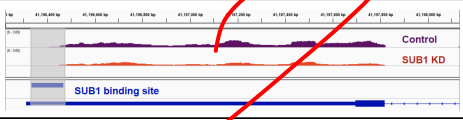
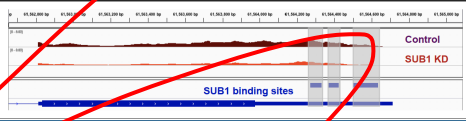
<ID>REF3.8 – novel oncogenes

<TYPE>$$$Annotation
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	Are there any novel oncogenes detected by the method?
Author Response	We than the referee to point out the novelty of discoveries. We have tried to make it clear that the main goal of this paper is to illustrate the value of ENCODE data and the usefulness of our deep annotations. We did find interesting genes that are associated with cancer, such as SUB1, ~~which is also mentioned by REF5 a potential novel oncogene.~~ To our knowledge, this is the first work to claim SUB1 to be associated with cancer as an RBP. ~~There are other work mentioning this gene, but not from the RBP aspect.~~ ~~We have added many follow up analysis on SUB1 in our revised version.~~
Excerpt 3.8-A (in Suppl.)	**Supplementary Figure X: eCLIP peaks of SUB1.** (a) The composition of SUB1 peaks over different gene regions is shown for each replicate. (b) For each gene region, the relative enrichment (fraction of SUB1 peaks / fraction of all peaks) of SUB1 peaks is shown. (c) The distribution of SUB1 peaks over 3'UTR regions is shown. The mean across all RNA binding proteins profiled by eCLIP experiments are shown as background with standard deviation as error bars. 
Excerpt 3.8-B (in Suppl.)	We found that SUB1 targets are enriched in cancer associated genes, such as genes in Cancer Gene Census (P=1.8e-16 by Fisher's exact test), and such genes showed larger down regulation upon SUB1 knockdowns. Among many of such genes, we have shown some IGV examples together with SUB1 binding sites on the 3' UTRs.

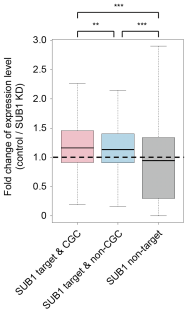
Deleted:
Deleted: 10
Formatted Table
Deleted: illustarate
Deleted: ... [61]
Deleted: 1 From ... [62]
Deleted: 2 From ... [63]



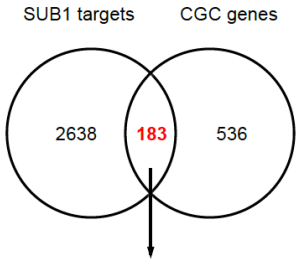
Gene	Functions	PMID	Expression profiles of the 3' UTR
BRCA1	The gene is involved in maintaining genomic stability	12677558, 17416853, 23620175, 16551709	
POLE	The gene is involved in DNA repair and replication	26133394, 28423643	
FEN1	The gene is involved in DNA repair and replication	20929870, 22586102	

Excerpt 3.8-C (in Suppl.)

Using ENCODE eCLIP data and TCGA tumor profiles, we applied RABIT framework to identify RNA binding proteins (RBP), whose target genes are differentially regulated in cancer. (a) The fractions of patients with target genes up or down regulated are shown for each combination of RBP and cancer type. (b) The patient fractions with target genes differentially regulated are shown for all cancer types and RBPs whose fraction values are larger than 50% in at least one cancer. (c) All lung adenocarcinoma patients are divided to two groups according to SUB1 activity predicted by RABIT. The overall survival was shown by KM plot. The association between SUB1 activity and survival was tested through Cox-PH regression. (d) In the left panel, the cumulative distributions of gene expression after SUB1 knock down in HepG2 cell are shown for predicted SUB1 targets and none targets. In the right panel, the cumulative distributions of mRNA decay rates in HepG2 cell are shown. The comparison between two categories is done through Wilcoxon rank-sum test.



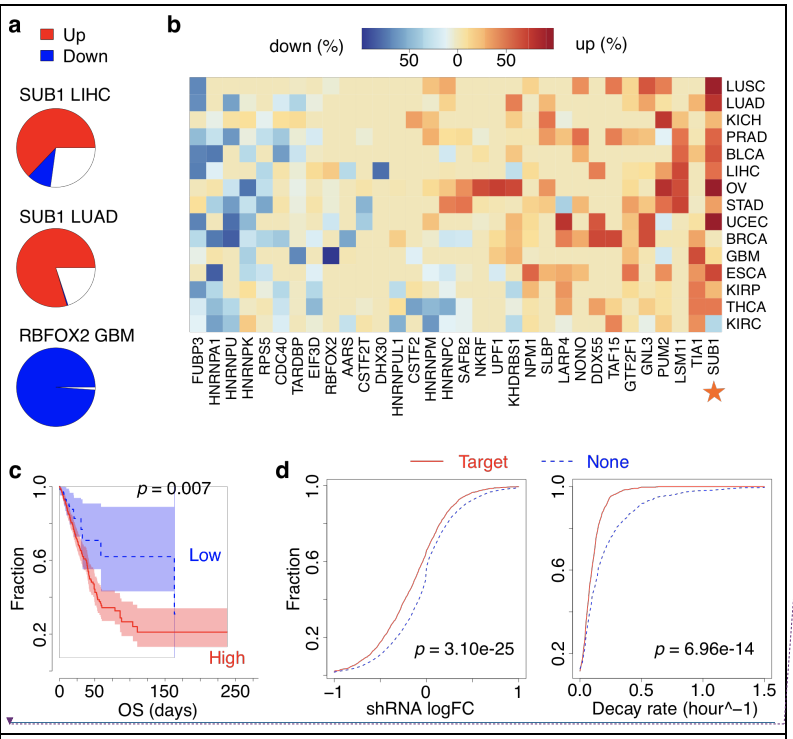
Deleted:

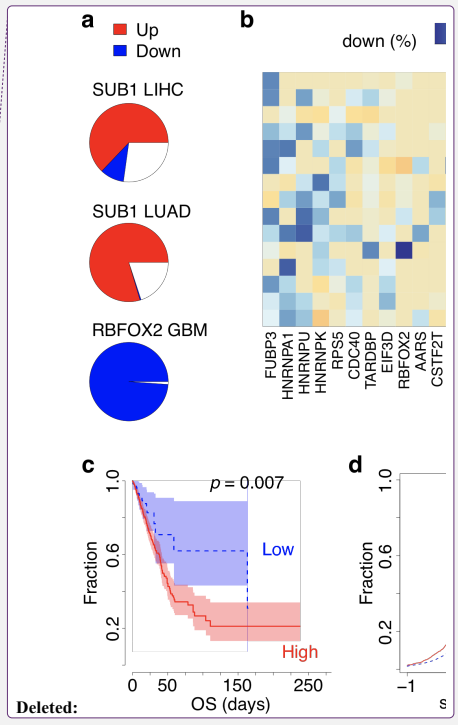


... [64]

Deleted: From

... [65]



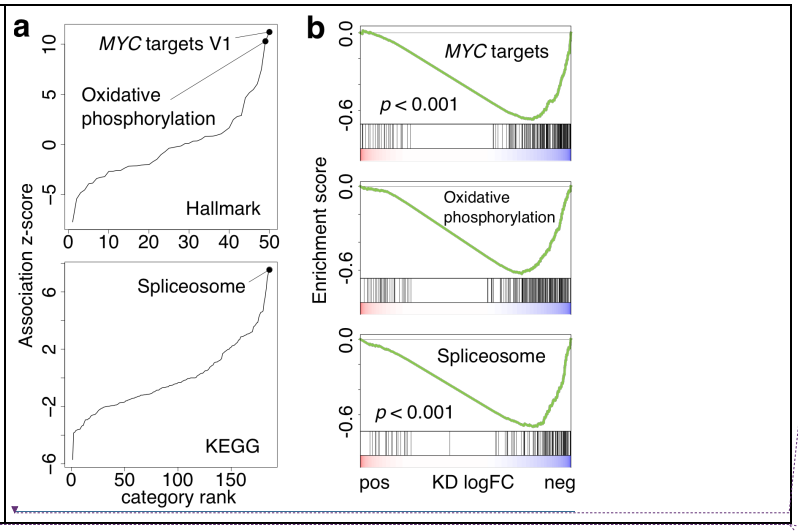


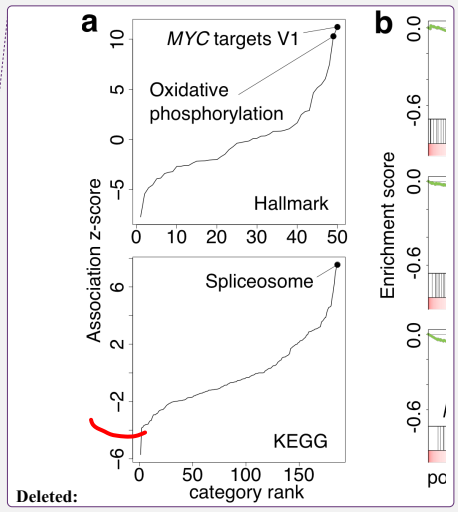
Deleted:

Excerpt 3.8-D (in Suppl.)

Among genes whose 3'UTR regions have *SUB1* eCLIP sites, we observed significant enrichment of functional categories including *MYC* targets and spliceosome. *MYC* activation induces an increase in total precursor messenger RNA synthesis, which increases the burden on the core spliceosome to process pre-mRNA [1]. Also, *MYC* activation can stimulate oxidative phosphorylation, which fulfills the bio-energetic demands of cancer cells [2]. These results together indicate that *SUB1* may stabilize the *MYC* target genes and pathways to promote the malignant growth of cancer cells.

Deleted: 4 From ... [66]





Deleted:

Deleted: Excerpt 5 From Feng's validations ... [67]

Deleted: Excerpt 5 From Feng's validations ... [68]

Deleted: 11

<ID>REF3.9 – Logic gates

<TYPE>$$$Network
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	Are circuit gates necessary for Fig 3B? There are OR, AND and NOT gates used. For Figure 3C(i), **what is the meaning of the values between the green and yellow dots (MYC and *)?** The figure legends are not explaining the figure very well and many details are omitted.
Author Response	• We have re-drawn the figure to make it clearer. • The circuit gates represent how MYC and NRF1 work together. • The value of green anad yellow means the number of genes under different situations. Specifically, <-113-> means in our network there are 113 genes regulate MYC and at the same time, are the target of MYC. <-1487- means there are 1487 genes regulating MYC, and -2135-> means there are 2135 genes being regulated MYC, but not regulate MYC.

Formatted Table

Formatted: Font:Bold

Formatted: Justified, Don't add space between paragraphs of the same style, Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Deleted: redrawn

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: In the orginal version

Formatted: Font:12 pt

SHUD IN CAPTN

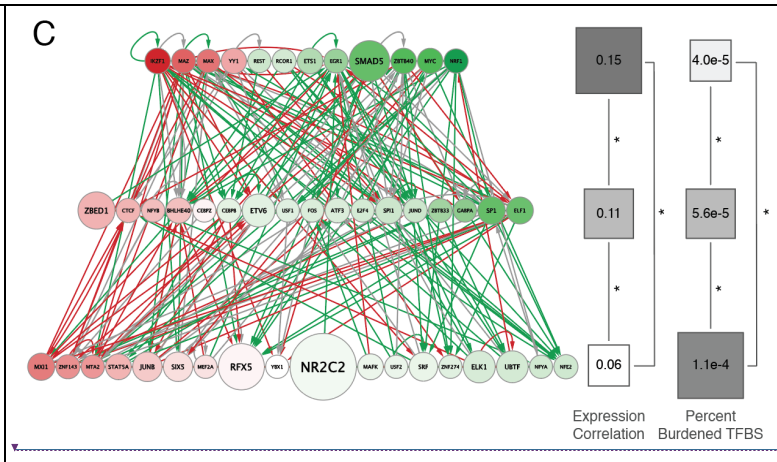
	• Figure legend have been updated
Excerpt 3.9-A (updated Fig and Legend)	Wait for Figure 2

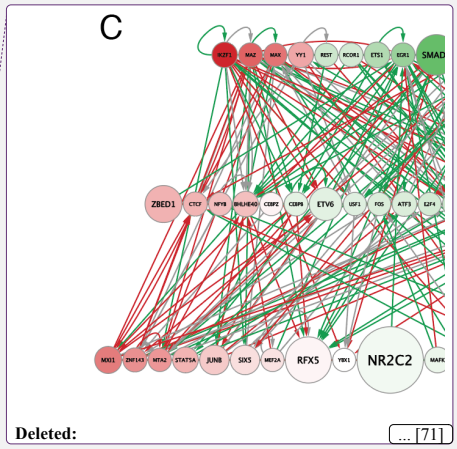
<ID>REF3.10 – Network hierarchy

<TYPE>$$$Hierarchy
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%99DONE

Referee Comment	For Figure 4, what does the star symbol (*) mean in the legend? Did the authors use a different grey color to show the connection between TFs? I'm not able to read the grey gradient for the edges.
Author Response	We thank referee for pointing out this issue. First, we have updated figure legend to make it clear what the star symbol (*) mean in the revised manuscript. In summary, we have performed Wilcoxon rank sum test to show the significance of regulators placed in different network hierarchy. Second, we also improved the presentation of the network hierarchy figure. For the cell type specific network, we highlighted gained and lost edges with green and red arrows, added labels colors to represent gainers and losers.
Excerpt 3.10-A (updated Fig)	**Figure 4. Regulatory network rewiring and hierarchies.** … … (C) Cell-type specific network using K562 and GM12878 … … If a p-value is less than 0.05, it is flagged with one star (*). If a p-value is less than 0.01, it is flagged with two stars (**). If a p-value is less than 0.001, it is flagged with three stars (***).

Deleted: From . ... [69]
Deleted: 12
Formatted Table
Deleted: we've
Deleted: to
Deleted: we've
Deleted: See excerpt for details.
Deleted: From . ... [70]





Deleted: ... [71]

<ID>REF3.11 – Network rewiring

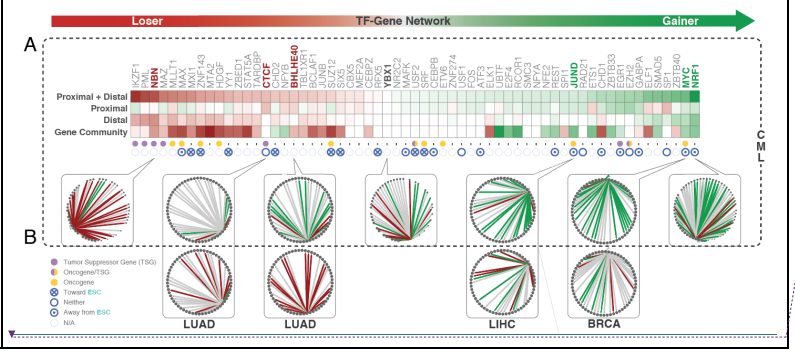
Deleted: 13

<TYPE>$$$Network
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%99DONE

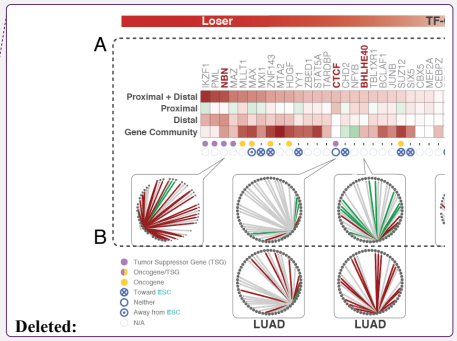
Referee Comment	For Figure 5B, what does the vertexes and edges represent? I guess they represent genes and their network connection, respectively? How did you select the genes and why are some of them "thick" while others "thin"?
Author Response	We thank referee for pointing this issue out. We have re-drawn the figure to make it clearer. Vertices represent genes (regulators) and edges represent regulatory linkage between TFs and genes. We have used colors and thickness to show regulatory rewiring between cell types. Thick edges are shown to highlight rewiring events while thin edges mean gene linkages are retained between cell types. We have redrawn the figure to make this clearer.

Formatted Table

Formatted: Font:12 pt
Deleted: In
Formatted: Font:12 pt
Deleted: rewiring analysis, vertices
Formatted: Font:12 pt

Excerpt From Revised Manuscript	**Figure 4. Regulatory network rewiring and hierarchies.** 

Deleted:



Referee #4 (Remarks to the Author):

<ID>REF4.1 – Strengths of the Paper

<TYPE>$$$NoveltyPos
<ASSIGN>@@@MG,@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	I fully acknowledge that the manuscript proposes a very important approach from detecting the mutations that are most relevant for each specific type of cancer, integrating epigenome data, transcription factor binding, chromatin looping to focus on key regions: ultimately, this work demonstrates the importance of functional data beyond the primary sequence of the genome. Other important aspects include the comprehensiveness and breadth of the data, the analysis and ultimately the whole integrated approach, which goes beyond commonly seen genomics analysis. However the manuscript is not trivial to read and digest in the first round: anyway I believe that the message, including the importance of the integration multiple types of data, is very important.
Author Response	We thank the referee for the positive comments.

Formatted Table

<ID>REF4.2 – Changing the presentation of the supplement

<TYPE>$$$Text,$$$Presentation
<ASSIGN>@@@DC,@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	Yet, efforts to make the manuscript more readable will be quite important. For instance, I could understand several sections of the manuscript after reading carefully the not so short supplementary part. The strategy of sample selection was easier to understand after seeing the first figure of the supplementary information, as well as fig S1-3 regarding the number of normal vs cancer cell lines. I'm not sure what the space limitation for this manuscript will be, but clarity should be an important component of a Nature paper.
Author Response	We thank the referee for pointing out that it is sometimes hard to see the full documentation of our methods in the main text -- one has to look at the

Formatted Table

	extensive supplements. We have tried our best to re-organize our analysis to better illustrate the value of the ENCODE data and our annotations. The very large scale of the supplement is typical for large genomic paper. We, in fact, have been actively discussing with Nature Publishing and other companions about the supplement with regard to the main text. We have attempted to put important contents in the supplement and to structure it very carefully. We are prepared to work very hard to make the structure of the supplement understandable. We have tried to revise it to make these clearer and also to move more into the main text, though we think given the current main text limitations of a typical paper in Nature and the scale of the results in the data in this paper, it is not easy to put everything into the main text. We are preparing to work constructively with the referees and the others to make this clear.

Deleted: ... [72]

<ID>REF4.3 – Trimming and editing parts of the manuscript

<TYPE>$$$Text,$$$Presentation
<ASSIGN>@@@DC,@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

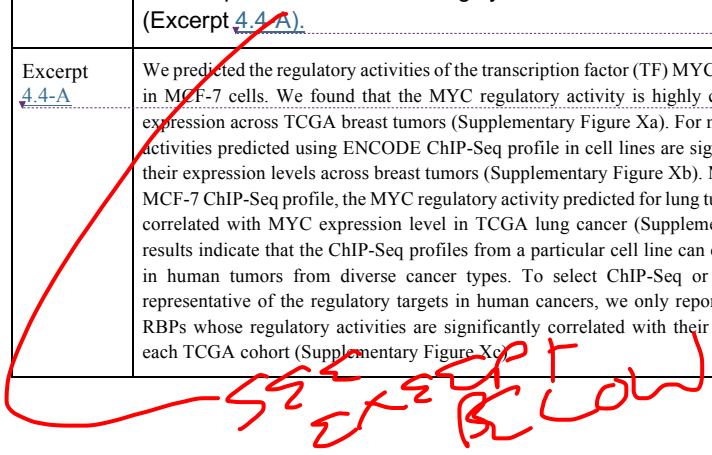
Referee Comment	1) **The manuscript is quite complex and efforts are needed to improve clarity.** Some of the text can seem to be somehow redundant or not needed (for instance, general comments about the ENCODE project; or the Step-Wise prioritization scheme (page7; other parts at page 7, for instance).
Author Response	As the reviewer has suggested, we have revised these sections in our revised manuscript for length and clarity.

Formatted Table
Formatted: Font:Bold
Deleted: ... [73]

<ID>REF4.4 – Validate the cell line results using tissue data

<TYPE>$$$CellLine,$$$Validation
<ASSIGN>@@@JZ,@@@DL,@@@Peng,@@@DC
<PLAN>
<STATUS>%%%85DONE

Referee Comment	One of the limitations of the analysis are the cells that are central in the ENCODE, that are immortalized, including cancer cells and "normal" immortalized counterparts. Most of these cell lines have been kept in culture for decades and further selected for cell growth very extensively. Many of the cell lines may have/have accumulated further mutation and rearrangements, if compared to what cancer cells are at the moment that they leave the human body. The authors accurately acknowledge, in the discussion, stating that it is difficult to match cancer cells with the right normal counterpart; it may also be even more difficult to define what are they really ... **It would be appropriate to (computationally) verify at least a small part of the data in other systems**, taking from published studies including normal cells control and primary cancers.
Author Response	We agree that it is important to verify the discoveries from cell lines in primary cancers. We have added such comparisons in our revised version. Specifically, we added a supplementary section to show that TF regulatory activities predicted from ENCODE TF regulatory networks compared with their expression levels are highly correlated in breast and lung cancer (Excerpt 4.4-A).
Excerpt 4.4-A	We predicted the regulatory activities of the transcription factor (TF) MYC using a ChIP-Seq profile in MCF-7 cells. We found that the MYC regulatory activity is highly correlated with the MYC expression across TCGA breast tumors (Supplementary Figure Xa). For most TFs, their regulatory activities predicted using ENCODE ChIP-Seq profile in cell lines are significantly correlated with their expression levels across breast tumors (Supplementary Figure Xb). Moreover, using the same MCF-7 ChIP-Seq profile, the MYC regulatory activity predicted for lung tumors is also significantly correlated with MYC expression level in TCGA lung cancer (Supplementary Figure Xa). These results indicate that the ChIP-Seq profiles from a particular cell line can capture regulatory targets in human tumors from diverse cancer types. To select ChIP-Seq or eCLIP profiles that are representative of the regulatory targets in human cancers, we only reported the results of TFs or RBPs whose regulatory activities are significantly correlated with their gene expression level in each TCGA cohort (Supplementary Figure Xc).



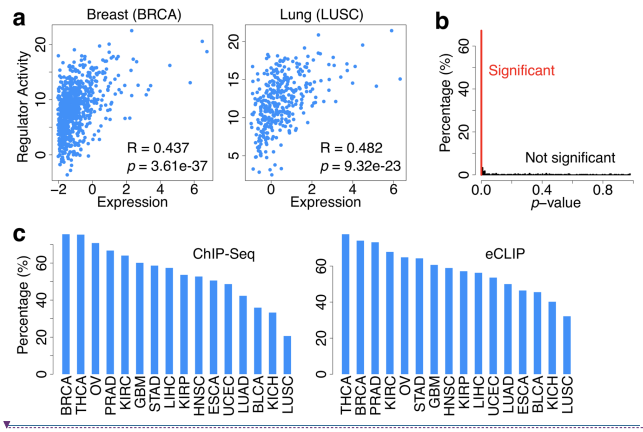
Formatted Table

Formatted: Font:Bold, Not Italic, No underline

Deleted: analysis to address this question, includ ... [74]

Deleted: 1 below). ... [75]

Deleted: From ... [76]



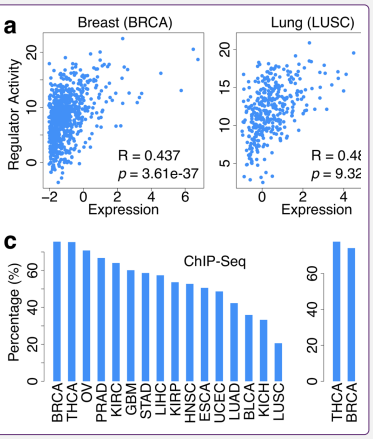
Supplementary Figure X. The clinical relevance of ENCODE cell line data in human primary tumors.

(a) The correlation between *MYC* expression level and regulatory activity across tumors. The MYC regulatory activity in each tumor was predicted using the ChIP-Seq profile in the MCF-7 cell line. The Pearson correlation between MYC gene expression levels and regulatory activity were computed across tumors in each cancer type. The statistical significance of the Pearson correlation was tested by the two-sided student t-test. BRCA: breast carcinoma. LUSC: lung squamous cell carcinoma.

(b) The distribution of correlation *p*-values in TCGA breast cancer. For each TF, we tested the statistical significance of Pearson correlation between TF expression levels and regulatory activities predicted across tumors through two-sided student t tests as for panel a). For the TCGA breast cancer cohort, most *p*-values are very significant with few non-significant values.

The fraction of regulators with statistically significant correlations in different cancer types for ChIP-Seq and eCLIP networks. In each TCGA cancer type, we computed the correlations between regulator expression levels and regulatory activities across tumors for all regulators (TFs, or RBPs). We selected regulators with statistically significant correlations through a two-sided student t test (FDR < 0.05).

Deleted:



<ID>REF4.5 – Loss of diversity in cancer cells

<TYPE>$$$CellLine
<ASSIGN>@@@JZ,@@@DL
<PLAN>&&&MORE
<STATUS>%%%95DONE

Referee Comment	I have seen data in other studies, showing that many of cancer cell transcriptome are quite similar to each other, if compared to

Formatted Table

	initial or primary cells, showing that **in particular cancer cells lose diversity**
Author Response	We agree with the referee that many cancer transcriptomes de-differentiate and lose diversity during tumorigenesis. We aimed to highlight this point using deep integration of the ENCODE resources. In relation to this and other points, we have expanded our analysis on stemness in the revised manuscript and made a new figure, which is shown in the response to the Excerpt 4.6-A.

Formatted: Font:Bold

Deleted: point REF4

<ID>REF4.6 – Relationship of H1 to other stem cells

<TYPE>$$$Stemness$$$Calc
<ASSIGN>@@@DL,@@@PE,@@@DC
<PLAN>&&&AgreeFix,&&&MORE
<STATUS>%%%75DONE

Referee Comment	3) One of the conclusions, deriving from the analysis of H1-hESC is the some cancer are "moving away from stemness". However, while it is true that the cancer cells pattern diverge from the H1 cells, H1 is a human embryonic stem cells: although interesting, **H1 may not necessarily be the best cells to compare with tumor phenotype.** Authors should discuss/defend of further elaborate on this approach. I believe that a key analysis should be done against **other stem cells** (like tissutal stem cells, etc.).
Author Response	We thank the referee for this comment, which we found insightful. In fact, one of the virtues of ENCODE is the large number of different tissues and cell types available. Thus, we have responded to the referee's comment and actually expanded on this point by showing all the cancer types in relation to a number of stem cells available within ENCODE. We have now included an additional figure. We initially focused on H1 because it is one of the top-tier ENCODE cell lines with broadest cell type coverage. In developing this figure, we were able to use the ENCODE knockdown data as a validation to observe overall pattern from the effect of oncogenes. Overall, we think this was a great

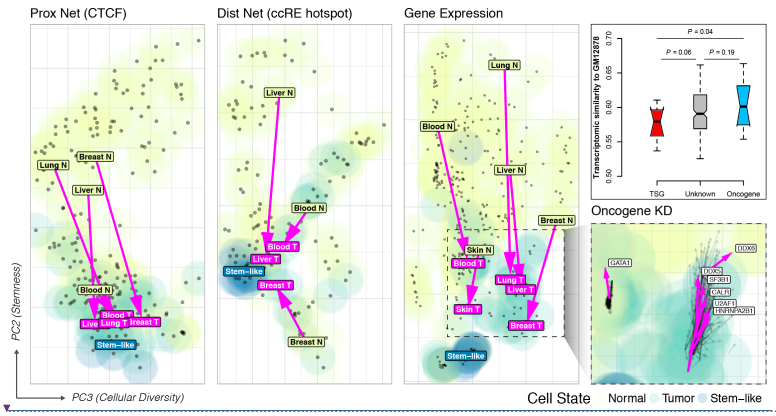
Formatted Table

Formatted: Font:Bold, No underline

Formatted: Font:Bold, No underline

Deleted: Furthermore, in

Moved (insertion) [5]

	comment, and we thank the referee very much for it. See excerpt for more details.
Excerpt 4.6-A	We have highlighted the de-differentiation of cancerous cell types into stem-like cell types using proximal regulatory network (CTCF ChIP-seq) and distal regulatory network (ccRE ELS hotspots), and we show that our findings are in agreement with previous findings using gene expression (RNA-seq). We performed PCA analysis (reference component analysis (RCA) for gene expression; {\cite: Li, Huipeng, et al. "Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors." Nature Genetics 49.5 (2017): 708.}) using uniformly processed poly A long RNA-seq, CTCF ChIP-seq, and candidate cis-regulatory element from ENCODE encyclopedia. We consistently found that cancer cells tend to cluster together, closer to the stem-like cell cluster, in contrast to their normal counterparts. Figure 5. PCA (RCA) of regulatory networks and gene expression. 
Excerpt 4.6-B (in suppl.)	We find that stem-like cells in ENCODE, including top-tier H1-hESC cell line, form a cluster and their regulatory patterns and expression profiles are distinct from differentiated normal cell types and tissues. This highlights that pluripotent embryonic stem cells like H1-hESC maybe not far distinct from other stem-like cells and cell-of-origin. For the proximal network, we built a simple regulatory network based on CTCF binding peaks. Our preliminary network consists of 14,536 TSS (2.5kb up/downstream) with CTCF peaks across 207 cell types. We filtered for recurrent CTCF binding in at least 20 different cell types to subset the network, and finally, we used 9,506 CTCF hotspots near TSS across 207 cell types to perform PCA analysis. For distal network, we built 990,079 merged ccRE ELS sites across 609 ccRE annotation. We used two filters to select recurrent distal element. First, we selected ccRE ELS sites that are 100kb away from TSS, and second, we selected ccRE ELS sites seen in more than 20 different cell types. We finally used 13,497 ccRE ELS hotspots across 134 cell types and performed PCA analysis. For the gene expression, we simply used replicate-merged FPKM of 20,345 protein coding genes across 329 cell types to run RCA (reference component analysis).

