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

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

Formatted: Font color: Auto

Deleted: Keep it more compact, and mentioned as what we mentioned in our email ... [1]

Formatted Table

<STATUS>%%%TBC

Format:

Referee Comment: Courier New, 10pt

Author Response: Helvetica Neue, 12pt

Excerpt From Revised Manuscript: Times New Roman, 10pt

Formatted: Font:10 pt

Formatted: Font:10 pt

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

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 referee's concerns. In the attached sheets which have a point by point response.

We corresponded a bit about this manuscript before so I will be brief here and simply say that we consider this paper as 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.

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

We hope you like the manuscript and we look forward to hearing from you.

Yours sincerely,
marK

Deleted:

Formatted: Normal, Justified, Space Before: 0 pt, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Deleted: orli

Formatted: Font:Helvetica Neue, 12 pt

Formatted: Font:Helvetica Neue, 12 pt

Moved (insertion) [1]

Formatted: Font:Helvetica Neue, 12 pt

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

Moved (insertion) [2]

Formatted: Font:Helvetica Neue, 12 pt

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

Deleted: ... [2]

Editor:

<ID>REF 0.1 - Overall comments on the paper

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

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-

Formatted: Font:10 pt
Formatted Table
Deleted: respond to extensively
Deleted: in
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: new version. In summary, we have answered most of these
Formatted: Font:12 pt
Deleted: were
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: in large
Formatted: Font:12 pt
Deleted: , particularly the
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: The
Formatted: Font:12 pt
Deleted: - Extended gene
Formatted: Font:12 pt
Deleted: both
Formatted: Font:12 pt
Deleted: analysis ... [3]
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted:
Formatted: Font:12 pt
Deleted: SVs
Formatted: Font:12 pt
Deleted:
Formatted: Font:12 pt
Deleted:
Formatted: Font:12 pt
Deleted: One area that we wish to clarify a little on is to ask us to compare our calculations to that for driver identification. We think that the value of our paper was misunderstood by some of the reviewers. The point of this paper is not to develop a novel method of dri ... [4]
Formatted: Font:12 pt
Formatted: Font:12 pt, Bold
Formatted: Font:12 pt
Deleted: For example, usually a tumor sample cor ... [5]
Formatted: Font:12 pt

	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 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

Formatted ... [6]
Deleted: .
Formatted ... [7]
Deleted: to
Formatted ... [8]
Deleted: as
Formatted ... [9]
Deleted:
Formatted ... [10]
Deleted: really
Formatted ... [11]
Deleted: in our analysis
Formatted ... [12]
Deleted: improved the
Formatted ... [13]
Deleted: estimationnt
Formatted ... [14]
Formatted Table ... [15]
Formatted ... [16]
Deleted: . We want
Formatted ... [17]
Deleted: the prior studies have been cited
Formatted ... [18]
Deleted: , such as
Formatted ... [19]
Deleted: came out
Formatted ... [20]
Deleted: Campbell's lab
Formatted ... [21]
Deleted:), two and half months after we submitte ... [22]
Formatted ... [23]
Deleted: the Martincorena et al 2017 paper is
Formatted ... [24]
Deleted: about
Formatted ... [25]
Deleted: as
Formatted ... [26]
Deleted: .
Formatted ... [27]
Deleted: cited
Formatted ... [28]
Deleted: made
Formatted ... [29]
Deleted: is quite unfair for us
Formatted ... [30]
Deleted: make
Formatted ... [31]
Deleted: it. Actually
Formatted ... [32]
Deleted: came out linking
Formatted ... [33]

	al 2018. We strongly believe that our ENCODEC resource would benefit such analyses, so we have updated our reference list in this revised version. *"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.*

Deleted: analysis and hence

Deleted: the

Formatted: Font:12 pt

Formatted: Font:12 pt

Formatted: Font:12 pt

<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:***

Formatted: Font:10 pt

Formatted Table

Moved (insertion) [3]

Formatted: Font:12 pt

Formatted: Add space between paragraphs of the same style, No bullets or numbering, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Font:12 pt, Bold, Italic, Underline

Deleted: Have more validations than other paper, tons of unique validations ... [34]

- this paper should be 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 is focused on broad and universal annotations (for all cell types) based on 4 assays [JZ2MG: do you say >=20 assays?]
- this is the only paper in ENCODE that provides comprehensive networks from ENCODE3 (JZ2MG: can we say we are the only paper representing the functional charaterization centerl? Or some PI from there? Is this confidential to Orli, can the reviewers see it?)

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

Formatted: Font:12 pt

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"

Deleted:

Formatted: Font:12 pt

Deleted: (2)

Formatted: Font:12 pt

Deleted: current Encyclopedia *package is not meant to be structured like previous packages* (i.e. '12 ENCODE). The

Formatted: Font:12 pt

Deleted: analysis is meant to be spread over a number of papers and not centered on a single one. This

Formatted: Font:12 pt

Deleted: is in fact meant to be

Formatted: Font:12 pt

Deleted: integrative analysis paper of

Deleted: package

Formatted: Font:12 pt

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: (3) note that the ENCODE 3 "data" is

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"

Formatted: Font:12 pt

Deleted: roll-outs

Formatted: Font:12 pt

Deleted: and make use of these data contingent on that paper's publication

Formatted: Font:12 pt

Formatted: Font:12 pt, Bold, Italic, Underline

Deleted: novelty of

Formatted: Font:12 pt, Bold, Italic, Underline

Deleted: ,, and its analysis of networks.

Formatted: Font:12 pt

Deleted: . ... [35]

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 noted 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 our main goal and clearly organizing our analysis to illustrate the value of the resources in this paper. Specifically, we would like to emphasize two points.

[JZ2DL: please fill in xxx, only focus the data we used]

1. The goal of this paper and its distinct role in the whole 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

- 2017 histone ChIP-Seq data (1339 from tissues/primary cells; in contrast to 169 in Marticorena et al. 2017)
- 52 replication timing data sets from xx tissues (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 knockdows 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

Deleted: We would like to appreciate the referee's feedback.
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: the
Formatted: Font:12 pt
Deleted: We want to specifically point out
Formatted: Font:12 pt
Deleted: ... [36]
Formatted: Font:12 pt
Deleted: main
Formatted: Font:12 pt
Deleted: resource
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: We want to make it clear and emphasize that the goal of this paper is to build a new annotation "*resource*", not to discover novel biology in cancer. We feel that cancer is the best application to illustrate certain key aspects of ENCODE data and analysis - particularly the deep annotations and network changes. Thus, where the referee asks for novelty in cancer gene discovery - we strongly feel that this is out of scope. We have listed some more details about the resource of this paper as below (Table R1 and Figure. R1).
Annotation type ... [37]

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 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 utility and value of this resource to prioritize both key regulators and genomic variations (SNVs and SVs) using various techniques, such as luciferase assay, CRISP, and knockdowns. Collectively, we believe that all of these illustrate the value of our resource to cancer genomics.

2. Regarding the the BMR section

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 (see Table R1 below). 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; see Table R1).

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. 2.

Third, we want to point out that the BMR application is just **one out of many** potential ENCODE data applications. Even for Figure 2, we also include SV and GWAS germline SNV analyses. There are many other ENCODE applications, such as regulatory activity, rewiring, and stemness, which are also key to interpreting and prioritizing variants effects in cancer genomics.

Formatted ... [41]
Formatted ... [39]
Deleted: annotation ... [40]
Deleted: We are the depth paper
Formatted ... [42]
Deleted: Specifically for
Formatted ... [43]
Deleted: mentioned
Formatted ... [44]
Deleted: references
Formatted ... [45]
Deleted: like
Formatted ... [46]
Deleted: First, we
Formatted ... [47]
Deleted: to
Formatted ... [48]
Deleted: lot
Formatted ... [49]
Deleted: references and
Formatted ... [50]
Deleted: did cite many
Formatted ... [51]
Deleted: them
Formatted ... [52]
Deleted: initial submission (table R2
Formatted ... [53]
Deleted: However,
Formatted ... [54]
Deleted: of the
Formatted ... [55]
Deleted: (more details in the following table). We ... [56]
Formatted ... [57]
Deleted: ... [58]
Formatted ... [59]
Deleted: make a
Formatted ... [60]
Deleted: paper
Formatted ... [61]
Deleted: help
Formatted ... [62]
Deleted: , as
Formatted ... [63]
Deleted: shown
Formatted ... [64]
Deleted:
Formatted ... [65]
Formatted ... [66]
Deleted: the variant investigations
Formatted ... [67]
Deleted: have
Formatted ... [68]
Deleted: analysis from GWAS studies in this paper.
Formatted ... [69]
Deleted: interpret
Formatted ... [70]
Formatted ... [71]

Table R1. status of the related references

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: R2

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:

<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 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.

Formatted: Font:10 pt
Formatted Table
Deleted: the
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt

<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. 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. 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

Formatted: Font:10 pt
Formatted Table
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: reviewer
Formatted: Font:12 pt
Deleted: proposing
Formatted: Font:12 pt
Deleted: .
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: did recognize
Formatted: Font:12 pt
Deleted: genomic features were used to estimate BMR and improve
Formatted: Font:12 pt
Deleted: mutation
Formatted: Font:12 pt
Deleted: . We have cited ALL. Our aim here was

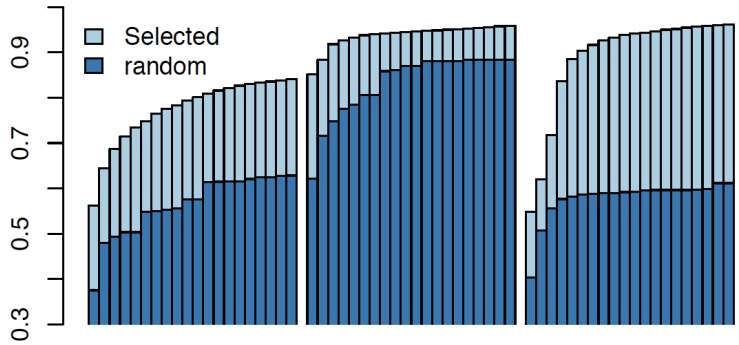
	the revised version, we have tried to make it more clear that we are not claiming to have developed a new model for BMR estimation for driver detection, or presenting a new discovery that "matched" features are better correlated with BMR. Instead, we explicitly clarified how the new ENCODE data can be 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 -- our work includes data on 2017 histone modification and 52 replication time. We have shown that this larger scale of data can benefit many models described in previous works to better characterize BMR.
Excerpt From Revised Manuscript	Wait for main text

<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 point is that the ENCODE3 rollout dramatically expands the number 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 ... [72]
Deleted: claim
Deleted: better
Deleted: model nor to propose
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: novel
Formatted: Font:12 pt
Deleted: performs
Deleted: . We have made it more apparent in our ... [73]
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted ... [74]
Deleted: help
Deleted: in
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: . ... [75]
Formatted: Font:10 pt
Deleted: We cited everyone, and we have made it
Formatted: Font:10 pt
Formatted Table
Deleted: were *misunderstood* at this point becau ... [78]
Formatted: Font:12 pt
Formatted ... [76]
Formatted: Font:12 pt
Deleted: pointing
Formatted: Font:12 pt
Deleted: out. We agree that "matched"
Formatted: Font:12 pt
Deleted: "more" features performs better in BMR ... [77]
Formatted: Font:12 pt
Formatted: Font:Helvetica Neue, 12 pt
Deleted: this. ... [79]
Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear
Deleted: developed
Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear
Deleted: or its application to cancer genomics. We p ... [80]
Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear
Deleted: of available cell types in tissues
Formatted ... [81]
Deleted: XXX to YYY. Furthermore, we show in our fi ... [82]
Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear
Deleted: The implication here is t ... [83]

	This data is useful from two aspects: • It provides a significantly larger pool to find the best match for a given cancer type • More data is 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 below).
Excerpt From Revised Manuscript	The 2017 uniformly processed histone modification and 52 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.  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: , and that's the main point we're trying to do in highlighting the ENCODE data. The reason we believe that the large amount of

Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear

Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear

Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear

Deleted: that while it's

Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear

Deleted: matching a

Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear

Deleted: cell

Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear

Deleted: , as has well been described by the referees and others,

Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear

Deleted: , and

Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear

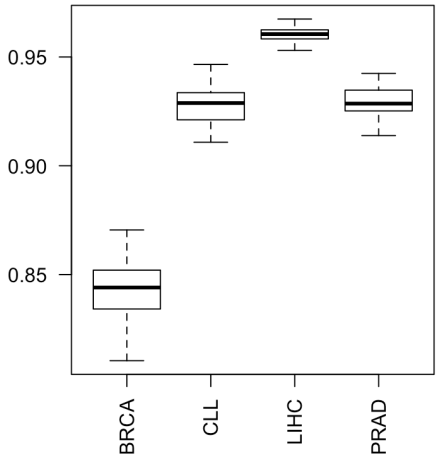
Deleted: single match is maybe not the best one, and a variety

Formatted: Font:Helvetica Neue, 12 pt, Pattern: Clear

Deleted: rate.

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Moved (insertion) [4]



Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

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

Deleted: Tissues

Deleted: Cell lines

<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.

Formatted: Font:10 pt

Formatted Table

Formatted: Font:10 pt

Deleted: ... [84]

Deleted: We thank the referee for pointing out the comparison of cell line vs. tissues and we feel this is a good suggestion. In our revised manuscript, we further investigated it in detail by extending our analysis to many new data types, such as RNA-seq and distal/proximal TF ChIP-Seq data. We think slightly differently with the referee on value of ENCODE data. Several points we want to emphasize are ... [85]

	• We have added more discussions that accurate cell-of-origin definitions are challenging. Distinct subtypes within an organ may derive from different 'cells of origin' \cite{21248838}. (see excerpt 1) 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). •
Excerpt From Revised manuscript	Newly added to the discussion section: 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.

<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 as if clarifying that ENCODE is not just about cell lines is a great suggestion. In our revised manuscript, we have extensively discussed the use different data from multiple aspects in both the main manuscript and the supplements: *Regarding the cell line data in the BMR part* • We added a table to clarify that the data we used in is not just from cell lines. The majority are from tissues or primary cells (excerpt 1). • We added figures (in the supplement) to demonstrate how cell line data can show comparable performance (excerpt 2).

Formatted: Add space between paragraphs of the same style, No bullets or numbering, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Moved up [3]: <#>***Regarding the***

Formatted: Font:12 pt

Deleted: <#>"ENCODE cell lines can be problematic", we want to highlight that ENCODE is not just about cell lines. There are many ENCODE tissue data for histones (339 cell line vs **818** tissue, details see excerpt 2 below). We have added a supplementary table on this point. Again, for the BMR part, we select the best possible features for prediction (no matter it is from cell line or tissue), instead of manually find a matching. ... [86]

Formatted: Font:12 pt, Bold, Italic, Underline

Deleted: 1. Comparison of mutation rate vs features in tissue/cell lines. We provided 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 even better) correlation with mutation counts. ... [87]

Formatted: Font:(Default) Helvetica Neue, (Asian) Helvetica Neue

Deleted: 1

Deleted: Supplementary file

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Moved down [6]: To avoid overfitting problem, we performed 5 fold cross validation using the selected model for each cancer type and listed the performance as below.

Moved down [6]: 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: Excerpt 2 From
2. Summary of ENCODE histone ChIP-seq data
Excerpt 3 From
At 1mb bin resolution, we compared the performance of models using random features vs. computationally selecting best features sequential (forward selection). It has sho ... [89]

Deleted: Excerpt 2 From
2. Summary of ENCODE histone ChIP-seq data ... [88]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<table>
<tr><td></td><td>

- 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).

Regarding the global comparison of cell lines and tissues

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

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 5).

</td></tr>
<tr><td>Excerpt 1 From Revised Supplementary file</td><td>

In total there are 2017 histone ChIP-seq and 52 Replication timing 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.

Summary of ENCODE histone ChIP-seq data

<table>
<tr><th>Cell Type</th><th># histone marks</th></tr>
<tr><td>tissue</td><td>818</td></tr>
<tr><td>primary-cell</td><td>521</td></tr>
<tr><td>cell-line</td><td>339</td></tr>
<tr><td>in-vitro-differentiated-cells</td><td>179</td></tr>
<tr><td>stem-cell</td><td>114</td></tr>
<tr><td>induced-pluripotent-stem-cell-line</td><td>46</td></tr>
</table>

[JZ2DL: please add the table of replication timing data]

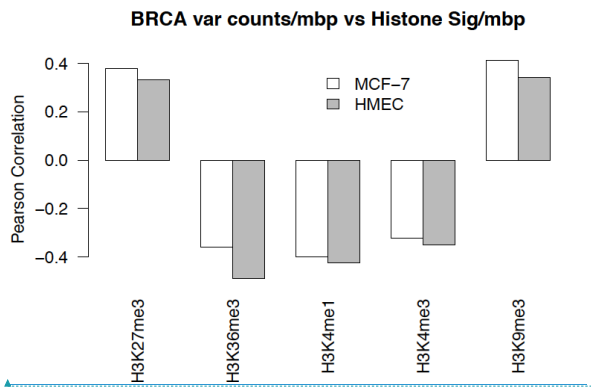
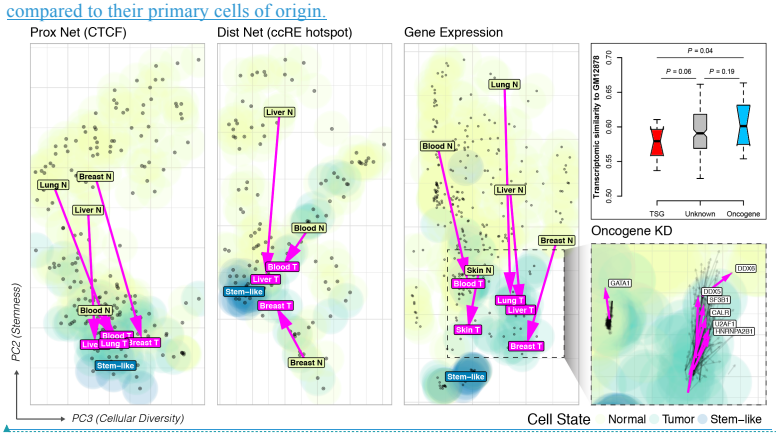
</td></tr>
<tr><td>Excerpt 2 From Revised Supplementary file</td><td>

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

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.

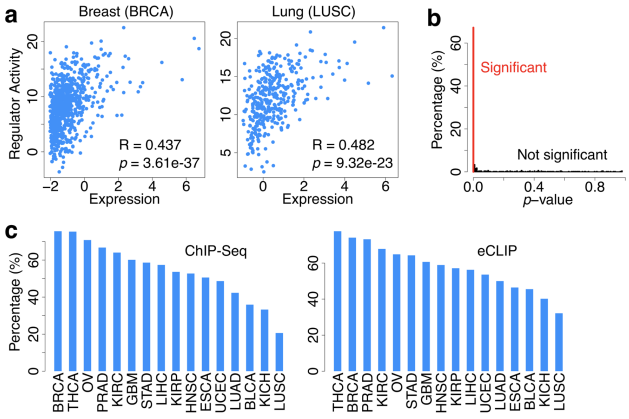
</td></tr>
</table>

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

	
Excerpt 3 From Revised Discussion	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.
Excerpt 4 From Revised Supplementary file	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. 

Formatted: Font:(Default) Helvetica Neue, (Asian) Helvetica Neue

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt

Excerpt 5 From Revised supplement	Regarding the validation of cell line conclusions on real patient 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.  **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).

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<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 referee pointed out that we and others have tried to prioritize non-coding elements before. This is definitely true, and we have tried to make it more clear in our revision that we are not claiming to be among the first to attempt this. 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) and performed targeted validations on key regulators (Excerpt 2). • 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 3). The fact that we coupled this with successful validation demonstrates the considerably greater value of the integrated ENCODE data.
Excerpt 1 From Revised Manuscript	New legend of figure 3. Figure to put here Ask Feng's group to write up here! [JZ2MG: wait]

Formatted: Font:10 pt
Deleted: The rest of the sections (and their corresponding supplement sections) are variable in significance and quality.
Formatted: Font:10 pt
Formatted Table
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: . However, we believe
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: we used here
Formatted: Font:12 pt
Deleted: new
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: .
Formatted: Font:12 pt
Deleted: important aspect is that it takes advantage of many new
Formatted: Font:12 pt
Deleted: and integrates over many different aspects. Detailed changes please see the Excerpt blow
Formatted: Font:12 pt
Moved (insertion) [5]

Excerpt 2 from Revised figure and supplement	Feng's validation to come here
Excerpt 3 From Revised Manuscript	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: it

Formatted Table

Deleted: 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 incurred dat ... [90]

Referee #2 (Remarks to the Author):

<ID>REF2.0 – Preamble

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

[JZ2DL: please fill in the xxx here]

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 make it more clear that this is the main integrative paper in ENCODE3 to provide deep and accurate annotation focusing on several data-rich cell types. 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). 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

Formatted: Font:12 pt

Deleted: would like to

Formatted: Font:12 pt

Deleted: about

Formatted: Font:12 pt

Deleted: on

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: Regarding the novelty point

Formatted: Font:12 pt

Deleted: want

Formatted: Font:12 pt

- extended gene definitions that incorporate numerous types of regulatory elements in a gene-centric way
- universal and tissue-specific regulatory network 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 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: its
Formatted: Font:12 pt
Deleted: of
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: We have listed some more details about the novelty of this paper as below.
Deleted:
Annotation type ... [91]

<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: Font:10 pt
Formatted Table
Formatted: Font:12 pt

<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?

Formatted: Font:10 pt
Formatted Table

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. There are three reasons to explain why we did not directly applied available methods: • the Marticorena et al. paper came out in Nov 2017, which was almost three months after our initial submission, and it is more about positive selection in coding regions than BMR estimation. • the main focus of 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 for 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).
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

Deleted: We thank referee for the suggestion. The referee is pointing out that negative binomial regression has been used before. We also feel that the fact that other papers also used negative binomial regression bolsters the underlying technical validity of our argument. While we admit it does slightly undercut a claim of novelty in this regard, that is not central to our work. (reference all these papers) ... [92]

Deleted: of

Formatted: Font:12 pt

Deleted: using

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: the scheme in that paper.