Moved up [5]: We initially focused on H1 because it is one of the top-tier ENCODE cell lines with broadest cell type coverage.

Deleted: ... [77]

Deleted:

Deleted: 1 From …

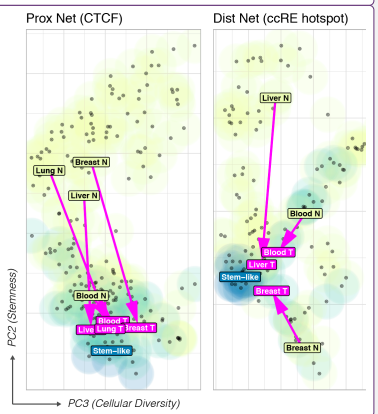
Deleted: …

Deleted: ...

Deleted: We have not used PC1, instead used PC2 and PC3 to highlight, because PC1 may contain potential batch effect given we are making a comparison of data generated from different labs. Removing PC1 removed outliers and provided cleaner separation of clusters. We have chosen CTCF ChIP-seq since it provided broadest coverage of cell types in ENCODE. ...

Deleted: ...

Deleted:



Deleted:

Deleted: 2 From … [78]

Deleted: … [79]

<ID>REF4.7 – Fixes for Figure 1

<TYPE>$$$Presentation,$$$Later
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

Referee Comment	4) I have **difficulties to fully understand Fig.1**, in particular the patient cohort (PC) at the bottom of the "depth approach" (just above the green box of cell –specific analysis). The two rows are at the bottom of the columns report mutation and expression, but they belong to the columns of the cell lines (K562, HepG2, etc). I just simply do not understand that part of the figure, in particular the relation between cell lines and the patient cohort (the figure legend does not help, and also supplementary material did not help).
Author Response	In the revised manuscript, we have modified the figure 1 to make it more clear. We understand that numbers at the mutation and expression rows can be misleading, so we have moved cohort-based data matrix out of cell-type data matrix to the supplement. In addition, we have attempted to emphasize the value of ENCODEC as a resource in this overview schematic.
Excerpt 4.7-A (updated Fig. 1)	(to be continued for fig 1)

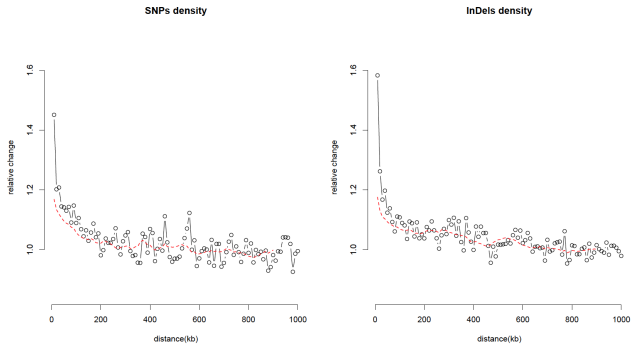
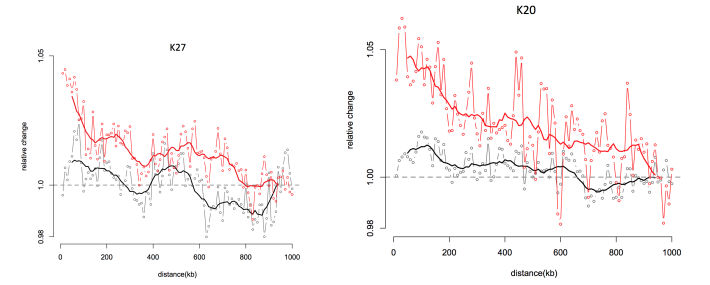
Formatted: Font:Bold
Formatted Table
Deleted: From ... [80]

<ID>REF4.8 – SVs affecting BMRs & Network

<TYPE>$$$BMR,$$$Network,$$$Calc
<ASSIGN>@@@DL,@@@XK, @@@TG,@@@STL
<PLAN>&&&AgreeFix,&&&MORE
<STATUS>%%%30DONE

Referee Comment	5) The analysis assumes that genomes of all the cells discussed are essentially the same. However, for many of the cancer genomes, there have been rearrangements, often dramatic like Chromothripsis. How is this affecting the BMR and the linking of non-coding elements to

Deleted: [JZ2DL, XM, TG, STL: would you please help to fill in the stuff?]
Formatted Table

	the target genes? How many of the cells analyzed were dramatically rearranged?
Author Response	The referee asked us to comment on the relationship of structural variants, BMR, and network wiring. We think these are very useful suggestions. We would have benefitted? highlighting SVs more. And in the revision, we have responded to and extended the referee's suggested in multiple respects as listed below.
Excerpt 4.8-A (call SNV and SV in top-tier cell lines, in suppl.)	We have called SV and SNVs from multiple ENCODE cell lines by integrating various assays as shown in the following table. JZ2DL: add Feng's table
Excerpt 4.8-B (SNV density around SVs, in suppl.)	We compared the SNV/InDel density near the SV boundaries in strictly matched ENCODE cell lines and found that there are noticeably elevated SNV/InDel rates around SVs. 
Excerpt 4.8-C (SV vs. histone modification)	We extracted SV events in K562 and compared them with several histone modification marks. We found clear patterns as below. [JZ2STL: please add more text and the exact procedure below] 

Deleted: , including (JZ2DL: please fill in xxx)
Excerpt 1 From ... [81]

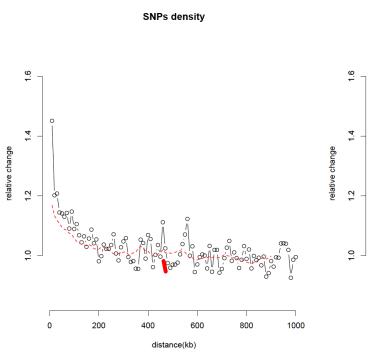
Deleted: Excerpt 1 From ... [82]

Deleted: variousassays

Deleted: JZ2JZ

Formatted: Highlight

Deleted: 2 From ... [83]

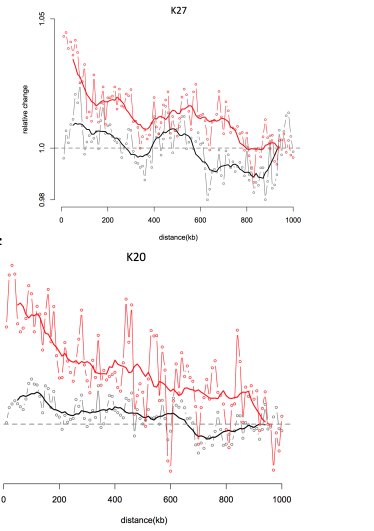


Deleted:

Deleted: 3 From ... [84]

Moved down [6]: [JZ2STL: please add more text and the exact procedure below]

Moved (insertion) [6]

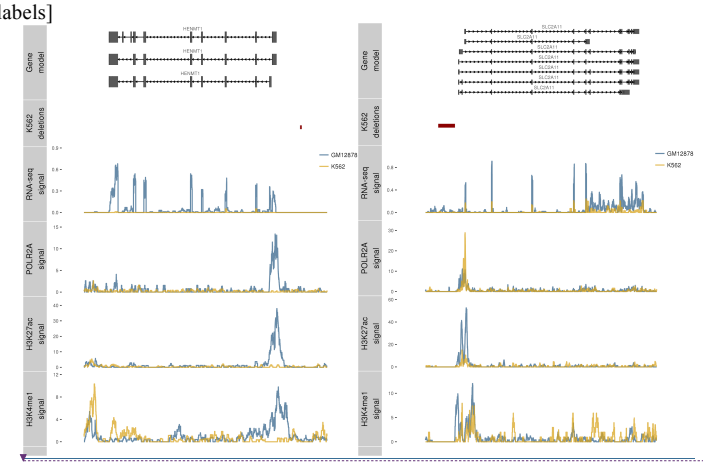


Deleted:

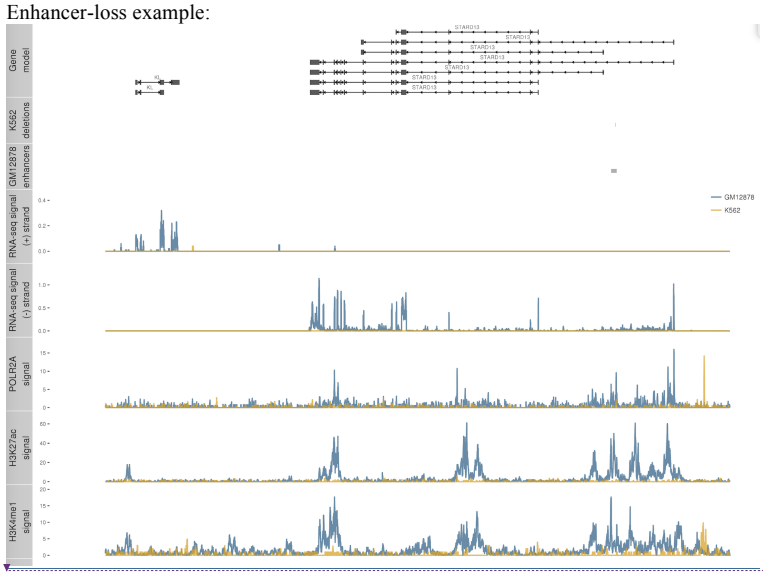
Excerpt 4.8-D (SV vs. gene expressions)

We have shown in the follow figure several examples of SVs near promoter regions that may affect gene expression.

[JZ2TG: please add more text to describe your procedure here. Also please add x axis labels]



Enhancer-loss example:

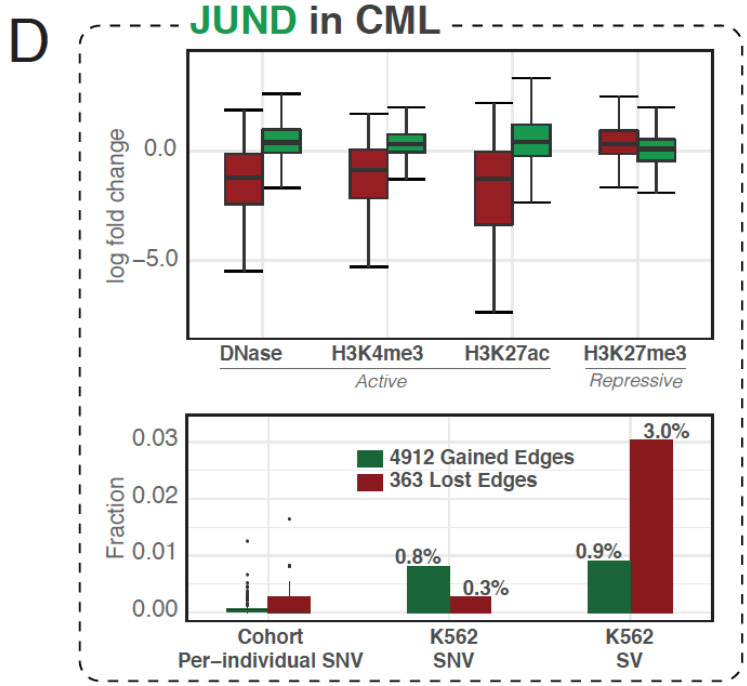


Deleted: figureseveral

Deleted: From ... [85]

Deleted: ... [86]

Deleted:

Excerpt 4.8-E (SV and rewiring)	Figure 4. Rewiring panel D (JZ2DL: pls describe what you have done here) We examined the fraction of rewired edges affected by SNVs and SVs. Larger fraction of gained edges were affected by SNVs while larger fraction of lost edges were affected by SVs. 
Excerpt 4.8-F (SV and oncogene activation)	Ask Feng to write a text

Deleted: 5

Deleted: sub-

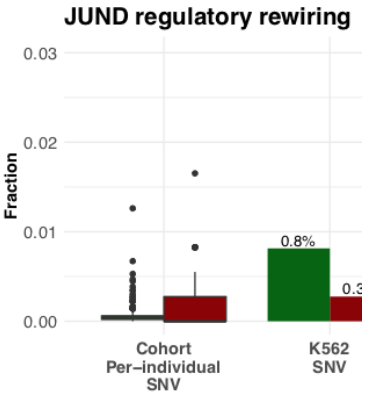
Deleted: update

Deleted: 5 From ... [87]

Comment [12]: done

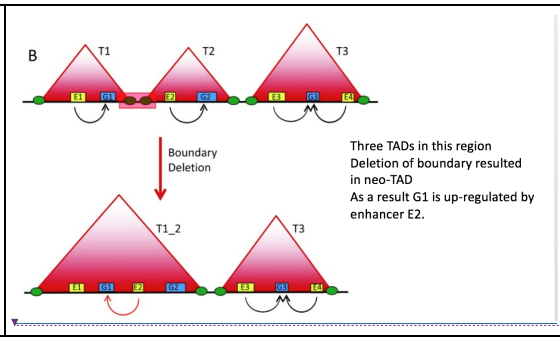
Formatted: Highlight

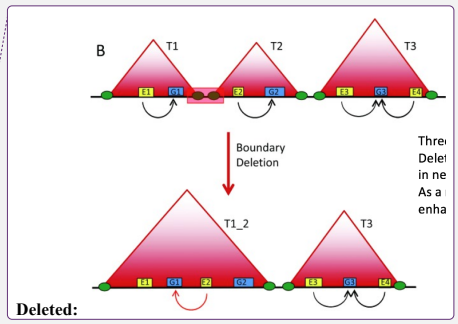
Formatted: Highlight



Deleted:

Deleted: 6 From ... [88]

Deleted: 

<ID>REF4.9 – Aspects of heterogeneity related to cell lines

<TYPE>$$$CellLine,$$$Text
<ASSIGN>@@@WM,@@@JZ,@@@MRS
<PLAN>&&&AgreeFix
<STATUS>%%%65DONE

Referee Comment	6) Most cancers are not necessarily represented by a single cell type used to obtain genomics data in this study, but contains numerous types of cells with different mutations, as well as normal cells, infiltrating cells, all in a three dimensional structure, often producing metastatic colonizing other organs. However, this study focuses only on comparisons between cells. These limitations should be better discussed, also to put in perspective future studies on single cells.
Author Response	We thank the referee for bringing this up and we completely agree with the referee that genomic and epigenomic heterogeneity in tumor cells, as well as heterogeneity in the tumor microenvironment (e.g., immune cell infiltrates, hormonal factors, normal cell populations, etc.) are significant factors in tumor growth and development. In our revised manuscript, as suggested we have tried to • Added more discussion in main text about the limitation and how future technique can help (Excerpt 1) • Specifically for the BMR part, clearly point out that most cancers can not be represented by a single cell type and that is exactly why we used multiple genomic features to characterize BMR. ENCODE data

Formatted Table

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

	expanded features by more than a factor of 10 as compared to other related work published recently). • Regarding the rewiring part, better introduce the concept of composite normal and discussed the limitation of current technique
Excerpt 4.9-A (new text aboug single-cell sequencing in discussion)	One limitation of the current ENCODE data is that most of the current release of data is performed over a small number of cells. However, genomic and epigenomic heterogeneity in tumor cells, as well as heterogeneity in the tumor microenvironment (e.g., immune cell infiltrates, hormonal factors, normal cell populations, etc.) are significant factors in tumor growth and development. We believe that the development of single-cell sequencing technologies may capture important tumor biology present and provide new insights in cancer.
Excerpt 4.9-B (Heterogeneity & BMR in main text)	While it is valuable to match cancer to its cell of origin, tumors are highly heterogeneous and there are usually multiple normal cell types are around and inside tumor cells, so a combination of different data sets provide the best overall fit to mutation rate.
Excerpt 4.9-C (Heterogeneity & BMR in Suppl.)	The ENCODE3 rollout dramatically expands the genomic data available for this type of regression by more than a factor of 10 (2069 vs. 169), many of which are from tissue or primary cells. In total there are 2,017 histone ChIP-seq and 51 replication timing Repli-chip and Repli-seq features to predict BMR. We did a PCA of the signals from these features and selected the best combination of 20 PCs for BMR prediction. It is worth pointing out that the majority of our data is from tissue or primary cells. A summary of cell types for these features is given below.

Table S1. Summary of ENCODE histone ChIP-seq data

Cell Type	# histone marks
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

Table S2. Summary of ENCODE3 Replication timing data

Cell Type	Repli-seq	Repli-chip
cell line	101	10
in vitro differentiated cells	0	35

Deleted: Excerpt From ... [89]

Moved down [7]: 169), many of which are from tissue or primary cells.

Deleted: From
In the main text: ... [90]

Deleted: In the main text: ... [91]

Comment [13]: Are we defending not having perfect cell line matches?

It's not clear that using different data sets provides a best overall fit to mutation rate. Perhaps one cell type dominates the tumor mutation rate or is most relevant. It's also not clear that data should be combined i ... [92]

Deleted: ... [93]

Moved down [8]: We did a PCA of the signals from ... [94]

Deleted: fromtissue or primary cells. A summary of c ... [95]

Moved down [9]: Summary of ENCODE histone Ch ... [96]

Moved down [10]: JZ2DL: please add

Formatted: Font:10 pt

Cell Type	# hist mark
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

Deleted: ... [97]

Formatted: Font:Arial, 12 pt

Deleted: the table of replication timing data]

Deleted: From ... [98]

Moved (insertion) [7]

Moved (insertion) [8]

Formatted: Left

Moved (insertion) [9]

Formatted: Font:10 pt

Moved (insertion) [11]

Formatted: Justified

Moved (insertion) [12]

Formatted: Font:10 pt

Formatted: Font:Times New Roman, 10 pt

Formatted Table

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:Times New Roman, 10 pt

primary cell	12	5
stem cell	6	11
induced pluripotent stem cell line	0	2

Table S3. Summary of 51 replication timing features from Repli-chip and Repli-chip

Cell State	Repli-chip/Repli-seq
Pluripotent	8
DE	3
Liver/Pancreas	6
Neural crest/Early mesoderm	7
Late mesoderm	6
NPC	2
Myeloid/Erythroid	5
Lymphoid	5
Cancer	9

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:10 pt

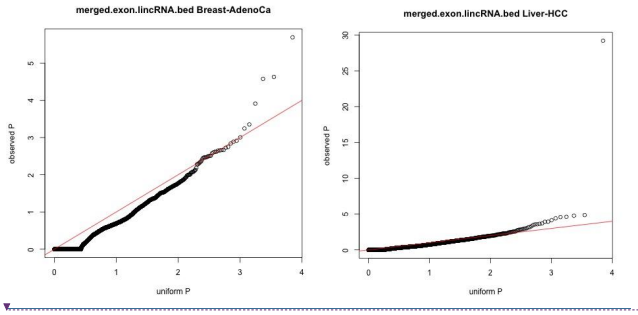
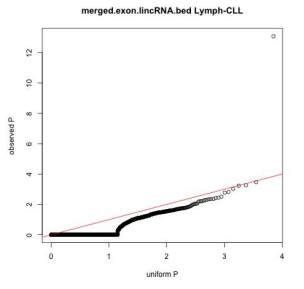
Deleted: One limitation of the current ENCODE data is that most of the current release of data is performed over a number of cells. However, genomic and epigenomic heterogeneity in tumor cells, as well as heterogeneity in the tumor microenvironment (e.g., immune cell infiltrates, hormonal factors, normal cell populations, etc.) are significant factors in tumor growth and development. We believe that the development of single-cell sequencing technologies may capture important tumor biology present and provide new insights in cancer.

<ID>REF4.10 – lncRNAs and BMR

<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%90DONE

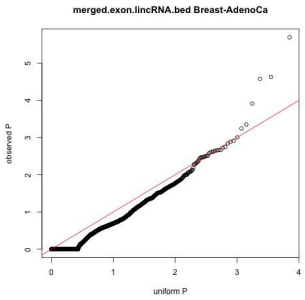
Referee Comment	7) When analyzing the BMR in cancer, did the author estimate the mutation rate in the lncRNAs? Is there any other interesting lesson from the analysis of the non-coding regions and their mutations rate?
Author Response	We thank the referee to point out this. Our BMR model captures the mutation rate over the whole genome.Thus, we are able to calculate the

Formatted Table

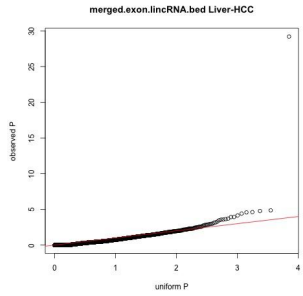
	mutation burden of lincRNAs. We have added results on lincRNAs in our revised supplements.
Excerpt 4.10-A (burden test on lincRNAs)	We also calculated the mutation burden on lincRNAs. We have found well-known cancer associated lncRNAs to be burdened, such NEAT1 in liver cancer, MALAT1 in breast cancer. Results and QQ-plots were given in Supplementary Table X.  

Deleted: (see excerpt below).

Deleted: From ... [99]



Deleted:



... [100]

<ID>REF4.11 – (Minor) updates to figure numbering in supplementary

<TYPE>$$$Minor,$$$Presentation
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

Referee Comment	In the supplementary material, there is room to improve figures (some numbers are too small).

Formatted Table

Author Response	We thank the referee for pointing this out and we have made revisions to the supplementary figures in our revised manuscript to improve interpretability.

<ID>REF4.12 – (Minor) Figure legends

<TYPE>$$$Minor,$$$Presentation
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

Referee Comment	Figure legends. Figure legends are essential but I struggled to understand the figures based on the legends only.
Author Response	We thank the referee for this comment and we have revised our figure legends to improve.

Formatted Table

Referee #5 (Remarks to the Author):

<ID>REF5.0 – Preamble

<TYPE>$$$Text
<ASSIGN>@@@MG,@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

We appreciate the referee's feedback. We found many comments quite valuable. It was particularly useful to receive the authors comments on further power analyses, the false positive rate of rewiring, comparisons with other networks, additional validation using external data, and further exploration of SUB1 biology. As suggested, we have addressed all the comments and significantly expanded our analysis. We have tried to better clarify our main goal and clearly organize our analysis to illustrate the value of the resources in this paper. Specifically, we want to emphasize two points:

1. The goal of this paper and its distinct role in the whole ENCODE package.
We have tried to make clear that this is the only paper in ENCODE3 to provide deep and accurate integrative annotation focusing on several data rich cell types. The breadth and accuracy of our annotation extends far beyond the encyclopedia paper in this regard. We feel that cancer is an excellent application to illustrate certain key aspects of ENCODE data and analysis - particularly the deep and integrative annotations, regulatory potentials of key TF/RBPs, network rewirings, and normal-tumor-stem comparisons. We have tried to clarify that we have developed many new methods in this paper to deeply annotate several cancer associated cell types , including:

- Multi-level compact and accurate enhancer predictions.
- Integrative gene-enhancer linkages.
- Extended gene definitions that incorporate numerous regulatory elements in a gene centric way.
- Universal and tissue-specific regulatory networks built on ChIP-Seq and eCLIP data for xxx TFs and xxx RBPs.
- Matched TF regulatory profiles and their rewiring status.
- Normal-tumor-stem distance quantifications based on expression and network profiles.

We have also tried to illustrate the usefulness of the above resource to prioritize both key regulators and genomic variations (single nucleotide and structural variations) using

Comment [14]: Unsure about the use of the word 'goal' in this context, given that it is a scientific study.

Perhaps 'main results' in substitution.

Deleted: The main encyclopedia paperprovidesannotations for all cell types based on just 4 assays.

Deleted: encylopedia

Deleted: For instance, the new ENCODE3 data used in this paper includes: ... [101]

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Comment [16]: Just a general comment that there are very few acronyms that are defined on first use throughout this supplement. Not sure if this is a problem or not.

various techniques, such as luciferase assays, CRISPR, and knockdowns. We hope that all the above aspects serve as good examples to illustrate the value of our resource to cancer genomics.

2. Regarding the BMR part

With respect to the BMR estimation part in particular, the reviewer noted that there had been many existing publications focusing on applications such as cancer driver detection.

We thank the referee for pointing out a body of related work. As suggested, we have tried to provide better context of previous work in our revised manuscript. We would also like to point out that some references were either published after our initial submission (such as Marticorena et al. 2017) or with a different focus (i.e., other than BMR estimation).

Second, we would also like to emphasize that the main goal of our paper is not to present novel methods of driver discovery, but rather to illustrate that the richness of the ENCODE data can be leveraged to noticeably improve the accuracy of BMR estimation. Hence, we feel it is slightly outside the scope for our ENCODE resource paper to make detailed comparisons with driver gene discovery. In the revised version, we have clearly highlighted the value of ENCODE data in our updated Fig. 1.

Third, we want to point out that the BMR application is just **one out of many** potential ENCODE data applications. Given that most of the comments focussed on the BMR, we assume that a number of other points were valuable (e.g. the networks rewiring, stemness measure, and regulator/SNV/SV prioritization) and based on this we have further emphasized this in the manuscript).

Deleted: the

Deleted: Specifically related toBMR

Deleted: mentioned

Deleted: are

Deleted: prior studies

Deleted: like

Deleted: First, we

Deleted: these related references and we haved cited many of them

Deleted: initial submission (table R2 below).

Deleted: want

Deleted: of the

Comment [17]: Although this is true, and there is some unfairness if we are criticized for not recognizing these studies, it's not necessarily true that the reviewers will recognize this unfairness.

It seems they feel the published studies have similar content to our study, regardless of when they were published.

Deleted: afocus

Deleted: (more details in the following table).

Deleted:

Comment [18]: Again, not sure about the word goal in this context.

Suggest perhaps 'main result' instead.

Deleted: want

Deleted: the BMR part in

Deleted: make

Deleted: discoveries

Deleted: how

Deleted: , as

Deleted: attempted to showin our

Deleted: 2.

Deleted: thatBMR estimation

Formatted: Font:Bold, Underline

Deleted: ofmany

Deleted: of ENCODE data. Even for the variant investigation part alone

Deleted: also have germline

Deleted: SV analysis in

Deleted: paper. There are many other ENCODE applications, such as regulatory activity, rewiring, and stemness, which are also key to investigate in cancer genomics.

Reference	Initial	Revised	Main point	Comments
Lawrence et al, 2013	Cited	Cited	Introduce replication timing and gene expression as covariates for BMR correction	Replication timing in one cell type
Weinhold et al, 2014	Cited	Cited	One of the first WGS driver detection over large scale cohorts.	Local and global binomial model
Araya et al, 2015	No	Cited	Sub-gene resolution burden analysis on regulatory elements	Fixed annotation on all cancer types
Polak et al (2015)	Cited	cited	Use epigenetic features to predict cell of origin from mutation patterns	Use SVM for cell of origin prediction, not specifically for BMR
Martincorena et al (2017)	No (out after our submission)	Cited	Use 169 epigenetic features to predict gene level BMR	No replication timing data is used
Imielinski (2017)	No	Yes	Use ENCODE A549 Histone and DHS signal for BMR correction	Limited data type used from ENCODE
Tomokova et al. (2017)	No	Yes	8 features (5 from ENCODE) for BMR prediction and mutation/indel hotspot discovery	Expand covariate options from ENCODE data
huster-Böckler and Lehner (2012)	Yes	Yes	Relationship of genomic features with somatic and germline mutation profiles	NOT specifically for BMR
Frigola et al. (2017)	No	Yes	Reduced mutation rate in exons due to differential mismatch repair	NOT specifically for BMR
Sabarinathan et al. (2016)	No	Yes	Nucleotide excision repair is impaired by binding of transcription factors to DNA	NOT specifically for BMR
Morganella et al. (2016)	No	Yes	Different mutation exhibit distinct relationships with genomic features	NOT specifically for BMR
Supek and Lehner (2015)	No	Yes	Differential DNA mismatch repair underlies mutation rate variation across the human genome.	NOT specifically for BMR

Deleted:

Reference	Initial	Revised
Lawrence et al, 2013	Cited	Cited
Weinhold et al, 2014	Cited	Cited
Araya et al, 2015	No	Cited
Polak et al (2015)	Cited	cited
Martincorena et al (2017)	No (out after our submission)	Cited
Imielinski (2017)	No	Yes
Tomokova et al. (2017)	No	Yes
huster-Böckler and Lehner (2012)	Yes	Yes
Frigola et al. (2017)	No	Yes
Sabarinathan et al. (2016)	No	Yes
Morganella et al. (2016)	No	Yes
Supek and Lehner (2015)	No	Yes

<ID>REF5.1 – Positive comment of the paper

<TYPE>$$$Text
<ASSIGN>@@@MG,@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	the resources provided in this manuscript are potentially interesting for the cancer genomics community and comprise an extensive body of work

Formatted Table

Author Response	We thank the referee for the positive comment.