Formatted: Font:12 pt

Deleted: 1. 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"

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: instead of

Formatted: Font:12 pt

Deleted: 2. The

Formatted: Font:12 pt

Deleted: that

Formatted: Font:12 pt

Deleted: about

Formatted: Font:12 pt

Deleted: they

Formatted: Font:12 pt

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: there is no data related with the noncoding regions. Also

Formatted: Font:12 pt

Deleted: has been released

Formatted: Font:12 pt

Deleted: 3. The Marticorena et al. paper has 169

Formatted: Font:12 pt

Deleted: included -

Formatted: Font:12 pt

Deleted: comparison to our 2067 features. ... [93]

Formatted: Font:12 pt

Deleted: . paper. We are not aiming to make a ne ... [94]

Formatted: Font:12 pt

Deleted: very large amount of data and is able to ... [95]

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

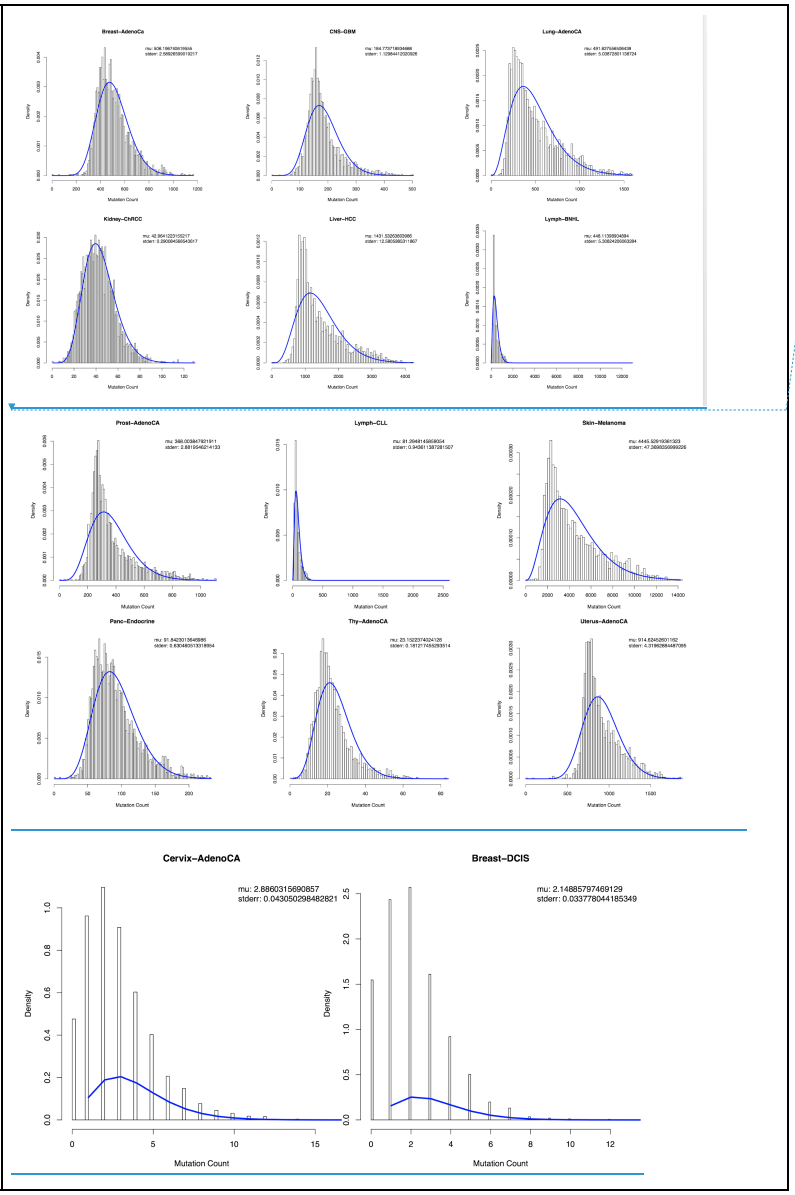
<ID>REF2.3 – Questions about the Goodness of fit of the Gamma-Poisson Model

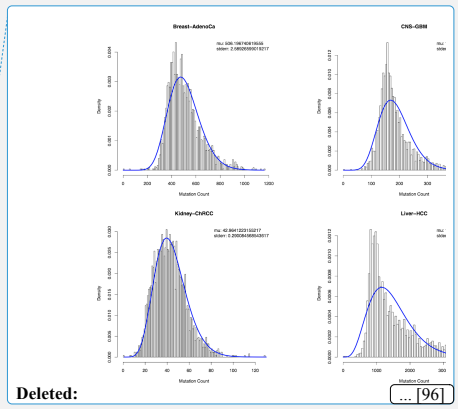
<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix,&&&OOS
<STATUS>%%%100DONE

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. 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.

Formatted: Font:10 pt
Formatted Table
Deleted:
Deleted:
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: provided
Formatted: Font:12 pt
Deleted: supplementary file
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: of the
Formatted: Font:12 pt
Deleted: the
Formatted: Font:12 pt
Deleted: of
Formatted: Font:12 pt
Deleted:)..
Formatted: Font:12 pt
Deleted: made a mention of
Formatted: Font:12 pt

Excerpt From Revised Supplementary file





Deleted: ... [96]

<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, (details please see excerpt 1) • 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. • 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 1 From Revised Supplementary file	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: Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Not Highlight

Formatted: Font:10 pt

Formatted Table

Deleted: To answer this question, we plotted the overall

Deleted: count under different 3mer context

Deleted: estimated

Deleted: (using the AER package) in R in the following

Deleted: . On one side

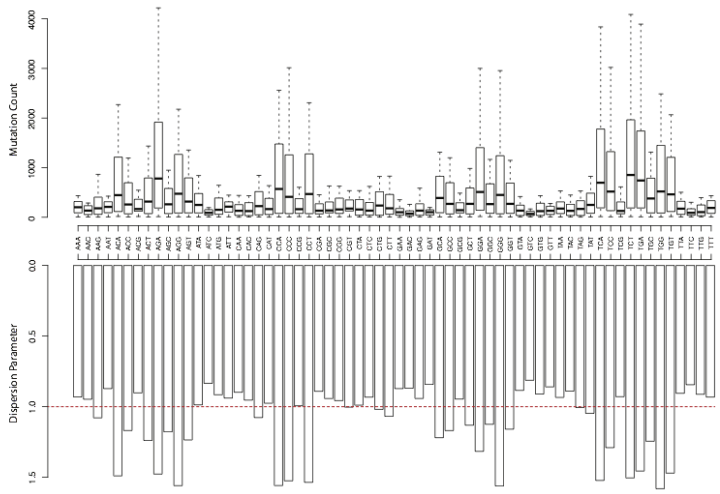
Deleted: obvious that

Comment [5]: think about this - could we say this better

Comment [6]: consider revising, needs to directly answer the question (or sound like we are answering)

Deleted: those 3mers with more variants, there is a tendency to introduce overdispersion and accept the Gamma-Poisson model. It could be either the power issue, or

Deleted: level of heterogeneity among samples, or even both. We have put more in supplementary file.

	
Excerpt 2 From Revised Supplementary file	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. JZ2JZ: remember to put the figure in.

Formatted Table

Deleted: We also want to point out that the overdispersion problem on count data is also confounded by omitting related covariates. That is the main reason why we want to introduce more feature candidates from ENCODE and at the same time avoid overfitting. Many other methods (such as Marticorena, 2017) directly use Negative Binomial regression without checking whether it is necessary. It is simpler to not introduce additional parameters. However, we think it is better to check how heterogeneous the count data is even after correcting enough covariate effects ... [97]

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

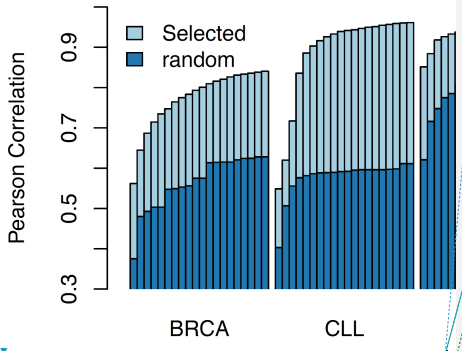
<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE,%%%CalcDONE
Add the cross validation in this response section

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

Formatted Table

	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.
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 (details in excerpt 1), and we have moved the PCA part into the supplement.

added a supplementary figure of cross validations (details in excerpt 1)

•	
Excerpt 1 From Revised supplement	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. 

revised version, we wanted…ave tried to bring out this point,. The point of…etter illustrate our approach ... [98]

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"

Deleted: , but …(details in excerpt 1), and we realized that actually did not get across very clearly, so we have replotted this figure and now simply show ... [99]

Deleted:

Deleted: Yes - we do cross validation as we better describe now in the suppl

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

Moved up [2]:
We hope

Formatted: Font:Helvetica Neue, 12 pt

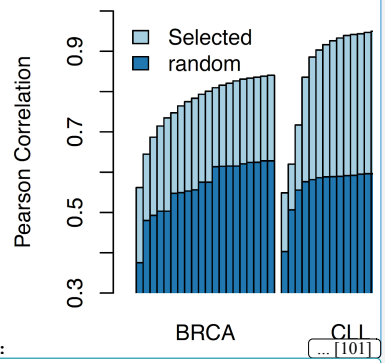
Formatted Table

Deleted: this gets the point across. The aim here is not to highlight a complicated mathematical method but just simply to get across the idea that the exten ... [100]

Formatted: Font:10 pt

Formatted: Add space between paragraphs of the same style, No bullets or numbering

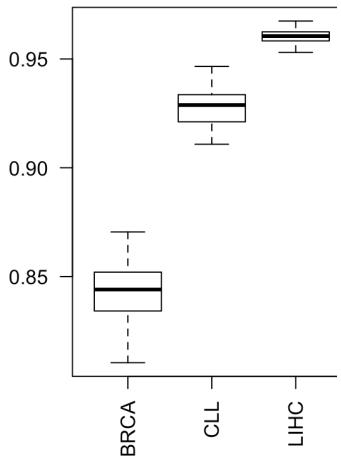
Deleted: Manuscript



Deleted: ... [101]

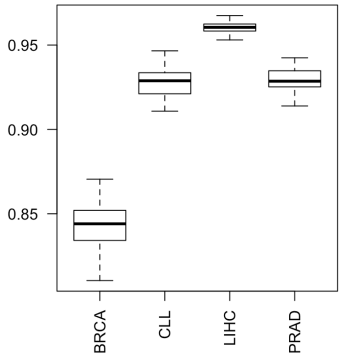
Formatted ... [102]

Moved up [4]: 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: Font:(Default) Times New Roman, (Asian) Times New Roman

Excerpt 2 From Revised supplement	To avoid overfitting problem, we performed 5 fold cross validation using the selected model for each cancer type and listed the performance as below. 

Moved (insertion) [6]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<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]

Deleted: wait for the GWAS

Deleted: be added here, are still working to refine the results

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.

Formatted: Font:10 pt

Formatted Table

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 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 (see supplement and excerpt 1 below) • more discussion (in the main text) about the pros and cons of merging test regions (see in excerpt 2) • real examples in supplement (see in excerpt 3) • a new section of quality metrics of the compact annotations to catpure functional sites and rm noise(see in excerpt 4)
Excerpt 1 From Revised Supplementary file	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 hypotheis, the proability to observe at least one mutation per patient is $p_0 = 1-(1-\mu_i)^{\frac{l_i^n+l_i^r}{c}}$ Under the alternative hypotheis, $p_1 = 1-(1-\mu_i)^{l_i^n}(1-\lambda_i\mu_i)^{l_i^r}$ 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 quicktly.

Deleted: The

Deleted: is correct

Formatted: Font:12 pt

Formatted: Font:12 pt

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: calculation in our revised manuscript.

Formatted: Font:12 pt

Deleted: our initial submission, the assumption is that we were trimming off the nonfunctional sites while preserving the

Formatted: Font:12 pt

Deleted: ones. Two examples can explain

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"

Formatted: Font:12 pt

Deleted: motivation

Formatted: Font:12 pt

Deleted: this assumption

Formatted: Font:12 pt

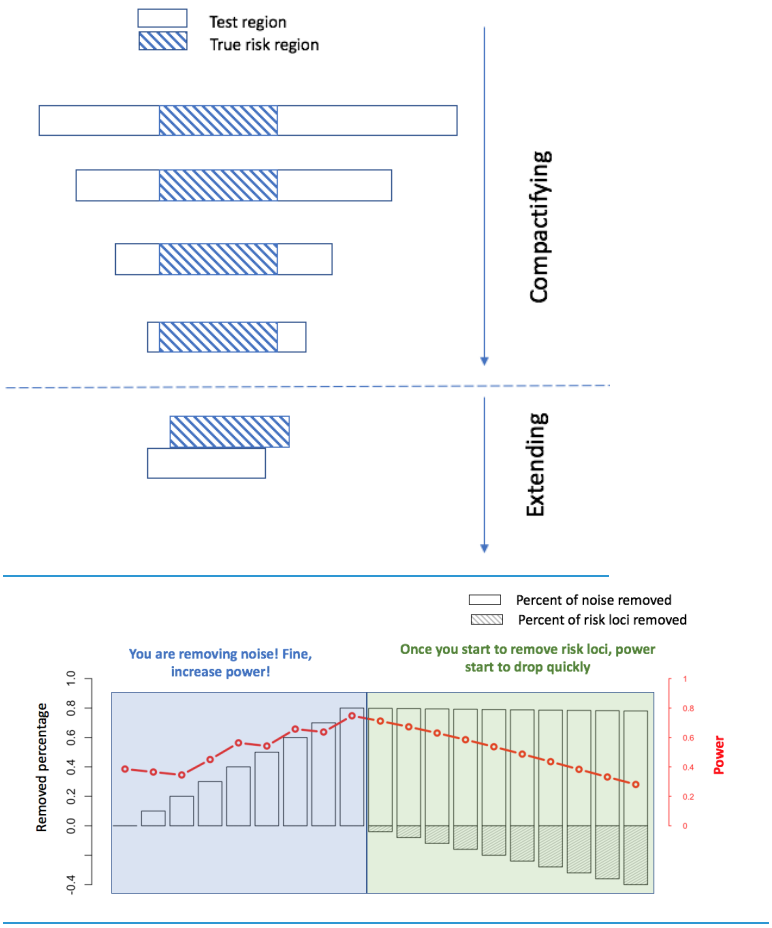
Deleted: details

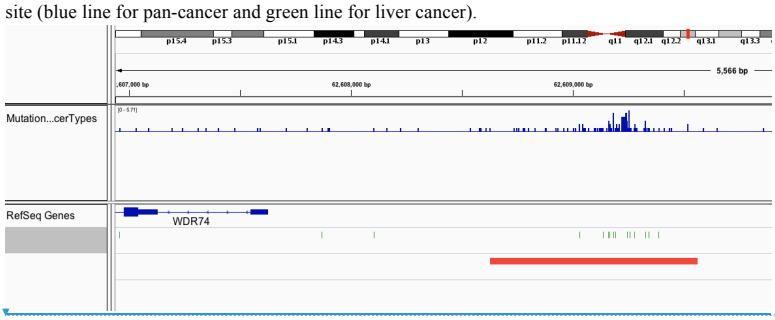
Formatted: Font:12 pt

Deleted: 1 below).

Formatted: Font:12 pt

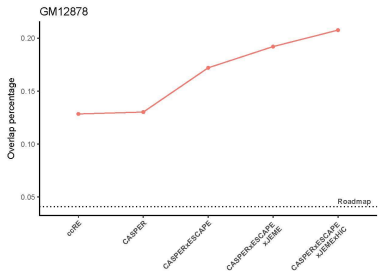
Deleted: ... [103]

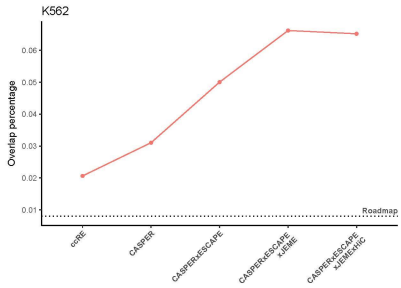
	
Excerpt 2 From Revised main text	In summary, our claim is that first we provide compact annotations to pick up functional nucleotides and remove noisy ones through the guidance of many functional characterization assays. Then we hope to join the distributed functional sites together to increase statistical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

Excerpt 3 From Revised Supplementary file	We provided two examples to explain the motivation of our compact and extended gene annotations and why we feel our assupmptions 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 4 From Revised Supplementary file	• *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: 1
Formatted Table
Formatted: Font:Times New Roman, 10 pt
Deleted: Two
Deleted: can
Formatted: Font:Times New Roman, 10 pt
Deleted: this assumption.
Formatted: Font:Times New Roman, 10 pt
Deleted: that we are making
Formatted: Font:Times New Roman, 10 pt
Deleted: assumption
Formatted: Font:Times New Roman, 10 pt
Deleted:
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
Deleted:


Deleted: 2
Deleted: Manuscript

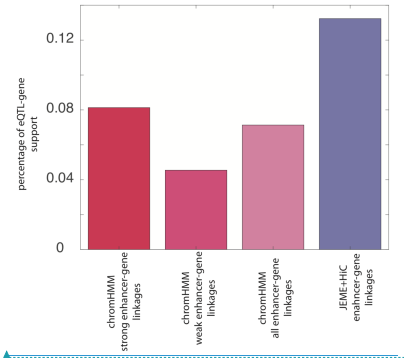




- *Regarding the quality of enhancer-gene linkages:*

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: Font:(Default) Times New Roman, (Asian) Times New Roman



<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

Formatted: Font:(Default) Helvetica Neue, (Asian) Helvetica Neue

Moved down [7]: 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.

Deleted: ... [104]

Formatted: Font:16 pt

Deleted: – Q-Q plots

Moved down [8]: <TYPE>$$$BMR,$$$Calc <ASSIGN>@@@JZ <PLAN>&&&Defer

Moved down [9]: Referee Comment ... [105]

Deleted: <STATUS>%%%70DONE

Formatted: Font:10 pt

Formatted: Justified

Formatted Table

Moved down [10]: <ID>REF2.8

Deleted: Author
We thank the referees for this comment. We have updated the QQ-plots in our revised manuscript and they look fine. ... [107]

Deleted: Author
We thank the referees for this comment. We have updated the QQ-plots in our revised manuscript and they look fine. ... [106]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:11 pt

Formatted: Font:10 pt

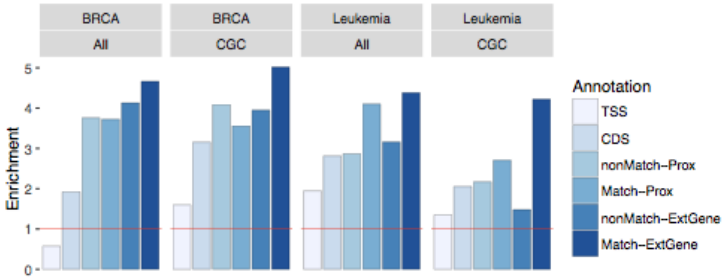
Formatted Table

Formatted: Font:12 pt

Deleted: We

Formatted: Font:12 pt

Deleted: and added several new sections to highlight

	all our related analysis in the orginal supplement to the main text to better illusrate the value of our extended gene resource, which includes • GWAS germline variant enrichment analysis across different annotations in the main figure (see in excerpt 1) • A new figure panel to stratift patient expression levels based on the mutation status from various annotations. We found that extended genes perfromed better than others (see in excerpt 2) • A new figure in the supplement to show variant effect in extended gene regions on regulator activities (see in excerpt 3) • A CRISPR based validation of onco-gene activation based on extended genes (see in excerpt 4)
Excerpt 1 From Revised 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 From Revised 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. Specifically, we found 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.

Deleted: extended genes, such as
2. We showed that by using the extended gene, we can better stratify the gene expressions and regulat ... [108]

Formatted: Font:12 pt

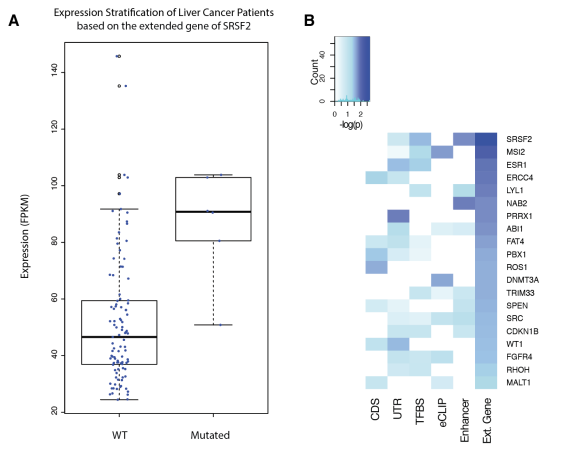
Deleted: analysis (as we pointed up in the response to <ID>REF2.6 – Comments on the power analysis and compact annotations)

Formatted: Font:12 pt

Formatted: Font:12 pt

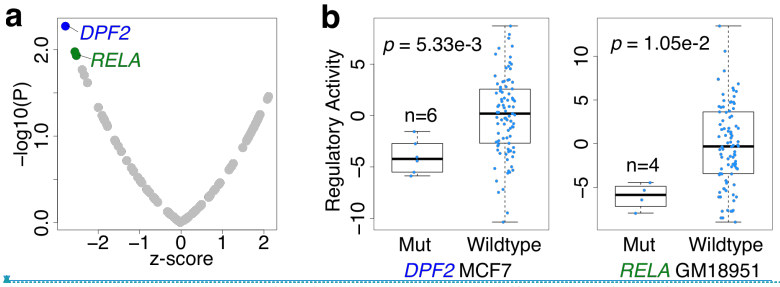
Deleted: 2. We showed that by using the extended gene, we can better stratify the gene expressions and regulations ... [109]

Moved (insertion) [7]



Excerpt 3 From Revised Supplement

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.



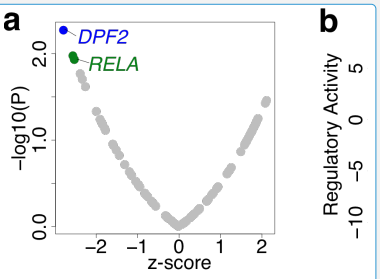
Supplementary Figure X. Mutations in level one enhancers affects the activity of nearby TFs. (a) The association between TF regulatory activity and mutation in enhancer regions. For each cancer type, the association between TF regulatory activity computed using ChIP-seq data and mutation status of nearby 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: 1

Formatted Table

Formatted: Font:10 pt

Deleted: Manuscript



Deleted:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

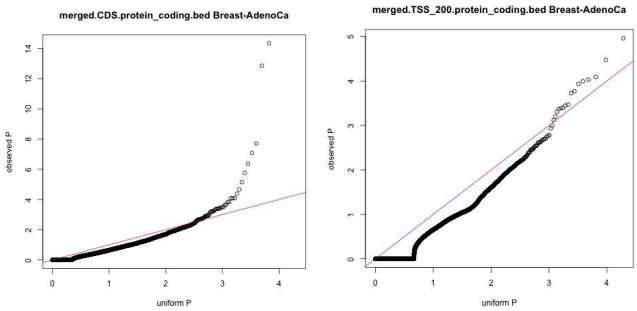
Formatted: Font:10 pt

Deleted:

Excerpt 4 From Revised Manuscript	Ask Feng's group for text and wait for figure to come in

<ID>REF2.8 – Q-Q plots

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

Referee Comment	5) 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 and they look fine. 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 From Revised Manuscript (in supplement)	

Moved (insertion) [10]

Formatted: Font:11 pt

Moved (insertion) [8]

Moved (insertion) [9]

Formatted: Font:10 pt

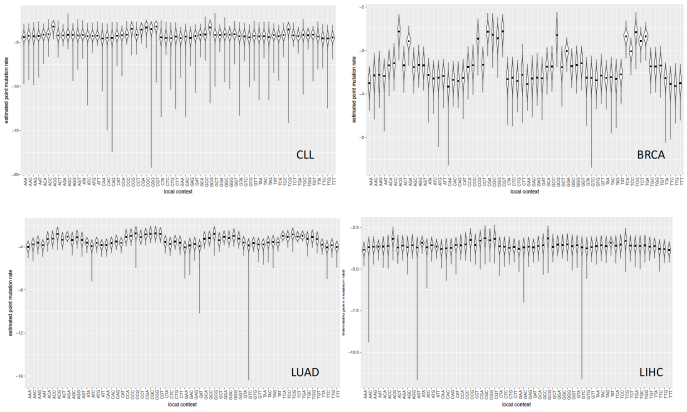
Formatted: Justified

Formatted Table

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<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 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 From main text and supplementary file	*The newly added sentence in the 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. *From original supplement:* 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, ⋕) Violin plot of estimated BMR over local context and genomic locations 

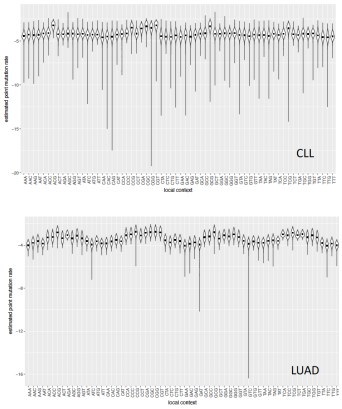
Formatted: Font:10 pt

Formatted Table

Deleted: In the main figure, we did not show how local context effect may affect BMR in order to highlight the effect of accumulating features. However, in the supplementary file where we described our method, we separate the 3mers to run negative binomial regression. We showed that in Supplementary figure xxx that local context effect is huge - usually up to several order of effect on BMR (Please see details in the following excerpt). [[we have tried to make clearer]]

Deleted: Original Supplementary

Figure S 2-2 (TL, ⋕) 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 From Revised Manuscript	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: Font:10 pt
Formatted Table
Formatted: Font:12 pt
Deleted: incorporate
Formatted: Font:12 pt
Formatted: Font:10 pt

<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.