<ID>REF5.2 – BMR: novelty compared to previous work

<TYPE>$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%85DONE

Referee Comment	1. The manuscript does not clearly state **innovation and novelty over previously published data and methods.** Several published studies have used epigenomic data types, including replication time and histone modifications from ENCODE and other sources, to model background mutational background density and define genomic elements of interest. The use of the Negative Binomial/gamma-Poisson distributions to model mutational background in cancer has also been published (Imielinski et al 2016; Martincorena et al, 2017).
Author Response	We thank the reviewer for identifying relevant references. In the revised manuscript, we have tried to provide a better context of related work. We have also tried to make it clear that BMR accuracy can be improved by using ENCODE3 data. Negative binomial regression is a standard statistical technique that serves this goal. We have made the following changes to attempt to fully address the reviewer's comments. [JZ2MG: this is a key question they are looking for, so I prefer to summerize it in the following bullet points. Other questions, I can put them into Excerpt 5.2-A (about xxx) for a more concise doc. Pls comment] • A new supplementary table to summarize our 2069 features (vs. 169 in Martincorena et al., 2017) (Excerpt 5.2-A) • We added several references, and tried to provide a better context for previous work (Excerpt 5.2-B). • We have showed how more features with careful feature selection can improve BMR estimation (Excerpt 5.2-C). • We have stated clearly in the main text: more data are helpful, and we have added discussions about the motivation for this - a single matched cell line is not enough due the heterogeneous nature of a tumor (Excerpt 5.2-D).

Formatted: Font:Bold
Formatted Table
Deleted: bacdkground
Moved (insertion) [13]
Deleted: .
Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"
Deleted: 1) This is the reason why we did not directly use these approaches (Imielinski et al 2016; Martincorena et al, 2017).
Deleted: 3
Deleted: about our goal clearly in the main text
Deleted: is
Deleted: 4
Deleted: .
Moved up [13]: We thank the reviewer for identifying relevant references.
Deleted: In the revised manuscript, we have tried to make it clear that our goal in this section is to demonstrate the value of the data - the ENCODE3 rollout dramatically expands the number of features by more than a factor of 10. Negative binomial regression is a standard statistical technique that serves our goal.In the revised manuscript we clearly stated that we are not claiming to be the first to apply it to BMR estimation. In summary,

Excerpt 5.2-A (more features in ENCODE3, in Suppl)

Table S1. Summary of ENCODE3 histone ChIP-seq data

Cell Type	Histone ChIP-seq
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

Table S2. Summary of ENCODE3 Replication timing data

Cell Type	Repli-seq	Repli-chip
cell line	101	10
in vitro differentiated cells	0	35
primary cell	12	5
stem cell	6	11
induced pluripotent stem cell line	0	2

Excerpt 5.2-B (better context of previous work)

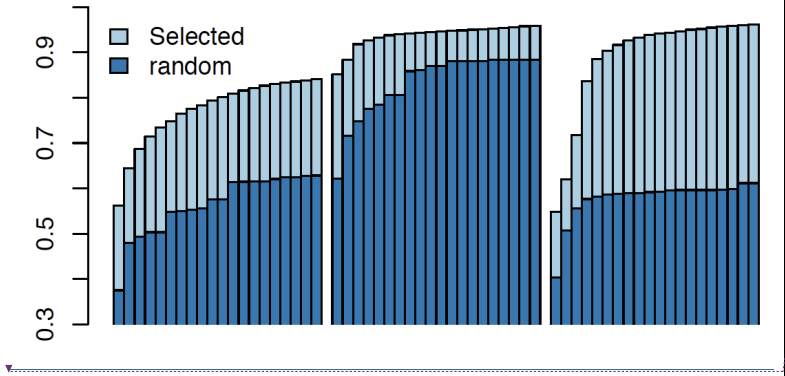
Many methods have incorporated effects from multiple genomic features by techniques such as negative binomial regression and poisson regression.

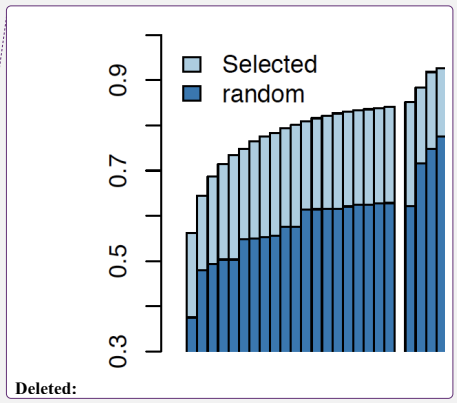
Excerpt 5.2-C (updated main text and Fig.)

The 2,017 uniformly processed histone modification signal tracks and 51 replication timing data may serve as a resource to significantly improve BMR estimation accuracy.

We also found that BMR estimation can be improved dramatically by selecting an appropriate combination of multiple features from ENCODE.

Deleted: 1 From ... [102]
Deleted: Seq
Formatted Table
Deleted: [JZ2DL: pls make such table and put it here] DL: done JZ: to disc on Tuesday
Formatted Table
Deleted: From ... [103]
Deleted: 3 From ... [104]
Formatted: Font:10 pt
Formatted: Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)
Deleted: 2017
Formatted: Font:10 pt
Deleted: 52
Formatted: Font:10 pt

	
Excerpt 5.2-D (more text in discussion)	Recent work has focused on the effect of cell-of-origin on tumor attributes such as mutational process and tumor classifications. However, to accurately define tumor cell-of-origin is sometimes challenging. For example, even different subtypes of tumor from the same organ may originate from different cell types. The richness of ENCODE data provides a larger pool from which to draw the most representative cell of origin.



Deleted:

Deleted: 4 From ... [105]

<ID>REF5.3 – BMR: TCGA benchmark

<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ,@@@WM
<PLAN>&&&MORE
<STATUS>%%%40DONE,%%%CalcDONE

Referee Comment	Throughout, the main manuscript lacks data and statistics supporting the claims made. For example, the performance of tissue-specific background mutation models applied to TCGA data needs to be evaluated against known results and benchmarks from TCGA. It seems that some of these are presented in the extensive supplement and should be moved to the main manuscript.
Author Response	We thank the referee for this comment. As suggested, we have added detailed explanations for every claim of significance by moving a lot of results from the supplement to the main text. Specifically for the BMR part, we fully agree with the referee that it is useful to compare our BMR to established benchmarks. In our revised manuscript,

Deleted: 2.

Formatted Table

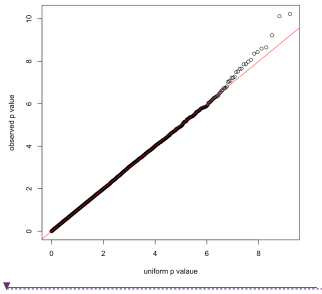
Deleted: and

and tried to benchmark our BMR to other datasets as suggested. We are aware of community efforts and are very involved with the PCAWG effort to do whole genome cancer analysis. One of our authors is the co-leader of the non-coding driver group. In fact, in PCAWG, which is a hybrid of TCGA and ICGC, has not developed specific BMR benchmark. Instead, what they have done is to develop several randomization schemes accepted by multiple groups. Hence, we tried to compare our estimated BMR with such randomizations. Please note that the TCGA Pancan paper is not appropriate here since it is the whole exome and we focus on noncoding.

Please also note that this work is comparing to accepted PCAWG benchmarks, which are not fully published yet, so we only include them in this response. If these papers come out before the ENCODE package, we can certainly move sections of this response to the text of the paper.

1. Using a permuted breast cancer dataset, we performed BMR estimation and calculated somatic mutation burden on the CDS regions of ~20k protein coding regions. We found no gene burdening in this randomized dataset (QQ plot given below).

Figure R 2. QQ plot of observed vs. uniform p values from permuted breast cancer data set. Diagonal shown in red.



2. We downsampled the simulated dataset. We used half of the data for training and compared the rest with our predictions in the promoter regions. The reason why we picked this particular comparison is because most of other published TCGA benchmarks interrogated protein coding regions, where the relative rates of synonymous and nonsynonymous mutations can be used to calibrate BMRs. This particular calibration is not possible in noncoding regions.

Deleted: we have benchmarked

Deleted: data sets

Deleted: . ... [106]

Deleted: annotation

Deleted: any explicit

Deleted: , which are supposed to measure the BMR rate to calibrate driver detection.

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

Deleted: are

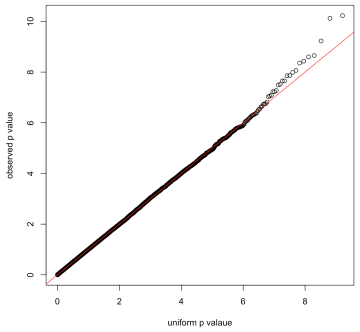
Deleted: including

Formatted: Font:Arial, Pattern: Clear

Deleted: data set

Deleted: plotgiven

Deleted: cancerdata



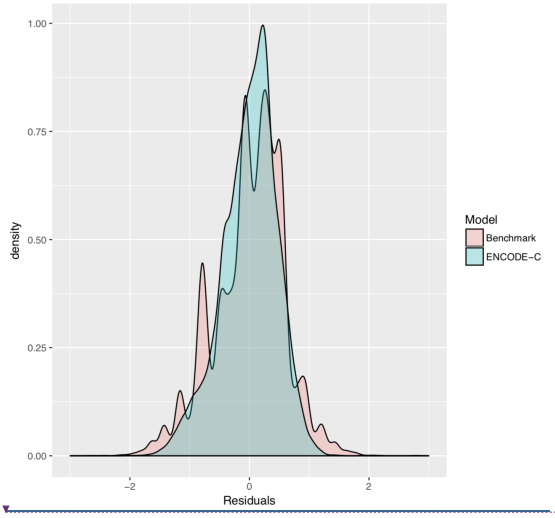
Deleted:

Deleted: only

Deleted: innoncoding

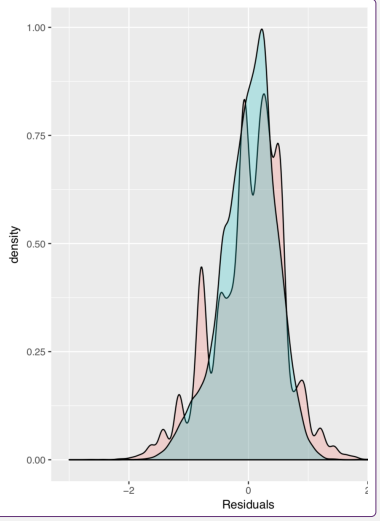
Specifically, we split the PCAWG Liver-HCC somatic SNV set equally into training and testing sets. We applied the Sanger permutation approach used in PCAWG on the training set and used this to predict mutation rates for each of 14,000 promoters, and calculated the residuals between these predictions and the withheld testing data. Similarly, we calculated predicted mutation rates for those same promoters using the ENCODEC model for liver tissue, and calculated the residuals of these predictions from the testing set promoter mutation rates. Overall, the residuals from the ENCODEC predictions are comparable to the PCAWG-derived predictions.

Figure R X. Down sampling of PCAWG data on promoter regions



Deleted: ENCODE-C

Deleted:



<ID>REF5.4 – Power analysis

<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&MORE
<STATUS>%%%75DONE

JZ2JZ: add more

Deleted: JZ2MG: wait, not yet updated. Equations to come in

Referee Comment	4. How do the new "compact annotations" lead to improved results over traditional annotations?
Author Response	We thank the referee for this feedback, and we certainly agree with the referee. We have updated Fig. 2. In short, we integrated multiple assays to compactify the size of annotation without sacrificing accuracy. In short, previous power analysis assumes that all functional sites are within the test regions, which is not practical in noncoding regions due to the resolution and accuracy of annotations. We assume that by removing non-functional sites in the annotations, we can improve statistical power in somatic burden tests. More details are in the excerpts below. • As suggested, we have largely expanded our somatic burden power discussions under various assumptions. In summary, we have now included: • an entirely new section on power analysis and the effect of test region functional site ratios (Except 5.4-A) • more discussion (in the main text) about the pros and cons of merging test regions (Except 5.4-B) • real examples in the supplement (Except 5.4-C) • a new section of quality metrics of the compact annotations to capture functional sites and rm noise(Except 5.7-A) •
Excerpt 5.4-A (power analysis on compact annotations)	Suppose that we define the following parameters. l_i^n : noise region length for region i l_i^r : noise region length for region i μ_i : BMR in region i λ_i : effect size in risk region i $\rho_i = \frac{l_i^r}{l_i^r + l_i^n}$ Then under the null hypothesis, the probability to observe at least one mutation per patient is $p_{i_0} = 1-(1-\mu_i)^{\frac{l_i^r+l_i^n}{G}}$ Under the alternative hypothesis, $p_{i_1} = 1-(1-\mu_i)^{l_i^r}(1-\lambda_i\mu_i)^{l_i^r}$ We did a simulation by starting from a very noisy test region with pretty low true risk loci percentage. We have showed that by trimming the nosie loci, statistical power can be increased. But after we have removed the noise and start to trim the true functional loci, the statistical power drops quickly.

Formatted Table

Moved down [14]: The power considerations for selecting genomic elements are valuable. "Increased" power of the combined strategy is suggested in the manuscript, yet comparison to prior work is missing.

Deleted: .

Formatted: Font:12 pt

Deleted: recognizing

Formatted: Font:12 pt

Deleted: value ofselecting genomic elements. Following the reviewer's suggestions, in our revised manuscript we

Formatted: Font:12 pt

Deleted: completed a formal

Formatted: Font:12 pt

Deleted: . he most important contribution to power comes from including additional

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: supports the extended gene concept. Secondary and lesser, contributions

Deleted: power come from

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: . The core assumption of our compacting

Formatted: Font:12 pt

Deleted: is that

Formatted: Font:12 pt

Deleted: accurately distinguish the more important

Formatted: Font:12 pt

Deleted: nucleotides from

Formatted: Font:12 pt

Deleted: less important ones through

Formatted: Font:12 pt

Deleted: guidance

Formatted: Font:12 pt

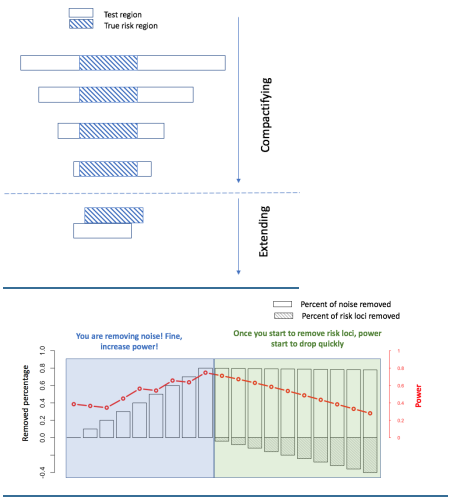
Deleted: many

Formatted: Don't add space between paragraphs of the same style, Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5", Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Font:12 pt

Deleted: characterization assays.

Formatted: Font:12 pt

	
Excerpt 5.4-B (added in disc. sect)	In summary, our claim is that first we provide compact annotations to pick up functional nucleotides and remove noisy ones through the guidance of many functional characterization assays. Then we hope to join the distributed functional sites together to increase statistical power.
Excerpt 5.4-C (more examples in Suppl on compact annotation)	We provided two examples to explain the motivation of our compact and extended gene annotations and why we feel our assumptions for the power analysis is reasonable. 1) Enhancers: Traditionally, enhancers were called as a 1kb peak regions, which admittedly introduced a lot of obviously nonfunctional sites. We believe we can get functional region more accurately by trimming the enhancers down using the exact shapes of many histone marks and further integration with STARR-seq and Hi-C data. 2) TFBS hotspots around the promoter region of WDR74. Instead of testing the conventional up to 2.5K promoter region, we can trim the test set to a core set of the promoter region where many TFs bind, which perfectly correlates with the mutation hotspots (red block) for this well-known driver site (blue line for pan-cancer and green line for liver cancer).

Comment [21]: This does not appear to be an excerpt from the manuscript. It is unclear to me what is an excerpt from the manuscript.

Deleted: Excerpt 1 From
Regarding compact annotation:

Formatted: Font:10 pt

Formatted Table

Deleted: Regarding compact annotation:

Deleted: can

Formatted: Font:10 pt

Deleted: this assumption.

Formatted: Font:10 pt

Deleted:

Formatted: Font:10 pt

Formatted: Font:10 pt

Deleted: may introduce

Comment [22]: Do we actually have some evidence for this? Or is it just a hypothesis? What is the basis for the hypothesis?

Formatted: Font:10 pt

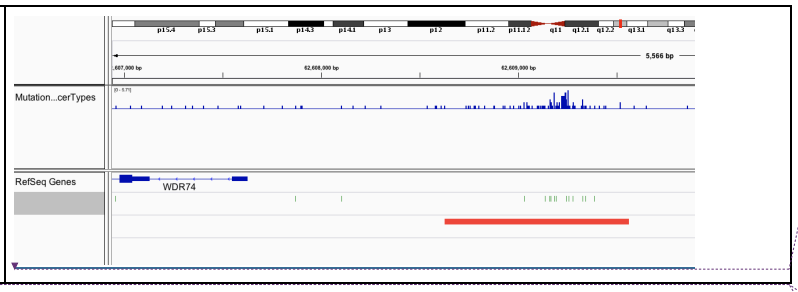
Formatted: Font:10 pt

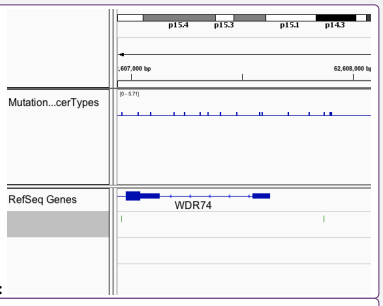
Formatted: Font:10 pt

Formatted: Font:10 pt

Comment [23]: Is this text part of the supplement?

Formatted: Font:10 pt





Deleted:

Deleted: Excerpt 2 From
Regarding extended genes ... [108]

Deleted: Excerpt 2 From
Regarding extended genes ... [107]

Formatted: Heading 2, Space Before: 14 pt

<ID>REF5.5 – Power analysis: adding more reference

<TYPE>$$$BMR,$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&MORE
<STATUS>%%%75DONE

Referee Comment	The power considerations for selecting genomic elements are valuable. "Increased" power of the combined strategy is suggested in the manuscript, yet comparison to prior work is missing. … The power considerations … Prior efforts to address this problem with restricted hypothesis testing for cancer genes should be cited (Lawrence et al, 2014; Martincorena, 2017).
Author Response	We thank the referee for identifying these previous efforts. We have added citations to these papers to our revised manuscript.
Excerpt 5.5-A from main manuscript	Excerpt to be added here JZ2JZ

Formatted Table

Moved (insertion) [14]

Deleted: 4 .

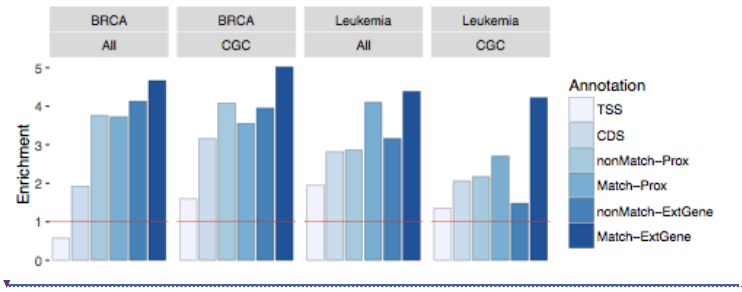
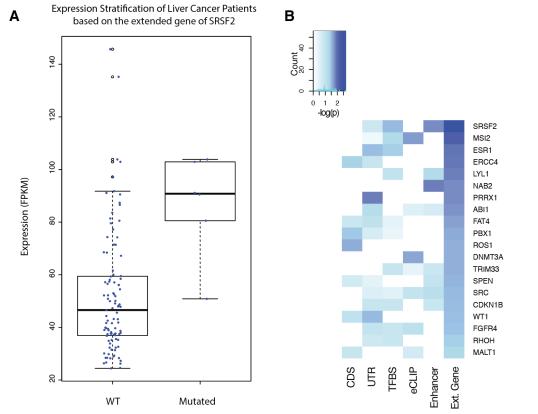
<ID>REF5.6 – BMR & Power analysis: detailed driver detection comparison

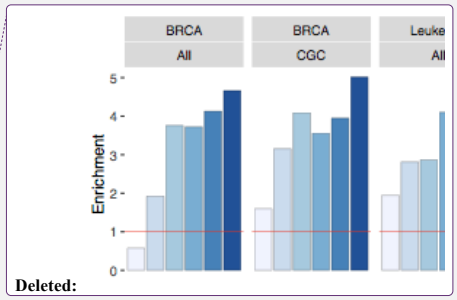
<TYPE>$$$Power,$$$Text

<ASSIGN>@@@JZ
<PLAN>&&&MORE,&&&OOS
<STATUS>%%%25DONE

Referee Comment	Again, sensitivity/specificity analyses of driver discovery with large sets, or long vs. reduced element size need to be added. An improvement of background mutation rate is suggested in the manuscript. But concrete comparisons of discovered drivers with previous work, highlighting how the presented approach is more sensitive or improves specificity, are missing.
Author Response	We thank the referee for this comment, and we have made extensive revisions to address it thoroughly. For the driver discovery part, we have now labeled known driver genes in our calculations with supporting literature and further compared our results with established methods. We have also tried to make it clear that the main purpose of our BMR analysis is not to make novel driver discoveries but to test the hypothesis that the richness of the ENCODE data can noticeably improve BMR estimation accuracy. We feel it is out of the scope of this paper to make a detailed comparison of cancer driver discovery rates. The main goal of Fig.2 is to demonstrate the usefulness of the extended gene annotations. Hence, we have also tried to re-organize all of our related analysis from the supplement to serve this goal, which includes • Better annotation disease associated germline variants (Excerpt 5.6-A). • Better stratify gene expression level by mutational status (Excerpt 5.6-B). • CRISPR based validation of oncogene activation by SV events (Excerpt 5.6-C).
Excerpt 5.6-A (extended gene in GWAS SNPs)	We extracted all breast cancer and leukemia GWAS variants from the EMBL-EBI GWAS Catalog. We removed studies with irrelevant phenotypes such as BMI after chemotherapy and only kept studies with European ancestry. Then we extracted all LD SNPs within 500kb of the GWAS SNP with r2>0.8 in 1000 Genomes Phase 3 data to calculate variant enrichment in different annotations categories. The R package VSE was used (https://cran.r-project.org/web/packages/VSE/vignettes/my-vignette.html). We found that • Adding more associated annotations significantly improved the GWAS SNP enrichment (Distal+Proximal+CDS > Proximal+CDS> CDS). • Tissue specific annotations work better then annotations from distant cell types (for breast cancer MCF-7 > K562, and for leukemia K562 > MCF7)

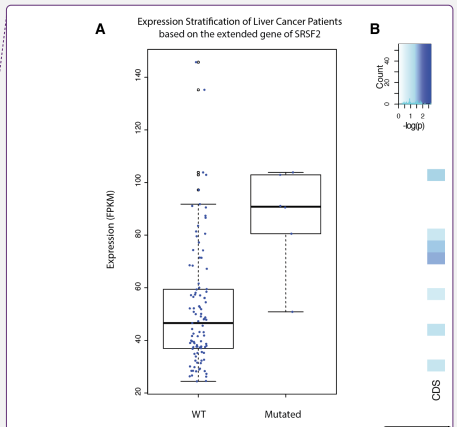
Formatted Table
Deleted: . We
Formatted: Not Highlight
Deleted:
We nonetheless hope
Deleted: Hence, we
Formatted: Not Highlight
Deleted: We nonetheless hope
Formatted: Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)
Deleted: illustrate how
Deleted: concept can be used in cancer. We
Deleted: organized allrelated
Deleted: to better demonstrate our idea in the revised manuscript. In summary, we have used extended genes
Deleted: :
Deleted: see
Deleted: 1
Deleted: see
Deleted: 2
Comment [25]: Is this correct?
Deleted: allbreast
Deleted: 1 From ... [109]
Deleted: allLD
Deleted: main figure and supplement text
Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

	
Excerpt 5.6-B (extended gene in expression analysis)	For a given gene, separated patients into groups with or without mutations in certain annotations, such as CDS, UTR, TF/RBP binding sites, enhancers, and our extended gene. We then tested differences in gene expression (FPKM) between groups based on a two-sided Wilcoxon rank sum test. We found that our extended gene annotation provides better expression separation between these groups. Specifically, we found a well-known splicing factor SRSF2, which has been recently reported contribute to liver cancer development \cite{28082404}, gives the strongest p-value for stratifying expression out of all genes in liver cancer.  JZ2JL: please update using DL's new figure
Excerpt 5.6-C (extended gene in oncogene activation)	Feng's Figure (come around Friday night)



Deleted:

Deleted: 2 From ... [110]



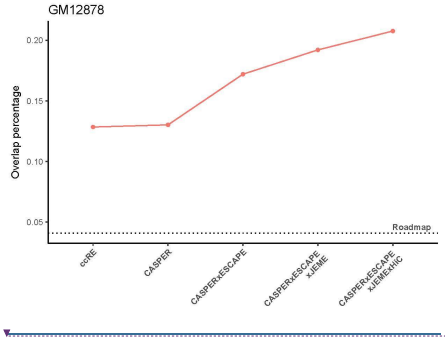
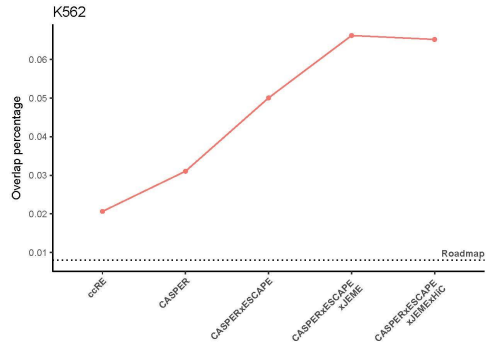
Deleted: ... [111]

<ID>REF5.7 – Annotation: false positive rates of enhancers

<TYPE>$$$Power,$$$Text
<ASSIGN>@@@JZ,@@@MTG
<PLAN>&&&AgreeFix
<STATUS>%%%95DONE

Referee Comment	6. The authors claim that reduction of functional elements increases power to discover recurrently mutated elements. This point needs quantitative support in the main manuscript (some analysis is given in the supplemental). **For example, in the enhancer list derived from the ensemble method, what fraction of enhancers are estimated to be false positives?**
Author Response	We thank the referee for raising this issue of quality metrics of our annotations, such as the enhancers and we feel this is a great opportunity to demonstrate some of the key aspects of ENCODE - quality and standard. As suggested, we have revised our manuscript to discuss the quality of annotations, including: • Enhancers (Excerpt 5.7-A) • Enhancer-gene linkages (Excerpt 5.8-A) • TF regulatory networks (Excerpt 5.14-A,B,C) It is worth mentioning that one of the authors in our paper is co-leading the ENCODE enhancer challenge in mouse. We have done extensive performance comparisons and FDR rate calibration using various assays. Although it is not completely suitable here, we have added further internal comparisons of relative performance after incorporating additional novel assays, and we now include FDRs for our methods as below. This data are unpublished data from the functional characterization group in ENCODE, so we just added this part in the response letter instead of putting it into the supplementary file. [JZ2MTG: pls help find figures, numbers and tables here]
Excerpt 5.7-A (enhancer QC)	With the ensemble method, we can get more accurate annotation and pin-point to sequences where transcription factors would bind to. To estimate the false positive rate is challenging as there is no gold-standard experiment that could assert that a predicted enhancer is negative. Here we took the FANTOM enhancer dataset and assessed the overlap percentage of our enhancer annotation in each ensemble step. We showed that each ensemble step indeed increases the percentage of overlap between our annotation and the FANTOM enhancer set. The overlap

Formatted Table
Deleted:
Formatted: Font:Bold
Formatted: Font:12 pt
Moved (insertion) [15]
Formatted: Font:12 pt
Deleted: .
Formatted: Font:12 pt
Deleted:
Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"
Moved up [15]: As suggested, we have revised our manuscript to discuss the quality of annotations, including:
Formatted: Font:12 pt
Deleted: ... [112]
Formatted: Font:12 pt
Deleted: details in
Formatted: Font:12 pt
Deleted: 1 below
Formatted: Font:12 pt
Deleted: details in
Formatted: Font:12 pt
Deleted: 1 to REF
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: details in
Formatted: Font:12 pt
Deleted: 1-3 to REF
Formatted: Font:12 pt
Deleted: 12
Formatted: Font:12 pt
Deleted: We
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: Through the process of this revision, we noticed that there is no gold standard to define enhancers in human, so it is difficult to directly call false positives. ... [113]
Formatted: Highlight
Deleted: ... [114]
Deleted: As for the enhancer part, with
Deleted: 1 From
As for the enhancer part, with
Deleted: actually
Deleted: data set

	percentage for our annotation is higher than that of the Roadmap annotation and is also higher than the main encyclopedia enhancer annotation (ccRE).  