Formatted: Font:10 pt
Formatted Table

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 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: Font:12 pt
Formatted: Font:12 pt
Deleted: . In our revised manuscript,
Formatted: Font:12 pt
Deleted: still
Formatted: Font:12 pt
Deleted: the word burden but made it clear
Formatted: Font:12 pt
Deleted: variant analysis
Formatted: Font:12 pt
Deleted: about
Formatted: Font:12 pt
Deleted: , but also include germline variants,
Formatted: Font:12 pt
Deleted: the
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: . ... [110]
Formatted: Font:12 pt
Deleted: of a lot of confusion. We're using burden here b/c we do intend this is useful
Formatted: Font:12 pt
Deleted: the case-control . see ref GWAD for ref2
Formatted: Font:10 pt
Formatted Table
Formatted: Font:12 pt

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 commitee, 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. We admit that maybe this construction is not that intuitive. 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>%%%TBC

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 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

Deleted: and genomics
Deleted: it's
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: . We
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: have
Formatted: Font:12 pt
Formatted: Font:12 pt

Deleted: We've
Formatted: Font:12 pt
Deleted: these
Deleted: appointives
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: nature
Formatted: Font:12 pt
Deleted: the
Formatted: Font:12 pt
Deleted: the data in
Formatted: Font:12 pt
Deleted: it's simply
Formatted: Font:12 pt
Formatted: Font:10 pt
Formatted Table
Formatted: Justified

	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 orginal thinking was some of the contents are distributed in multiple sections. For example, each step in the final prioritization scheme are corresponding to a separate section 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: We tried the new way of citing supplementary info.

Deleted: Excerpt From ... [112]

Deleted: Excerpt From ... [111]

<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 orginal quetion, Figure 2A supports the claim becuase replication timing from MCF-7 outperforms that from HeLa to predict BMR. We have added a sentence in the supplent and moved this panel to supplement.
Excerpt From Revised Manuscript	Wait for new figure 1

Formatted: Font:10 pt

Formatted Table

Formatted: Justified

Deleted: We have changed figure ... [113]

<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, xxx, and xxx, we called the SNV and SVs for serveral 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. JZ2DL: would you pls check Feng's email (you were cced) to double check what assays they used for the SV calling?
Excerpt From Revised Manuscript	Wait for updated Fig 1
Excerpt From Revised Supplementary file	JZ2DL: could you pls make a table from Feng's data and deposit it to our resource?

Formatted: Font:10 pt
Formatted: Justified
Formatted Table
Deleted: We have changed the fighure ...
Deleted: WE have

<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 From Revised Manuscript	JZ2MTG: may need something more about CRASPER, Please add here

Formatted: Font:10 pt
Formatted: Justified
Formatted Table
Deleted: The desxribeion of this is in the suppl. We have made this clearer in the revised version... see the exerpt below

<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 (details in the excerpt below).
Excerpt From Revised Supplementary file	Our model incorporated many genomics features. Here features only means one set of functional genomics data, such as H3K27ac and DHS. The absolute value of regression coefficient is closely related with 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,

Formatted: Font:10 pt
Formatted Table
Formatted: Justified
Deleted: We
Formatted: Justified
Formatted: Font:12 pt
Deleted: And
Formatted: Font:12 pt
Deleted: be descbie ... [114]
Formatted: Font:12 pt
Deleted:
Formatted: Font:12 pt
Deleted: Manuscript

	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.

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

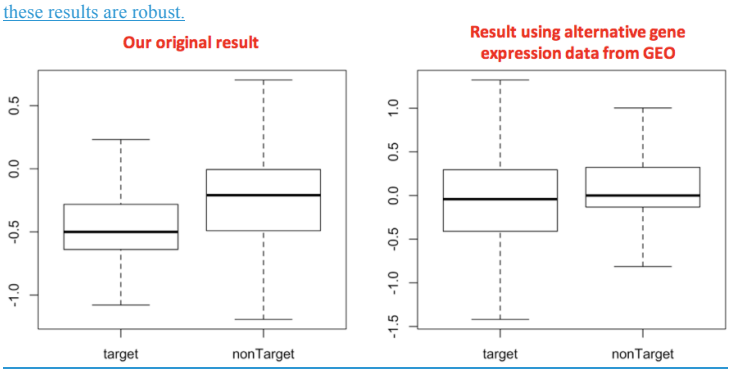
<TYPE>$$$Annotation
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

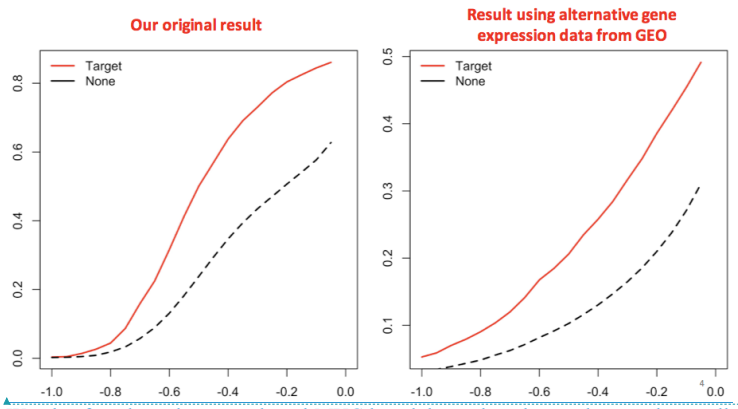
Referee Comment	For Figure 2C, more explanation is needed on how to form an extended gene.
Author Response	We thank the referee for this comment and we have added a paragraph in the supplement to better describe how we generated the extended genes. (see excerpt below)
Excerpt From Revised Manuscript	There are four important basic elements in our extended gene definitoin: 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.

<ID>REF3.7 – validations

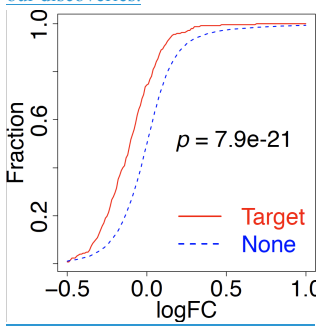
<TYPE>$$$Annotation
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

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?

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Moved (insertion) [11]

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	

Moved (insertion) [12]

Formatted Table

<ID>REF3.9 – Quality 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:

Deleted: For Figure 2C, more explanation is needed on how to form an extended gene.

Formatted Table

Deleted: ... [115]

Comment [10]: break up

Deleted: ... [116]

Deleted: Think about how we should responsed ... [117]

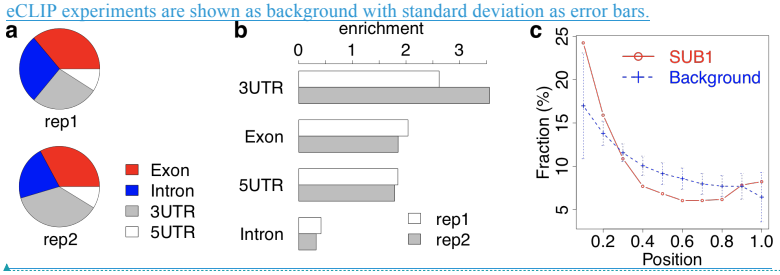
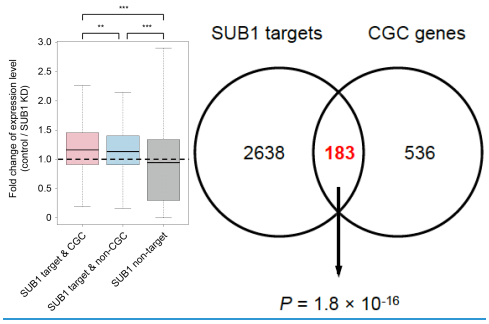
Excerpt From Revised Manuscript	

<ID>REF3.10 – novel oncogenes

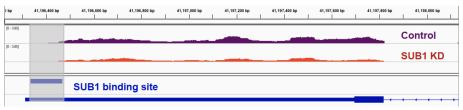
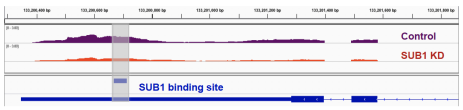
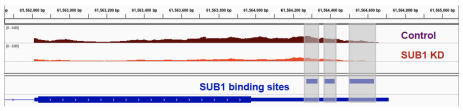
Deleted: 7

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

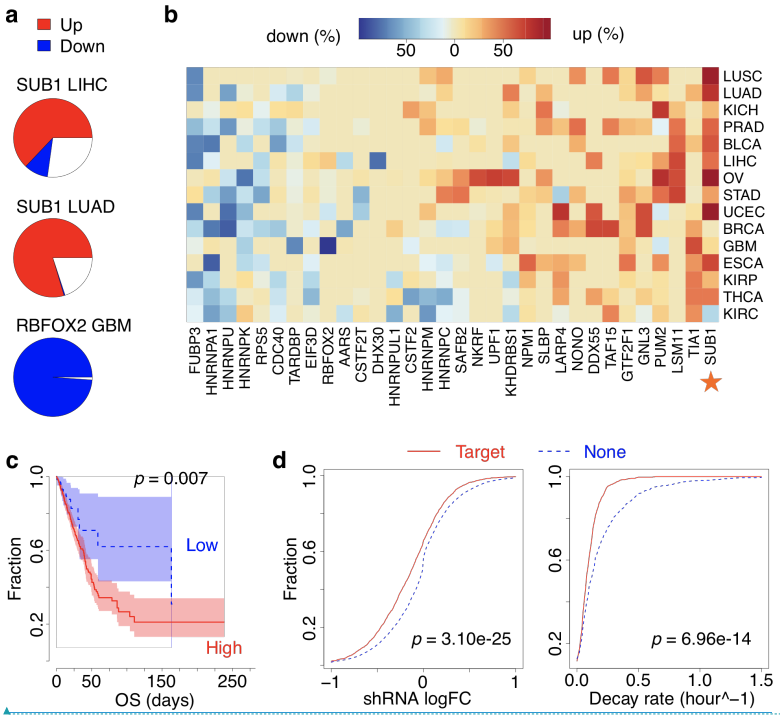
Referee Comment	Are there any novel oncogenes detected by the method?
Author Response	We than the referee to point out the novelty of discoveries. We have tried to make it clear that the main goal of this paper is to illustarate the value of ENCODE data and the usefulness of our deep annotations. We 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 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)
Excerpt 1 From Revised Supplement	**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 2 From Revised Supplement	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. 

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

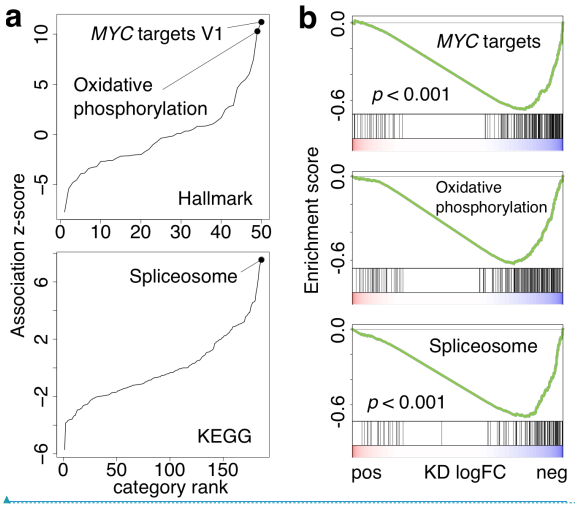
Gene	Functions	PMID	Expression profiles of the 3' UTR
BRCA1	The gene is involved in maintaining genomic stability	12677558, 17416853, 23620175, 16551709	
POLE	The gene is involved in DNA repair and replication	26133394, 28423643	
FEN1	The gene is involved in DNA repair and replication	20929870, 22586102	

Excerpt 3 From Revised Supplement	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 4 From Revised Supplement	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:(Default) Times New Roman, (Asian) Times New Roman, Highlight

	
Excerpt 5 From Revised Supplement	Feng's validations

<ID>REF3.11 – 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 redrawn the figure to make it clearer. In the orginal version, <-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

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted Table

	genes regulating MYC, and -2135-> means there are 2135 genes being regulated MYC, but not regulate MYC.
Excerpt From Revised Manuscript	Wait for Figure 2

<ID>REF3.12 – 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've 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 to show the significance of regulators placed in different network hierarchy. Second, we've 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. See excerpt for details.
Excerpt From Revised Manuscript	**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: 8

Deleted: 90DONE

Formatted: Font:10 pt

Formatted Table

Deleted: Author . ... [118]

Formatted Table

Deleted: We thank referee for point out this issue. We have updated the figure

Formatted: Font:Times New Roman, 10 pt, Bold

Deleted: to show the significance testing of

Formatted: Font:Times New Roman, 10 pt, Bold

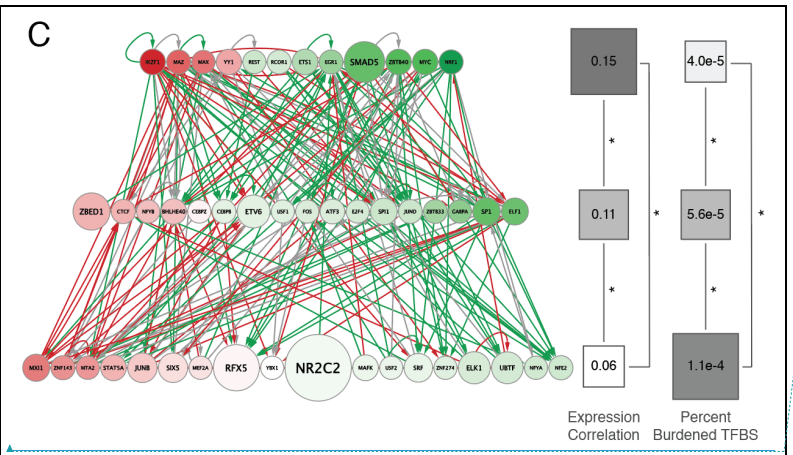
Deleted: hierarchy analysis.

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:Times New Roman, 10 pt

<ID>REF3.13 – Network rewiring

<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. In the rewiring analysis, vertices represent genes (regulators) and edges represent regulatory linkage between TFs and genes. We have used colors and thickness to show regulatory rewiring

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt

Moved up [12]: Excerpt From Revised Manuscript ... [119]

Formatted Table

Deleted: 9

Moved up [11]: <PLAN>&&&AgreeFix

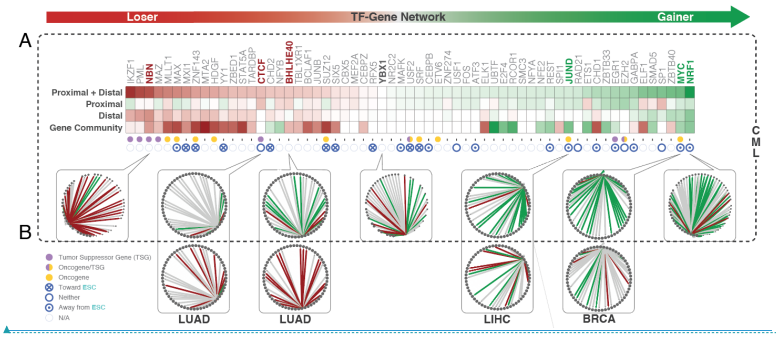
Deleted: 100DONE

Formatted: Justified

Formatted Table

Deleted: First of all, you are correct that vertexes are representing

Deleted: are representing

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

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

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: Font:10 pt
Formatted Table
Formatted: Font:12 pt

<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
Formatted: Font:10 pt
Formatted: Font:12 pt
Deleted: part and
Formatted: Font:12 pt

	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 admit that maybe this construction is not that intuitive. 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: We are well aware

Formatted: Font:12 pt

Deleted: this fact.

Formatted: Font:12 pt

Deleted: Excerpt From
[JZ2MG: is there an excerpt here?] ... [121]

Deleted: Excerpt From
[JZ2MG: is there an excerpt here?] ... [120]

<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: Font:10 pt

Formatted Table

Deleted: We thank

Formatted: Font:12 pt

Deleted: referee for his/her suggestions on our presentations. As requested

Formatted: Font:12 pt

Deleted: trimmed and edited

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: We have revised the manuscript.

<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 analysis to address this question, including

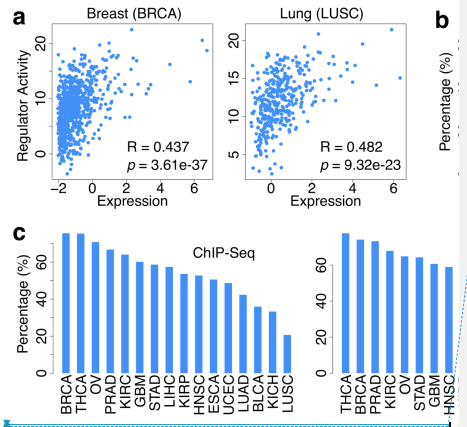
- A supplementary section to show that TF regulatory activities predicted from ENCODE TF regulatory networks compared with their expression levels are highly correlated in breast and lung cancer (Excerpt 1 below).
- JZ2DL: imputed vs imputed network?

Excerpt From Revised Manuscript	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

Moved up [5]: ... [123]
Deleted: [JZ2MG: ongoing] ... [122]
Deleted: Peng for more analysis
Formatted Table
Formatted: Font:10 pt
Formatted Table

Deleted: Mention that theres a lot of tissue in ENCODE
Formatted: Font:12 pt
Deleted: take the referee's comment to heart and we agree with the reviewer
Formatted: Font:12 pt
Deleted: from
Formatted: Font:12 pt
Deleted: ... [124]
Moved up [1]: In the revision
Formatted: Font:Helvetica Neue, 12 pt
Formatted: Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)
Formatted Table
Deleted: , we compared the concordance level of our conclusions made from ENCODE cell line data to observations from patients with primary cancers. And we clarified that although ENCODE data are profiled in cell culture models, the regulatory targets are still representative of the gene regulations in human cancers.
Formatted: Font:12 pt
Deleted: a new section in the revised supplementary file for more discussions.
Formatted: Font:12 pt
Deleted: In addition, we built an imputed network from a published dataset outside ENCODE and evaluated the rewiring of regulatory network. We used ATAC-seq dataset from the paper {\cite: Philip, Mary, et al. "Chromatin states define tumour-specific T cell dysfunction and reprogramming." Nature 545.7655 (2017): 452.} and show that the rewiring from ChIP-seq based network can be recapitulated using T cell ATAC-seq data. ... [125]
Formatted: Font:12 pt
Deleted: ... [126]
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt

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).

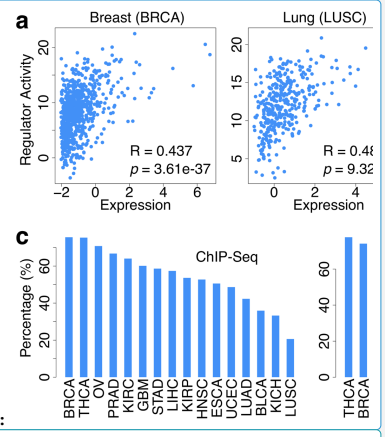


Supplementary Figure X. The clinical relevance of ENCODE cell line data in human primary tumors.

(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*-

Formatted: Font:10 pt



Deleted:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:10 pt

Formatted: Justified

Formatted: Font:10 pt

Deleted: level

Formatted: Font:10 pt

Formatted: Font:10 pt

Deleted: invasive

Formatted: Font:10 pt

Formatted: Font:10 pt

Deleted: sides

Formatted: Font:10 pt

Formatted: Font:10 pt

Deleted: .

Formatted: Font:10 pt

Formatted: Font:10 pt

	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: a

Formatted: Font:10 pt

Formatted: Font:10 pt

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

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

Formatted: Normal, Space Before: 0 pt

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 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 point REF4.6.

Formatted: Font:10 pt

Formatted Table

Deleted: Author

We thank referee for bringing this point and we feel it is a good comment. Actually, the referee is correct many of the cancer transcriptome is similar to each other and

In relation to this & other points

Excerpt 1 From ... [127]

Formatted Table

Deleted: One of the strengths of ENCODE release 3 is massive expansion of functional genomic data into various primary cells and tissue types. In this revision, we have extensively explored the chromatin landscape and expression patterns across all of available ENCODE primary cells and tissues, and compared them

Formatted: Font:Helvetica Neue, 12 pt

Deleted: existing immortalized cell lines with deep annotations. ... [128]

Formatted: Font:Helvetica Neue, 12 pt

Deleted: . We performed RCA/PCA

Formatted: Font:Helvetica Neue, 12 pt

Deleted: RNA-Seq, shRNA RNA-Seq, and TF ChIP-seq data and found that cancer cells tend to cluster together and stay away from their normal counterparts. ... [129]

Formatted: Font:Helvetica Neue, 12 pt

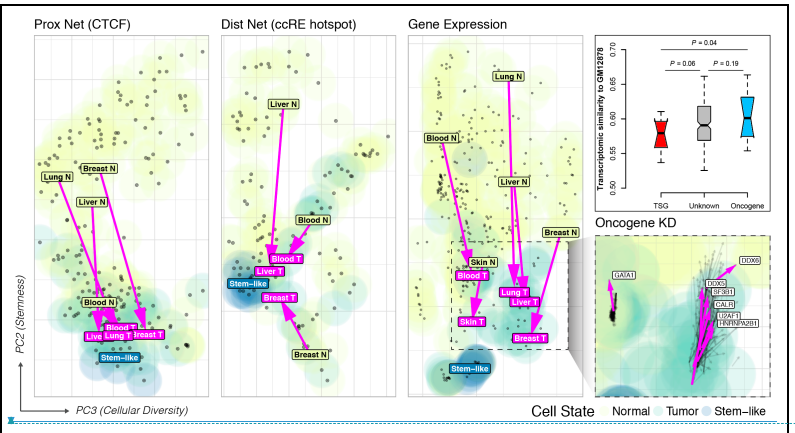
Formatted: Font:Helvetica Neue, 12 pt

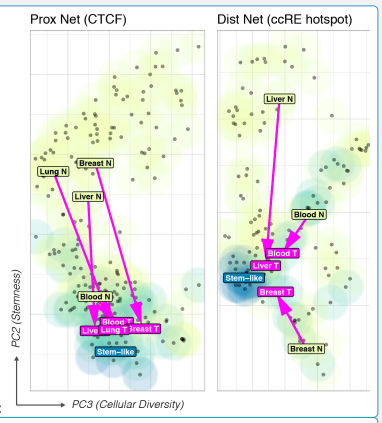
<ID>REF4.6 – Relationship of H1 to other stem cells

<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. Furthermore, 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 comment, and we thank the referee very much for it. See excerpt for more details. We initially focused on H1 because it is one of the top-tier ENCODE cell lines with broadest cell type coverage.
Excerpt 1 From Revised Main Manuscript	… 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 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. … … 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.

Formatted ... [130]
Formatted Table ... [131]
Deleted: think was very good. We initially focused o ... [136]
Formatted ... [132]
Deleted: We thank the referees for bringing this ... [133]
Formatted ... [134]
Formatted ... [135]
Formatted ... [137]
Deleted: end code
Formatted ... [138]
Deleted: very
Formatted ... [139]
Deleted: stem cells
Formatted ... [140]
Deleted: . ... [141]
Formatted ... [142]
Deleted: are responding
Formatted ... [143]
Deleted: expanding
Formatted ... [144]
Deleted: now
Formatted ... [145]
Deleted: the
Formatted ... [146]
Deleted: with an end code. This makes for a very ni ... [147]
Formatted ... [148]
Deleted: main text
Formatted ... [149]
Deleted: .
Deleted: we were able
Formatted ... [152]
Deleted: show how
Formatted ... [150]
Formatted ... [153]
Formatted ... [151]
Deleted: maps perfectly on
Formatted ... [154]
Deleted: this. So
Formatted ... [155]
Deleted: We also want to highlight here that there ... [156]
Formatted ... [157]
Formatted ... [158]
Comment [11]: Peng made a useful comment h ... [159]
Formatted ... [160]
Formatted ... [161]
Deleted: X
Formatted ... [162]

	
Excerpt 2 From Revised Supplement	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).



Deleted:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt

<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

Formatted: Font:10 pt

Formatted Table

	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 1 From Revised Main Manuscript	(to be continued for fig 1)

Deleted: We thank referee for the suggestion.
Formatted: Font:12 pt
Deleted: .
Deleted: revision
Deleted: extensively revised
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: separated
Formatted: Font:12 pt
Deleted: .
Formatted: Font:12 pt
Deleted: more emphasis was put into the overview schematic to highlight
Deleted: .
Formatted: Font:12 pt
Formatted: Font:10 pt
Formatted: Font:10 pt
Formatted: Font:10 pt

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

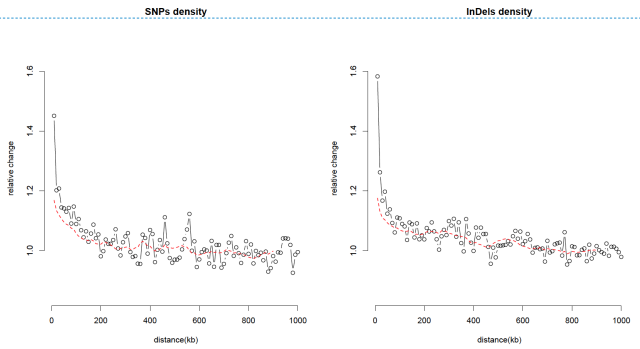
<TYPE>$$$BMR,$$$Network,$$$Calc
<ASSIGN>@@@DL,@@@XK, @@@TG,@@@STL
<PLAN>&&&AgreeFix,&&&MORE
<STATUS>%%%30DONE
[JZ2DL, XM, TG, STL: would you please help to fill in the stuff?]

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 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. In the revision, we have responded to and extended the referee's suggested in multiple respects, 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)

Formatted: Font:10 pt
Formatted Table

Deleted: good suggestions and we wished we had taken that more in this mission.

Deleted: . ... [163]

	• 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)
Excerpt 1 From Revised Supplement	We have called SV and SNVs from multiple ENCODE cell lines by integrating variousassays as shown in the following table. JZ2JZ: add Feng's table
Excerpt 2 From Revised Supplement	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 3 From Revised Manuscript	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: .(Excerpt 2) ... [164]

Deleted: . we showed

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"

Deleted: .(Excerpt 3) ina new main figure panel

Deleted: . We explored the SV events on the long range interactions (Excerpt 4)

Deleted: -

Deleted: . We estimated how much

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

Deleted: edges were

Formatted Table

Deleted: 1

Deleted: Regarding the relationship of SNV to SV ... [165]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

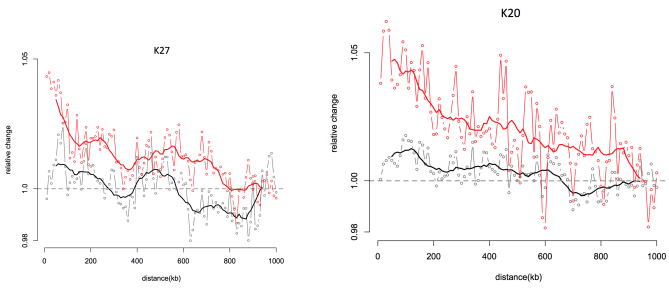
Deleted: Manuscript

Deleted: 2

Deleted: Relating SVs to histone modification marks.

Deleted: the

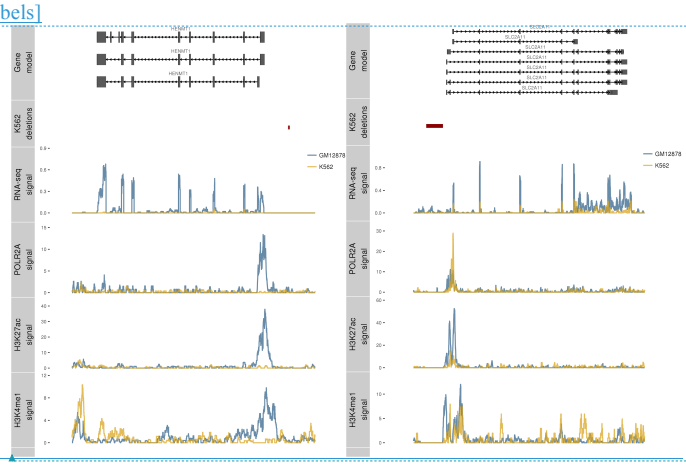
Deleted: it



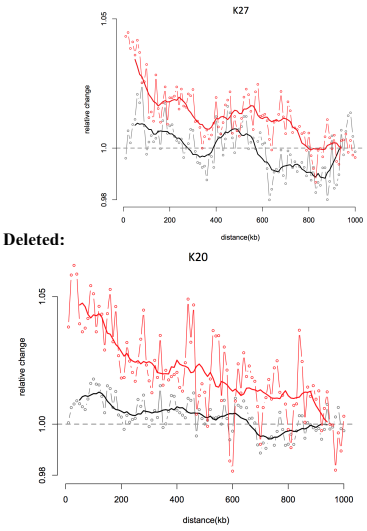
Excerpt 4 From Revised Supplement

We have shown in the follow figureseveral 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:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

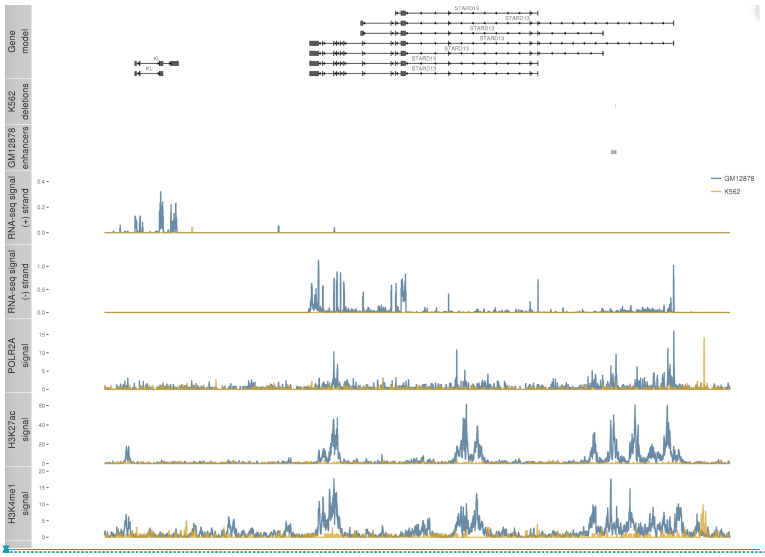
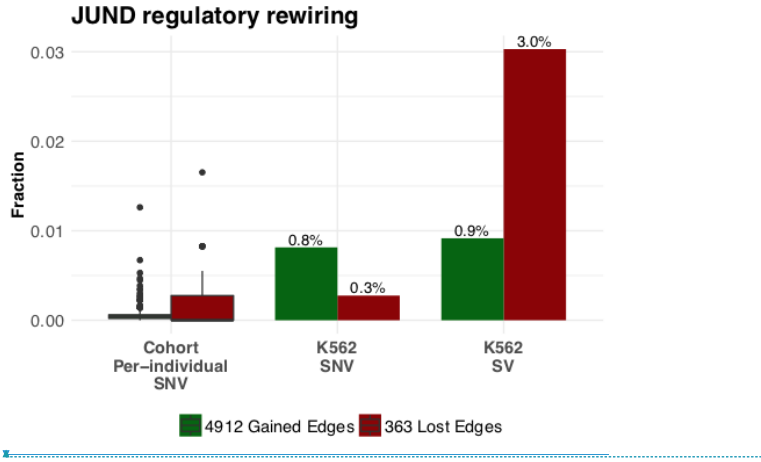
Deleted: 3

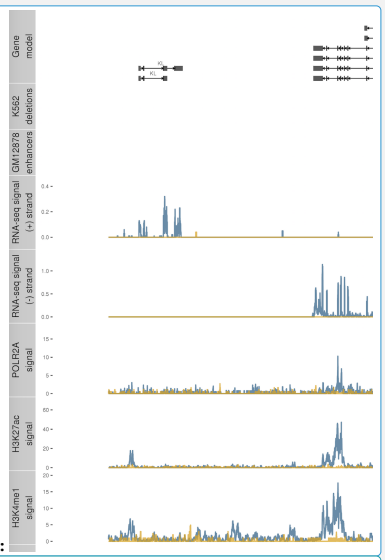
Deleted: Promoter and SV examples: ... [166]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Deleted: Manuscript

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

	
Excerpt 5 From Revised Manuscript	Figure 5. Rewiring sub-panel update (JZ2DL: pls describe what you have done here) 



Deleted:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

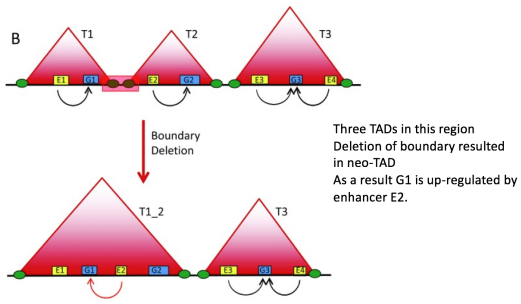
Deleted: Excerpt 4 From Feng's figure ... [167]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted Table

Deleted: ... [168]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Excerpt 6 From Revised Manuscript	Ask Feng to write a text 

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<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: Font:10 pt

Formatted Table

Formatted: Font:12 pt

Deleted: This is a limitation of the current technique, which we now discuss with greater emphasis (more details in the excerpt below). Thanks - this is exactly why we need so many data sets to model BMR. mention the factor of 10 or ENCODE data

Formatted: Font:12 pt

	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 From Revised Manuscript	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 From Revised Manuscript and supplement	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. 169), many of which are from tissue or primary cells. 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. In supplement: In total there are 2017 histone ChIP-seq and 52 Replication timing 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 fromtissue or primary cells. A summary of cell types for these features is given below. 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 [JZ2DL: please add the table of replication timing data]

Comment [12]: 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.

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Excerpt From Revised Manuscript	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.

Formatted Table

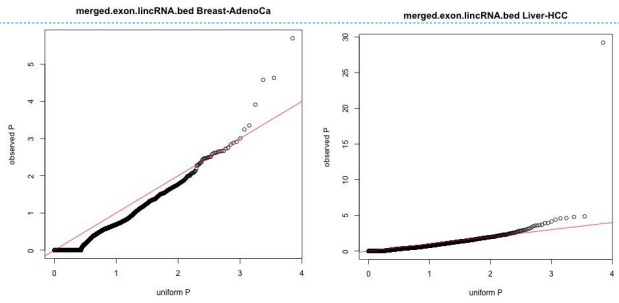
Deleted: in the further,

Deleted: ... [169]

<ID>REF4.10 – lncRNAs and BMR

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

Deleted: 75DONE

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 mutation burden of lincRNAs. We have added results on lincRNAs in our revised supplements (see excerpt below).
Excerpt From Revised Supplement	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. 

Formatted: Font:10 pt

Formatted Table

Formatted: Font:12 pt

Formatted: Font:12 pt

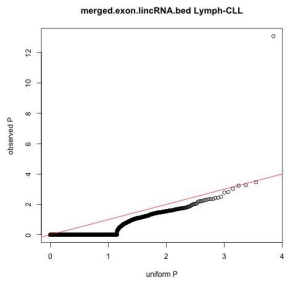
Deleted: the analysis of lncRNA by comparing BMRs

Formatted: Font:12 pt

Deleted: genes and lncRNAs.

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Deleted: Manuscript

<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).
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: Font:(Default) Times New Roman, (Asian) Times New Roman

Deleted: supplemantary

Formatted: Font:10 pt
Formatted Table
Deleted: to point out
Formatted: Font:12 pt
Deleted: fixed
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: Excerpt From ... [171]
Deleted: Excerpt From ... [170]

Formatted Table
Formatted: Font:10 pt
Deleted: to point out
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: fixed in our
Formatted: Font:12 pt
Deleted: manuscript
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: Excerpt From ... [172]

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 main encyclopedia paperprovidesannotations for all cell types based on just 4 assays. The breadth and accuracy of our annotation extends far beyond the encylopedia paper in this regard. 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

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

Deleted: would like to
Formatted ... [173]
Deleted: that
Deleted: , such as
Formatted ... [174]
Formatted ... [175]
Deleted: analysis
Formatted ... [176]
Formatted ... [177]
Deleted: comparison
Formatted ... [178]
Deleted: , are quite valuable.
Formatted ... [179]
Formatted ... [180]
Deleted: in our revised manuscript.
Formatted ... [181]
Comment [13]: Unsure about the use of the wor ... [183]
Deleted: main
Formatted ... [182]
Formatted ... [184]
Deleted: . ... [185]
Formatted ... [186]
Deleted: be considered as a "resource"
Formatted ... [187]
Deleted: it
Formatted ... [188]
Deleted: in our revised version
Formatted ... [189]
Deleted: our
Formatted ... [190]
Deleted: is
Formatted ... [191]
Formatted ... [192]
Deleted: instead
Comment [14]: Just flagging that numbers need ... [195]
Formatted ... [193]
Deleted: a
Formatted ... [194]
Deleted: biology paper.
Formatted ... [196]
Deleted: the best
Formatted ... [197]
Deleted: regulation
Formatted ... [198]
Deleted: listed some more details about the resource of

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 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 the BMR part

Specifically related toBMR estimation, the reviewer mentioned that there are many prior studies focusing on applications like cancer driver detection.

First, we thank the referee for pointing out these related references and we haved cited many of them in our initial submission (table R2 below). We want to point out that some of the references were either published after our initial submission (such as Marticorena et al. 2017) or with afocus other than BMR estimation (more details in the following table).

Second, we want to emphasize that the main goal of the BMR part in our paper is not to make novel driver discoveries but to illustrate how the richness of the ENCODE data can noticeably improve the accuracy of BMR estimation, as we have attempted to showin our updated Fig. 2.

Third, we want to point thatBMR estimation is just one out ofmany potential applications of ENCODE data. Even for the variant investigation part alone, we also have germline and SV analysis in this paper. There are many other ENCODE applications, such as regulatory activity, rewiring, and stemness, which are also key to investigate in cancer genomics.

Formatted ... [199]
Formatted ... [200]
Deleted: Table R1. Summary of annotation types
Formatted ... [201]
Deleted: example applications
Formatted ... [202]
Comment [15]: Just a general comment that the ... [203]
Formatted ... [204]
Formatted ... [205]
Deleted: paper
Formatted ... [206]
Deleted: ... [207]
Formatted ... [208]
Deleted: for the BMR
Formatted ... [209]
Deleted: part
Formatted ... [210]
Deleted: had been
Formatted ... [211]
Deleted: existing references
Formatted ... [212]
Deleted: to a lot of
Formatted ... [213]
Deleted: did cite
Formatted ... [214]
Comment [16]: Although this is true, and there i ... [215]
Formatted ... [216]
Deleted: a different focus
Formatted ... [217]
Deleted: We updated our reference as suggested.
Formatted ... [218]
Comment [17]: Again, not sure about the word g ... [219]
Formatted ... [220]
Deleted: a
Formatted ... [221]
Deleted: discovery
Formatted ... [222]
Deleted: that
Formatted ... [223]
Deleted: clearly shown in
Formatted ... [224]
Deleted: that the BMR application
Formatted ... [225]
Deleted: *of many*
Formatted ... [226]
Deleted: applications.
Formatted ... [227]
Deleted: investigations
Formatted ... [228]

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

Comment [18]: This image can't be modified, but it switches from using 'cited' to using 'yes' as the positive in the intial/revised column. Also, the reference formatting varies.

It might also be risky to provide a 'main point' for some of these papers. There is substantial room for disagreement about what the 'main point' of a paper is (if it can even be said to have a main point). It's also not clear what content is in the 'comments' column.

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:

<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: Font:10 pt

Formatted Table

Author Response	We thank the referee for the positive comment.

Formatted: Font:10 pt

Formatted: Font:12 pt

<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 bacdkground 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 have made the following changes to attempt to fully address the reviewer's comments. • A new supplementary table to summarize our 2069 features (vs. 169 in Martincorena et al., 2017) (Excerpt 1) This is the reason why we did not directly use these approaches (Imielinski et al 2016; Martincorena et al, 2017). • We added several references, and tried to provide a better context for previous work (Excerpt 2). • We have showed how more features with careful feature selection can improve BMR estimation (Excerpt 3). • We have stated clearly in the main text about our goal clearly in the main text: more data is 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 4). We thank the reviewer for identifying relevant references. 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,

Formatted: Space Before: 18 pt
Formatted: Font:10 pt
Formatted: Font:10 pt
Formatted Table
Deleted: background
Formatted: Font:10 pt
Deleted: We thank the reviewer for bringing out these references. In our revised manuscript, we tried to make it clear that we are not claiming to have developed negative binomial regression or to be the first to apply it to cancer genomics. We want to point out that negative binomial regression is a very standard statistical technique that has been used in many contexts in genomics. In fact, some of the references, such as Martincorena et al. 2017, came out after our initial submission in Aug 2017, and some of them have diverse focuses such as positive selection patterns instead of BMR estimation in noncoding regions. We have tried to give a better context of existing work in our revised manuscript. ... [229]
Formatted: Font:10 pt
Formatted ... [230]
Formatted: Font:12 pt
Deleted: . The ENCODE3 rollout dramatically ... [231]
Formatted: Font:12 pt
Deleted: and 20 PCs from 169 features in
Formatted: Font:12 pt
Deleted: Regarding this data,
Formatted: Font:12 pt
Deleted: <#>They are released in a ready to u ... [232]
Formatted: Font:12 pt
Deleted: accuracy either using the features dire ... [233]
Formatted: Font:12 pt
Deleted: . Our implication is that
Formatted: Font:12 pt
Deleted: . While it's valuable matching a canc ... [234]
Formatted: Font:12 pt, Not Italic, No underline
Deleted: *others*, are highly heterogeneous and hence
Formatted: Font:12 pt
Deleted: match
Formatted: Font:12 pt
Deleted: . A variety
Comment [19]: Having 'a goal of demonstrating ... [235]
Deleted: different
Formatted: Font:12 pt
Deleted: sets provide the best overall fit to estim ... [236]
Formatted: Font:12 pt

Excerpt 1 From Revised 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

Excerpt 2 From Revised Manuscript

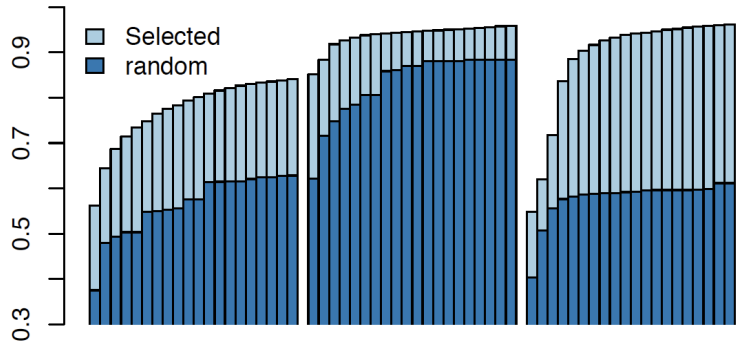
Many methods have incorporated effects from multiple genomic features by techniques such as negative binomial regression and poisson regression.

Excerpt 3 From Revised Manuscript

The 2017 uniformly processed histone modification signal tracks and 52 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.

Formatted: Font:10 pt
Formatted: Font:10 pt
Formatted: Font:10 pt
Formatted Table
Deleted: # histone marks
Formatted: Font:Times New Roman, 10 pt
Deleted: Manuscript (in
Formatted: Font:10 pt
Deleted:)
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
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
Formatted: Font:Times New Roman, 10 pt
Formatted Table
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt

	
Excerpt 4 From Revised Manuscript	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.

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<ID>REF5.3 – BMR: TCGA benchmark

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

Referee Comment	2. 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 and we fully agree with the referee that it is useful to compare our BMR to established benchmarks. In our revised manuscript, we have benchmarked our BMR to other data sets as suggested.

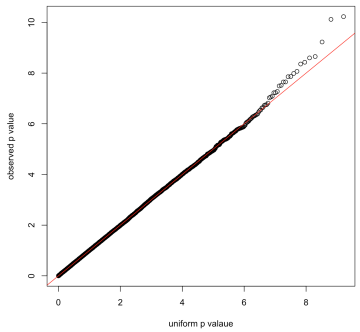
Formatted: Font:10 pt
Formatted: Font:10 pt
Formatted Table
Deleted: . ... [237]
Deleted: . We
Formatted: Font:10 pt
Formatted: Font:12 pt
Formatted: Font:12 pt

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, what they have done is to develop several randomization schemes accepted by multiple groups, which are supposed to measure the BMR rate to calibrate driver detection. Hence, we tried to compare our estimated BMR with such randomizations.

Please note that this work is comparing to accepted PCAWG benchmarks, which are not fully published yet, so we are only including 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.

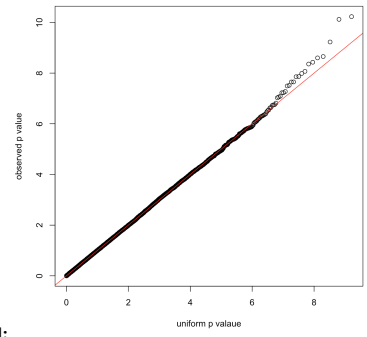
1. Using a permuted breast cancer dataset, we performed BMR estimation and calculated somatic mutation burden on the CDS regions of ~20k protein coding regions. We found no gene burdening in this randomized data set (QQ plotgiven below).