Deleted: ,

Deleted: annotation

Deleted: . ... [115]

<ID>REF5.8 – Assessing quality of enhancer gene linkage annotation

<TYPE>$$$Annotation,$$$Text
<ASSIGN>@@@KevinYip,@@@SKL,@@@GG
<PLAN>&&&MORE
<STATUS>%%%95DONE

Referee Comment	7. The authors claim superior quality of gene-enhancer links and gene communities derived from their machine learning approach. The **method should at least be outlined in the main text, and accompanied**

Formatted Table

Formatted: Font:Bold

	by data supporting its accuracy and better performance compared to existing approaches.
Author Response	We thank the referee for his/her comments, and we totally agree that it is important to provide quality comparison of annotations. We have tried to fully address the referee's comment by • Adding a section to the supplement to show our JEME+Hi-C enhancer targets are better than the chromHMM ones (Excerpt 5.8-A) • Adding a comparison of our gene community method with others such as NMF showing that our method improves preservation of the original data structure of ChIP-seq experiments (Excerpt 5.8-B)
Excerpt 5.8-A (QC of enhancer-gene linkage)	Previously, we developed a computational approach JEME to predict enhancer-gene linkages. We have done extensive benchmark against other methods, such as IM-PET, Prestige, and Targetfinder. Details can be found in \cite JEME. In this paper, we used a 2-step approach of finding enhancer-target gene linkages. First, we used our previously published JEME algorithm to find the linkages. We then filtered the enhancer-target gene linkages using the significant Hi-C interactions that are found using the method FitHiC (ref Fithic). This 2-step filtering provides confidence that our enhancer-target gene linkages are likely to have physical interactions between them. To show how our JEME+Hi-C approach captures more accurate enhancer-gene linkages compared to existing linkages, we used published chromHMM derived enhancer-gene linkages (cite chromhmm) as the comparison dataset and GTEx whole blood eQTLs as the benchmark. We found the linkages, which the enhancer has an eQTL that changes the expression of the target gene significantly. After finding all the eQTL supported linkages for chromHMM and JEME+Hi-C, we calculated the fraction of enhancer-gene linkages that has eQTL support for various types of linkages in chromHMM and in JEME+Hi-C. As can be seen in figure below, JEME+Hi-C has higher fraction overlapped with eQTL-gene linkages. Figure R X. Overlapping the gene-target linkages with GTEx eQTLs.

Deleted: Again we

Deleted: their

Deleted: addressed

Deleted:

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5", Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Deleted: compare

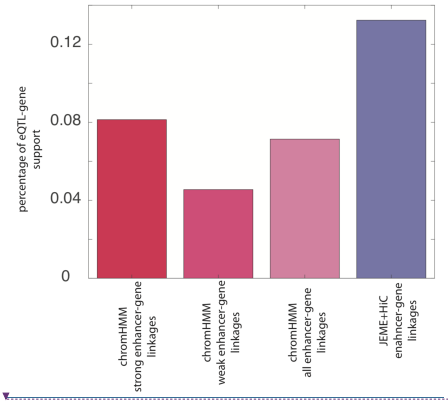
Deleted: targetsthan

Deleted: excerpt 1 below

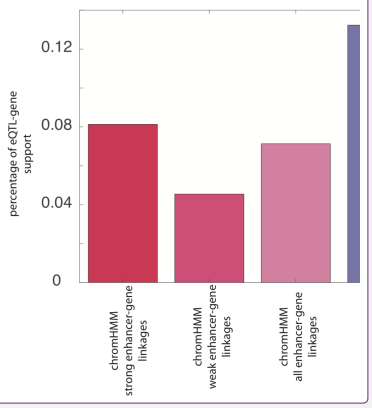
Deleted: excerpt 2 below)

Deleted: 1 From *1. Regarding the gene-enhancer linkages*

Deleted: *1. Regarding the gene-enhancer linkages*



Deleted:



Excerpt 5.8-B (gene community method comparison)

Mixed membership model is a hierarchical Bayesian topic model framework and can help to uncover the underlying semantic structure of a document collection. The core of topic models is Latent Dirichlet Allocation (LDA), which cast the mixed-membership (topics) problem into a hidden variable model of documents. The LDA model has been widely used to analyze a wide variety of data types, including but not limited to text and document data, genotype data, survey and voting data. The advantage of LDA over other algorithms (like SVD, PLSI) used in semantic analysis has been described in Blei 2003. In particular, this paper mentioned that LDA allow document to belong to multiple topics simultaneously, and the topic mixture weight was treated as k-hidden random variable to reduce overfitting problem rather than a set of individual parameters that explicitly link to the training set.

With regards to the referee's question, there is no ready-made answers since the data type (TF target network) and problem-definition of our study are both specific. Fundamentally the LDA method is an unsupervised, therefore there is no labels on the dataset and accuracy metrics is not applicable. If we treat the LDA mixed-membership analysis as a dimensionality reduction problem, it is possible to compare how well of a model can reproduce the information of original data, as described in paper (Guo, Y., & Gifford, D. K. (2017). Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. BMC Genomics, 18(1), 45.). The correlations of the original target gene vectors between two TFs are compared with those of dimension reduced vectors. The better method should be much close to original vectors correlations.

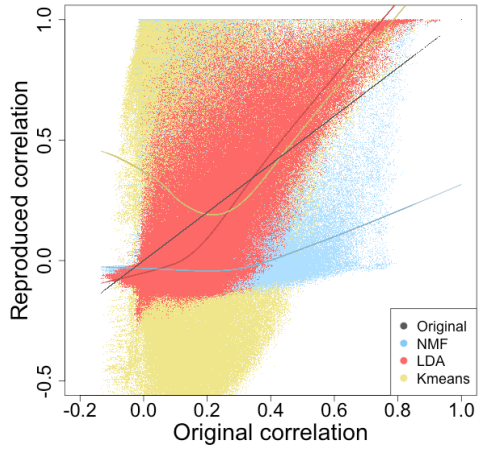
To explore how well the LDA mixed-membership analysis on TF regulatory network, we extend our dataset from 122 GM12878 and K562 samples to all the 862 TF ChIP-Seq assays included in ENCODE data portal. In order to get a reliable correlation, we also increase the number of topic to 50 as the number of TF sample increases. The non-negative matrix factorization (NMF) and Kmeans clustering are used for comparison because the nature of regulatory network requires a non-negative decomposition. The same target dimension K =50 was used to NMF and target number of clusters K=50 for Kmeans. The Euclidean distance between each data the centroids are used to calculated

Deleted: 2 From ... [116]

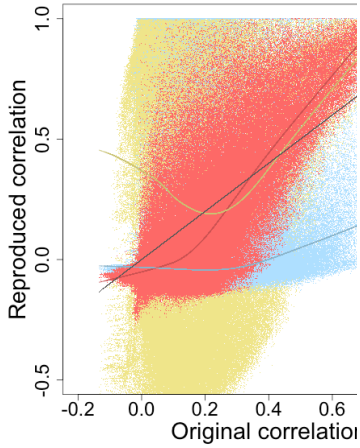
Deleted: GM

Deleted: K526

Deleted: centroidds

	the correlation. As shown in the figure, the x-axis is original correlation of two TF regulatory target, y-axis is reproduced correlation from LDA document to topic distribution and NMF decomposed matrix. The solid line is the 'loess' smoothing curve for the scattered dots. We can see the LDA method can reproduce the original correlation better than either NMF or Kmeans. Overall correlation between the reproduced pairwise correlation and the original correlation were 0.123 in Kmeans, 0.404 in NMF and 0.788 in LDA. 

Deleted:



<ID>REF5.9 – What data sets are used

<TYPE>$$$BMR
<ASSIGN>@@@JZ
<PLAN>&&&Defer
<STATUS>%%%75DONE

Referee Comment	8. From the main manuscript, it is not clear which cancer data sets were analyzed with the new background mutation rate estimates and functional regions. Datasets and sample size should be mentioned explicitly.
Author Response	We thank the referee for bringing out this point. We provide it here in the table and summarized it in a line in the main text.

Formatted Table

Excerpt 5.9-A	Wait for the main text JZ2JZ

Deleted: From ... [117]

<ID>REF5.10 – Mutational signatures

<TYPE>$$$BMR,$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%85DONE

Referee Comment	9. Do the authors take into account mutational signatures?
Author Response	We thank the reviewers for pointing this out. In the BMR calculation section, we did consider the local 3mer context effect. But we did not specifically looked into the mutational signatures otherwise. We have made this clear in the discussion section in the revised manuscript.
Excerpt 5.10-A (added in disc. sect.)	We hope that in the future new models that can incorporate, sequence coverage, mutational signatures, small scale features (TF and nucleosome binding), would further integrate the full potential of ENCODE data to better calibrate background mutation rates.

Formatted Table

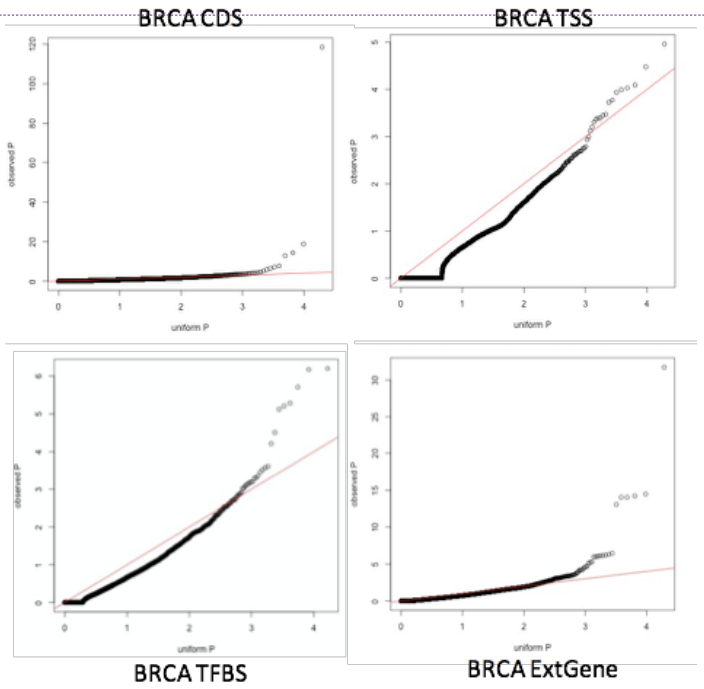
Deleted: From ... [118]

<ID>REF5.11 – Additional QQ plots

<TYPE>$$$BMR,$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	10. The significance analysis of cancer cohorts (Figure 2) should highlight known cancer genes versus those newly found in this study.

Formatted Table

	A QQ-plot should be included to confirm that the algorithm accurately models the background expectation.
Author Response	We thank the reviewers for pointing this out. We have updated Fig. 2 to label the known cancer genes (Except 5.11-A). Yes, we have provided the QQ plot in the supplementary file in our initial submission and we have extracted some of QQ-plot in the excerpt below. The QQ-plot below indicates no obvious P value inflation, which indicates our BMR estimation is should be OK.
Excerpt 5.11-A (updated Fig.2)	JZ2DL: please label known cancer genes on Fig.2
Excerpt 5.11-B (in suppl.)	QQ-plot for breast cancer on various annotations. 

Deleted:

Formatted: Not Highlight

Deleted: new genes we discovered? Too much

Formatted: Not Highlight

Deleted: driver detections then, or out of scope?

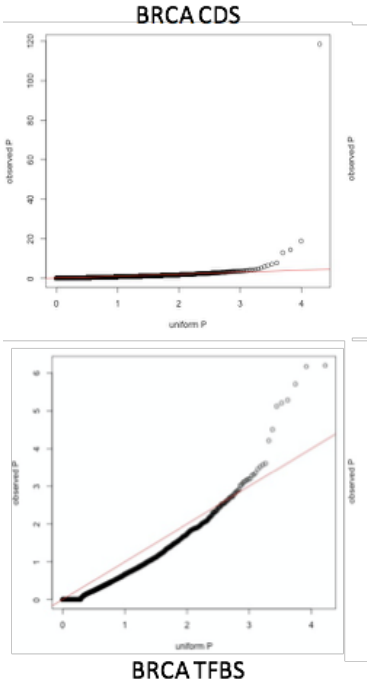
Formatted: Not Highlight

Deleted:

Formatted Table

Deleted: From

... [119]



Deleted:

<ID>REF5.12 – Sequence coverage

<TYPE>$$$BMR,$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	Do the authors include sequence coverage in their method?
Author Response	We did not consider sequence coverage but this is a good point. We included discussion of this point in our revised manuscript.
Excerpt 5.12-A	We hope that in the future new models that incorporate sequence coverage, mutational signatures, and small scale features (TF and nucleosome binding), will show the full potential of ENCODE data to better calibrate background mutation rates.

Formatted Table

Deleted: the

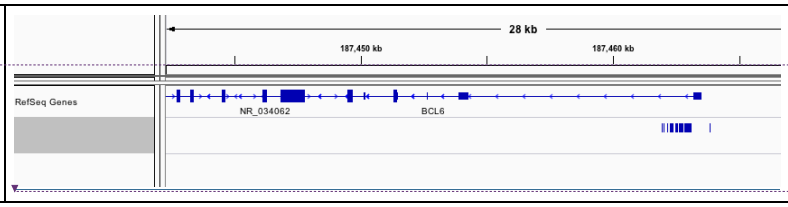
Deleted: From ... [120]

<ID>REF5.13 – BCL6 Questions

<TYPE>$$$Annotation,$$$Calc
<ASSIGN>@@@XK,@@@TG
<PLAN>&&&AgreeFix
<STATUS>%%%TBC
[JZ2JZ: more investigations]
JZ2MG: wait, not yet updated

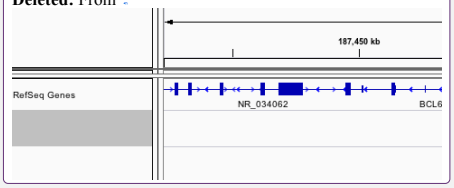
Referee Comment	11. The authors mention that BCL6 would have been missed in an exclusively coding analysis. In which part of the extended annotations were recurrent BCL6 mutations found? If near the promoter, is the BCL6 5' region a known AID off-target? Are BCL6 mutations in CLL associated with translocations?
Author Response	JZ2JZ: check We thank the referee for this comment. As suggested, we found that the there is a mutation hotspot near the first intron of BCL6.

Formatted Table

Excerpt 5.16-A	

Deleted:

Deleted: From



<ID>REF5.14 – ChIP-seq vs other computational based networks: FP of network

<TYPE>$$$Network,$$$Calc
<ASSIGN>@@@Peng,@@@JZ,@@@DL
<PLAN> &&&AgreeFix
<STATUS>%%%95DONE

Referee Comment	12. The manuscript notes that the new networks presented contain "more accurate and experimentally based" gene links. This claim should be supported with **comparisons with existing networks** and statistical evaluation. How many of the derived networks are false positives? How many networks are derived in total?
Author Response	We thank the referee for bringing this point up, and we find that this is the core strength of ENCODEC. We also feel that it is important to make comparisons with existing networks with more statistical evaluation. We have made the following revisions in the updated manuscript. **1. Regarding the proximal regulatory element network:** *1.1 Comparison with Biogrid and String:* our networks can capture a higher fraction of standard interactions than networks such as Biogrid and String (Excerpt 5.14-A). *1.2 Comparison with DHS-based imputed networks:* our networks provided better correlations with TF knockdown experiments than the DHS-based imputed network provided in Neph et. al. 2012 (Excerpt 5.14-B). *1.3 False positive rate:* ENCODE has always enforced a strict data quality standards for all ENCODE produced ChIP-seq experiments, which allow rigorous false positive control (Excerpt 5.14-C). **2. Regarding the distal regulatory element network:** With the ChIP-seq, DHS, STARR-seq, ChIA-PET, and Hi-C experiment, ENCODE has a distal TF-enhancer-gene network of high quality, which is less discussed

Formatted Table

Formatted: Font:Bold

Deleted:

Deleted: *experimental interactions.*

Deleted: (from manually curated

Deleted: from TTRUST) than protein physical networks, including

Deleted: experimental interactions (see details in excerpt 1

Deleted:

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

Deleted:

Deleted: . (see details in excerpt 2

Deleted:

Deleted: *estimation of the ChIP-Seq based net* ... [121]

Deleted: consortium

Deleted: transcription factor

Deleted: us to rigorously

Deleted: the false positives (see details in excerpt 3

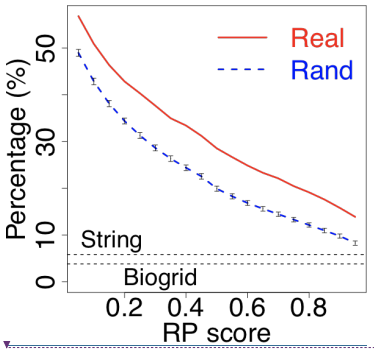
and investigated previously. We feel this is one of the unique aspects of our resource.

2.1 High quality of integrative enhancer definitions; (Excerpt 5.7-A).

2.2 High quality of enhancer-gene linkages; (Excerpt 5.8-A).

Excerpt 5.14-A (comparison with Biogrid and String network)

To evaluate the quality of ENCODE transcriptional regulatory networks, we utilized the TRRUST database, which manually curated transcriptional regulations from Pubmed articles (Han et al., 2018). We defined the TRRUST interactions as the standard and tested the fraction of standard interactions that other networks can recapitulate. The ENCODE network can capture a higher fraction of standard interactions than protein physical networks, including Biogrid and String experimental interactions (Supplementary Figure X). Moreover, the fraction of standard networks that ENCODE network recapitulated is consistently higher than random. These results supported the higher relevance of ENCODE networks on transcriptional regulation compared to other networks. We also constructed another post-transcriptional network between RBPs and target genes through linking the RBP binding sites on gene 3'UTR regions. To the best of our knowledge, the current study is the first one to study RBP-gene interactions systematically; thus we are not aware of any previous resources that can provide gold standard regulations for comparison.



Supplementary Figure X. ENCODE networks captured a higher fraction of curated regulations than other networks. The TRRUST database manually curated 8,412 transcriptional regulatory interactions from Pubmed articles (Han et al., 2018). We computed the fractions of TTRUST interactions that other networks can recapitulate. Since each ENCODE ChIP-Seq interaction has a regulatory potential (RP) score, we showed the fractions with different RP thresholds. The random fraction for ENCODE network was estimated through 100 perturbed TTRUST networks using the stub-rewiring method that preserved the gene network degrees (Milo et al., 2002).

Excerpt 5.14-B (comparison with imputed network)

Our new regulatory network edges are derived from ENCODE TF ChIP-seq experiments, and they provide more accurate gene linkages than imputed networks from other genomic features. To demonstrate the superiority of our new network, we have evaluated our experimentally derived ChIP-seq networks with DHS-based imputed networks from previous publications. We have used two types of ChIP-seq networks. The first one is based on proximity to TSS and the second one based on target identification from profiles (TIP) method. For imputed network, we used Neph et. al. 2012 (Neph, Shane, et al. "Circuitry and dynamics of human transcription factor regulatory networks." Cell 150.6 (2012): 1274-1286.) TF-to-TF network imputed from DNase I hypersensitive

Deleted: aspect

Deleted:

Deleted: *after integrating many histone ChIP-seq and DHS, and STARR-Seq data*

Deleted: – Annotation: false positive rates of enhancers".

Deleted:

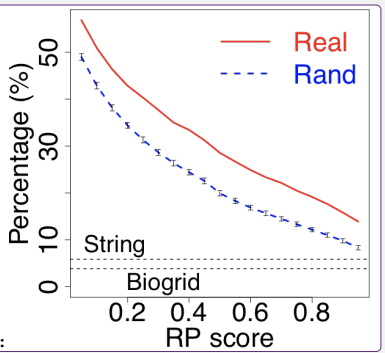
Deleted: ... [122]

Deleted: .

Formatted: Font:Arial

Deleted: *Regarding Comparison with Biogrid and String experimental interactions.*

Deleted: 1 From *Regarding Comparison with Biogrid and String experimental interactions.*



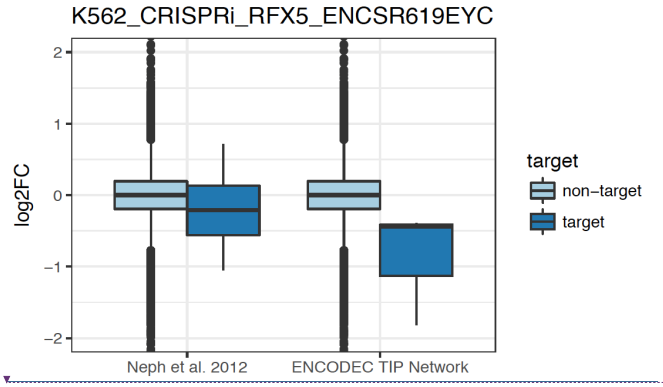
Deleted:

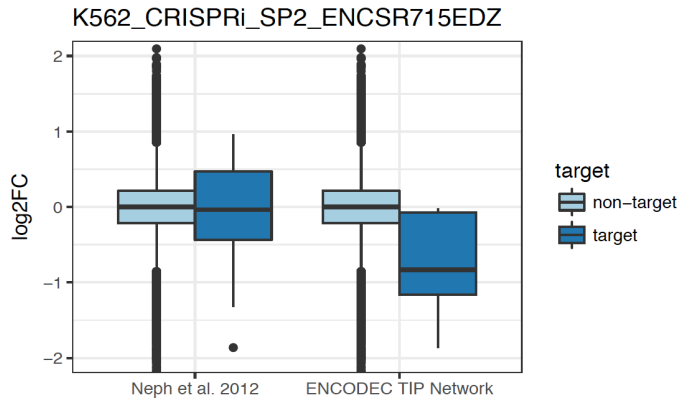
Deleted: *Regarding comparison with imputed network*

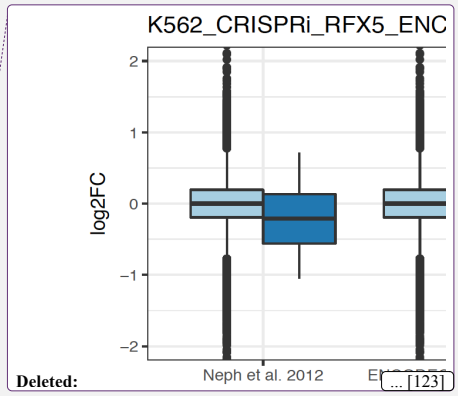
Deleted: 2 From *Regarding comparison with imputed network*

footprints. In addition to Neph et. al. DHS network, we also built our own version of similar DHS network by utilizing the ENCODE DNase-seq dataset. To test the gene linkages, we have utilized ENCODE RNAi based TF knockdown and CRISPR-based TF knockout datasets to test how the target gene linkages defined by various network definition are affected by after KD/KO. Overall, target genes of ENCODE ChIP-seq networks had larger differential expression after knocking down (Supplementary figure X). Moreover, DHS-imputed network derived from ENCODE DNase-seq performed better than the previously published method (not shown here, available in Supplementary document).

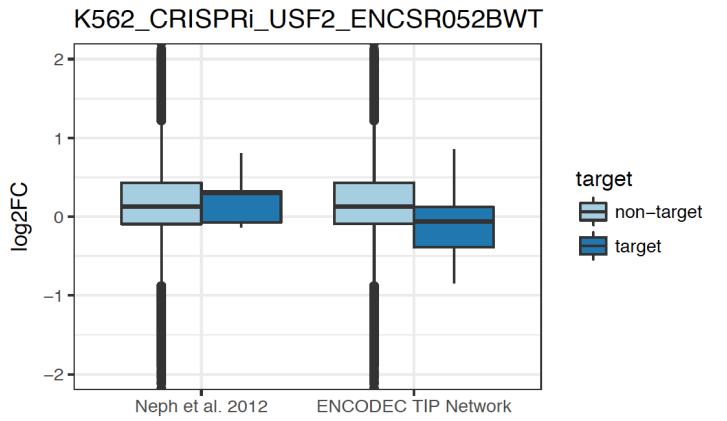
Supplementary figure X. Evaluation of ENCODEC network with previously published regulatory network using ENCODE CRISPRi knockdown data. Target genes of ENCODEC ChIP-seq based networks have larger expression differential after knocking down. Examples of RFX5, SP2, and USF2 shown. More details with full figures comparing all variants of ENCODEC networks can be found in supplementary document.







Deleted: ... [123]



1

Excerpt 5.14-C (false positives)

In order to ensure that experiments are reproducible, at least two replicates must be performed in either isogenic or anisogenic conditions (For more information about ENCODE 3 ChIP-seq experimental guidelines, please refer https://www.encodeproject.org/documents/ceb172ef-7474-4cd6-bfd2-5e8e6e38592e/@@download/attachment/ChIP-seq_ENCODE3_v3.0.pdf).

For transcription factor experiments, 1486 of 1863 (80%) ChIP-seq experiments we have used to compile ENCODEC resources have more than 2 replicates, which allows further quality control of the derived network. ENCODE used IDR (Irreproducible Discovery Rate) framework to ensure reproducibility of high-throughput experiments by measuring consistency between two biological replicates within an experiment. All processed experiments had both rescue and self consistency ratios are less than 2.

Self-consistency Ratio	Rescue Ratio	Resulting Data Status	Flag colors
Less than 2	Less than 2	Ideal	None
Less than 2	Greater than 2	Acceptable	Yellow
Greater than 2	Less than 2	Acceptable	Yellow
Greater than 2	Greater than 2	Concerning	Orange

After extensive quality controls for the concordance between replicates, peaks are called using macs2 {"Zhang et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* (2008) vol. 9 (9) pp. R137"} with p-value cutoff of 0.01.

Deleted: *Regarding False positive rate estimation of the ChIP-Seq based networks*

Deleted: 3 From *Regarding False positive rate estimation of the ChIP-Seq based networks*

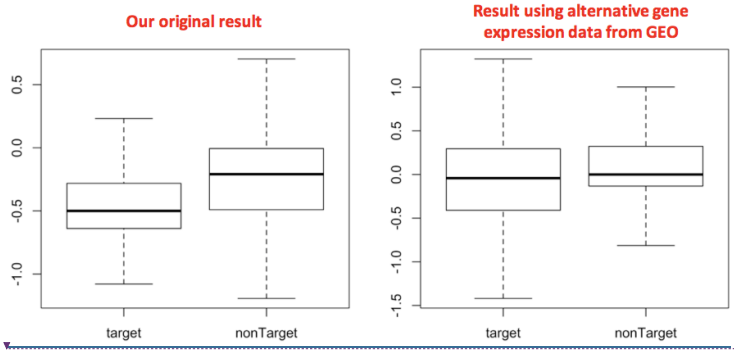
Deleted:

Self-consistency Ratio	Rescue Ratio
Less than 2	Less than 2
Less than 2	Greater than
Greater than 2	Less than 2
Greater than 2	Greater than

<ID>REF5.15 – MYC KD Validation

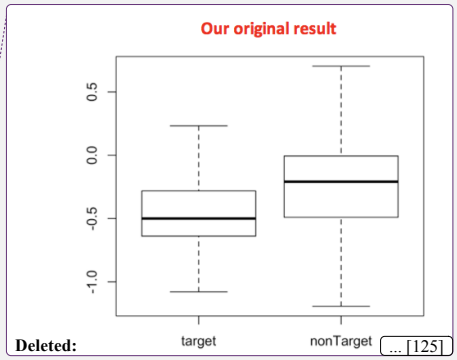
<TYPE>$$$Network,$$$Text
<ASSIGN>@@@DC
<PLAN>&&&AgreeFix

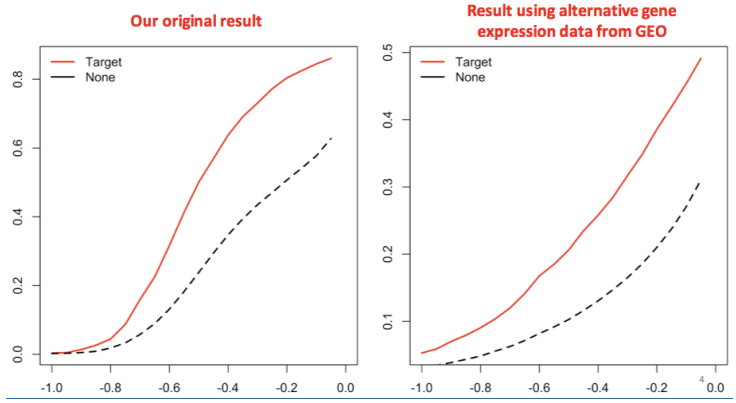
<STATUS>%%%100DONE

Referee Comment	13. MYC is known to have profound effects on gene networks. Have the authors considered comparing the results from their MCF7 knockdown experiment to existing data from similar MYC knockdowns to validate the behavior of the network?
Author Response	We thank the referee for this suggestion and we feel this is a good comment. As suggested we searched for external dataset from multiple platform and cell types and used them to compare with our discoveries. Both datasets confirmed our claims.
Excerpt 5.15-A (MYC KD validation)	We carried out these analyses after first identifying an alternative dataset. Specifically, we identified a dataset of gene expression for both MYC knockdowns (as well as a corresponding control) in Gene Expression Omnibus (GEO accession number GSE86504). For these alternative data, gene expression was measured by RNA-seq in the HT1080 cell line. We note that, even though these alternative analyses were conducted on a different cell line, the results we obtain (shown below in the right panels, and now made available in the supplementary materials) validate the behavior of the network, and they are consistent with our previous results (in which gene expression was measured in the MCF-7 cell line). These comparable results in an alternative cell line suggests that these results are robust. 