Figure R 2. QQ plot of observed vs. uniform p values from permuted breast cancerdata 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 other published TCGA benchmarks only interrogated protein coding regions, where the relative rates of synonymous and nonsynonymous mutations can

Formatted: Font:12 pt
Deleted: . In fact, we
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: and one
Formatted: Font:12 pt
Deleted: ... [238]
Formatted: Font:12 pt
Deleted: what
Formatted: Font:12 pt
Deleted: did in the response is
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: just putting
Formatted: Font:12 pt
Deleted: the
Formatted ... [239]
Deleted: part
Formatted: Font:12 pt
Deleted: into
Formatted: Font:12 pt
Deleted: actual supplement
Formatted: Font:12 pt
Deleted: ... [240]
Formatted: Font:12 pt
Deleted: on the permuted dataset for breast ca ... [241]
Formatted: Font:12 pt
Deleted: burdened
Formatted: Font:12 pt
Deleted: there.
Formatted: Font:12 pt
Deleted: plot was given
Formatted: Font:12 pt
Deleted: .
Formatted: Font:12 pt
Deleted: the …bserved vs. uniform p value…alues f ... [242]

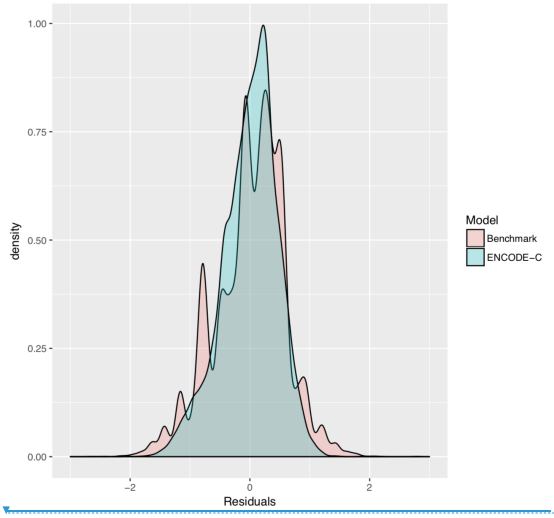


Deleted:
Formatted ... [243]
Formatted: Font:12 pt
Deleted: down sampled
Formatted ... [244]
Deleted: Results show that we have comparable ... [245]
Formatted ... [246]

be used to calibrate BMRs. This particular calibration is not possible innoncoding regions.

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 ENCODE-C 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: , which

Deleted: as straightforward in the noncoding

Formatted: Font:12 pt

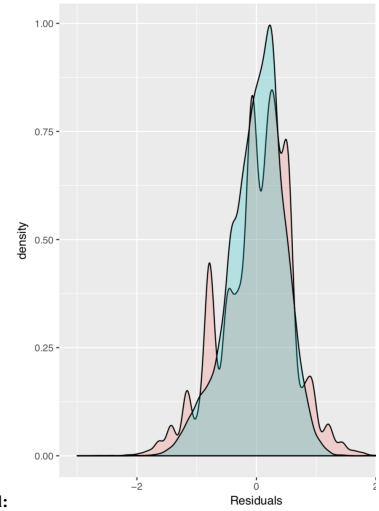
Formatted: Font:12 pt

Formatted: Font:12 pt

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

Formatted: Font:12 pt

Deleted:



<ID>REF5.4 – Power analysis

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

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

JZ2JZ: add

Referee Comment	4. How do the new "compact annotations" lead to improved results over traditional annotations? 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.
Author Response	We thank the referee for recognizing the value ofselecting genomic elements. Following the reviewer's suggestions, in our revised manuscript we have completed a formal power analysis, he most important contribution to power comes from including additional functional sites, which supports the extended gene concept. Secondary and lesser, contributions to power come from removing non-functional sites. The core assumption of our compacting annotations is that we can accurately distinguish the more important functional nucleotides from the less important ones through the guidance of many functional characterization assays.
Excerpt 1 From Revised Supplementary file	Regarding compact annotation: In our initial submission, the assumption is that we were trimming off the nonfunctional sites while preserving the functional ones. Two examples can explain the motivation of this assumption. 1) Enhancers: Traditionally, enhancers were called as 1kb peak regions, which may introduce 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 correlates with the mutation hotspots (red block) for this well-known driver site (blue line for pan-cancer and green line for liver cancer).

Formatted Table

Formatted: Font:10 pt

Deleted: his/her positive comment on

Deleted: of selecting

Deleted: element and suggestion on the power analysis.

Deleted: show in

Deleted: that the

Deleted: is of course by

Deleted: and then secondarily,

Deleted: , but to a lesser extent.

Deleted: in

Deleted: ... [247]

Comment [20]: This does not appear to be an excerpt from the manuscript. It is unclear to me what is an excerpt from the manuscript.

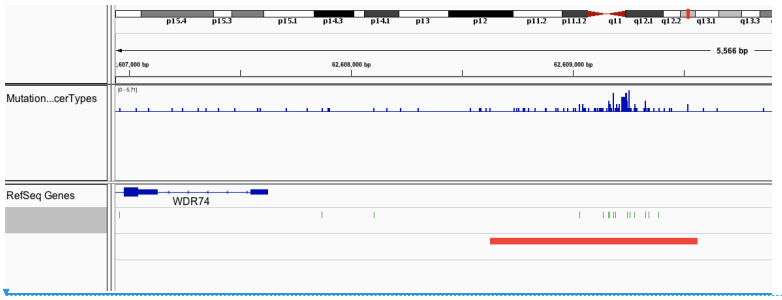
Deleted: a

Deleted: admittedly introduced a lot of obviously

Comment [21]: Do we actually have some evidence for this? Or is it just a hypothesis? What is the basis for the hypothesis?

Deleted: perfectly

Comment [22]: Is this text part of the supplement?

	
Excerpt 2 From Revised Supplementary file	Regarding extended genes

Deleted:

Comment [23]: Requires completion.

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

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

Referee Comment	4. 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.

Formatted: Font:10 pt

Formatted Table

Deleted: In fact, we cited the Lawrence et al, 2014 paper (and the paper before this one in the same group) in our initial submission. The Martincorena, 2017 was published after our submission for it is impossible for us to cite in the last round.

Formatted: Font:12 pt

Deleted: bring out

Formatted: Font:12 pt

Formatted: Font:12 pt

Deleted: it in

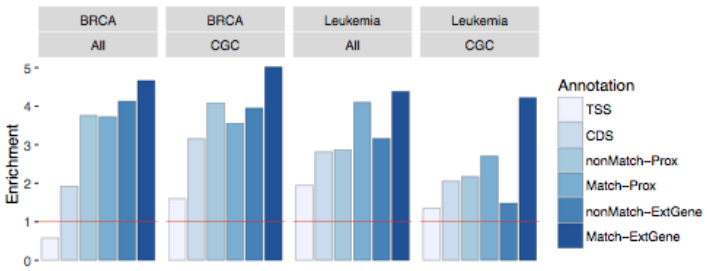
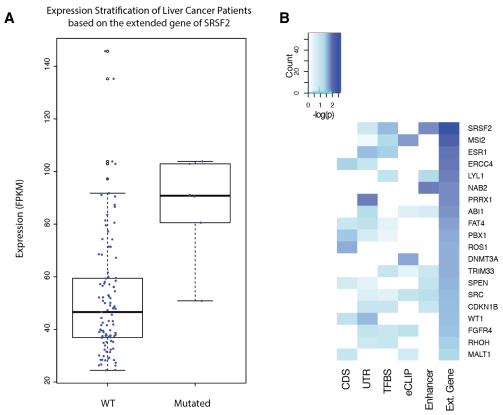
Formatted: Font:12 pt

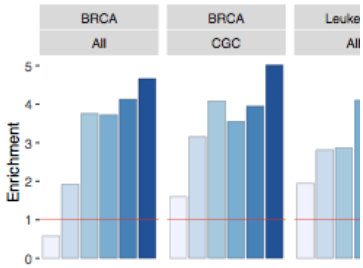
<ID>REF5.6 – BMR & Power analysis: detailed driver detection comparison

<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. We have now labeled known driver genes in our calculations with supporting literature and further compared our results with established methods. [JZ2MG: can we add the driver gene comparison with PCAWG, only in the resonse] 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. Hence, we feel it is out of scope of this paper to make a detailed comparison of cancer driver discovery rates. We nonetheless hope to illustrate how the extended gene concept can be used in cancer. We have re-organized allrelated analysis to better demonstrate our idea in the revised manuscript. In summary, we have used extended genes to: • Better annotation disease associated germline variants (see Excerpt 1). • Better stratify gene expression level by mutational status (see Excerpt 2).
Excerpt 1 From Revised Manuscript (in main figure and supplement text)	We extracted allbreast 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 allLD 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 ... [249]
Deleted: pointing
Formatted ... [251]
Deleted: out
Deleted: want to emphasize
Formatted ... [250]
Formatted ... [252]
Formatted ... [253]
Formatted ... [254]
Deleted: goal
Formatted ... [255]
Deleted: paper
Formatted ... [256]
Deleted: illustrate
Formatted ... [257]
Deleted: help the accuracy of
Formatted ... [258]
Deleted: . It
Formatted ... [259]
Deleted: the
Formatted ... [260]
Deleted: our
Formatted ... [261]
Deleted: discoveries. However, we did labeled t ... [262]
Formatted ... [263]
Formatted ... [264]
Deleted: have also tried
Formatted ... [265]
Deleted: emphasize
Deleted: is useful for much more than driver discover.
Formatted ... [266]
Formatted ... [267]
Deleted: added the following two aspects here. ... [268]
Formatted ... [269]
Deleted: gave us much better cancer associated ... [270]
Formatted ... [271]
Deleted: *2. Expression stratification analysis usi* ... [272]
Comment [24]: Is this correct?
Formatted ... [273]
Deleted: all the breast
Formatted ... [274]
Formatted ... [275]
Deleted: Catalogue
Formatted ... [276]
Deleted: irrelatvent
Formatted ... [277]
Deleted: BMR
Formatted ... [278]
Deleted: therapy
Formatted ... [279]
Deleted: those
Formatted ... [280]
Deleted: all the LD
Formatted ... [281]
Deleted: sites
Formatted ... [282]
Formatted ... [283]
Formatted ... [284]
Formatted ... [285]
Formatted ... [286]
Formatted ... [287]

	
Excerpt 2 From Revised figure and supplementary text	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. 



Deleted:

Deleted: For a given test region, we consider the expression (FPKM) of patients with a mutation or no mutation in that region to be separate distributions. By using a wilcoxon two-sided test, we test to see whether the expression of mutated patients versus non-mutated patients is different. The test regions we consider are the CDS, UTR, TFBS, eCLIP, Enhancer, and Extended Gene Definition. We find that in many genes, the p-value associated with expression stratification between the two groups is much more significant when using the extended gene than any of its individual parts, suggesting an advantage of the extended gene. Furthermore, when performing this analysis on liver cancer patients using the HepG2 annotations, we find that mutations in the extended gene of SRSF2 give the strongest p-value for stratifying expression of that gene. SRSF2 is a well known splicing factor involved heavily in driving hepatocellular carcinoma development. \cite{28082404}. The specific case of SRSF2 is shown in Panel A. Mutated samples in the extended gene definition are more likely to have higher expression of SRSF2 when compared to WT. Panel B below shows the -log p-value of stratifying expression of mutated and non-mutated patients in different genes using different test regions. ... [288]

Deleted: Manuscript (in main

Deleted:)

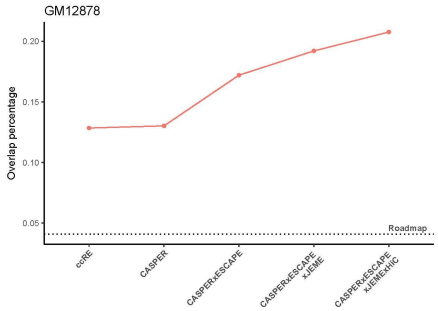
<ID>REF5.7 – Annotation: false positive rates of enhancers

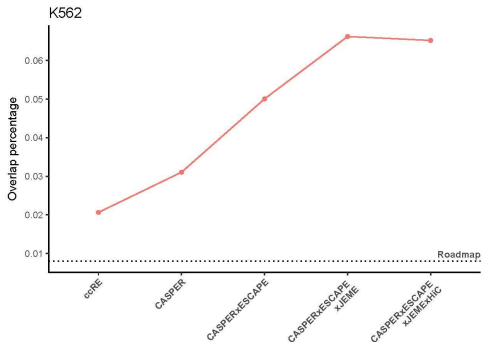
<TYPE>$$$Power,$$$Text
<ASSIGN>@@@JZ,@@@MTG
<PLAN>&&&AgreeFix
<STATUS>%%%95DONE

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. As suggested, we have revised our manuscript to discuss the quality of annotations, including: • Enhancers (details in Excerpt 1 below) • Enhancer-gene linkages (details in Excerpt 1 to REF 5.8) • TF regulatory networks (details in Excerpt 1-3 to REF 5.12) We have added further internal comparisons of relative performance after incorporating additional novel assays, and we now include FDRs for our methods. 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 find figures, numbers and tables here]
Excerpt 1 From Revised supplement	As for the enhancer part, with the ensemble method, we can get more accurate annotation and pin-point to sequences where transcription factors would actually 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 data set 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: Font:10 pt
Formatted Table
Deleted: pointing out the
Deleted: We fully agree with the referee that it is important to provide such information.
Deleted: added a whole section in our
Deleted: qualities
Deleted: our
Deleted: . We further extended such QC analysis from enhancers to our enhancer
Deleted: ,
Deleted: -gene
Deleted: the rewiring quantifications (details in updated supplement section xxx).
Deleted: Specifically for the enhancer part, we are actively involved with of the ENCODE enhancer challenge project and one of our authors co-leads that project. We
Deleted:
Deleted: , and they are
Comment [25]: There are sections here that were unclear, or that detracted from the response. Other sections are rearranged.
Comment [26]: What enhancer part does this refer to?
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
Deleted: , for example
Formatted: Font:Times New Roman, 10 pt
Deleted: would not be very practical at this stage
Formatted: Font:Times New Roman, 10 pt
Deleted: an
Formatted: Font:Times New Roman, 10 pt
Deleted: manuscript (in
Deleted:)
Deleted: definitely
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
Deleted: assess
Formatted: Font:Times New Roman, 10 pt

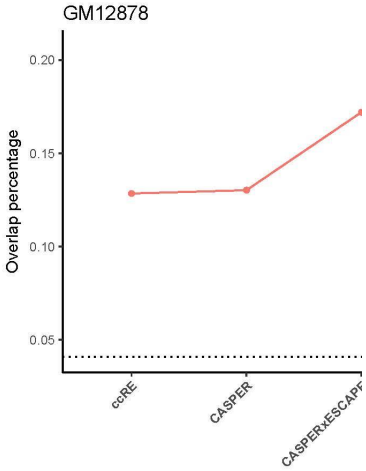
percentage for our annotation is higher than that of the Roadmap annotation, and is also higher than the main encyclopedia enhancer annotation annotation (ccRE).





Deleted: much

Formatted: Font:Times New Roman, 10 pt



Deleted: ... [289]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

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

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

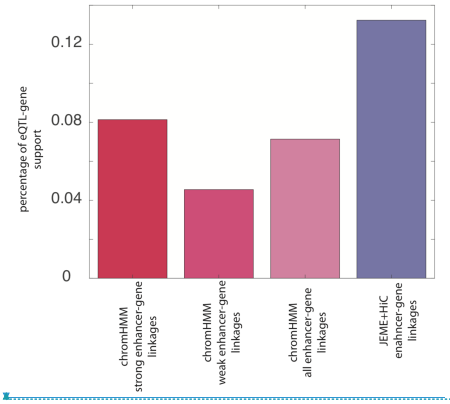
Referee Comment	7. The authors claim superior quality of gene-enhancer links and gene communities derived from their machine learning approach. The method should at least be outlined in the main text, and accompanied

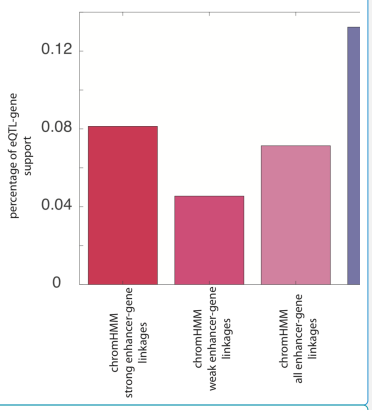
Formatted: Font:10 pt

Formatted Table

	by data supporting its accuracy and better performance compared to existing approaches.
Author Response	Again we thank the referee for their comments and we totally agree that it is important to provide quality comparison of annotations. We have tried to fully addressed the referee's comment by • Adding a section to the supplement to compare our JEME+Hi-C enhancer targetsthan the chromHMM ones (excerpt 1 below) • 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 2 below)
Excerpt 1 From Revised supplement	*1. Regarding the gene-enhancer linkages* 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: We
Deleted: the
Deleted: developed a method
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: . As suggested, we have briefly explained our method in the updated main text
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: added a whole new section to discuss these points in the revised supplementary file.
2. Regarding the
Formatted: Font:12 pt
Deleted: incorporates HiC data to our previous ... [290]
Formatted: Font:12 pt
Deleted: against published chromHMM enhanc ... [291]
Formatted: Font:12 pt
Deleted: *2. Regarding the*
Formatted: Font:12 pt, Not Italic, No underline
Deleted: *methods* ... [292]
Formatted: Font:12 pt
Deleted: other methods like
Formatted: Font:12 pt
Deleted: by extending our analysis from 122 GM ... [293]
Formatted: Font:12 pt
Deleted: after dimension reduction (Excerpt 2).
Formatted: Font:10 pt
Formatted ... [294]
Formatted: Font:Times New Roman, 10 pt
Deleted: Manuscript ... [295]
Formatted: Font:10 pt
Deleted:)
Formatted: Font:Times New Roman, 10 pt, Highlight
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt, Highlight
Formatted: Font:Times New Roman, 10 pt
Deleted: is done to make sure
Formatted: Font:Times New Roman, 10 pt
Deleted: all of
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt, Highlight
Formatted: Font:Times New Roman, 10 pt
Formatted: Font:Times New Roman, 10 pt

	
Excerpt 2 From Revised Manuscript (in supplement)	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 GM and K526 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 centroidds are used to calculated the correlation. As



Deleted:

Formatted: Font:(Default) Helvetica Neue, (Asian) Helvetica Neue

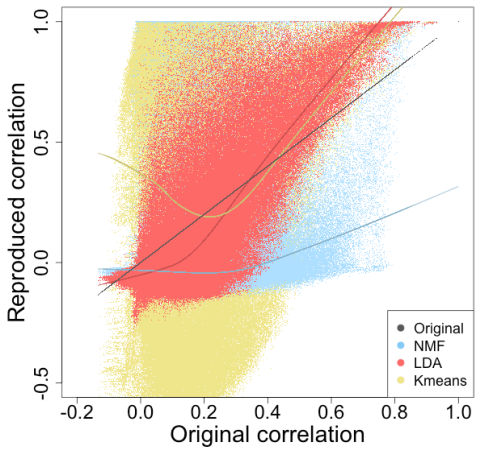
Formatted: Font:10 pt

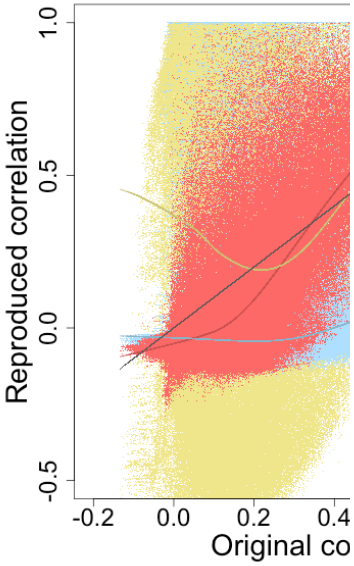
Formatted: Font:10 pt

Formatted: Font:10 pt

Deleted:

Formatted: Font:10 pt

	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:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<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: Font:10 pt

Formatted Table

Formatted: Font:12 pt

Excerpt From Revised Manuscript	Wait for the main text JZ2JZ

<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 From Revised Manuscript	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.

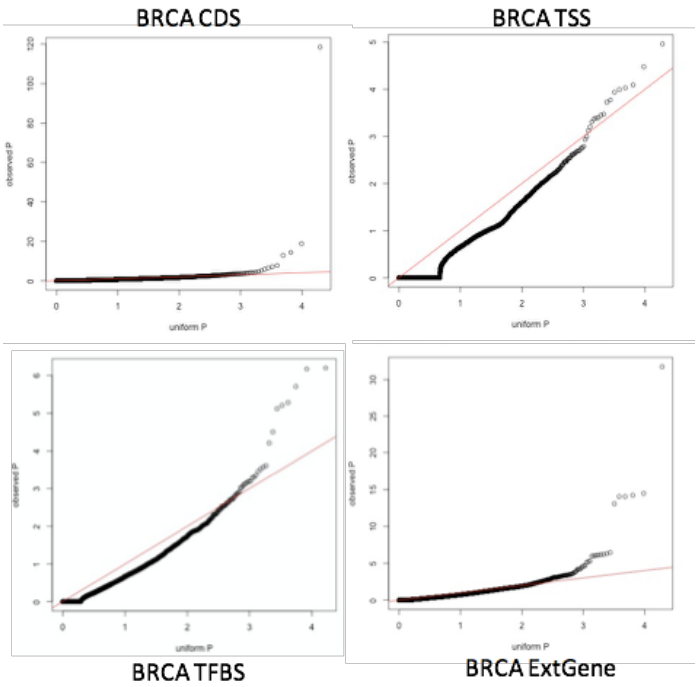
Formatted: Font:10 pt
Formatted Table
Formatted: Font:12 pt
Formatted: Font:10 pt

<ID>REF5.11 – Additional QQ plots

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

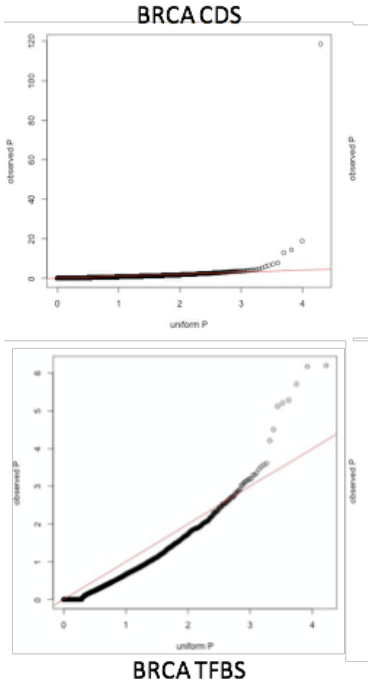
Referee Comment	10. The significance analysis of cancer cohorts (Figure 2) should highlight known cancer genes versus those newly found in this study.

Formatted Table
Formatted: Font:10 pt

	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. JZ2MG: should we add the new genes we discovered? Too much cancer driver detections then, or out of scope? 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 From Supplement	QQ-plot for breast cancer on various annotations. 

Formatted: Font:12 pt

Formatted: Font:12 pt



Deleted:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<ID>REF5.12 – Sequence coverage

<TYPE>$$$BMR,$$$Text

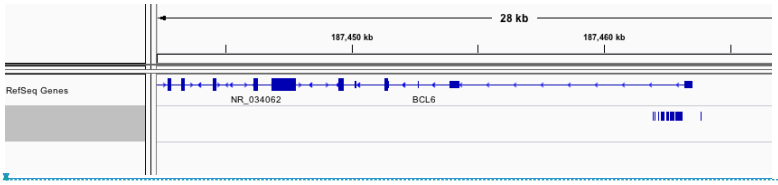
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

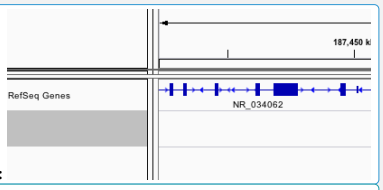
Referee Comment	Do the authors include sequence coverage in their method?
Author Response	We did not consider sequence coverage but this is a good point. We included discussion of this point in our revised manuscript.
Excerpt From Revised Manuscript	We hope that in the future new models that incorporate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

Formatted Table
Formatted: Font:10 pt
Deleted: Thanks for pointing this out.
Formatted: Font:12 pt
Formatted: Font:12 pt
Deleted: in the
Formatted: Font:12 pt
Formatted: Font:12 pt
Formatted: Font:10 pt
Deleted: can
Formatted: Font:10 pt
Deleted: ,
Formatted: Font:10 pt
Formatted: Font:10 pt
Deleted: would further integrate
Formatted: Font:10 pt

<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.
Excerpt From Revised Manuscript	

Formatted Table
Deleted: Check what is this question?


Deleted:
Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<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 experimental interactions.* We showed that the ENCODE ChIP-seq/eCLIP based networks can capture a higher fraction of standard interactions (from manually curated networks from TTRUST) than protein physical networks, including Biogrid and String experimental interactions (see details in excerpt 1). *1.2 Comparison with DHS-based imputed networks* We showed that ENCODE ChIP-seq based networks provided better correlations with TF knockdown experiments than the DHS-based imputed network provided in Neph et. al. 2012. (see details in excerpt 2). *1.3 False positive rate estimation of the ChIP-Seq based networks* 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 the false positives (see details in excerpt 3). **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: Font:10 pt

Formatted Table

Deleted: up this

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

2.1 High quality of enhancer definitions after integrating many histone ChIP-seq and DHS, and STARR-Seq data

We provide better enhancer definitions after integrating various assays. Please see details in response to "<ID>REF5.7 – Annotation: false positive rates of enhancers".

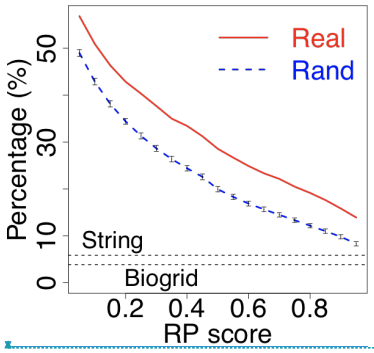
2.2 High quality of enhancer-gene linkages

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 5.8.

Excerpt 1 From Revised Supplement

Regarding Comparison with Biogrid and String experimental interactions

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

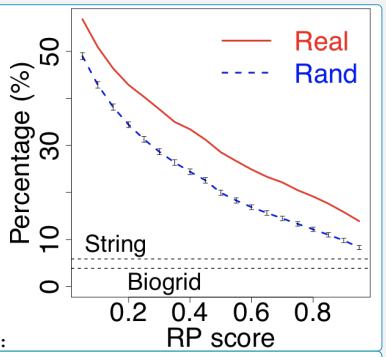
Deleted: 9

Formatted: Font:10 pt

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:10 pt

Deleted: Manuscript (in supplement)

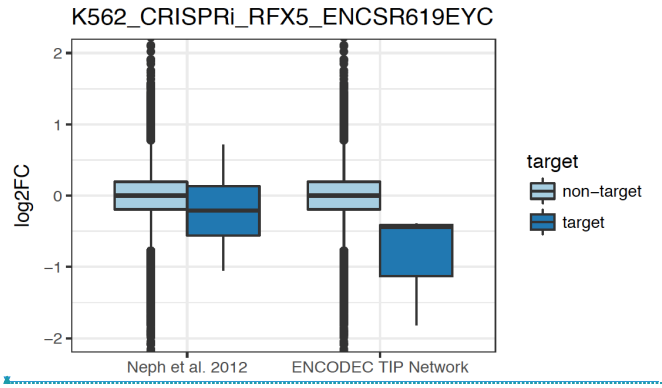


Deleted:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt

Formatted: Font:10 pt

	TTRUST networks using the stub-rewiring method that preserved the gene network degrees (Milo et al., 2002).
Excerpt 2 From Revised Manuscript	*Regarding 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 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. 

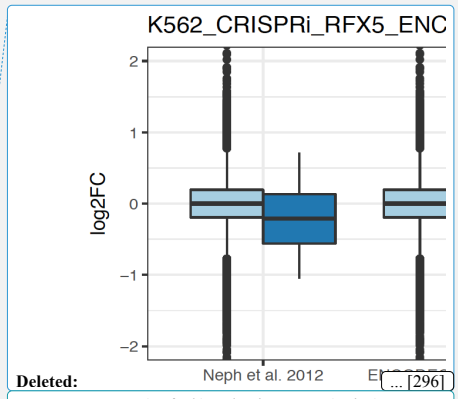
Formatted: Font:10 pt

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:10 pt

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:10 pt



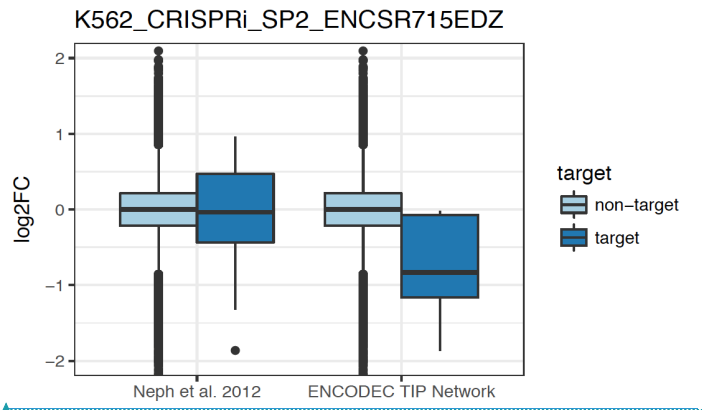
Deleted: ... [296]

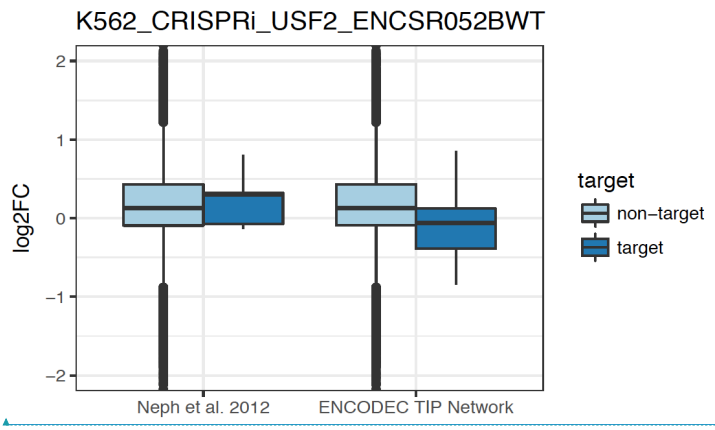
Formatted: Font:(Default) Helvetica Neue, (Asian) Helvetica Neue

Formatted: Font:(Default) Helvetica Neue, (Asian) Helvetica Neue

Formatted: Font:(Default) Helvetica Neue, (Asian) Helvetica Neue

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt





1

Excerpt 3 From Revised Manuscript

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

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.

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt

Formatted: Font:Times New Roman, 10 pt

Formatted: Font:10 pt

Formatted: Font:10 pt

Formatted: Font:10 pt

Formatted: Font:10 pt

Formatted: Font:10 pt

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:

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

Formatted: Font:(Default) Helvetica Neue, (Asian) Helvetica Neue

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt

Formatted: Font:10 pt

Formatted: Font:10 pt

Formatted: Font:10 pt

Formatted: Font:10 pt

<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 From Revised Manuscript	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.

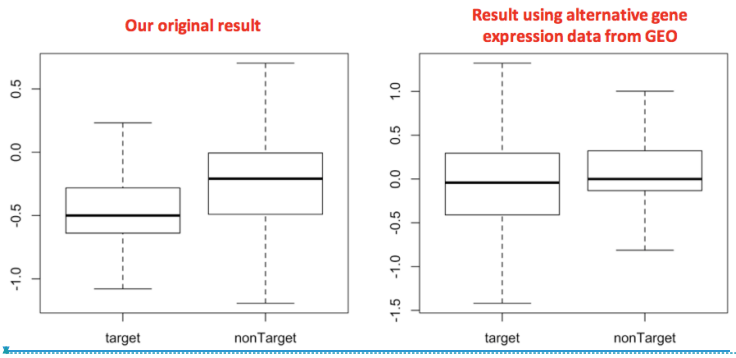
Formatted Table

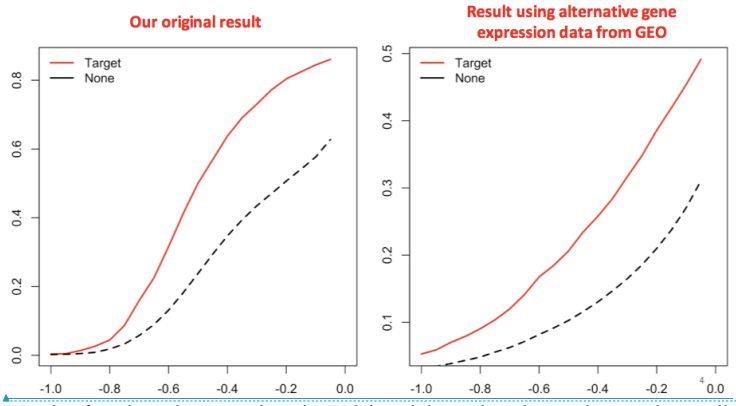
Formatted: Font:10 pt

Formatted: Font:12 pt

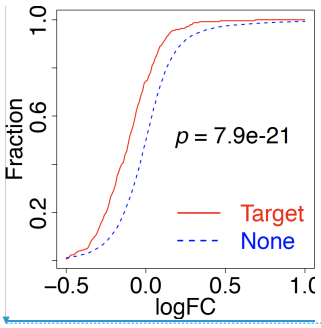
Deleted: 1.

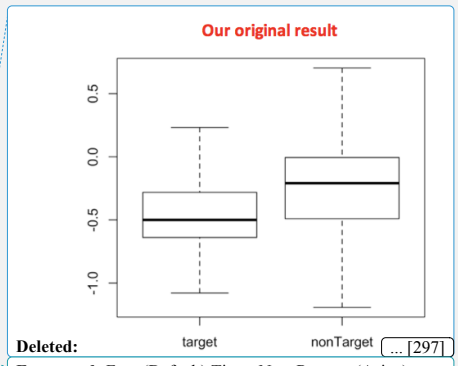
Formatted: Font:10 pt





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

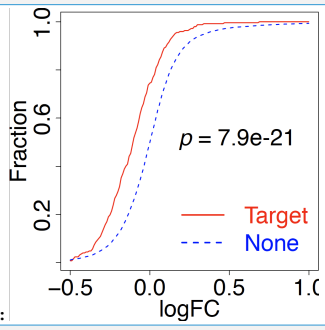




Deleted: ... [297]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman



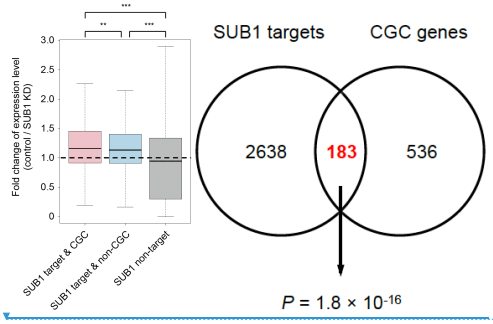
Deleted:

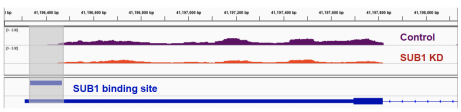
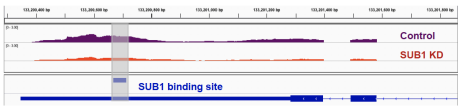
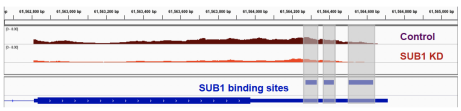
<ID>REF5.16 – SUB1 analysis

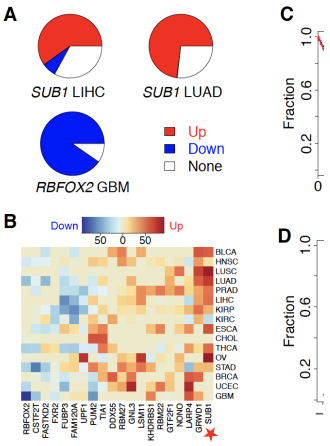
<TYPE>$$$NoveltyPos,$$$Calc
<ASSIGN>@@@MRS,@@@JL,@@@YY
<PLAN>&&&MORE
<STATUS>%%%95DONE

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 reasonable biological function in relation to cancer. We were able to figure out how it collaborates with other regulators, such as MYC, to really demonstrate how our multi networks, including the TF and RBP networks, really fit together to relate to biology. 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). ● 1. We investigated SUB1 regulation potential in different cancer types and found that they are consistent as below (see excerpt 1 below). 2. 3. 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. Finally, we did a new small scale validation experiment to drill into the SUB-1 MYC connection and validate it partially on several important oncogenes. While we do not think this is a 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. Sub1 regulated by myc

Formatted Table
Formatted: Font:10 pt
Formatted: Justified
Deleted: SUB-1
Deleted: , and actually, we were able to elaborate on this considerably. ... [298]
Deleted: found
Deleted: found
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:
Deleted: We checked the 3' UTR expression level of SUB1 target genes and found that the target genes are significantly down-regulated upon SUB1 KD. In addition, we found enrichment of SUB1 target genes for CGC (Cancer Gene Census) genes.
Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

	[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,]
Excerpt 1 From Revised Manuscript (in supplement)	
Excerpt 2 From Revised Manuscript (in supplement)	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. 

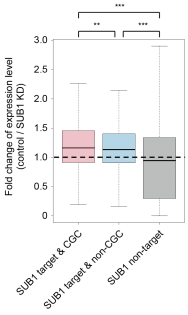
Gene	Functions	PMID	Expression profiles of the 3' UTR
BRCA1	The gene is involved in maintaining genomic stability	12677558, 17416853, 23620175, 16551709	
POLE	The gene is involved in DNA repair and replication	26133394, 28423643	
FEN1	The gene is involved in DNA repair and replication	20929870, 22586102	



Inference of RNA binding proteins that driv
ENCODE eCLIP data, we applied RABIT framev
target genes are differentially regulated in div
percentage of patients with target genes signif
not regulated (white) is shown for each cancer
to show the percentage of patients in each can
(red) or down regulated (blue). (C) All TCGA L
adenocarcinoma (LUAD) patients are divided t
predicted by RABIT. The overall survival was s
between RABIT regulatory activity and overall
cumulative distributions of gene expression af
predicted target genes and none-target genes.
expression changes is done through Wilcoxon
compared between predicted *SUB1* targets an

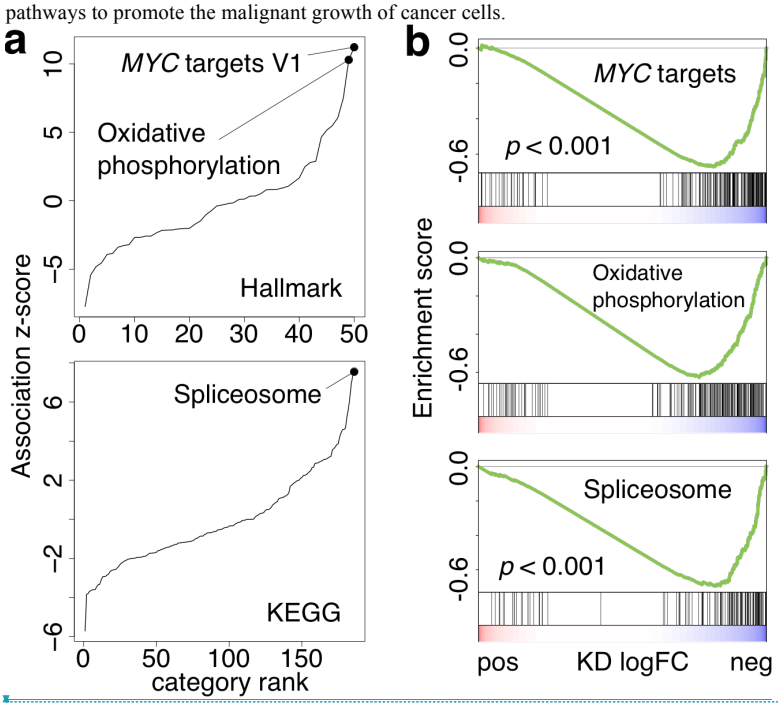
Deleted:

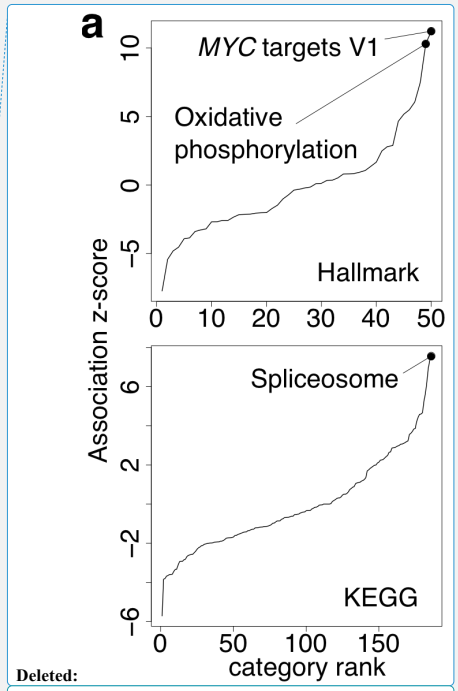
Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman



Deleted:

... [299]

Excerpt 3 From Revised Manuscript (in supplement)	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:

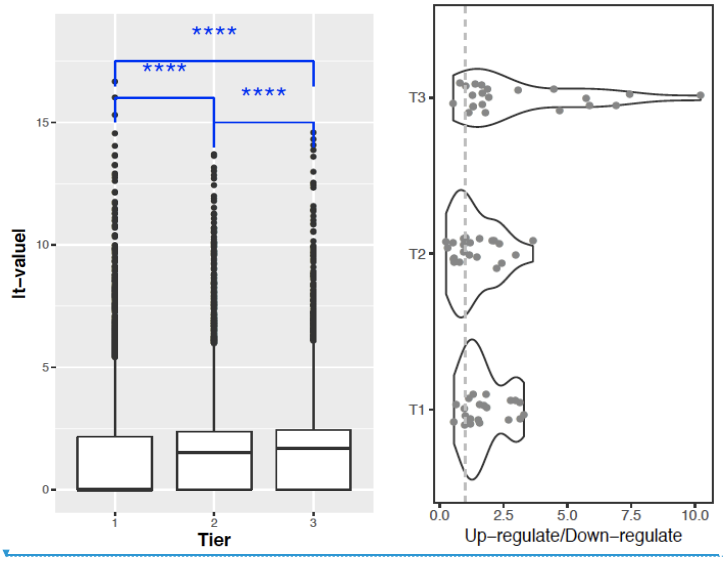
Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

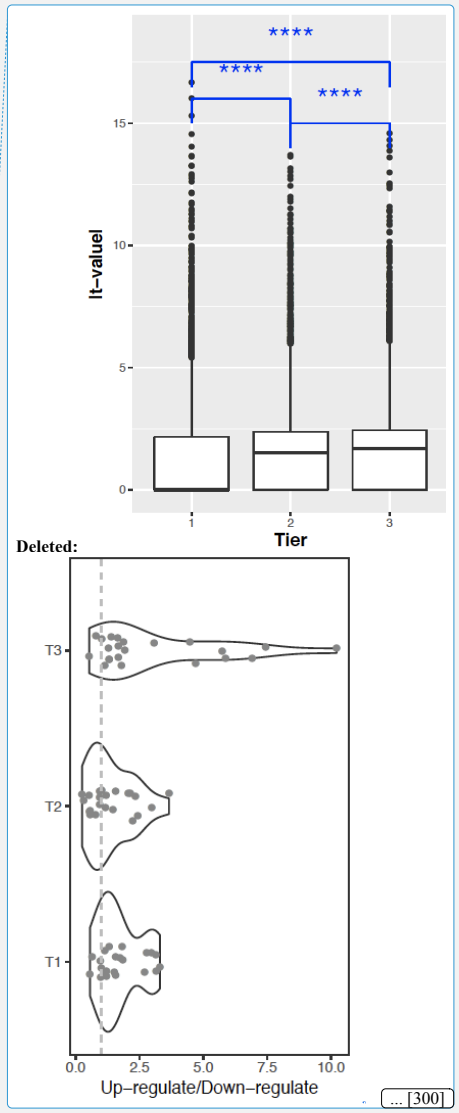
<ID>REF5.17 – Significance of regulatory network hierarchy

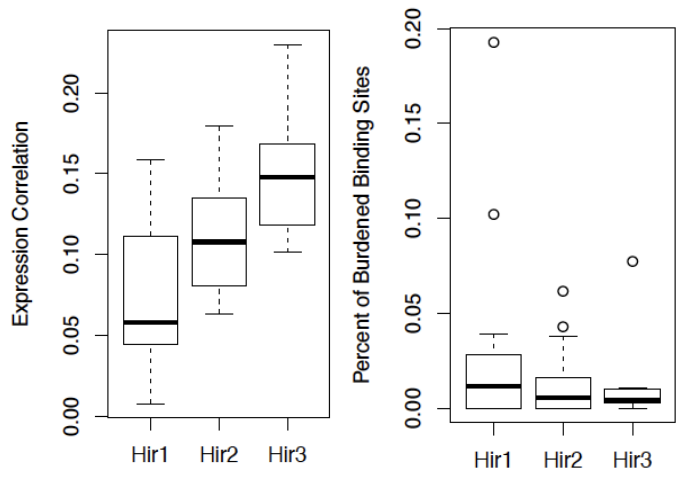
<TYPE>$$$Network,$$$Calc
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%99DONE

Referee Comment	15. The manuscript claims that transcription factors placed at the top level of the network hierarchy are enriched in

Formatted Table

	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? 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 From Revised Manuscript (in supplement)	Supplementary Figure X. 





<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?
Author Response	We thank 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 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. See excerpt for more details.

Formatted: Font:10 pt

Formatted Table

Deleted: ... [301]

Deleted: 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. ... [302]

Moved down [13]: 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.