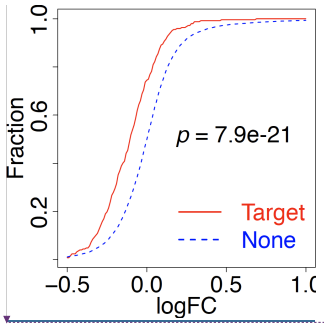
Formatted Table

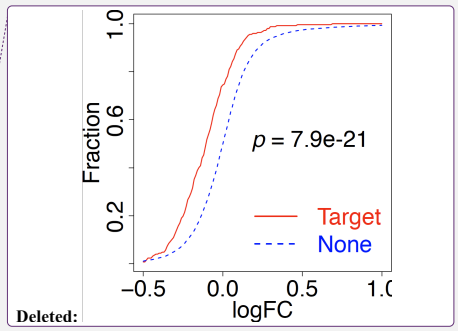
Deleted: From . ... [124]

Deleted:  ... [125]



We also found another array based MYC knockdown data the results correlate well with our discoveries.





Deleted:

<ID>REF5.16 – SUB1 analysis

<TYPE>$$$NoveltyPos,$$$Calc
<ASSIGN>@@@MRS,@@@JL,@@@YY
<PLAN>&&&MORE
<STATUS>%%%95DONE

Referee Comment	14. SUB1 is a potentially interesting new cancer gene. The authors should further explore the biology of this gene.
Author Response	We thank the referee for this comment about SUB1, and also the related previous comment about MYC. This spurred us to really think about the biology of these key factors. We found out that SUB-1 actually has quite a

Formatted Table

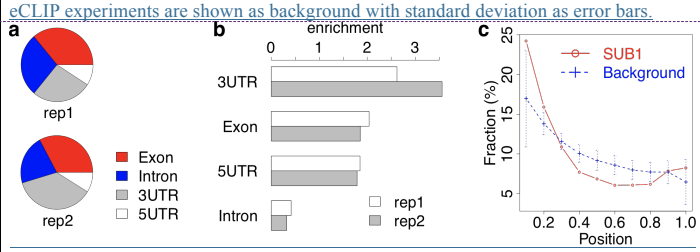
Formatted: Font:12 pt

reasonable biological function to cancer. We were able to figure out how it collaborates with other regulators, such as MYC, to demonstrate how our multi-networks, including the TF and RBP networks, really fit together to relate to biology. Finally, we updated Fig 3 by adding our new small-scale validation experiment to drill into the SUB-1 MYC connection and validate it partially on several important oncogenes.

Though it may not represent a complete novel finding in cancer biology, we do think it illustrates the way ENCODE networks are useful for highlighting the roles of certain key players and enabling follow-on drill down studies.

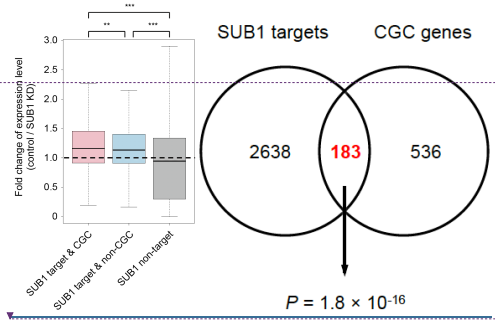
Excerpt 5.16-A (SUB1's enrichment in 3'UTR, in suppl.)

Supplementary Figure X: eCLIP peaks of SUB1. (a) The composition of SUB1 peaks over different gene regions is shown for each replicate. (b) For each gene region, the relative enrichment (fraction of SUB1 peaks / fraction of all peaks) of SUB1 peaks is shown. (c) The distribution of SUB1 peaks over 3'UTR regions is shown. The mean across all RNA binding proteins profiled by eCLIP experiments are shown as background with standard deviation as error bars.

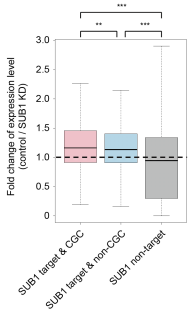


Excerpt 5.16-B (SUB1 target on famous oncogenes, in suppl.)

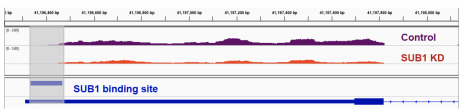
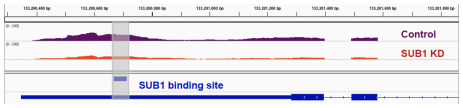
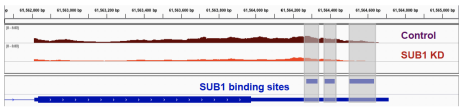
We found that SUB1 targets are enriched in cancer associated genes, such as genes in Cancer Gene Census (P=1.8e-16 by Fisher's exact test), and such genes showed larger down regulation upon SUB1 knockdowns. Among many of such genes, we have shown some IGV examples together with SUB1 binding sites on the 3' UTRs.



Deleted: in relation
Formatted: Font:12 pt
Deleted: really
Formatted: Font:12 pt
Deleted:
Formatted: Font:12 pt
Deleted: In summary, we were able to elaborate on this considerably in our revised version, including ... [126]
Formatted: Font:12 pt
Deleted: did a
Formatted: Font:12 pt
Deleted:
Formatted: Font:12 pt
Deleted: While we do
Formatted: Font:12 pt, Not Highlight
Deleted: think this is
Formatted: Font:12 pt, Not Highlight
Formatted: Font:12 pt, Not Highlight
Formatted: Font:12 pt
Deleted:
Formatted: Font:12 pt
Deleted: ... [127]
Formatted: Font:10 pt
Deleted: 1 From ... [128]
Formatted: Font:10 pt
Deleted: supplement)
Formatted: Font:10 pt
Deleted: 2 From ... [129]

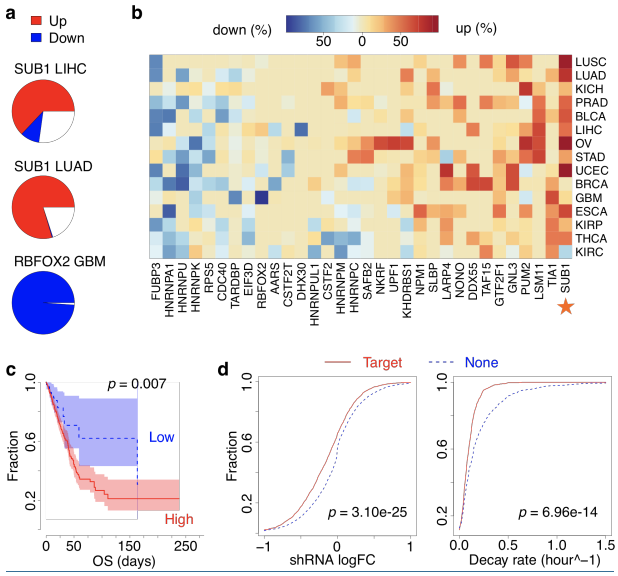


Deleted: ... [130]
Deleted: supplement)

Gene	Functions	PMID	Expression profiles of the 3' UTR
BRCA1	The gene is involved in maintaining genomic stability	12677558, 17416853, 23620175, 16551709	
POLE	The gene is involved in DNA repair and replication	26133394, 28423643	
FEN1	The gene is involved in DNA repair and replication	20929870, 22586102	

Excerpt 5.16-C (SUB1's regulatory potential in different cancer types, in suppl.)

Using ENCODE eCLIP data and TCGA tumor profiles, we applied RABIT framework to identify RNA binding proteins (RBP), whose target genes are differentially regulated in cancer. (a) The fractions of patients with target genes up or down regulated are shown for each combination of RBP and cancer type. (b) The patient fractions with target genes differentially regulated are shown for all cancer types and RBPs whose fraction values are larger than 50% in at least one cancer. (c) All lung adenocarcinoma patients are divided to two groups according to SUB1 activity predicted by RABIT. The overall survival was shown by KM plot. The association between SUB1 activity and survival was tested through Cox-PH regression. (d) In the left panel, the cumulative distributions of gene expression after SUB1 knock down in HepG2 cell are shown for predicted SUB1 targets and none targets. In the right panel, the cumulative distributions of mRNA decay rates in HepG2 cell are shown. The comparison between two categories is done through Wilcoxon rank-sum test.



Excerpt 5.16-D (SUB1-MYC co-regulation in suppl.)	Among genes whose 3'UTR regions have *SUB1* eCLIP sites, we observed significant enrichment of functional categories including *MYC* targets and spliceosome. *MYC* activation induces an increase in total precursor messenger RNA synthesis, which increases the burden on the core spliceosome to process pre-mRNA [1]. Also, *MYC* activation can stimulate oxidative phosphorylation, which fulfills the bio-energetic demands of cancer cells [2]. These results together indicate that *SUB1* may stabilize the *MYC* target genes and pathways to promote the malignant growth of cancer cells.

Formatted: Font:10 pt

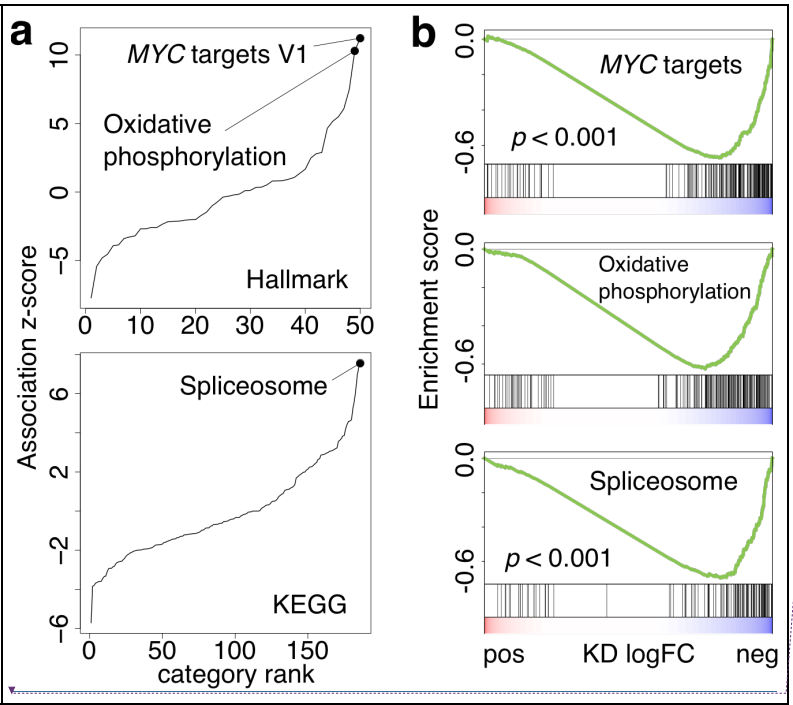
Formatted Table

Formatted: Font:10 pt

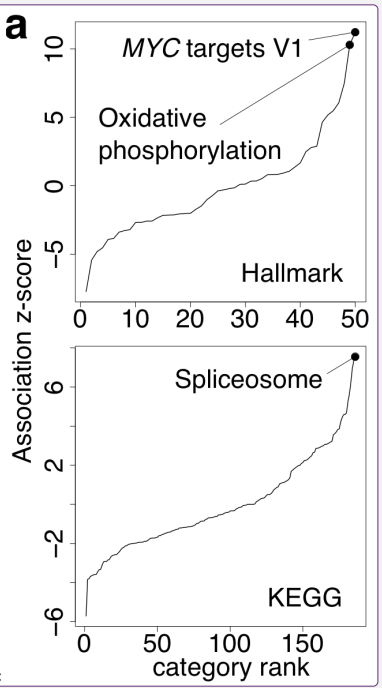
Deleted: 3 From ... [131]

Formatted: Font:10 pt

Deleted: supplement)

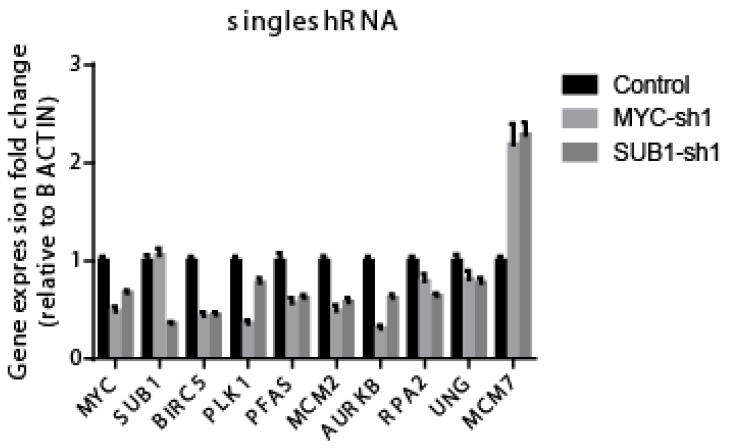


Deleted:



Excerpt 5.16-E (SUB1-MYC co-regulation validation, in suppl.)

To detect predicted common target gene of MYC and SUB1, shRNA plasmids containing 4 targets sites of each gene were used to transfected to HepG2 cell using LipofectamineTM 3000 following the manufacturer's instructions (Invitrogen) (target sites for each gene are listed in Sup table 1). Briefly, 0.12 M HepG2 cells were seeded in each well of one 24-well plates 24 hours before transfection. 500 ng plasmids containing either single shRNA or 4 shRNA plasmids as pool were mixed with 0.75 uL LipofectamineTM 3000 in Opti-MEM I medium (Invitrogen) and loaded to HepG2 cells in each well. Blank plasmids without shRNA target sequence was used as control. To improve transfection efficiency, 2 ug/mL puromycin was used to select successful transfected cells. 72 hours after transfection, total RNA was extracted using RNeasy Mini Kit (Qiagen) and followed by cDNA generation using SuperScript III (Invitrogen). Knockdown efficiency and target gene expression level were quantified and compared to BACTIN by qPCR using KAPA SYBR® FAST qPCR Master Mix (2X) Kit (Sigma). The qPCR primers were listed in Sup table X.

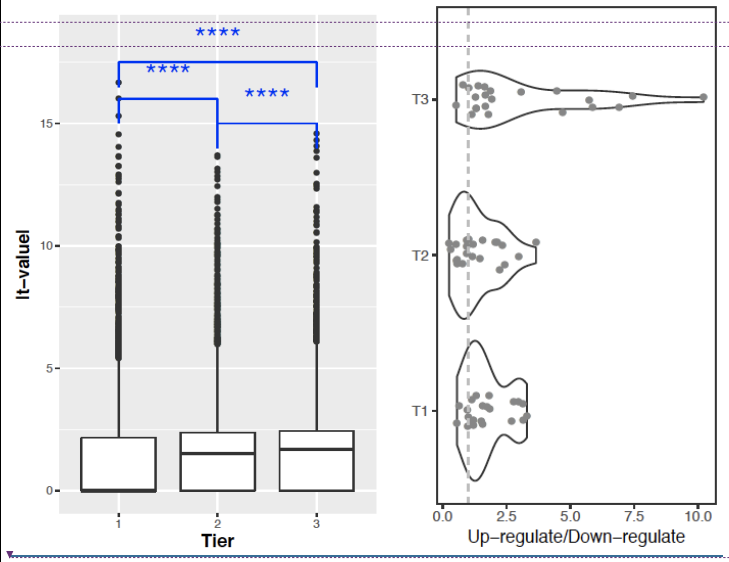
	
Excerpt 5.16-F (New Fig. 3)	New Figure 3, JZ2DL please add

<ID>REF5.17 – Significance of regulatory network hierarchy

<TYPE>$$$Network,$$$Calc
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%99DONE

Referee Comment	15. The manuscript claims that transcription factors placed at the top level of the network hierarchy are enriched in cancer-associated genes and drive expression changes. Both claims need to be supported with statistical tests.
Author Response	DL2JZ: can you fill in XXX below with the actual p-value from HierNet analysis? I tried to look up from old data, but I couldn't find exact pvals. Also could you add some descriptions to supplementary figures?

Formatted Table

	We would like to thank the referee for the comment. We actually have done statistical significance testings to support our claims in the original submission, however, it did not spell out. We do agree with the referee that statistical testings are important to support our claims, so we improved the presentation in the revised manuscript, and we provided additional statistical testings in the supplements to support our claims. The right panel of Figure 4 shows results from Wilcoxon signed-rank test. If a p-value is less than 0.05 it is flagged with one star (*). If a p-value is less than 0.01 it is flagged with two stars (**). If a p-value is less than 0.001 it is flagged with three stars (***). We find that the top-level of the generalized network was enriched with cancer-related TFs with p-value XXX and had larger correlation to drive target gene expression change (p-value XXX).
Excerpt 5.17-A (in suppl.)	Supplementary Figure X. 

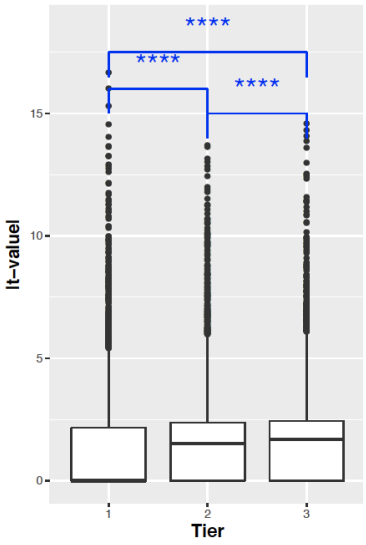
Deleted:

Formatted: Font:12 pt

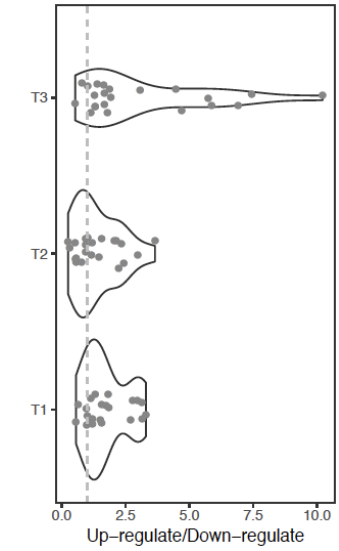
Deleted:

Deleted: From

... [132]

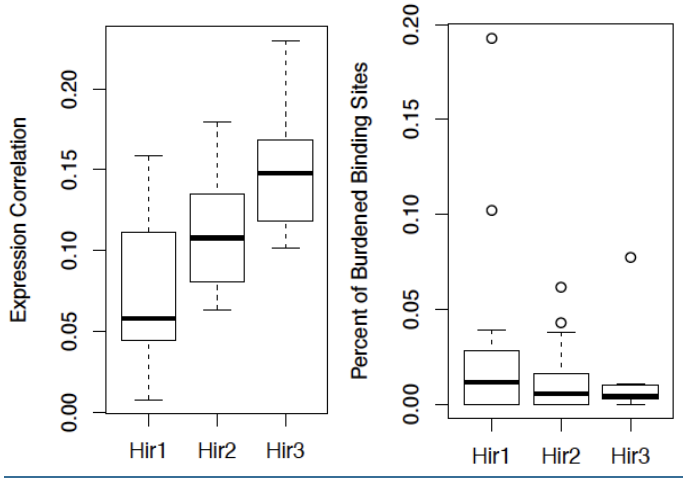


Deleted:



... [133]

Deleted: supplement)



<ID>REF5.18 – Rewiring of regulatory network: FP of rewring

<TYPE>$$$Network,$$$Calc
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	16. In the tumor-normal network comparison, is the fraction of edge changes related to the total number of edges for a given TF? This analysis should further clearly state its null hypothesis (what changes are expected?). What happens when edges are randomly permuted? [JZ2MG: we did not directly answer this question]
Author Response	We thank the referee for pointing out this issue. We agree with the referee that we need to be more clear about the analysis related to rewiring of the regulatory network in the revised manuscript. In short, we would like to clarify that the rewiring index is based on the fraction of regulatory edge changes between two cellular contexts. We have added more analysis in the revised supplement to estimate false positive rates of rewiring.

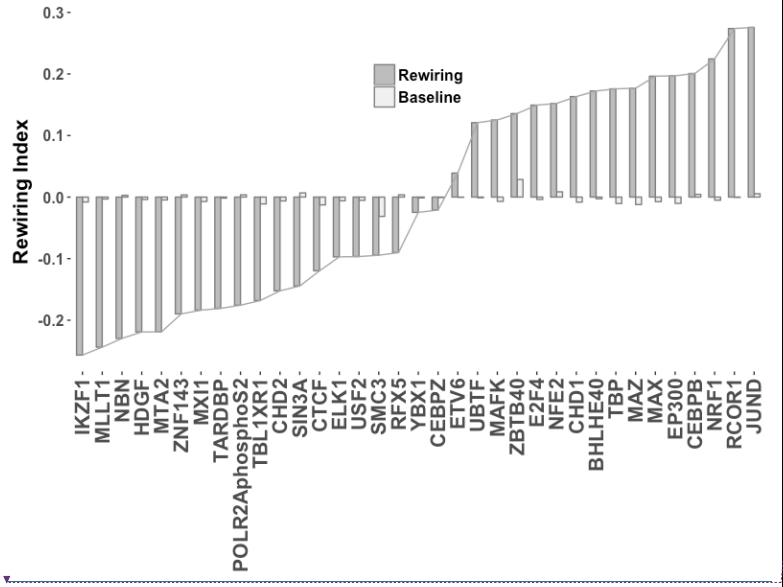
Formatted Table
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: See excerpt for more details.

Excerpt 5.18-A (in suppl.)

… The rewiring index is then normalized across all regulatory proteins, and the sign reflects the direction of rewiring. Details of rScore derivation can be found in Supplementary 5.3. Given this, we assume a null hypothesis to be no change in regulatory edge across cell types. We expect no or minimal change in edges when two cellular contexts are similar. To demonstrate, we selected all available GM12878 ChIP-seq experiments that have at least two replicates, and we then calculated the same rewiring index between isogenic replicates of the same cellular context. We expect very small rewiring score given they are the same cellular context, and the edge changes between two networks will be simply a noise from ChIP-seq experiments.

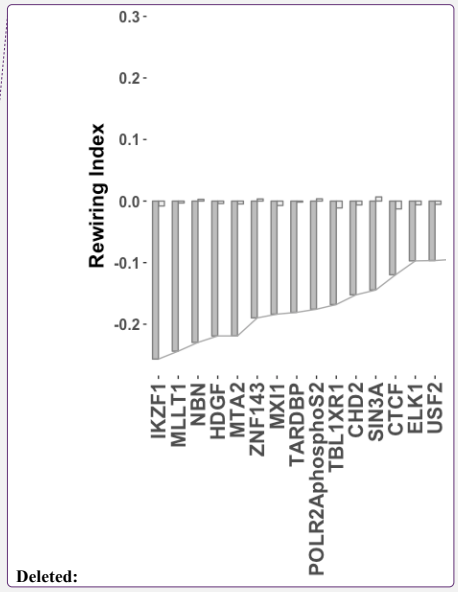
As expected, when two cellular context are similar, as shown in "baseline", minimal number of edges do change targets. However, in "rewiring", TF do change targets extensively when compared across cancerous (K562) to normal (GM12878) cell lines. To put this into perspective, we calculated the fraction of regulatory edges that are due to noise. We estimate that, on average, 1.36% of observed regulatory edges could be false positives.

Supplementary Figure X1.

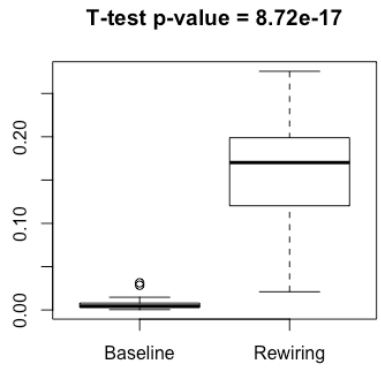


Supplementary Figure X2.

Deleted: From … [134]



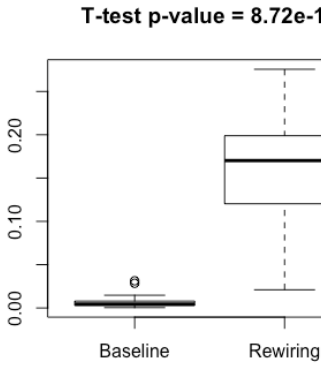
Deleted:



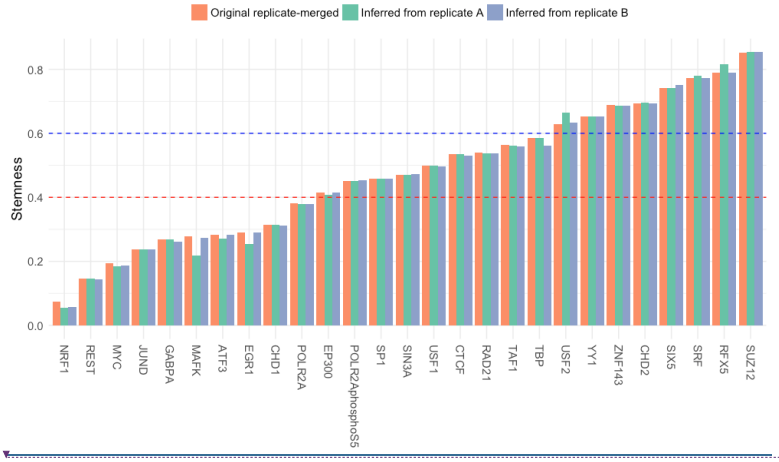
<ID>REF5.19 – Stemness in Rewiring analysis in the stem cells

<TYPE>$$$Stemness,$$$Calc
<ASSIGN>@@@DL,@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%25DONE

Referee Comment	17. The network change comparisons with the H1 stem cell models need **statistical testing for significance.** What fraction of the rewired edges are expected to be false positives?
Author Response	We thank the referee for pointing this out. We totally agree with the referee's suggestion and took this opportunity to significantly expand the statistical aspects of rewiring and stemness analysis, which includes *1. Regarding the false positives of the rewired edges*, approximately 1.36% of rewired regulatory edges are false positives (Excerpt 5.18-A). *2. Regarding the statistical testing in the normal-tumor-stem analysis*, a section in the supplementary file on our original rewiring analysis (Excerpt 5.19-A)



Deleted:
Formatted Table
Formatted: Font:Bold
Deleted: .
Formatted ... [135]
Comment [28]: put more in the suppl and summarize less
Deleted: regulatory network
Deleted: H1
Formatted: Font:12 pt
Deleted: model. In summary, we have done the ... [136]
Deleted: .
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: . ... [137]
Formatted ... [138]
Deleted: using examples from CML.
Formatted: Font:12 pt
Deleted: .
Formatted ... [139]
Formatted: Font:12 pt
Deleted: . ... [140]
Formatted: Font:12 pt, Not Italic
Deleted: *estimated by fractions* . ... [141]
Formatted: Font:12 pt
Deleted: 1 below.
Formatted: Font:12 pt

	3. *Regarding the new stemness analysis using PCA/RCA*: We ran Wilcoxon test to compare the tumor-stem and normal-stem distance (Excerpt 5.19-B,C) and found that tumor cells are more similar to stem cells, which is consistent with other findings \cite(TCGA i stemness).
Excerpt 5.19-A (in suppl.)	The H1 stem cell model uses fractional overlap of rewired edges between cancerous cell types vs. H1. Therefore we attempted to evaluate statistical significance of our model by measuring how much of H1 network changes are due to noise and use of other normal cell types to evaluate how much of rewired edges overlaps with H1.  Using replicates of H1-hESC ChIP-seq experiments, we made two independent H1 networks in addition to original replicate merged H1 network, and we made recalculated stemness of TF, whether they rewire toward or away from H1. We find that the results of all of stemness direction is reproduced using either replicate.
Excerpt 5.19-B (stemness in suppl.)	We performed PCA (RCA) analysis on RNA-seq, RNAi and CRISPR-based knockdown, and TF ChIP-seq data to demonstrate that clusters of cancerous cell types de-differentiate to a state that resemble more like stem-like cell types. We consistently found using different types of data that cancer cells' regulatory status as well as gene expression profiles are closer in euclidean distance to the stem state as compared to their primary cells of origin (Figure 5). We quantified and compared the L2 distance to stem-like clusters between cancerous cell types and normal cell types. We find that using both proximal network and gene expression profiles have statistically significant difference between normal-to-stem and cancer-to-stem distance (using Wilcoxon rank sum test, Suppl. Fig. A-B). We found observable difference in distal regulatory network but found no statistical significance.

Formatted: Font:12 pt, Underline

Deleted: *our*

Formatted: Font:12 pt, Underline

Formatted: Font:12 pt, Underline

Deleted: . ... [142]

Formatted: Font:12 pt

Comment [29]: supplement

Deleted: using multiple datasets. We

Formatted: Font:12 pt

Deleted: in general

Formatted: Font:12 pt

Deleted: the

Formatted: Font:12 pt

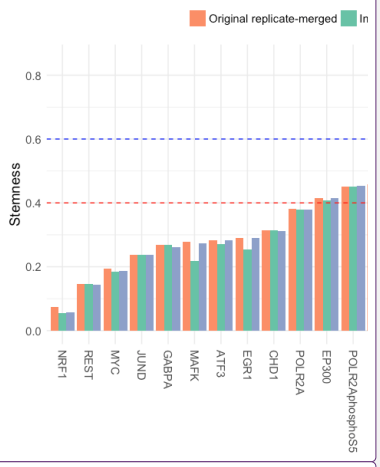
Deleted: in the recent TCGA paper

Deleted: Please see details in Excerpt 2 below.