Formatted: Font:Times New Roman, 10 pt

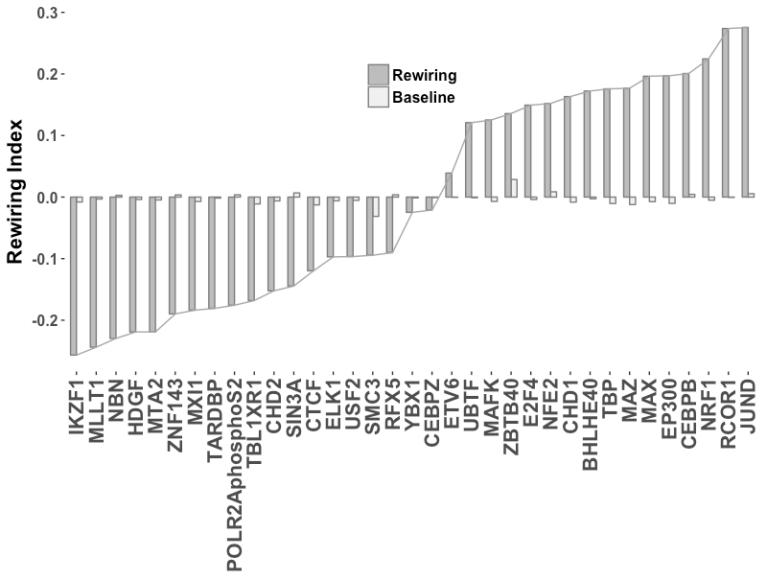
Deleted: We find that on average 1.36% of regulatory edges are false positives.

Excerpt From Revised Supplement

… 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.

Formatted: Font:10 pt

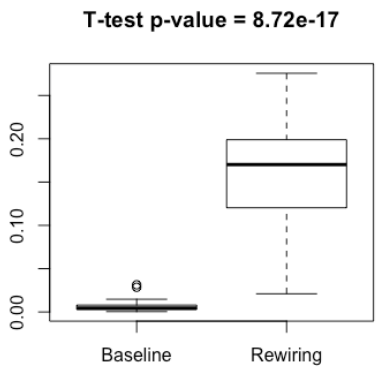
Deleted: Manuscript (in supplement)

Moved (insertion) [13]

Formatted: Font:Times New Roman, 10 pt

Deleted: ... [303]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt

<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 referee for pointing this out. We agree with the referee's suggestion and took this opportunity to significantly expand the statistical aspects of regulatory network rewiring and H1 stemness model. In summary, we have done the following analysis. *1. Regarding the false positives of the rewired edges* 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".

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 10 pt

Formatted Table

Comment [27]: put more in the suppl and summarize less

We then tried to measure the baseline of rewiring using replicates of ChIP-seq experiments, as we explored in REF5.18. We find that approximately 1.36% of rewired regulatory edges are false positives using examples from CML.

2. Regarding the statistical testing in the normal-tumor-stem analysis

- *Regarding our original rewiring analysis 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 Excerpt 1 below.
- *Regarding our new analysis using PCA/RCA*
  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 Wilcoxon test to compare the tumor-stem and normal-stem distance using multiple datasets. We found that tumor cells are more similar to stem cells in general, which is consistent with the findings in the recent TCGA paper \cite(TCGA i stemness). Please see details in Excerpt 2 below.

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

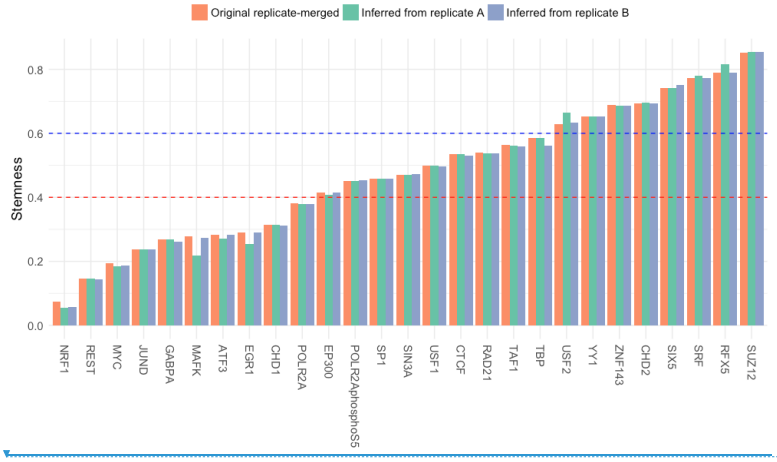
Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

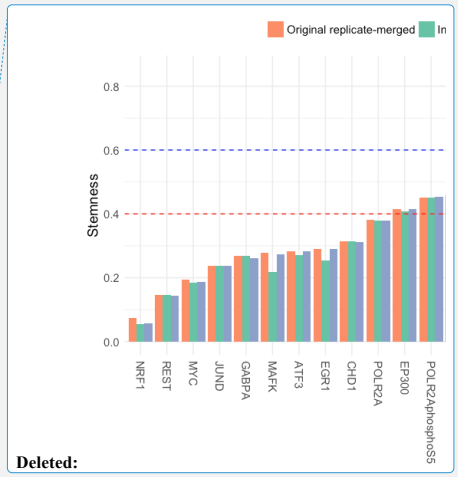
Comment [28]: supplement

Deleted: datasets

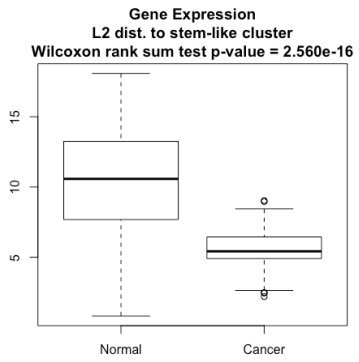
Excerpt 1 From Revised Manuscript (in supplement)

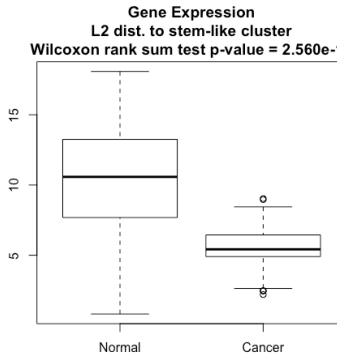
The H1 stem cell model uses fractional overlap of rewired edges between cancerous cell types vs. H1. Therefore we attempted to evaluate statistical significance of our model by measuring how much of H1 network changes are due to noise and use of other normal cell types to evaluate how much of rewired edges overlaps with H1.





Deleted:

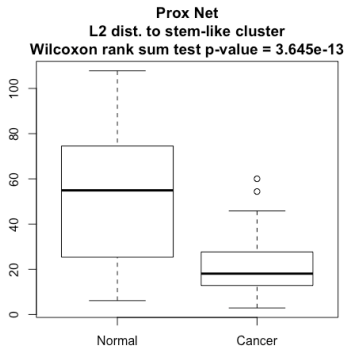
	Using replicates of H1-hESC ChIP-seq experiments, we made two independent H1 networks in addition to original replicate merged H1 network, and we made recalculated stemness of TF, whether they rewire toward or away from H1. We find that the results of all of stemness direction is reproduced using either replicate.
Excerpt 2 From Revised Manuscript (in supplement)	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.  Suppl. Fig. B



Deleted:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman



Suppl. Fig. C

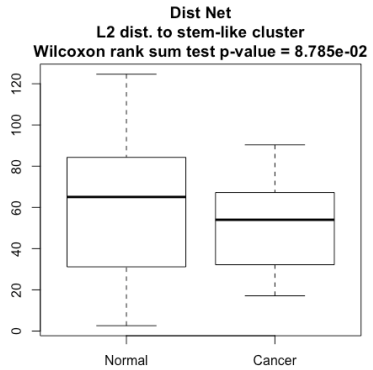
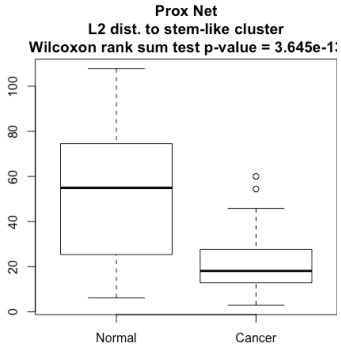


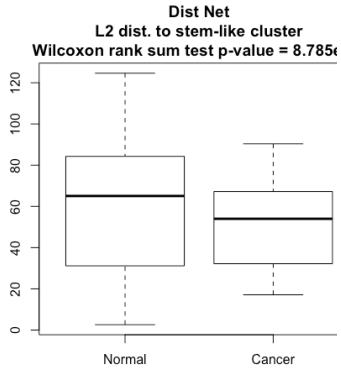
Figure 5. Proximal regulatory network, distal enhancer network, and gene expression profiles have been used to explore patterns across different cell types. As expected, stem-like cell types formed a cluster, suggesting stem-like cell types have a distinct regulatory profile from normal and cancerous 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:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

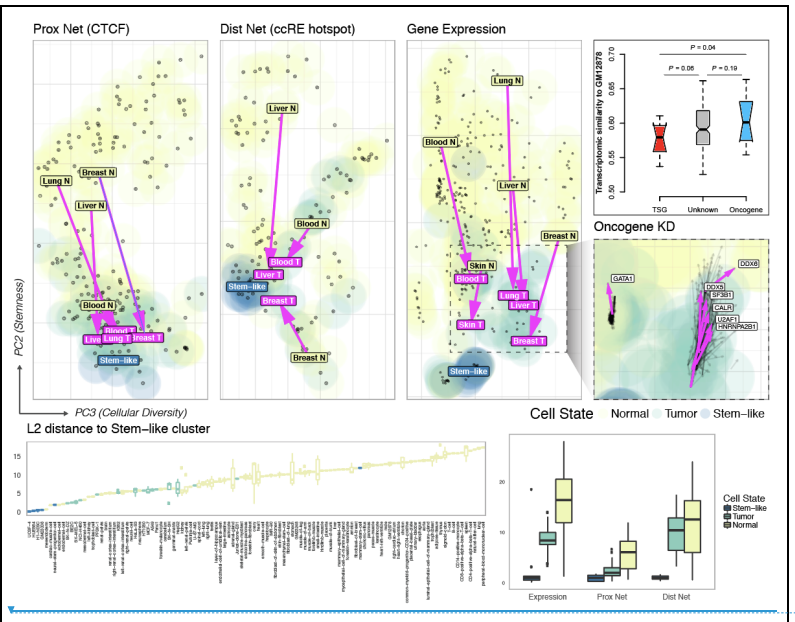


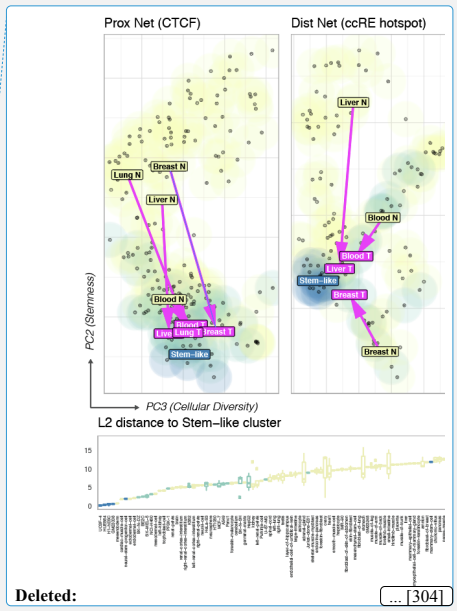
Deleted:

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Deleted: cance





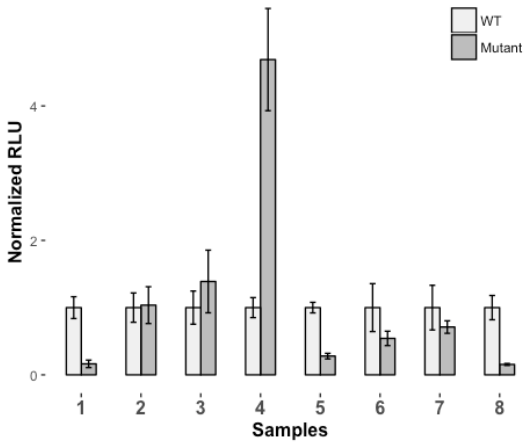
Deleted: ... [304]

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

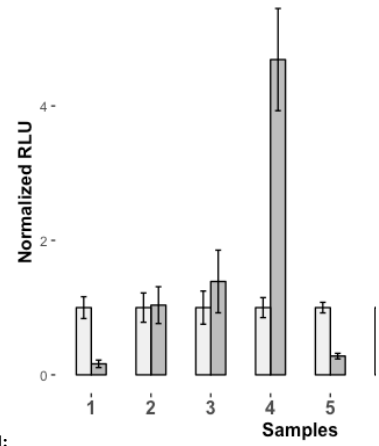
<TYPE>$$$Validation,$$$Text
<ASSIGN>@@@JZ,@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%85DONE

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 (see excerpt 1 below).

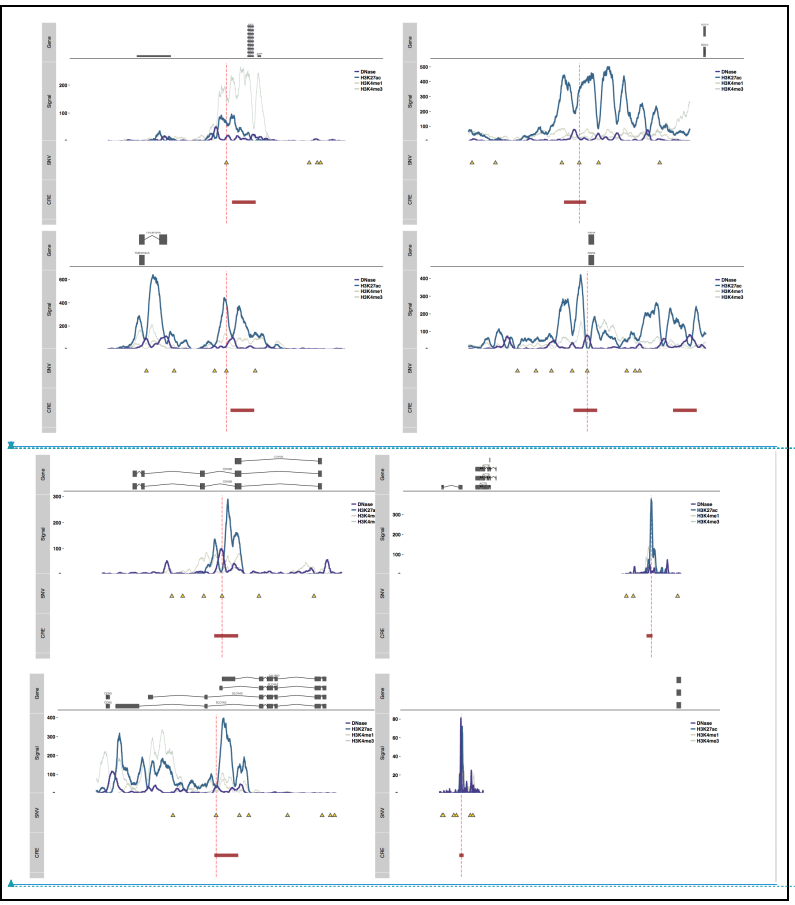
Formatted Table

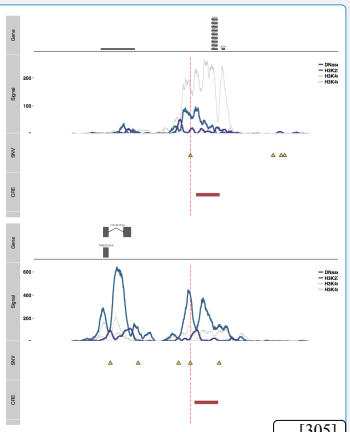
	JZ2MG: previously we mentioned that we selection these variants based on motif breaking but I feel that is not good. Could we say we do the prioritization based on procedures in figure 6? Is this dangerous? There are two individuals independently performed the experiment and each individual did three 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 the nearby genes (see excerpt 1 below).
Excerpt 1 From Revised Manuscript	We selected top ten regions with the highest motif breaking power 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 2 From Revised Manuscript	Details for all tested regions.

Comment [29]: supplement



Deleted:



Deleted: ... [305]

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman

<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

Formatted Table

	supplement, and in turn present data-rich figures from there with the main manuscript.
Author Response	We thank for the referee for this comments. We have tried to revise the figures as requested We have fixed figure XX & YY.
Excerpt From Revised Manuscript	

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

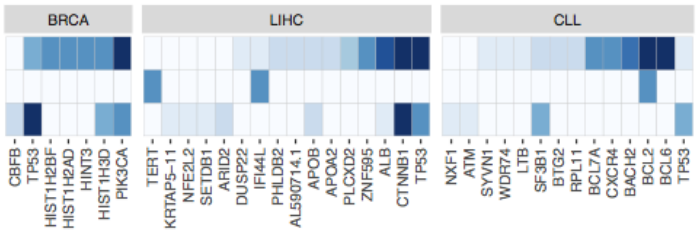
<TYPE>$$$Validation,$$$Text
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

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 incurred data.

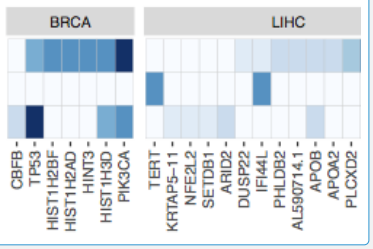
Formatted Table

<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).
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 From Revised Manuscript (in Main figure)	We have plotted the heatmap of p values for the recurrent analysis in three different cancer types. 

Formatted Table

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").

Formatted Table

Author Response	We thank the referee to point this out. References have been added in the new submission.

<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 (Figure 1 and supplementary table xxx).
Excerpt From Revised Manuscript	We highlighted the normal tissue data from the Roadmap (processed by ENCODE3) in our revised figure 1 as below.

Formatted Table

<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

Formatted Table

	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 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. As shown in excerpt 1, all stem cells tend to close to each other. Hence, we feel that H1 is a reasonable representative of stem cells. We also added a few sentence in the revised discussion section.
Excerpt 1 From Revised Manuscript	(Please refer REF5.19 for figure update.)

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

Formatted Table

	The reason we selected the candidate 5 instead of candidate 4 is that the candidate 5 had stronger motif breaking score when disrupted, had higher density of TF binding events, and aligned better with our integrative regulatory region calls. However, we feel that all other regions we tested are among the top prioritized regions 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 From Revised Manuscript	Please see figures in Excerpt 2 in response "to <ID>REF5.22 – Selection of regions for validation testing"

<ID>REF5.28 – Minor: SYCP2 and beyond

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

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 From Revised Manuscript	

Formatted Table

<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.
Excerpt From Revised Manuscript	

Formatted Table

<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.

Formatted Table

Excerpt From Revised Manuscript	

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

Formatted Table

<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.

Formatted Table

Author Response	We thank reviewer for the comment. We fixed the labeling of the network tiers in the revised manuscript.
Excerpt From Revised Manuscript	

<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 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 From Revised Manuscript	

Formatted Table

<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.
Excerpt From Revised Manuscript	

Formatted Table

Deleted: on

Deleted: made numerous improvements to the

Page 1: [1] Deleted	Author	5/4/18 9:05:00 PM

Keep it more compact, and mentioned as what we mentioned in our email

Page 3: [2] Deleted	Author	5/4/18 9:05:00 PM

Dear Orly. We're enclosing our revised version of the end code C, our manuscript. As you can see we have attempted to completely and definitively address all of the referee's concerns. In the attached sheets which have a point by point response. We've corresponded a bit about this manuscript before so I'll be brief here and simply say that we consider this an integral part of the end code package and the main manage group to do intuitive cross assay annotation and provide a network perspective on the annotation. We think cancer is the best application for this. But this, as we've said before is not a cancer genomics paper.

In the revision some of the highlights are we've done. We have new validation experiment to explain the effect of SV's on the extended gene. We also have a second validation experiment on some of our networks and we have additional highlights and data on the way the knockdowns are relate to the normal to cancer to stem transition and also how structure variance relate to functional genomics data.

We hope you like the manuscript and we look forward to hearing from you.

Yours sincerely, Mark

Page 4: [3] Deleted	Author	5/4/18 9:05:00 PM

analysis
- Network rewirings from various assays

Page 4: [4] Deleted	Author	5/4/18 9:05:00 PM

One area that we wish to clarify a little [1]on is to ask us to compare our calculations to that for driver identification. We think that the value of our paper was misunderstood by some of the reviewers. The point of this paper is not to develop a novel method of driver discovery or to find new cancer drivers, but to highlight the use of ENCODE3 data in cancer genomics, particularly related to understanding the overall patterns of mutations, network rewiring, and regulator and variant prioritization. To respond to previous comments, we have shown how in certain contexts, the ENCODE3 date can help with existing driver discovery measures.

Page 4: [5] Deleted	Author	5/4/18 9:05:00 PM

For example, usually a tumor sample contains a variety of cell types because the

Page 5: [6] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [7] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [8] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [9] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [10] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [11] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [12] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [13] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [14] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 5: [15] Formatted Table	Author	5/4/18 9:05:00 PM

Formatted Table

Page 5: [16] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [17] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [17] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [17] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [18] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [19] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [20] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [20] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [20] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [20] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [21] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [21] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 5: [22] Deleted **Author** **5/4/18 9:05:00 PM**

), two and half months after we submitted our paper in Aug 2017, so it is impossible us to cite in the initial submission.

Page 5: [23] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [24] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [25] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [25] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [26] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [27] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [28] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [28] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [29] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [29] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [30] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [31] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [32] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 5: [33] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 6: [34] Deleted **Author** **5/4/18 9:05:00 PM**

Have more validations than other paper, tons of unique validations
We are

We have more More validation target
Targeted validation

New methods - STARR-seq method

Unappreciated new methods disctinct from the encyl
1) Integative Annotation fusing different types of data
((: New integrative data-fusion methods computational method for fusing multi HM enhancer predictions w STARRseq & also activity correlations w/ Hic
))
2) New methods for analyzing network change & gene commun
Incl. metrics for positiining cancer cell relative norma & stem cells

Network clustering & Gene communities
Measuring rewriting
Prioritizing variantis

3) New methods for prioritzing regulatorsy (TFS or rbps) based on aggreg burdening & rewrigingh

We want to make it explicit that
(1) this paper is to

Page 7: [35] Deleted	Author	5/4/18 9:05:00 PM

Note also that while we do NOT feel ENCODEC is a cancer genomics paper, we feel that cancer is the best application to illustrate certain key aspects of ENCODE data and analysis - particularly deep annotations and network changes. We have listed some more details about novelty of this paper as below.

(1) Integrative annotation from various types of assays. ENCODEC is unique in its highlighting of a number of ENCODE assays (e.g. replication timing, TF knockdowns, STARR-seq and Hi-C), its deep, integrative annotations combining a wide variety of assays in specific cell types
(1) Networks. These are a core aspect of ENCODE, featured in the '12 roll out. None of the other papers highlight networks in the current package. In ENCODEC, in addition to looking at "universal" ChIP-Seq networks, merged across cell types, we also look at network changes ("rewiring") for specific cell-type comparisons in both proximal and distal networks. We feel that this is best exemplified in oncogenesis.

(2) Deep, integrative annotation – complementary to the Encyclopedia. While the encyclopedia paper considers broad, "universal" annotations across cell-types (currently the centerpiece of ENCODE), it focuses on data common to most cell types (DHS, 2 histone marks and 2 TFs). It does not take advantage of the cell types richer in assays -- the other dimension of ENCODE (diagrammed in ENCODEC's first figure). The ENCODEC paper takes a complementary approach, constructing a more accurate annotation using a large battery of

histone marks (>10), next generation assays such as STARR-seq and elements linked by ChIA-PET and Hi-C.