Formatted: Font:12 pt

Deleted: 1 From . ... [143]

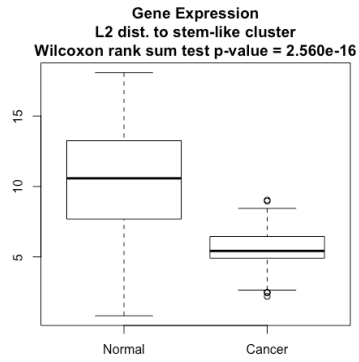
Deleted: supplement)



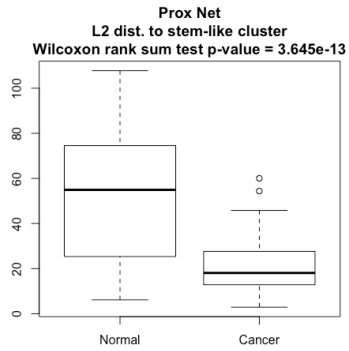
Deleted:

Deleted: 2 From . ... [144]

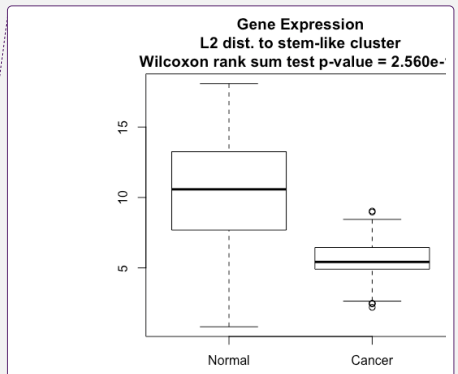
Deleted: supplement)



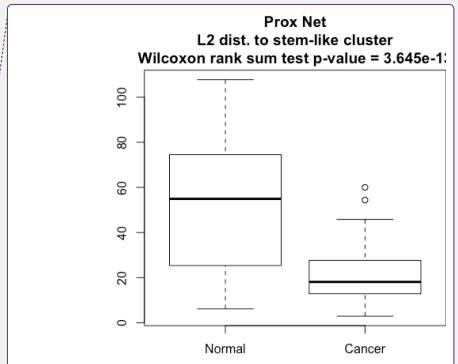
Suppl. Fig. B



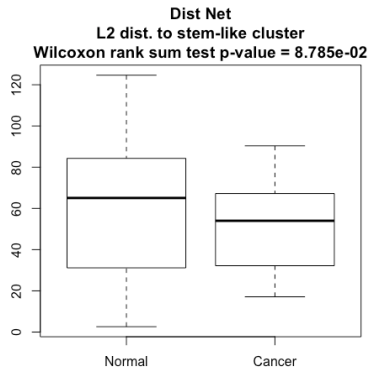
Suppl. Fig. C

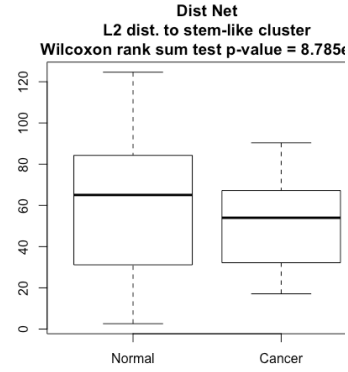


Deleted:



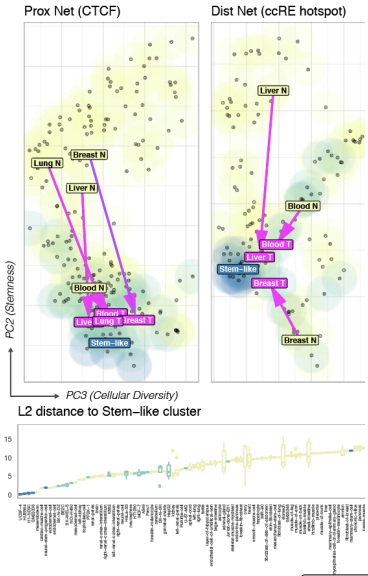
Deleted:

	
Excerpt 5.19-C (stemness in Fig.5)	Figure 5. Proximal regulatory network, distal enhancer network, and gene expression profiles have been used to explore patterns across different cell types. As expected, stem-like cell types formed a cluster, suggesting stem-like cell types have a distinct regulatory profile from normal and cancerous cell types, and stem-like cells including H1 and iPSCs have similar regulatory patterns. We find that cancerous cell types have closer distance toa state closer to stem-like clusters, suggesting cancer cells de-differentiate to a stem-like state both in their regulatory programs and gene expression profiles.



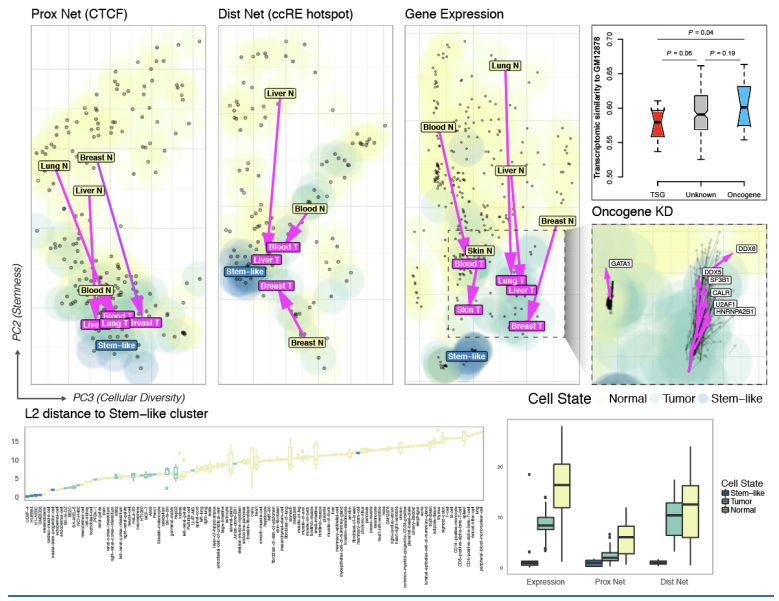
Deleted:

Moved down [16]: Figure 5. Proximal regulatory network, distal enhancer network, and gene expression profiles have been used to explore patterns across different cell types. As expected, stem-like cell types formed a cluster, suggesting stem-like cell types have a distinct regulatory profile from normal and cancerous cell types, and stem-like cells including H1 and iPSCs have similar regulatory patterns. We find that cancerous cell types have closer distance toa state closer to stem-like clusters, suggesting cancer cells de-differentiate to a stem-like state both in their regulatory programs and gene expression profiles.



Deleted: ... [145]

Moved (insertion) [16]



<ID>REF5.20 – Selection of regions for validation testing

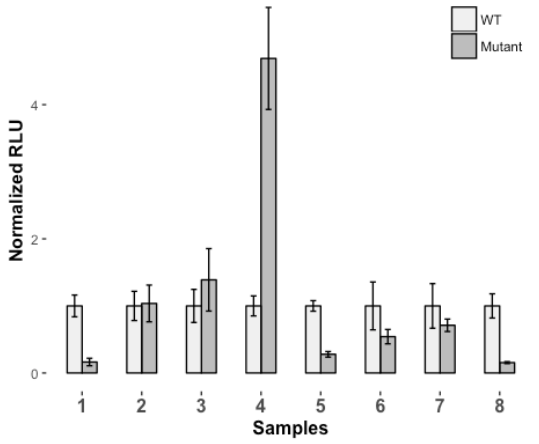
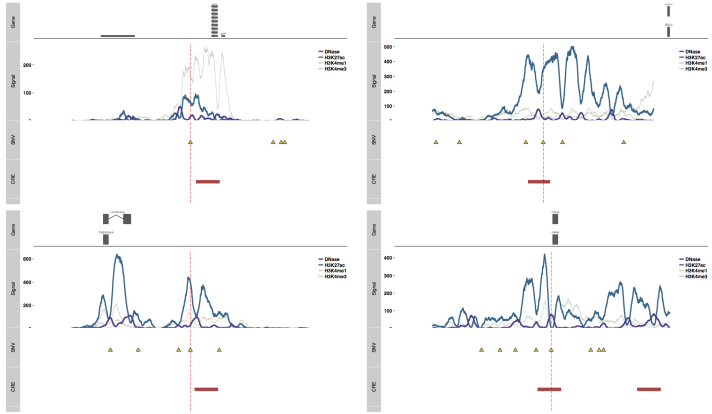
<TYPE>$$$Validation,$$$Text
<ASSIGN>@@@JZ,@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%85DONE

Referee Comment	18. How were the eight regions that were tested functionally selected? Where are these regions located in the genome, and with respect to neighboring genes? How many replicates were performed? What are the p-values?
Author Response	We thank the referee for this comment. The eight regions were selected from our integrative promoter and enhancer regulatory elements in MCF-7 cell lines. We prioritized these regulatory regions based on our integrative, stepwise variant prioritization as described in section 6.1 S. We have tried to make it more clear about the details of locations, surrounding genes, replicates and p values (Excerpt 5.20-A and Excerpt 5.20-B).

Formatted Table

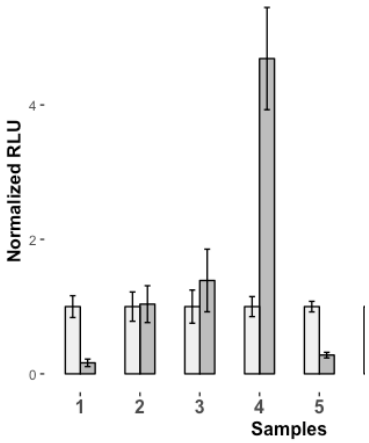
Formatted: Font:12 pt

Deleted: (see excerpt 1 below).
Excerpt 1 From ... [146]

Excerpt 5.20-A (selection, replicate, and pvalues, in suppl.)	We selected top ten regions from our proposed prioritization step and then tested their regulatory activities using luciferase assay as described in section 6.2 S. Two of ten regions we tested were failed due to issues with plasmid isolation. There were two biological replicates and three technical replicates for each biological replicate in designing luciferase assays validations. Error bar is representing 95% confidence interval across replicates. 
Excerpt 5.20-B (in suppl.)	We have provided details for the surrounding genes and genomic features of all tested regions as below. 

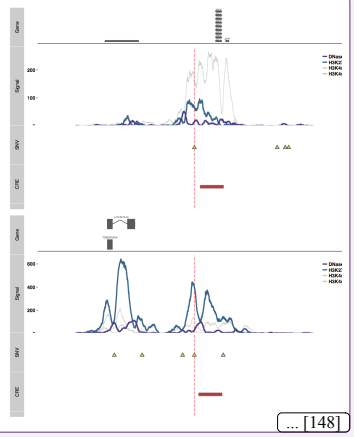
Deleted: Excerpt 1 From ... [147]

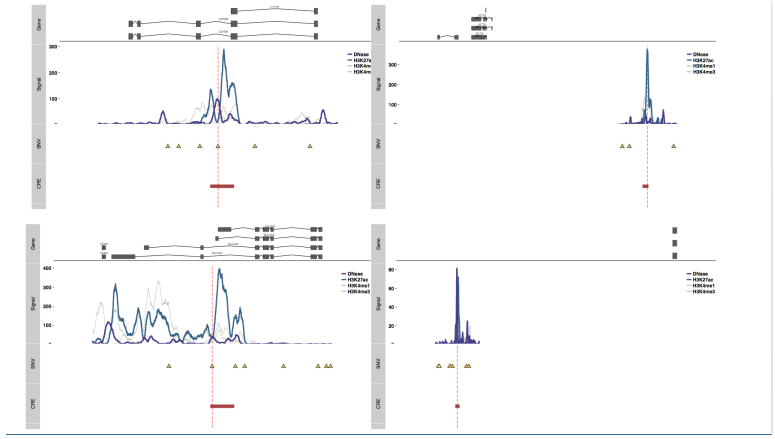
Deleted: with the highest motif breaking power

Deleted: 

Deleted: Details

Deleted: 2 From Details

Deleted:  ... [148]

<ID>REF5.21 – Presentation and revision to manuscript

<TYPE>$$$Minor,$$$Presentation,$$$Text
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	19. The authors should consider moving the general overview diagrams that constitute much of the main figures to the supplement, and in turn present data-rich figures from there with the main manuscript.
Author Response	We thank the referee for this comment. We have tried to revise the figures as requested. We have fixed figures 1 and xxx.
Excerpt 5.21-A	JZ2DL: please add new figure

Formatted Table

Deleted: for

Deleted: comments.

Deleted:

Deleted: figure XX & YY

Deleted: From

... [149]

<ID>REF5.22 – Difference between ENCODEC and existing prioritization methods

<TYPE>$$$Validation,$$$Text
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	20. It is not clear how variant prioritization differs or exceeds the variant prioritization method FunSeq published by the same group. Are they complementary approaches?
Author Response	We thank the referee to bring this up. We believe that the method that we used here is new and novel. The important aspect is that it takes advantage of many new ENCODE data and integrates over many different aspects. In particular, it takes into account the STARR-Seq data, the connections from Hi-C, the better background mutation rates, and the network wiring data, which is only possible in the context of the highly integrated and their data available on certain cell lines. We are showing this as an example of the best we can do with this level of integration. The fact that we coupled this with quite successful validation that we believe points to the great value of the integrated data.

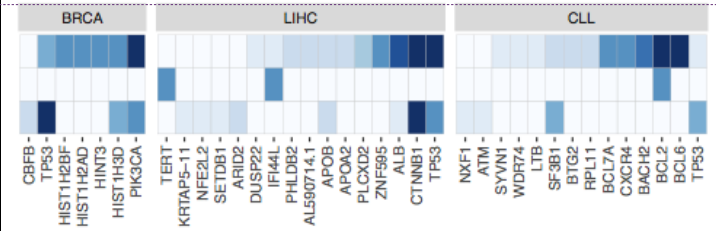
Formatted Table
Formatted: Font:12 pt
Deleted:
Formatted: Font:12 pt
Deleted: incurred
Formatted: Font:12 pt
Deleted:

<ID>REF5.23 – Minor: BMR: provide q-values

<TYPE>$$$Minor,$$$BMR
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

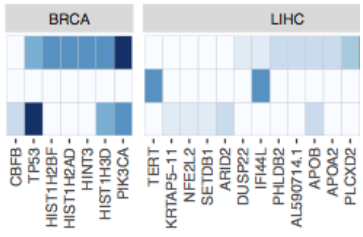
Referee Comment	21. When the authors describe recurrent events, are these significant? If so, please provide p-values (and q-values, when applicable).

Formatted Table

Author Response	We thank the referee to point this out. We have the values and q-values all deposited into our online resource and supplementary files. We have made this clearer in our revised manuscript.
Excerpt 5.23-A (in main text)	We have plotted the heatmap of p values for the recurrent analysis in three different cancer types. 

Deleted: From ... [150]

Deleted: Main figure



Deleted:

<ID>REF5.24 – Minor: Citation of previous work

<TYPE>$$$Minor,$$$Presentation
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	22. Prior work using ENCODE chromatin data to define regulatory regions and gene enhancers links should be cited (referred to in the manuscript as "Traditional methods").
Author Response	We thank the referee to point this out. References have been added in the new submission.

Formatted Table

Formatted: Font:12 pt

<ID>REF5.25 – Minor: Tumor normal comparison and composite model

<TYPE>$$$Minor,$$$CellLine
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	23. The use of a "composite normal" is not optimal for tissue or tumor-type specific analyses that the authors advocate. Although the described data resource (ENCODE) may not provide normal control data, normal tissue data from the Roadmap Epigenomics could be included instead (or in addition) to improve the quality of the tumor-normal comparisons.
Author Response	We thank the referee for bringing this out. We did noticed the Roadmap data. Actually, in the new release, ENCODE3 reprocess the complete set of roadmap data and we did include that in our data tables.
Excerpt 5.25-A	We highlighted the normal tissue data from the Roadmap (processed by ENCODE3) in our revised figure 1 as below. JZ2DL: pls add

Formatted Table

Formatted: Font:(Default) Arial Unicode MS

Deleted: (Figure 1 and supplementary table xxx).

Deleted: From ... [151]

<ID>REF5.26 –Use of H1 for stemness calculation

<TYPE>$$$Minor,$$$Stemness
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%50DONE

Referee Comment	24. The authors use the H1 embryonic stem cell line as model for "stemness" in cancer. Tumor "stemness" often resembles tissue progenitors, not embryonic stem cells. In the absence of reliable data for such progenitors the authors should note this caveat with their analysis.
Author Response	We thank the referees for bringing this point out. We mainly have chosen H1-hESC because it offers the broadest TF ChIP-seq coverage and also one of the top-tier cell lines with the most variety of experimental assays in ENCODE. We agree with the referee that the use of H1 embryonic stem cell for measuring "stemness" should be further discussed. We, therefore, have revised the manuscript with two additional analysis to show that use of H1-hESC maybe a suitable substitute for such analysis, especially in the absence of the proper progenitor cell data. In summary, we have included more stem-related samples in RNA-Seq, proximal TF network, and distal enhancer network to make the normal-tumor-stem comparisons (Excerpt 5.19-B&C). Hence, we feel that H1 is a reasonable

Formatted Table

Deleted: . As shown in excerpt 1, all stem cells tend to close to each other.

	representative of stem cells. We also added a few sentences in the revised discussion section.

<ID>REF5.27 – Minor: Validation of prioritized element

<TYPE>$$$Minor,$$$Validation
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%90DONE

Referee Comment	25. P-values should be given in Figure 6B for the luciferase reporter assay. The authors may also want to explain why candidate 5, rather than candidate 4 with a much larger expression fold difference was chosen for follow-up.
Author Response	We thank the referee for this comment. We now have added more details of how the validation of candidate regions we selected into the revised supplementary information (Excerpt 5.20-A&B). The reason we selected the candidate five instead of candidate 4 is that the candidate 5 had stronger motif breaking score when disrupted, had a higher density of TF binding events, and aligned better with our integrative regulatory region calls. However, we feel that all regions we tested are among the top prioritized ones and it is important to show these examples. In the revised manuscript, we have also included supplementary plots for all candidate regions tested in details, showing location of neighboring genes, cohort SNV data, histone marks and DHS signal tracks (Excerpt 5.20-B).

<ID>REF5.28 – Minor: SYCP2 and beyond

<TYPE>$$$Minor,$$$NoveltyPos
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%TBC
[JZ2JL: can you please do this quickly?]

Deleted: sentence

Deleted: Excerpt 1 From .
(Please refer REF5.19 for figure update.) ... [153]

Deleted: Excerpt 1 From .
(Please refer REF5.19 for figure update.) ... [152]

Formatted Table

Formatted: Font:12 pt

Deleted: please see Excerpt 2 in response to <ID>REF5.22 – Selection of regions for validation testing

Formatted: Font:12 pt

Deleted: 5

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: other

Formatted: Font:12 pt

Deleted: regions

Formatted: Font:12 pt

Deleted:

Formatted: Font:12 pt

Deleted: .

Deleted: Excerpt From .
Please see figures in Excerpt 2 in response "to <ID>REF5.22 – Selection of regions for validation testing" ... [155]

Deleted: Excerpt From .
Please see figures in Excerpt 2 in response "to <ID>REF5.22 – Selection of regions for validation testing" ... [154]

Referee Comment	26. The discovery of a previously unknown enhancer of SYCP2 is interesting. The authors should consider following up on this lead by integrating existing mutation and expression data from additional studies (e.g. 560 ICGC breast cancers from Nik-Zainal et al).
Author Response	TBC: add this quickly on Tuesday
Excerpt 5.28-A	

Formatted Table

Deleted: From ... [156]

<ID>REF5.29 – Minor: Utility of ENCODEC

<TYPE>$$$Minor,$$$Presentation
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%TBC
[JZ2MG: is it OK for the text?]

Referee Comment	27. The abstract mentions the usefulness of ENCODE data for interpretation of non-coding recurrent variants, yet this point is not explored much in the manuscript.
Author Response	We thank the referee for this comment. Actually, we tried to show in Fig 6 how each data type has been integrated to evaluate the function of variants. For example, the histone ChIP-seq, STARR-Seq, and DHS data helped to define function of surrounding element. The histone ChIP-seq, Replication timing, and Expression data help to calibrate local BMR to evaluate mutation rate and somatic burden. TF ChIP-seq/eCLIP data can help to investigate the local nucleotide effect. And Hi-C and ChIA-pet data can help to link noncoding variants to surrounding genes for better interpretation. We made this more clear in our revised manuscript.

Formatted Table

Excerpt 5.29-A	Wait for abs

Deleted: From ... [157]

<ID>REF5.30 – Minor: P-value of survival analysis

<TYPE>$$$Minor,$$$Presentation
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

Referee Comment	28. In Figure 2e, a p-value should be given with the analysis.
Author Response	We thank referee for the comment. We now have updated figure 2e with p-value.
Excerpt 5.30-A	JZ2DL: please add

Formatted Table
Formatted: Font:12 pt
Moved (insertion) [10]
Formatted: Font:Arial, 12 pt
Deleted: From

<ID>REF5.31 – Minor: Q-value of extended gene analysis

<TYPE>$$$Minor,$$$Presentation
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

Referee Comment	29. Figure 2d, q-values should be given for each identified driver gene.
Author Response	We thank referee for the suggestion. We would like to first point out that we were not focused in finding cancer drivers in this analysis. Figure 2d is to illustrate the utility of extended gene. However, we do agree with the referee that adding q-value to the figure would be important, so we have updated the figure in the revised manuscript (Excerpt 5.23-A).

Formatted Table
Deleted: .
Deleted: Excerpt From
Please see details in excerpt for REF5.23 ... [158]

<ID>REF5.32 – Minor: Presentation issue with network hierarchy

<TYPE>$$$Minor,$$$Presentation
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	30. Figure 4 would benefit from labeling of the network tiers.
Author Response	We thank reviewer for the comment. We fixed the labeling of the network tiers in the revised manuscript.
Excerpt 5.32-A	JZ2DL: please add

Formatted Table

Deleted: From . ... [159]

<ID>REF5.33 – Minor: Presentation

<TYPE>$$$Minor,$$$Presentation
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%95DONE

Referee Comment	31. In Figure 6b, it should be clarified whether "samples" refers to genomic locations, patients, or cell lines. The number of replicates for each experiment should be shown, and p-values between wt and mutant readings should be given.
Author Response	We thank the referee for pointing this issue out. We refer "samples" to the genomic locations in the submitted manuscript. We agree with the referee that this could be confusing to readers. We have updated the figure in the revised manuscript and we now refer them as candidates.
Excerpt 5.33-A	JZ2DL: please add

Formatted Table

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: From . ... [160]

<ID>REF5.34 – Minor: Supplementary document

<TYPE>$$$Minor,$$$Presentation
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

Referee Comment	32. The supplement contains multiple reference errors.
Author Response	We thank the referee for this comment and we have corrected reference errors in our supplementary document.

Formatted Table

Deleted: Excerpt From ... [162]

Deleted: Excerpt From ... [161]

Page 4: [1] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Page 8: [2] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

data we used]

1. The goal

Page 12: [3] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

From
Revised Manuscript

Page 13: [4] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

is useful from two aspects:

It

.

Page 13: [5] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

From
Revised Manuscript

Page 13: [6] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

2017

Page 13: [6] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

2017

Page 15: [7] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Newly added to the discussion section:

Page 15: [8] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

We added figures (in the supplement) to demonstrate how cell line data can show comparable performance (excerpt 2).

We added more discussion in the main text that some data types, like TF ChIP-seq, are only predominantly available in cell lines (excerpt 3).

Page 16: [9] Moved to page 46 (Move #2) jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Regarding the

Page 16: [10] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

global comparison of cell lines and tissues

We extended the normal-tumor-stem comparisons to both expression and regulatory networks (excerpt 4).

Page 16: [11] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

From
Revised Supplementary file

Page 16: [12] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Cell Type	# histone marks
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

[JZ2DL: please add the table

Page 17: [13] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Regarding the comparison of mutation rate vs features in tissue/cell lines:

Page 17: [14] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

3 From
Revised Discussion

Page 18: [15] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

4 From
Revised Supplementary file

Page 18: [16] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised supplement

Page 18: [17] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Regarding the validation of cell line conclusions on real patient data:

Page 20: [18] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 21: [19] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

3 From
Revised Manuscript

Page 21: [20] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

[JZ2DL: please fill in the xxx here]

Page 23: [21] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

:

the

Page 23: [22] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

it is more about positive selection in coding regions than BMR estimation.

the main focus

Page 23: [23] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

the coding regions, and no source code or software package is available for the whole genome.

ENCODE dramatically increased the available features from 169 (in Marticorena et al.) to 2069 (summarized in the table in supplement).

Page 23: [24] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Excerpt From Revised Manuscript (in supplement)

Table S1. Summary of ENCODE3 histone ChIP-Seq data

Cell Type	Histone ChIP-seq
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

Table S2. Summary of ENCODE3 Replication timing data

[JZ2DL: pls make such table and put it here] DL: done JZ: to disc on Tuesday

Cell Type	Repli-seq	Repli-chip
cell line	101	10
in vitro differentiated cells	0	35
primary cell	12	5
stem cell	6	11
induced pluripotent stem cell line	0	2

Page 23: [25] Moved to page 68 (Move #11) jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Table S2. Summary of ENCODE3 Replication timing data

Page 23: [26] Moved to page 68 (Move #12) jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Cell Type	Repli-seq	Repli-chip
cell line	101	10
in vitro differentiated cells	0	35
primary cell	12	5
stem cell	6	11
induced pluripotent stem cell line	0	2

Page 23: [27] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Excerpt From Revised Manuscript (in supplement)

Table S1. Summary of ENCODE3 histone ChIP-Seq data

Cell Type	Histone ChIP-seq
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

Table S2. Summary of ENCODE3 Replication timing data
[JZ2DL: pls make such table and put it here] DL: done JZ: to disc on Tuesday

Cell Type	Repli-seq	Repli-chip
cell line	101	10
in vitro differentiated cells	0	35
primary cell	12	5
stem cell	6	11
induced pluripotent stem	0	2

	cell line		

Page 23: [28] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt From Revised Manuscript (in supplement)

Table S1. Summary of ENCODE3 histone ChIP-Seq data

Cell Type	Histone ChIP-seq
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

Table S2. Summary of ENCODE3 Replication timing data
[JZ2DL: pls make such table and put it here] DL: done JZ: to disc on Tuesday

Cell Type	Repli-seq	Repli-chip
cell line	101	10
in vitro differentiated cells	0	35
primary cell	12	5
stem cell	6	11
induced pluripotent stem cell line	0	2

Page 23: [29] Moved to page 68 (Move #11) **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Table S2. Summary of ENCODE3 Replication timing data

Page 23: [30] Moved to page 68 (Move #12) **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Cell Type	Repli-seq	Repli-chip
cell line	101	10

in vitro differentiated cells	0	35
primary cell	12	5
stem cell	6	11
induced pluripotent stem cell line	0	2

Page 23: [31] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt From Revised Manuscript (in supplement)

Table S1. Summary of ENCODE3 histone ChIP-Seq data

Cell Type	**Histone ChIP-seq**
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

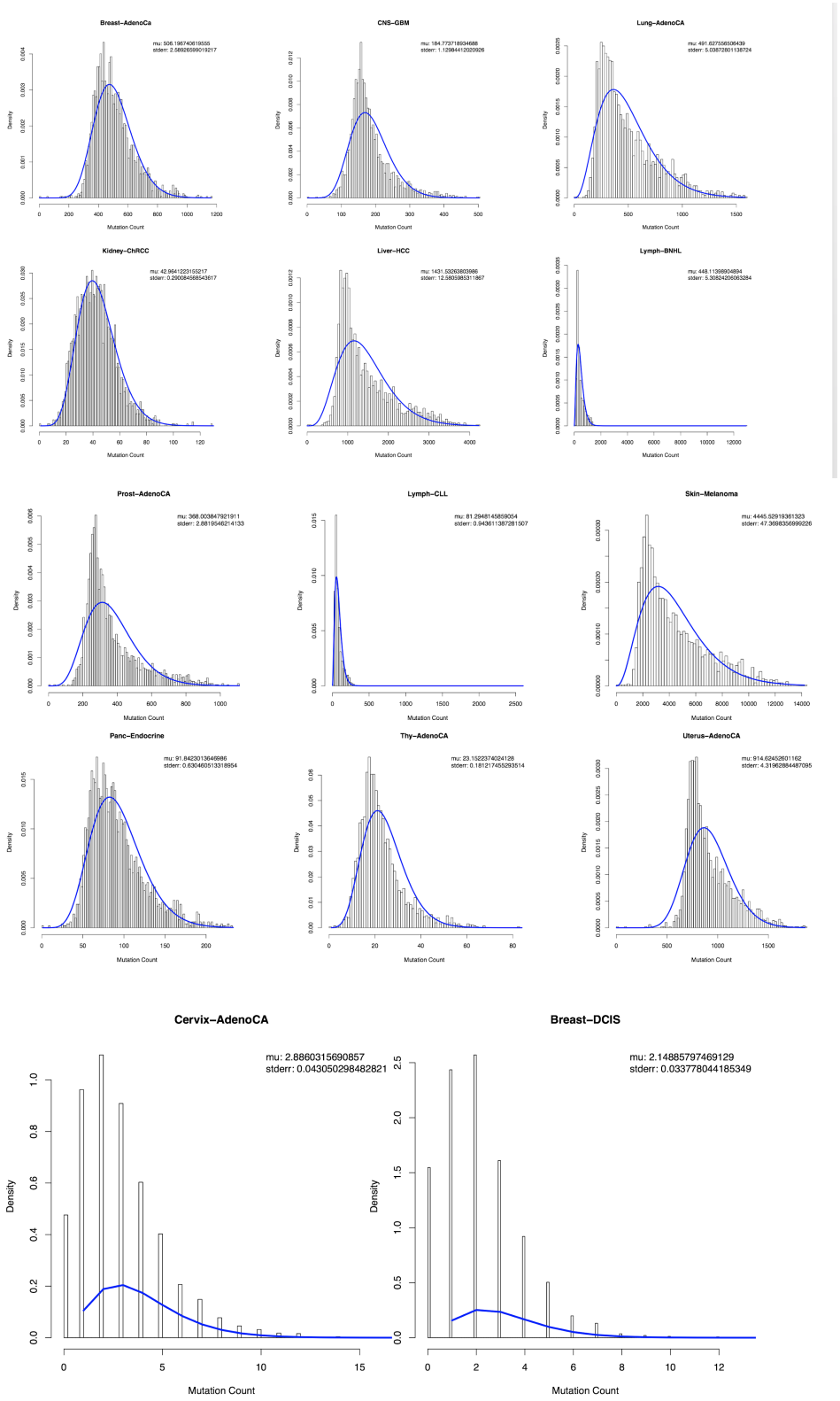
Table S2. Summary of ENCODE3 Replication timing data
[JZ2DL: pls make such table and put it here] DL: done JZ: to disc on Tuesday

Cell Type	**Repli-seq**	**Repli-chip**
cell line	101	10
in vitro differentiated cells	0	35
primary cell	12	5
stem cell	6	11
induced pluripotent stem cell line	0	2

Page 24: [32] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Supplementary file

Page 24: [33] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**



1 From
Revised Supplementary file

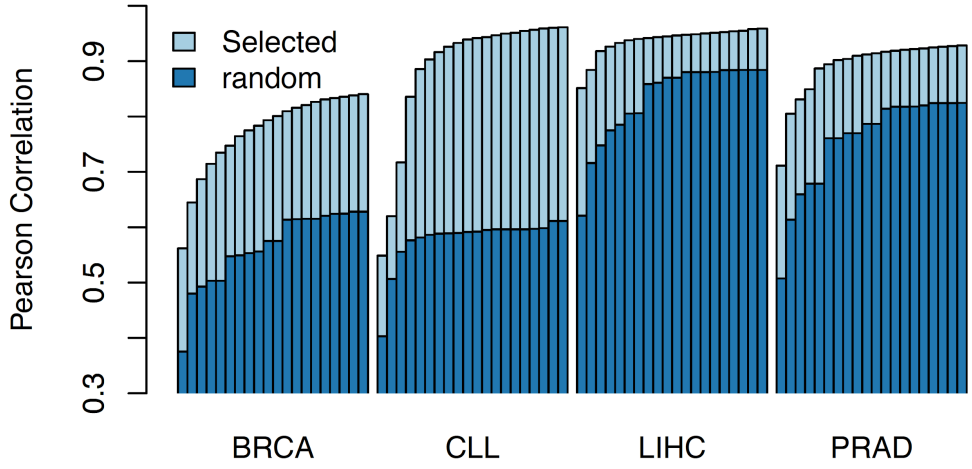
Page 26: [35] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Supplementary file

Page 28: [36] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

1 From
Revised supplement

Page 28: [37] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**



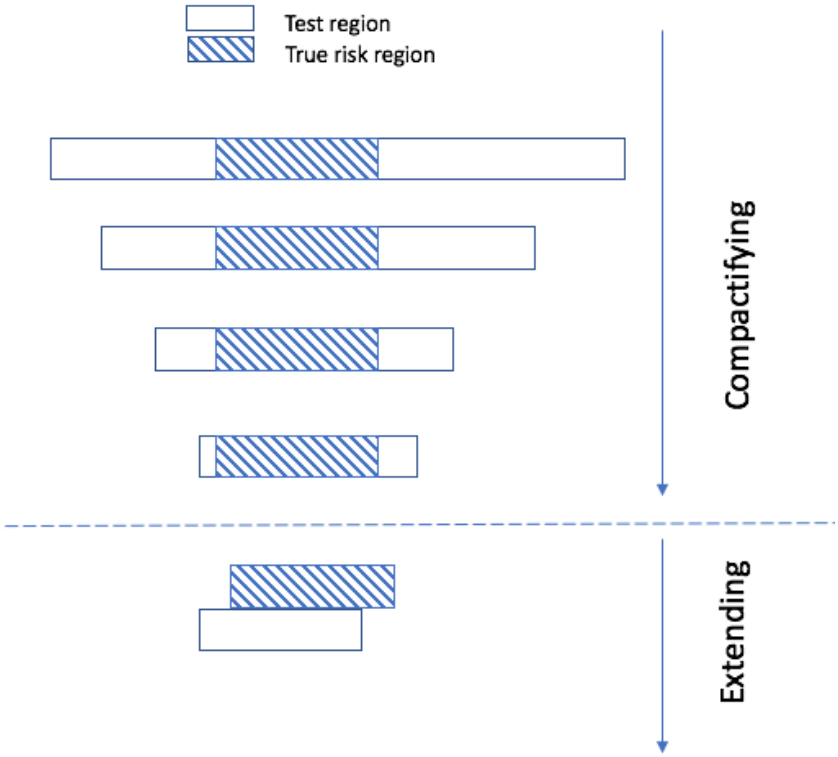
Page 28: [38] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

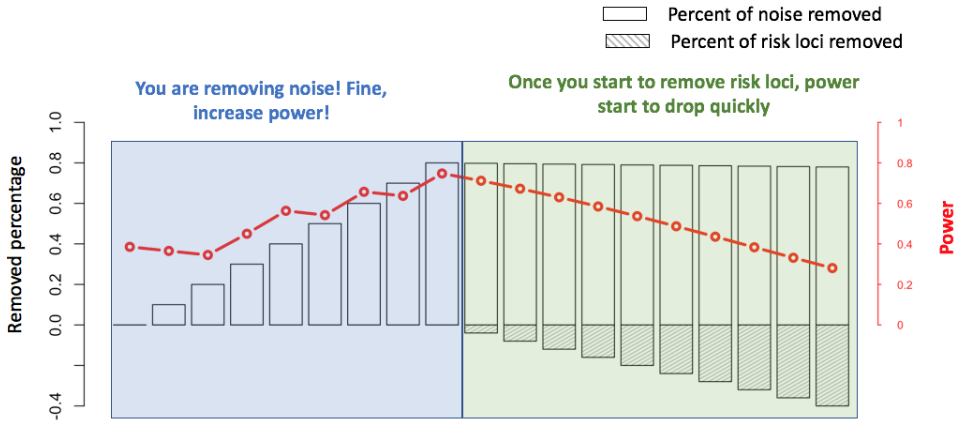
From
Revised supplement

Page 30: [39] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

1 From
Revised Supplementary file

Page 31: [40] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**





Page 31: [41] Deleted	**jingzhang.wti.bupt@gmail.com**	**5/12/18 6:25:00 AM**

From
Revised

Page 32: [42] Deleted	**jingzhang.wti.bupt@gmail.com**	**5/12/18 6:25:00 AM**

3 From
Revised Supplementary file

Page 32: [43] Deleted	**jingzhang.wti.bupt@gmail.com**	**5/12/18 6:25:00 AM**

4 From
Revised Supplementary file

Page 35: [44] Deleted	**jingzhang.wti.bupt@gmail.com**	**5/12/18 6:25:00 AM**

1 From
Revised

Page 35: [45] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised

Page 36: [46] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

3 From
Revised Supplement

Page 37: [47] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

4 From
Revised Manuscript

Page 37: [48] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript (in supplement)

Page 38: [49] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From original supplement:

Page 39: [50] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 43: [51] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 43: [52] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

JZ2DL: would you pls check Feng's email (you were cced) to double check what assays they used for the SV calling?

Page 43: [53] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 43: [54] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Supplementary file

Page 44: [55] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 45: [56] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Supplementary file

Page 45: [57] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

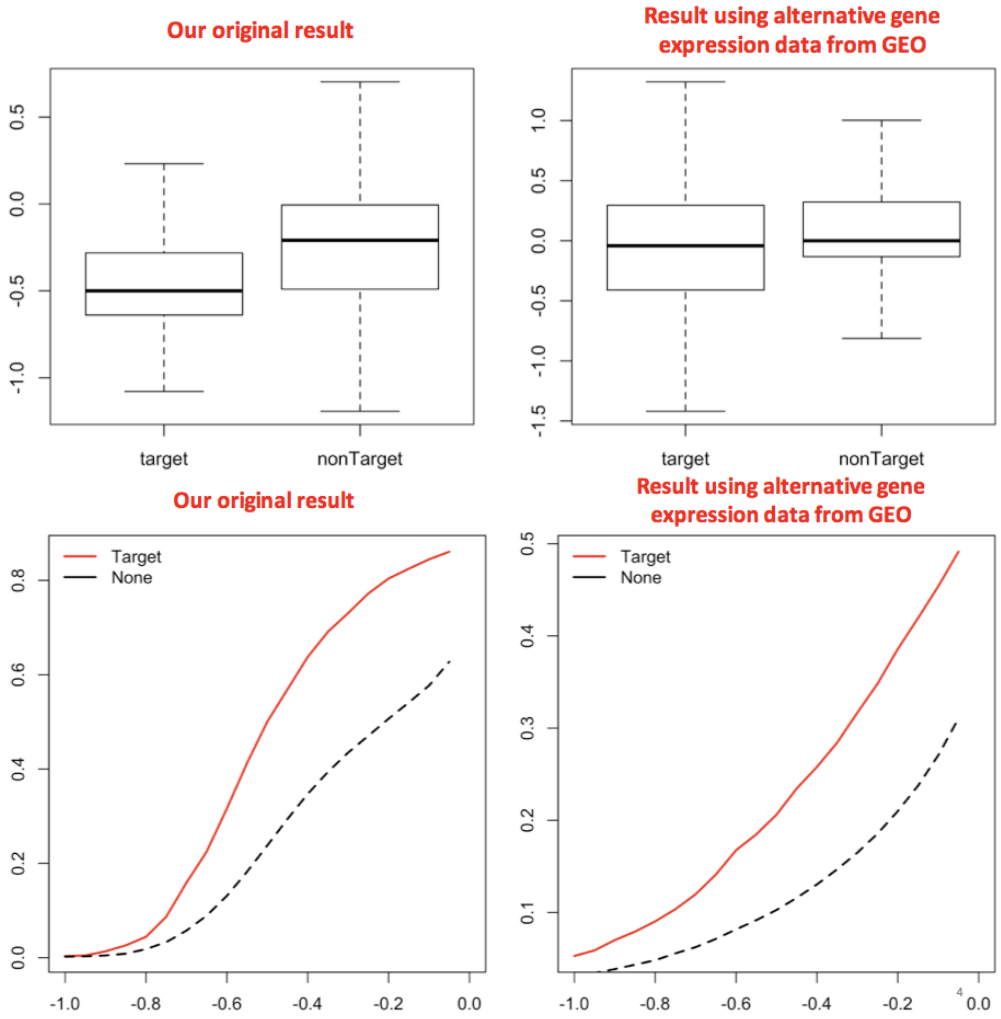
Page 45: [58] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

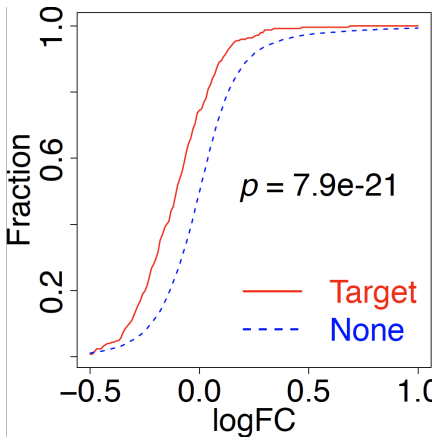
From
Revised Manuscript

Page 45: [59] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

<TYPE>$$$Annotation
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	For the Figure 2D and its description on the third paragraph of page 4 (as well as Figure 3A), did the authors validate all the genes systematically?
Author Response	We thank the referee for raising the question of validations. For Figure 2D, it is about the somatically burded genes. We fully agree with the referee that it is useful to compare our BMR to established benchmarks. We are aware of community efforts and are very involved with the PCAWG effort to do whole genome cancer analysis. One of our authors is the co-leader of the non-coding annotation group. PCAWG, which is a hybrid of TCGA and ICGC, has not developed any explicit BMR benchmark. Instead, we have provide literature support for our discovered genes and added them into a supplementary table (excerpt 1). For Fig. 3A, We have used TF/RBP knockdown experiments to validate sevral key regulators, such as MYC and SUB1. We have alse used external data to validate our conclusion. These analysis were added into our revised supplements (excerpt 2 below).
Excerpt 1 From Revised supplement	We have listed the literature supporting our discovered genes with higher than expected mutations. JZ2DL: please add the table here

Excerpt 2 From Revised supplement	We carried out these analyses after first identifying an alternative dataset. Specifically, we identified a dataset of gene expression for both MYC knockdowns (as well as a corresponding control) in Gene Expression Omnibus (GEO accession number GSE86504). For these alternative data, gene expression was measured by RNA-seq in the HT1080 cell line. We note that, even though these alternative analyses were conducted on a different cell line, the results we obtain (shown below in the right panels, and now made available in the supplementary materials) validate the behavior of the network, and they are consistent with our previous results (in which gene expression was measured in the MCF-7 cell line). These comparable results in an alternative cell line suggests that these results are robust.  We also found another array based MYC knockdown data the results correlate well with our discoveries.

<ID>REF3.8 – Quality and Validation of extended gene

<TYPE>$$$Annotation
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Referee Comment	For the Figure 2D and its description on the third paragraph of page 4 (as well as Figure 3A), did the authors validate all the genes systematically? Is there any validation rate showing the precision rate of the method?
Author Response	We thank the referee for raising this issue of quality metrics of our annotations, such as the enhancers. We fully agree with the referee that it is important to provide such information. We have struggled hard to explain the much greater accuracy of our annotations than previous effort, such as the chromHMM based enhancers purely from computation and imputed network based on DHS only. As suggested, we have added a whole section in our revised our manuscript to discuss the qualityies of annotations, including: Xxxxxxxxxx [JZ2MG: it is easy to add the QC section from other referees. However, do you think the referee is actually asking for the precision rate of variant prioritization? I am confused.]

Excerpt From Revised Manuscript	

<ID>REF3.9 – Quality of extended gene

Page 46: [60] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 50: [61] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

We found that SUB1 tends to bind to further end of 3'UTR side of a transcripts to upregualte its target gene expression in many cancer types. The regulatory activity level of SUB1 is significantly associated with patient survival. In our revision, we have investigate deep into the biology of SUB1, including

We investigated SUB1 regulation potential in different cancer types and found that they are consistent as below (excerpt 1 below).

We added several examples of keys SUB1 target oncogenes using SUB1 knockdowns (excerpt 2 below).

We also hyposize that SUB1 tends to bind to the 3'UTRs to stabilize its target mRNA. The decay rate of SUB1 is slower than non-targets (excerpt 3 below).

We found SUB1 is a direct target of MYC in various cancer types. These factors showed significant co-regulation, even after correcting several covariates. We suspect that that SUB1 may stabilize the MYC target genes and pathways to promote the malignant growth of cancer cells. (excerpt 4 below).

We performed SUB1 and MYC knockdowns and validated their regulation effects on key oncogenes using qPCRs (excerpt 5 below)

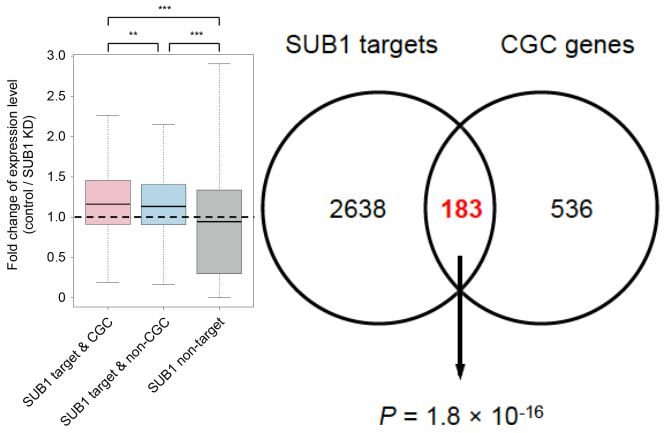
Page 50: [62] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

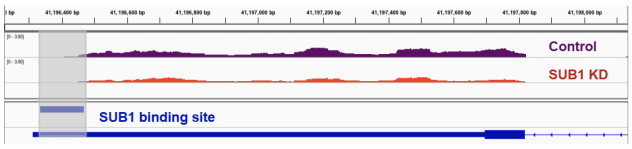
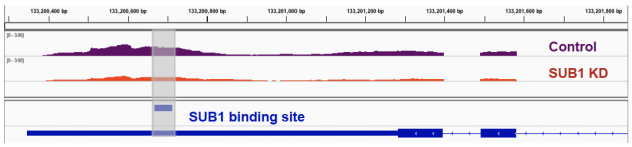
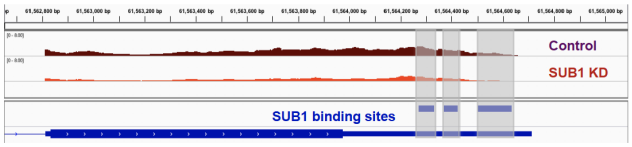
1 From
Revised Supplement

Page 50: [63] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

2 From
Revised Supplement

Page 51: [64] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**



Gene	Functions	PMID	Expression profiles of the 3' UTR
BRCA1	The gene is involved in maintaining genomic stability	12677558, 17416853, 23620175, 16551709	
POLE	The gene is involved in DNA repair and replication	26133394, 28423643	
FEN1	The gene is involved in DNA repair and replication	20929870, 22586102	

Page 51: [65] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Supplement

Page 52: [66] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

4 From
Revised Supplement

Page 53: [67] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt 5 From Revised Supplement	Feng's validations

Page 53: [68] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt 5 From Revised Supplement	Feng's validations

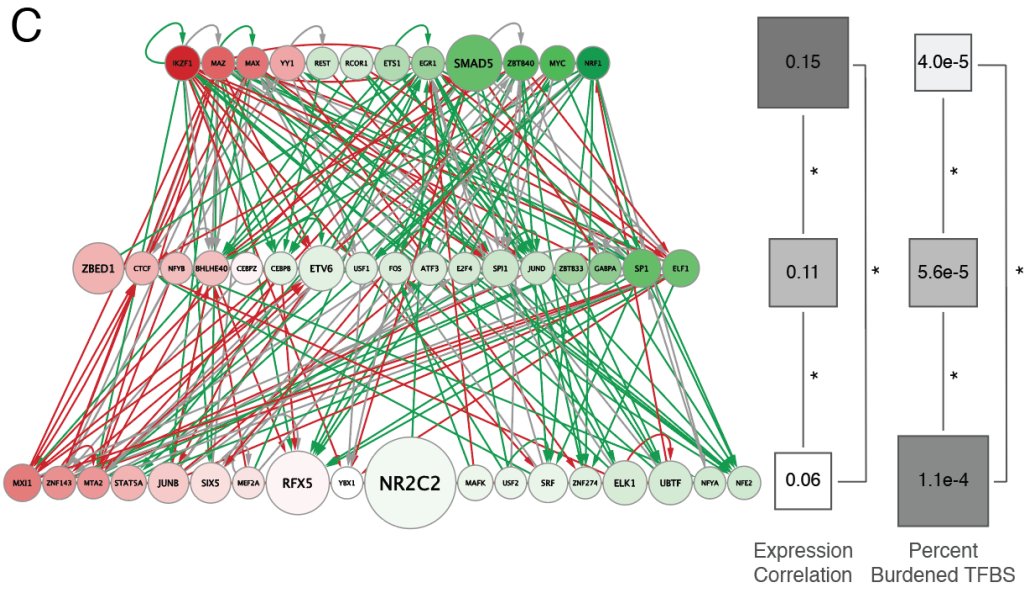
Page 54: [69] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 54: [70] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 55: [71] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**



Page 58: [72] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

We admit that maybe this construction is not that intuitive.

Page 58: [73] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Page 59: [74] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

analysis to address this question, including

A

Page 59: [75] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

1 below).

JZ2DL: imputed vs imputed network?

Page 59: [76] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 62: [77] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Page 62: [78] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

2 From
Revised Supplement

Page 62: [79] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Page 63: [80] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Main Manuscript

Page 64: [81] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

, including (JZ2DL: please fill in xxx)

Called SNV and SVs in xxx top-tier cell lines using integrative data, including WGS, Hi-C, and others (excerpt 1)

A supplementary figure to relate SNV to SVs to examine effect of SVs on SNV inmatched cell lines (excerpt 2)

A figure panel in updated Fig.2 regarding the relationship between SVs and several histone modification marks (excerpt 3)

Highlighted several examples in supplementary files to show the SV introduced enhancer gain/loss events and relate them to gene expression changes (excerpt 4)

A new figure panel in Figure 5 to estimate the number of rewiring regulatory edge affected by SV events (Excerpt 5)

A new CRISPR based validation on SV effects onlong range interactions activating the well-known oncogene ERBB4 (Excerpt 6)

Page 64: [82] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt 1 From
Revised Supplement

Page 64: [83] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

2 From
Revised Supplement

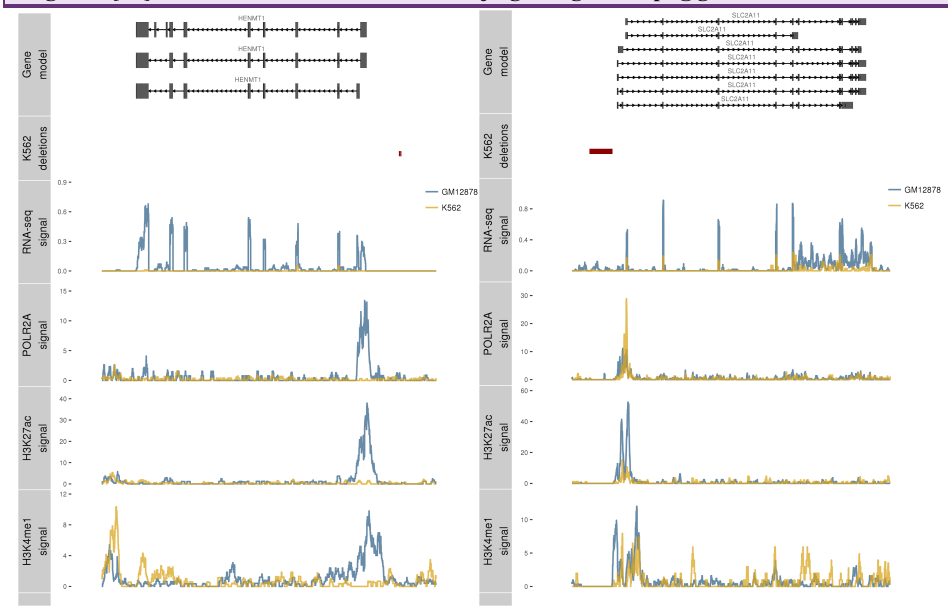
Page 64: [84] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

3 From
Revised Manuscript

Page 65: [85] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Supplement

Page 65: [86] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**



Page 66: [87] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

5 From
Revised Manuscript

Page 66: [88] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

6 From
Revised Manuscript

Page 68: [89] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt From
Revised Manuscript

Page 68: [90] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript and supplement

Page 68: [91] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

In the main text:
Instead, our key point is that the ENCODE3 rollout dramatically expands thegenomic data available for this type of regression by more than a factor of 10 (2069 vs.

Page 68: [92] Commented **Patrick McGillivray** **5/4/18 10:40:00 PM**

Are we defending not having perfect cell line matches?

It's not clear that using different data sets provides a best overall fit to mutation rate. Perhaps one cell type dominates the tumor mutation rate or is most relevant. It's also not clear that data should be combined into an overall fit, rather than each cell type treated individually.

Page 68: [93] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

[1]

In supplement:

In total there are 2017 histone ChIP-seq and 52 Replication timing features to predict BMR.

Page 68: [94] Moved to page 68 (Move #8) jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

We did a PCA of the signals from these features and selected the best combination of 20 PCs for BMR prediction. It is worth pointing out that the majority of our data is

Page 68: [95] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

fromtissue or primary cells. A summary of cell types for these features is given below.

Page 68: [96] Moved to page 68 (Move #9) jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Summary of ENCODE histone ChIP-seq data

Page 68: [97] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Cell Type	# histone marks
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

[

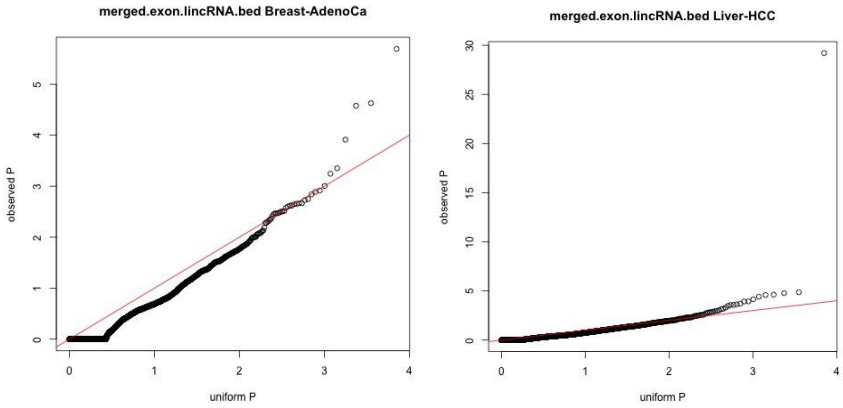
Page 68: [98] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

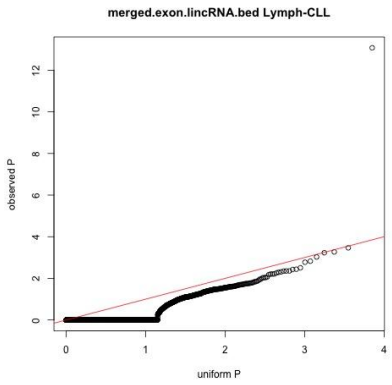
From
Revised Manuscript

Page 70: [99] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

From
Revised Supplement

Page 70: [100] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM





Page 72: [101] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

For instance, the new ENCODE3 data used in this paper includes:

2017 histone ChIP-Seq data (1339 from tissues/primary cells vs. 169 in Marticorena et al. 2017)
52 replication timing datafrom xx tissues (as compared with 16 in Polak et al. 2015)
Xxx TF ChIP-Seq from xxx cell types (vs. xx in ENCODE2)
Xxx tumor-normal matched TF ChIP-Seq for xxx cancer types (vs. xxx for only K562 in ENCODE2)
Xxx TF knockdown data to xxx in xxx cell types (vs. xx in ENCODE2)
A number of novel assays, such STARR-Seq, Hi-C, ChIA-PET, and eCLIP[2]

Page 77: [102] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

1 From
Revised supplement

Page 77: [103] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

From
Revised Manuscript

Page 77: [104] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

3 From
Revised Manuscript

Page 78: [105] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

4 From
Revised Manuscript

Page 79: [106] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

Page 83: [107] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

Excerpt 2 From Revised Supplementary file	Regarding extended genes [3]

Page 83: [108] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Excerpt 2 From Revised Supplementary file	Regarding extended genes [4]

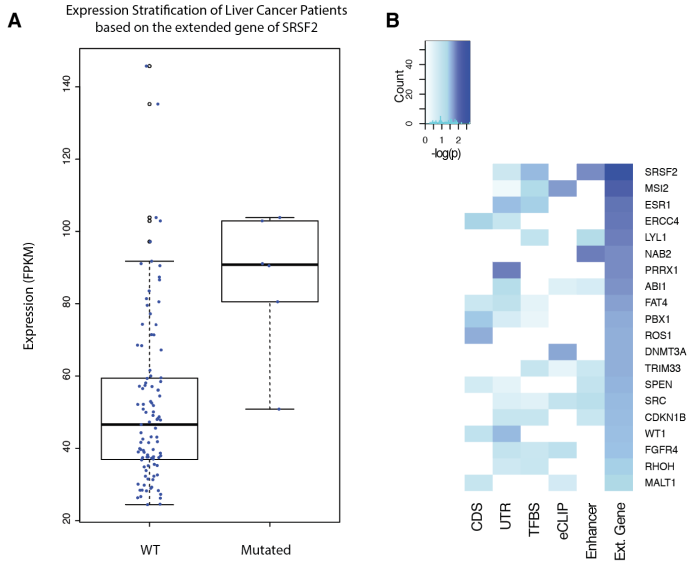
Page 84: [109] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

1 From
Revised Manuscript (

Page 85: [110] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

2 From
Revised figure and supplementary text

Page 85: [111] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM



Page 86: [112] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Page 86: [113] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

Through the process of this revision, we noticed that there is no gold standard to define enhancers in human, so it is difficult to directly call false positives.

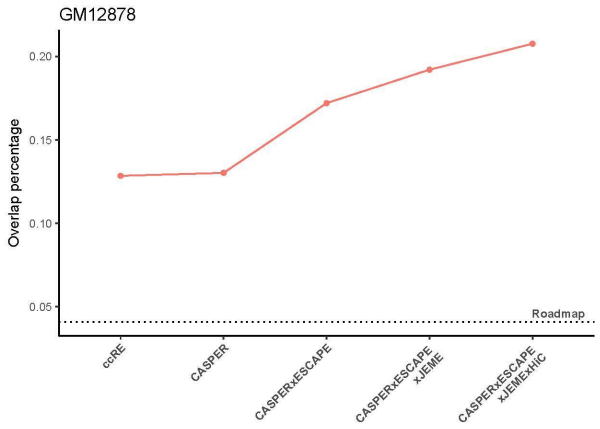
Instead, we calculated the overlapping percentage with the FANTOM enhancers using our annotations and showed that by incorporating more assays, the overlapping percentage increases significantly -- consistently higher than those from the Roadmap and the main encyclopedia enhancers. Please see details in the following excerpt for more information.

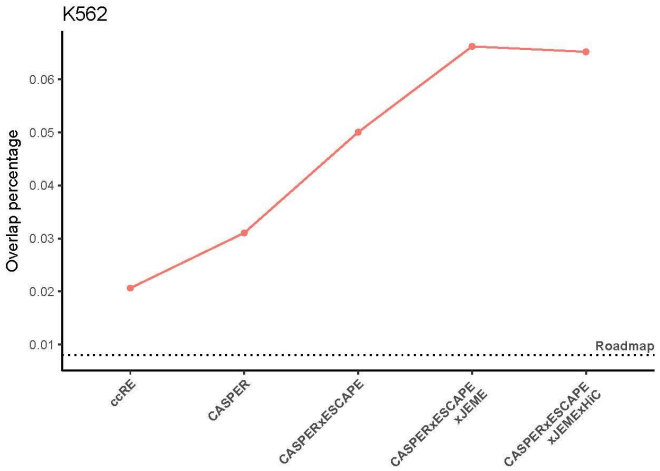
[JZ2JZ: talk to MTG to

Page 86: [114] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

[5]

Page 87: [115] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM





Page 89: [116] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

2 From
Revised Manuscript (in supplement

Page 91: [117] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

From
Revised Manuscript

Page 91: [118] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

From
Revised Manuscript

Page 92: [119] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

From
Supplement

Page 93: [120] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

From
Revised Manuscript

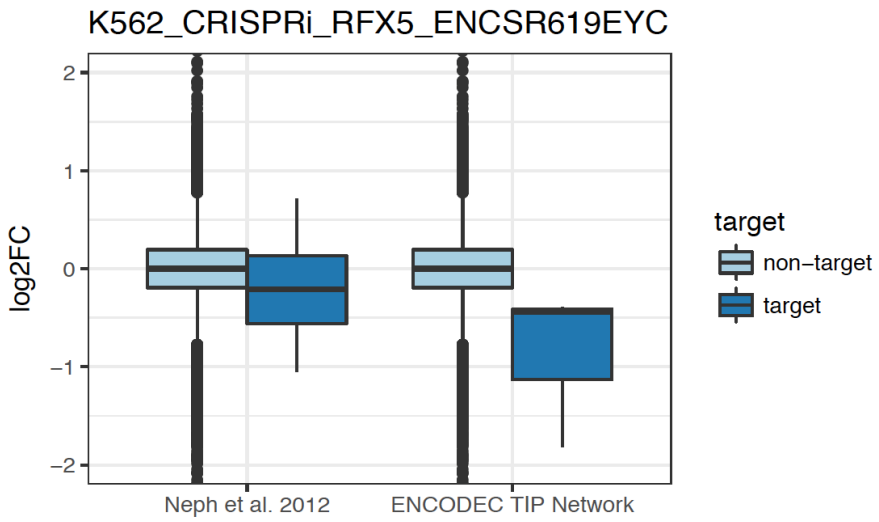
Page 94: [121] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

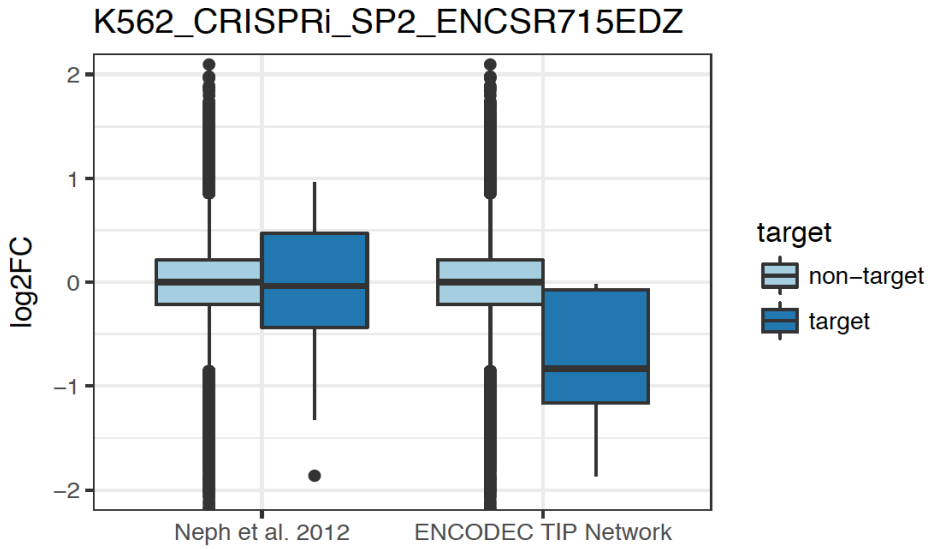
estimation of the ChIP-Seq based networks
The

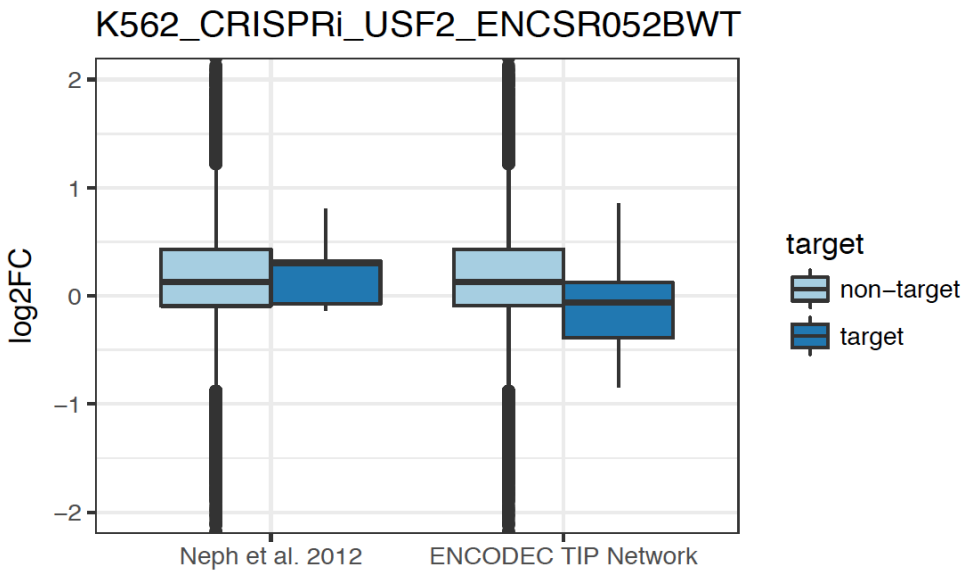
Page 95: [122] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

We have compared the quality of our enhancer target prediction linkages with other computational based methods and our results showed superior quality. Details please see REF

Page 96: [123] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM


K562_CRISPRi_RFX5_ENCSR619EYC
log2FC
2
1
0
−1
−2
Neph et al. 2012
ENCODEC TIP Network
target
non−target
target

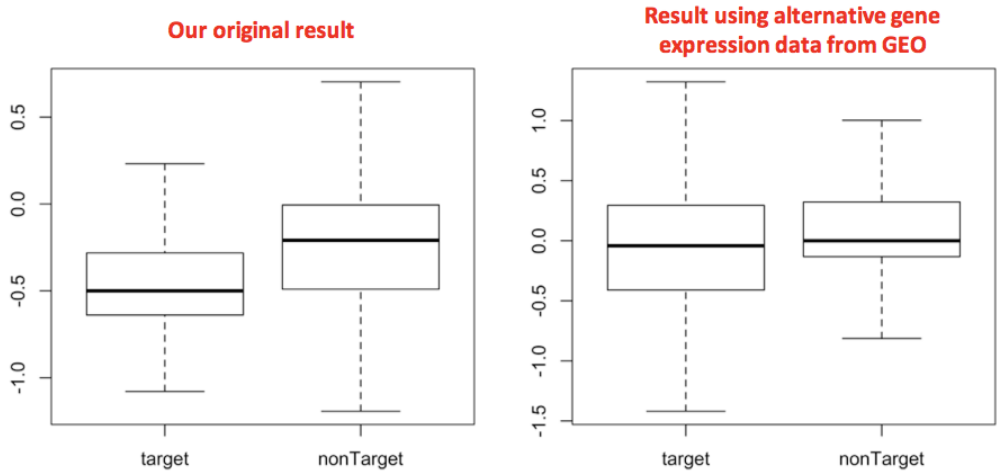

K562_CRISPRi_SP2_ENCSR715EDZ
log2FC
2
1
0
−1
−2
Neph et al. 2012
ENCODEC TIP Network
target
non−target
target

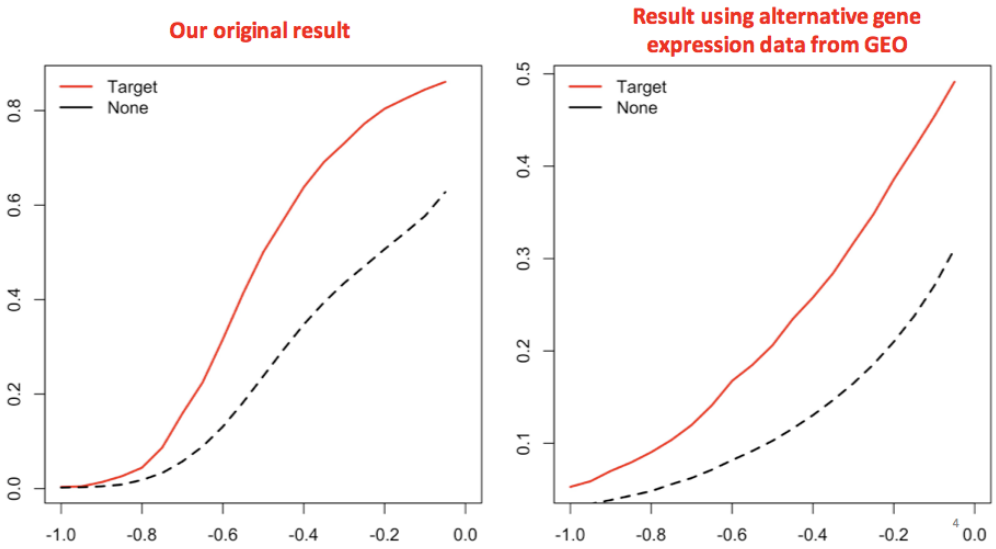


Page 98: [124] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 98: [125] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**





Page 100: [126] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

In summary, we were able to elaborate on this considerably in our revised version, including

We investigated SUB1 regulation potential in different cancer types and found that they are consistent as below (see excerpt 1 below).
We added several examples of keys SUB1 target oncogenes using SUB1 knockdowns (see excerpt 2 below).
We also hyposize that SUB1 tends to bind to the 3'UTRs to stabilize its target mRNA. The decay rate of SUB1 is slower than non-targets (p value=1.91e-10).

We investigated SUB1 regulation potential in different cancer types and found that they are consistent as below (see excerpt 1 below).

We compared the SUB1 targets with other TFs and found that MYC showed significant co-regulation, even after correcting several covariates. Details please see excerpt 3 below. We suspect that that SUB1 may stabilize the MYC target genes and pathways to promote the malignant growth of cancer cells.

Page 100: [127] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Sub1 regulated by myc

[JZ2MG: the highlighted part is way too strong, and I would like not to be that negative about ourselves. Suggested change, Though it may not represent a complete novel finding in cancer biology,]

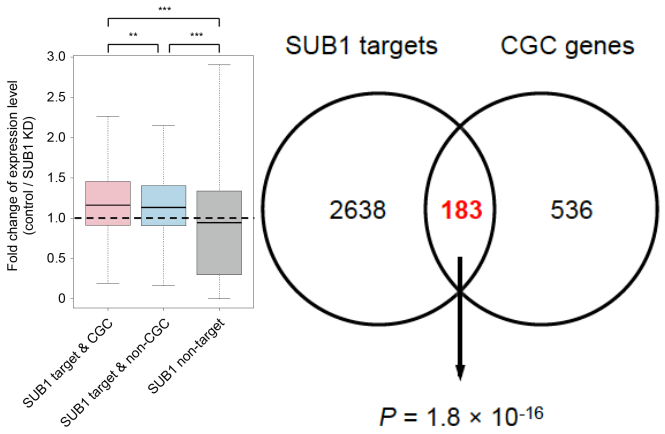
Page 100: [128] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

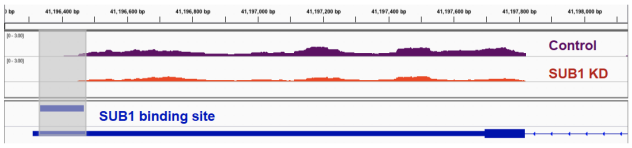
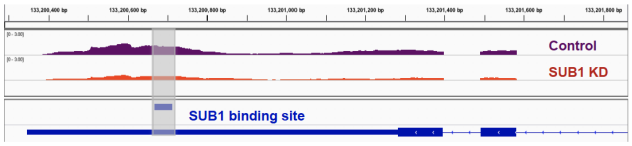
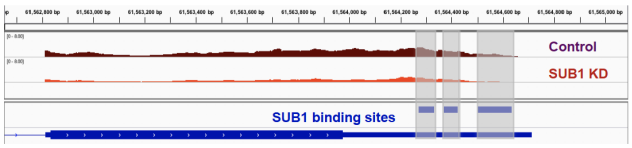
1 From
Revised Manuscript (

Page 100: [129] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

2 From
Revised Manuscript (

Page 100: [130] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM



Gene	Functions	PMID	Expression profiles of the 3' UTR
BRCA1	The gene is involved in maintaining genomic stability	12677558, 17416853, 23620175, 16551709	
POLE	The gene is involved in DNA repair and replication	26133394, 28423643	
FEN1	The gene is involved in DNA repair and replication	20929870, 22586102	

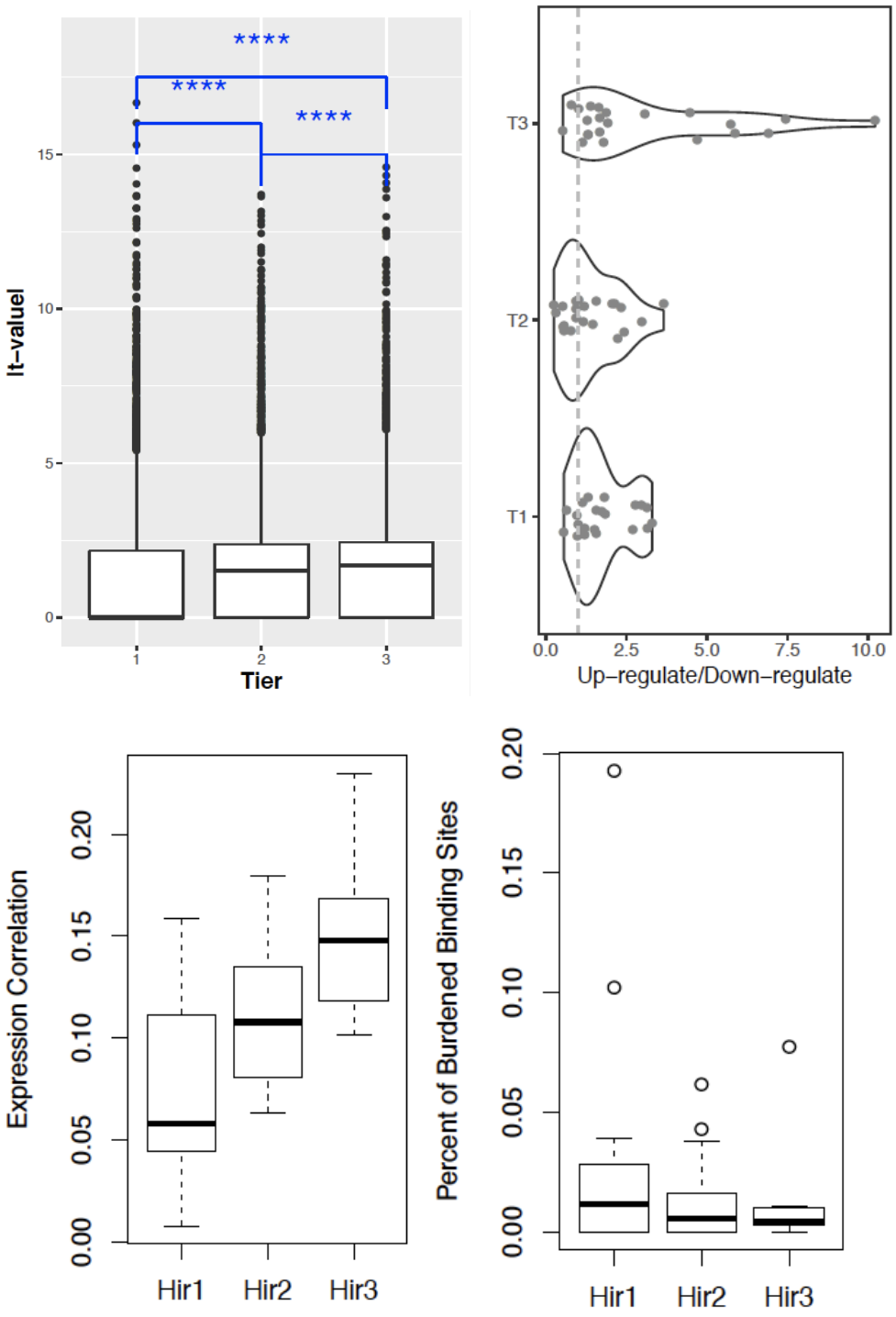
Page 102: [131] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

3 From
Revised Manuscript (

Page 105: [132] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM

From
Revised Manuscript

Page 105: [133] Deleted jingzhang.wti.bupt@gmail.com 5/12/18 6:25:00 AM



Page 107: [134] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

From
Revised Supplement

Page 108: [135] Formatted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

Font:12 pt

Page 108: [135] Formatted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

Font:12 pt

Page 108: [135] Formatted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Font:12 pt

Page 108: [136] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

model. In summary, we have done the following

Page 108: [137] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

As we answered earlier in REF5.14, we derived our TF networks from ChIP-seq experiments. The ENCODE consortium has always enforced a strict data quality standards for all ENCODE produced transcription factor ChIP-seq experiments, which allow us to rigorously control for the false positives. Please refer to Excerpt 3 in response to "REF5.14 – ChIP-seq vs other computational based networks".

We then tried to measure the baseline of rewiring using replicates of ChIP-seq experiments, as we explored in REF5.18. We find that

Page 108: [138] Formatted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Font:Helvetica Neue

Page 108: [138] Formatted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Font:Helvetica Neue

Page 108: [139] Formatted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Indent: Left: 0", Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 108: [140] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Regarding

Page 108: [141] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

estimated by fractions

Using replicates of H1-hESC ChIP-seq experiments, we made two independent H1 networks in addition to original replicate merged H1 network, and we made recalculated stemness of TF, whether they rewire toward or away from H1. We find that the results of all of stemness direction is reproduced using either replicate. Please see details in

Page 109: [142] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

We extended our analysis of H1 to RNA-Seq, TF ChIP-Seq (proximal and distal), and TF knockdown data (details in the Excerpt below). We were able to run

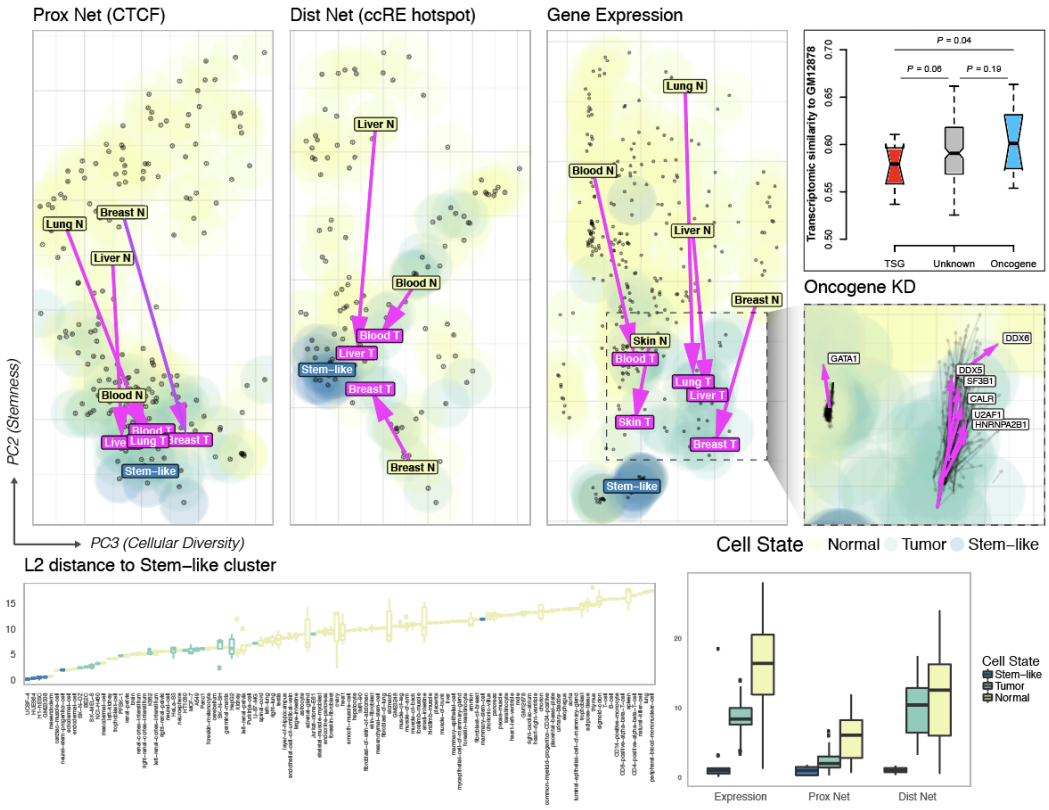
Page 109: [143] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

1 From
Revised Manuscript

Page 109: [144] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

2 From
Revised Manuscript(

Page 111: [145] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM



Page 112: [146] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

(see excerpt 1 below).

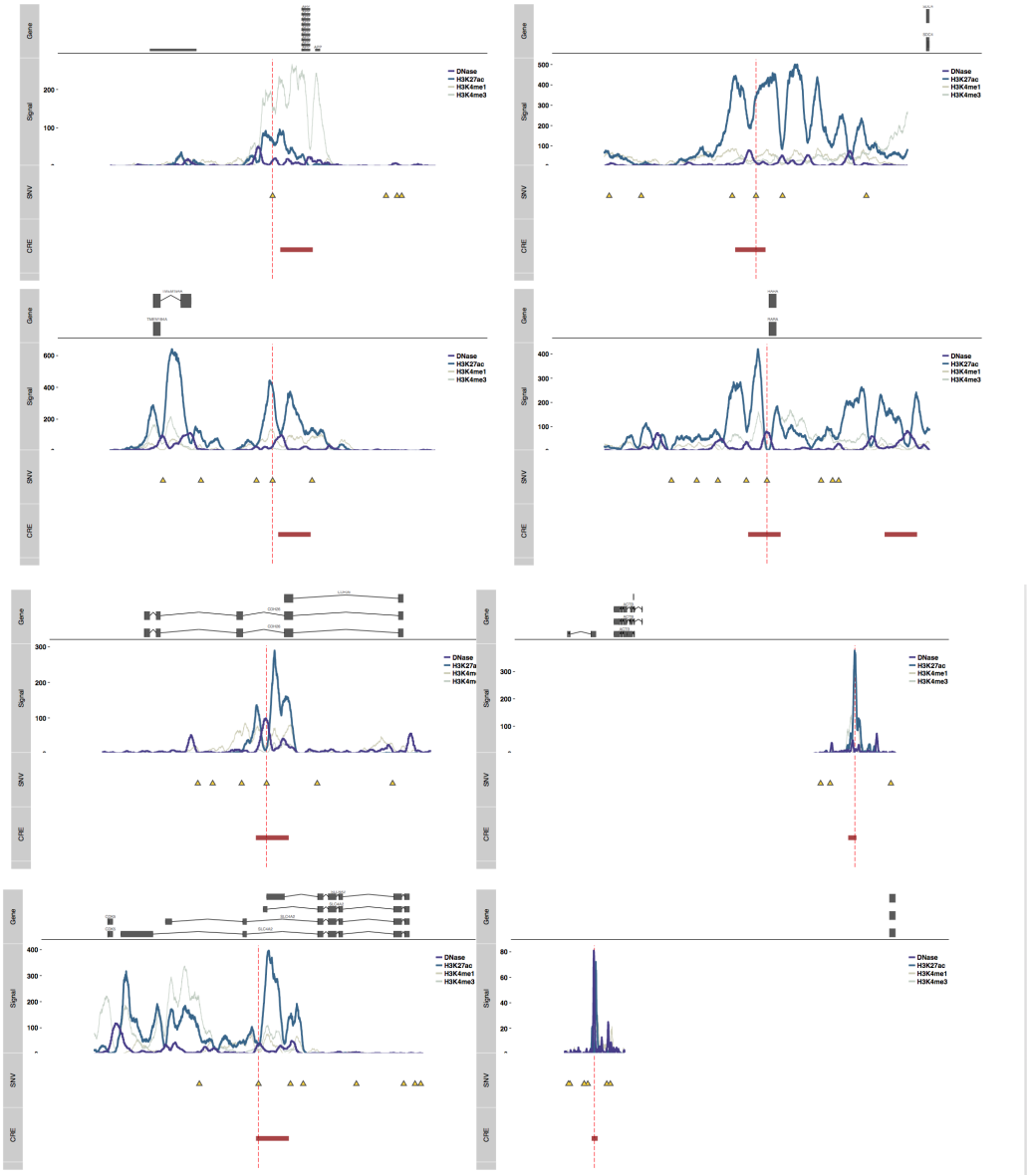
JZ2MG: previously we mentioned that we selection these variants based on motif breaking but I feel that is not good. Could we say we do the prioritization based on procedures in figure 6? Is this dangerous?

There are two individuals independently performed the experiment and each individual did three replicates for each region. So there are 6 replicates for each tested region. We provided the error bar with 95% confidence interval after merging the replicates. All the raw data are in the supplementary file in our initial submission. We also IGV plots for all the other regions in the supplementary file showing the genomic features and [6]the nearby genes (see excerpt 1 below).

Page 113: [147] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

Excerpt 1 From
Revised Manuscript

Page 113: [148] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM



Page 114: [149] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

From
Revised Manuscript

Page 116: [150] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

From
Revised Manuscript

Page 117: [151] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

From
Revised Manuscript

Page 118: [152] Deleted	jingzhang.wti.bupt@gmail.com	5/12/18 6:25:00 AM

Excerpt 1 From Revised Manuscript	(Please refer REF5.19 for figure update.)

Page 118: [153] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt 1 From Revised Manuscript	(Please refer REF5.19 for figure update.)

Page 118: [154] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt From Revised Manuscript	Please see figures in Excerpt 2 in response “to <ID>REF5.22 – Selection of regions for validation testing”

Page 118: [155] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt From Revised Manuscript	Please see figures in Excerpt 2 in response “to <ID>REF5.22 – Selection of regions for validation testing”

Page 119: [156] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 120: [157] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 120: [158] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt From Revised Manuscript	Please see details in excerpt for REF5.23

Page 121: [159] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 121: [160] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

From
Revised Manuscript

Page 122: [161] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt From Revised Manuscript	

Page 122: [162] Deleted **jingzhang.wti.bupt@gmail.com** **5/12/18 6:25:00 AM**

Excerpt From Revised Manuscript	