(3) Replication Timing. Although a major feature of ENCODE is replication timing, none of the other papers feature it. Previous work on mutation burden calculation usually selects replication timing data from the HeLa cell line due to the limited data availability. The wealth of the ENCODE replication timing data greatly helps to parametrize somatic mutation rates.

(4) SVs. One unappreciated aspect of ENCODE is that next-generation assays, in addition to characterizing functional elements in the genome, enable one to determine structural variations.
(5) Knockdowns. ENCODE has 222 TF knockout/knockdown experiments, which are not explored systematically in other papers.

Page 9: [36] Deleted	Author	5/4/18 9:05:00 PM

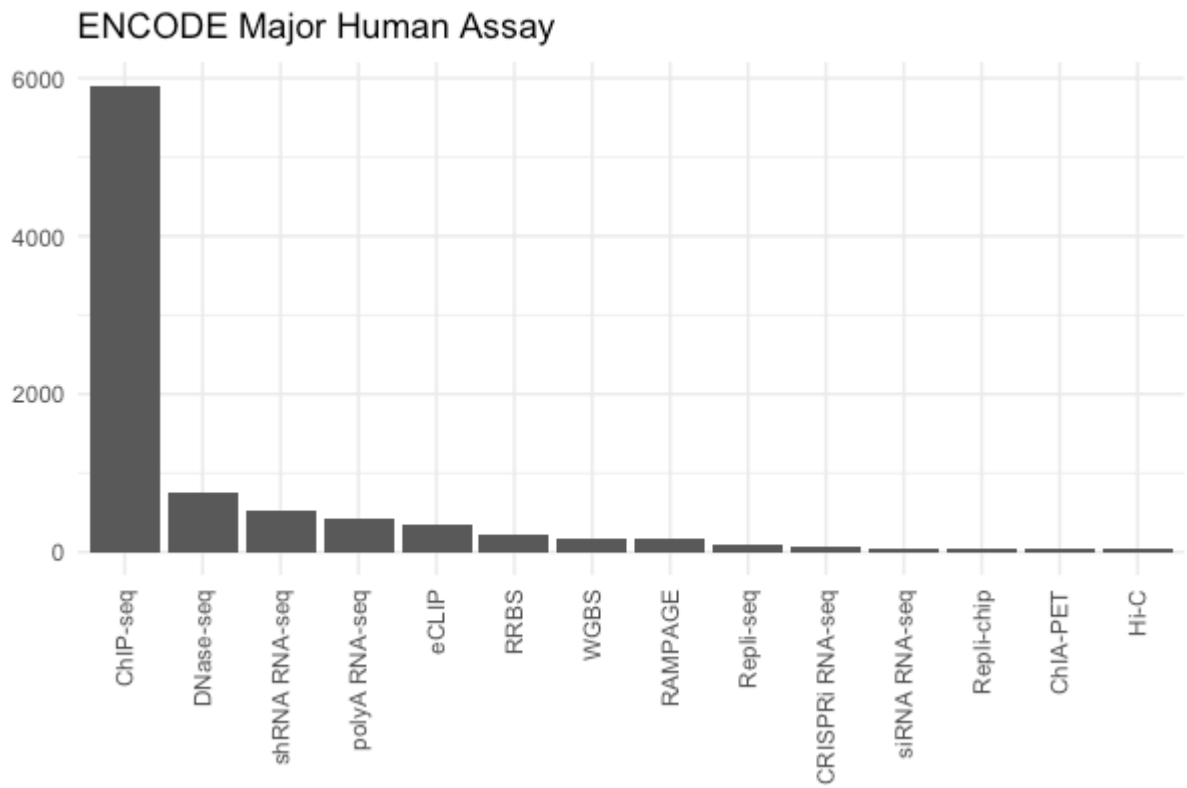
Our key point is that the new encode really dramatically expands the amount of data useful for cancer genomics the number of ways in particular it expands the amount of data for a mutation rate are prediction by more than a factor of 10 allowing for more than 2008 assess as opposed to 159 before and heard 3 second of all it dramatically expands the scale of data available for Network comparison in the ENCODE to roll out about if one was looking for instance a network change one could look at the maximum of maybe 30 Ts that changed between say G & K that number has more than doubled or tripled or something like that and now it also we have a tremendous amount of histone mark data on a number of key cell lines allowing us to do accurate enhancer prediction using multi histone mark data and STARR-Seqon XXX cells.

Page 9: [37] Deleted	Author	5/4/18 9:05:00 PM

We want to make it clear and emphasize that the goal of this paper is to build a new annotation "*resource*", not to discover novel biology in cancer. We feel that cancer is the best application to illustrate certain key aspects of ENCODE data and analysis - particularly the deep annotations and network changes. Thus, where the referee asks for novelty in cancer gene discovery - we strongly feel that this is out of scope. We have listed some more details about the resource of this paper as below (Table R1 and Figure. R1).
(put these figures into the supplementary files, the new ENCODE has a lot of cancer cell line data, more than previous.) scale of data to build thes models changes to a factor of 10. The scale of regulatory network has go up a x factor

It matters what's relevant to us

Figure R1. Summary of the raw signal tracks released in this paper



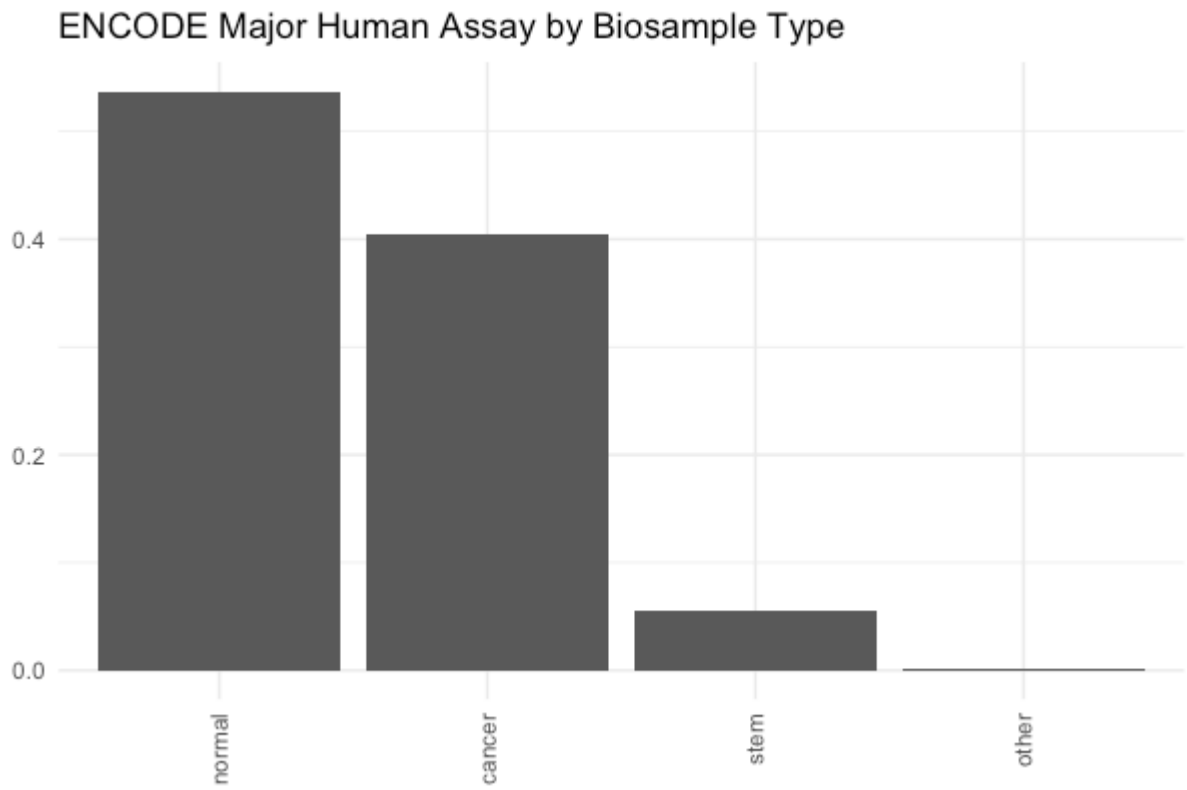


Table R1. Summary of annotation types and example applications in our paper

The big thing is the integrative annotation, includes histone marks and starr-seq, and the linkcatges. Fair difference between us and the main encyclopedia paper.

Level	Annotation type	Example Applications
Element	- TF/RBP binding peaks & motifs - DHS peaks - Replication timing profiling - Enhancers level 1-3 - Hi-C TADs and ChIA-pet loops - SV and SNV in cell lines	- BMR estimation (Fig. 2) - Genome annotation (Fig. 6) - Variant prioritizations (Fig. 6)
Gene	- Extended genes definitions - RNA-Seq expressions (dangerous) - Expression changes after knockdowns	- Somatic & germline burdens (Fig.1) - Stemness analysis (Fig. 5) - Variant prioritizations (Fig.3 & Fig. 6)
Network	*Distal network:* - Enhancer-gene (computational) - Enhancer-gene (computational + Hi-C) - TF-Enhancer-gene *Proximal network:* Experimental based: - TF/RBP Universal networks (strong & weak) - TF/RBP tissue specific networks (binary & probabilistic) Imputed: - DHS imputed tissue specific TF networks	- TF/RBP Regulatory Activities (Fig.3) - Network rewiring (Fig. 4) - Network Hierarchies (Fig. 5) - TF binding disruptiveness (Fig. 5) [2]

Page 10: [38] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [39] Formatted **Author** **5/4/18 9:05:00 PM**

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

Page 10: [40] Deleted **Author** **5/4/18 9:05:00 PM**

annotation
Integrative

Page 10: [41] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [42] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [43] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [43] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [44] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [45] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [46] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [47] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [48] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [49] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [50] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [51] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [52] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [53] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [54] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [55] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [55] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [56] Deleted	Author	5/4/18 9:05:00 PM

(more details in the following table). We updated our reference as suggested but we do feel it is a bit unfair to make a direct comparison for papers with such different focuses.

Page 10: [57] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 10: [58] Deleted **Author** **5/4/18 9:05:00 PM**

Second, we want

Page 10: [59] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [60] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [60] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [61] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [61] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [61] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [62] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [63] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [64] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [65] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [66] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [66] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [66] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [67] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [68] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [68] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [69] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [69] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [70] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 10: [71] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 13: [72] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 13: [72] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 13: [73] Deleted **Author** **5/4/18 9:05:00 PM**

. We have made it more apparent in our revised manuscript that our purpose is to showcase

Page 13: [74] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 13: [74] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 13: [75] Deleted **Author** **5/4/18 9:05:00 PM**

.

With the wealth data available through ENCODE data, we had a much larger pool of features to choose from to potentially improve BMR estimation. There are 2017 of histones modification marks that are released into a ready to use format (see details in the table below), and 818 of which are from real tissues..

Also, we have provided other data types, such as replication timing, that has been proven to affect BMR but has not been widely by others. We believe that such data, when released into a ready to format, can help BMR estimation through many existing models.

Page 13: [76] Formatted **Author** **5/4/18 9:05:00 PM**

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

Page 13: [77] Deleted **Author** **5/4/18 9:05:00 PM**

“more” features performs better in BMR prediction is not a novel discovery. We believe that

Page 13: [78] Deleted **Author** **5/4/18 9:05:00 PM**

were *misunderstood* at this point because this conclusion is served as an illustration of the value of the new annotation “resource” using the richness of ENCODE data. Here, we are not trying to reproduce the claims on how epigenomic features affect BMR but rather to show how the richness of ENCODE data can make improved BMR estimations.

We made following changes in the main text to

Page 13: [79] Deleted	Author	5/4/18 9:05:00 PM

this.

Key idea is data up by xxx%, we are doing this large scale regression is the key. NG regression is more stable. The # of dataset if 10 times factors, and it make a differnce to use this scale of data

Makes it difference to have one full order of magnitude more data

MARK'S DICTATION: BMR Insert.

We wish to be

Page 13: [80] Deleted	Author	5/4/18 9:05:00 PM

or its application to cancer genomics. We point out that a number of references have used this. A negative binomial regression is a very standard statistical technique that has been used in many contexts in genomics.

Our main point that we wish to make clear here is that the N code

Page 13: [81] Formatted	Author	5/4/18 9:05:00 PM

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 13: [81] Formatted	Author	5/4/18 9:05:00 PM

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 13: [82] Deleted	Author	5/4/18 9:05:00 PM

XXX to YYY. Furthermore, we show in our figure that this expansion is quite significant. One does not get most of the modeling of background mutation rate by including one or two features, but actually, including up to 20 or 30, or even more, does continue to incrementally give further improvement, and this is either using the features directly or principal components.

Page 13: [83] Deleted	Author	5/4/18 9:05:00 PM

The implication here is that more data is actually

Page 15: [84] Deleted	Author	5/4/18 9:05:00 PM

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.

Page 15: [85] Deleted	Author	5/4/18 9:05:00 PM

We thank the referee for pointing out the comparison of cell line vs. tissues and we feel this is a good suggestion. In our revised manuscript, we further investigated it in detail by extending our analysis to many new data types, such as RNA-seq and distal/proximal TF ChIP-Seq data. We think slightly differently with the referee on value of ENCODE data. Several points we want to emphasize are

<u>- On a large scale (up to mbp)</u>

First, the Polak 2015 paper did not perform large-scale comparison across various cancer cell lines. As seen from Except 1 below, cell line data provides comparable, sometimes even better, correlation with mutation counts. We have added a new section in the supplementary file to discuss this. We feel that due to the heterogeneous nature of cancer data, it does not hurt to computationally search the best features that explains the mutational landscape in a tumor specific way.

As compared to cell line data, there are way less functional characterization data in tissues. For example, there are no prostate tissue data from the REMC. We have updated supplementary table 1 for a comparison of data richness in ENCODE3.

Page 16: [86] Deleted	**Author**	**5/4/18 9:05:00 PM**

"`ENCODE cell lines can be problematic`", we want to highlight that ENCODE is not just about cell lines. There are many ENCODE tissue data for histones (339 cell line vs **818** tissue, details see excerpt 2 below). We have added a supplementary table on this point. Again, for the BMR part, we select the best possible features for prediction (no matter it is from cell line or tissue), instead of manually find a matching.

Our purpose in the BMR section is not to find the best matching cell type, but to better use the ENCODE data to improve estimation accuracy. The bulk tumor samples from a patient usually contains diverse collection of cells harbouring distinct molecular signatures. As we have shown in Excerpt 3 below, the addition of more features usually can introduce noticeable accuracy improvement. T Actually some of the recent papers, such Martincorena et al. (2017), also used the top 20 PCs of 169 histone features in their model. On this point, we uniformly processed thousands of features in a ready-to-use format. Many of them are not mentioned in other literature, such as replication time from 51 tissue/cell lines. [3][4]They have proven useful but are less frequently matched probably due to the lack of data incorporated into previous BMR models. We believe that this is quite useful for cancer genomics.

Just to say it better and use the dictation text instead, highlight more the 169 vs. 2067

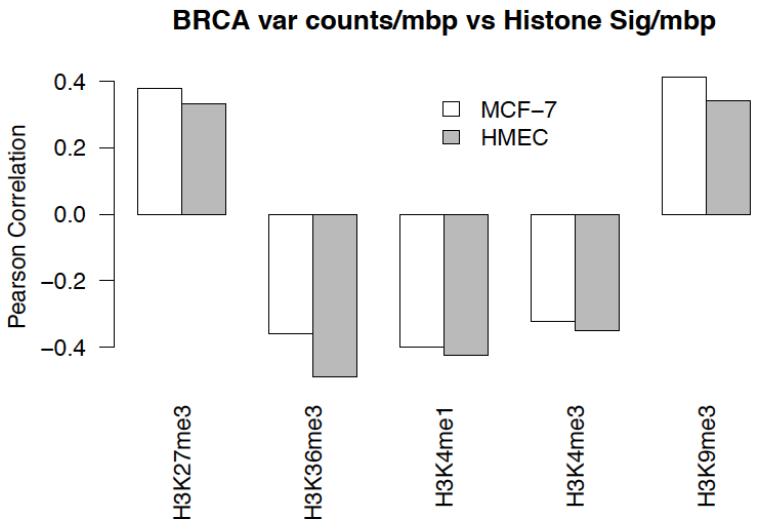
<u>- On a small scale cancer cell lines might be a better source to use for cancer data</u>

Features, like expression levels and TF binding events, have been used widely to affect somatic mutation rates. As suggested by the referee, 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. We have added this result into the main figure and supplementary files.

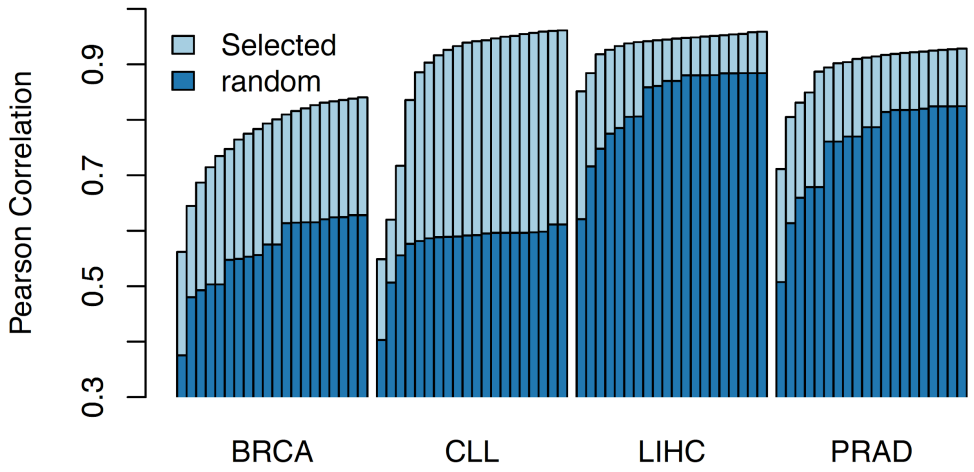
Page 16: [87] Deleted	Author	5/4/18 9:05:00 PM

1. Comparison of mutation rate vs features in tissue/cell lines. We provided 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 even better) correlation with mutation counts.

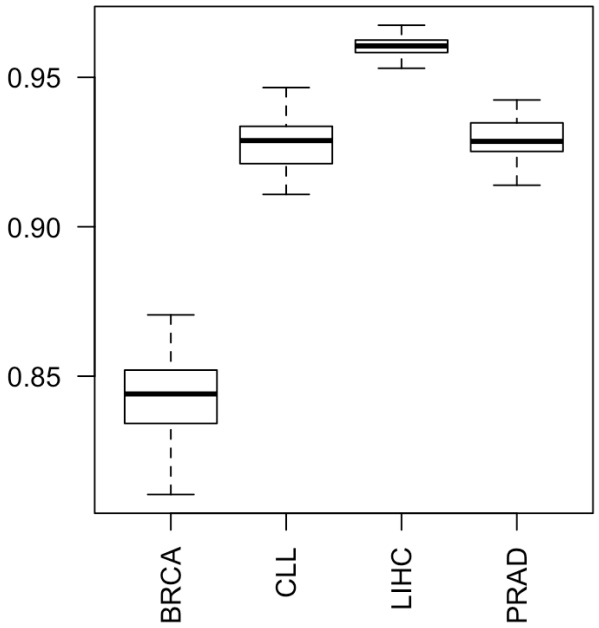


Page 16: [88] Deleted	Author	5/4/18 9:05:00 PM

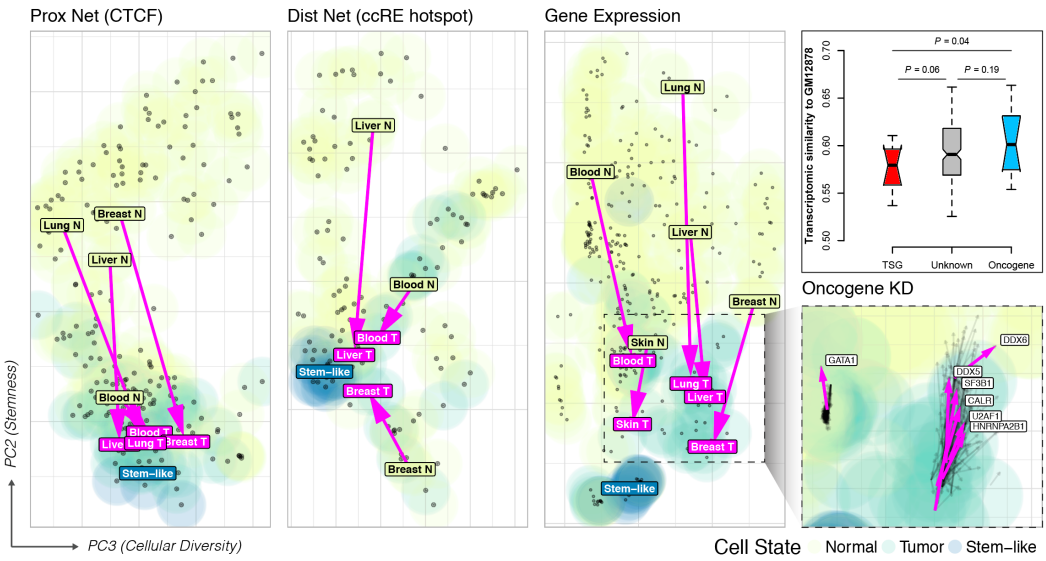
Excerpt 2 From Revised Supplementary file	2. Summary of ENCODE histone ChIP-seq data **Cell Type** — **# histone marks** tissue — 818 primary-cell — 521 cell-line — 339 in-vitro-differentiated-cells — 179 stem-cell — 114 induced-pluripotent-stem-cell-line — 46
Excerpt 3 From Revised Supplementary file	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.



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



Excerpt 4 From Revised Supplementary file	3. 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.



Page 16: [89] Deleted **Author** **5/4/18 9:05:00 PM**

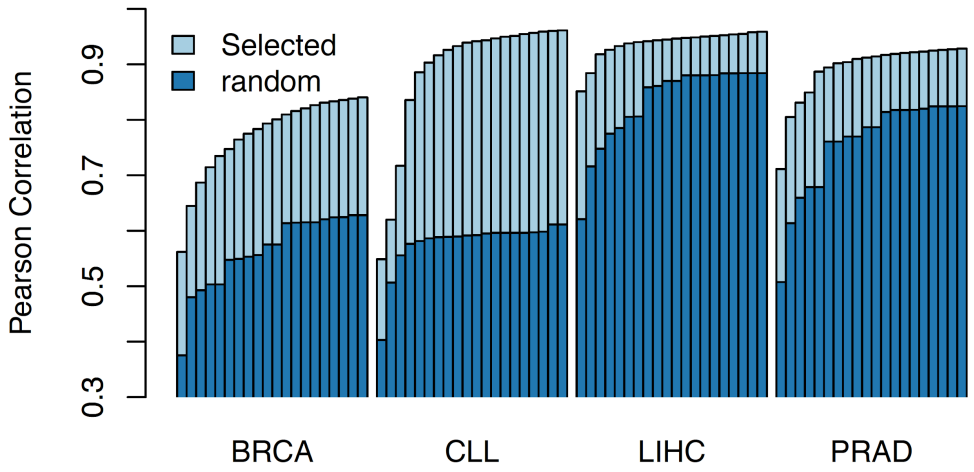
Excerpt 2 From Revised Supplementary file

2. Summary of ENCODE histone ChIP-seq data

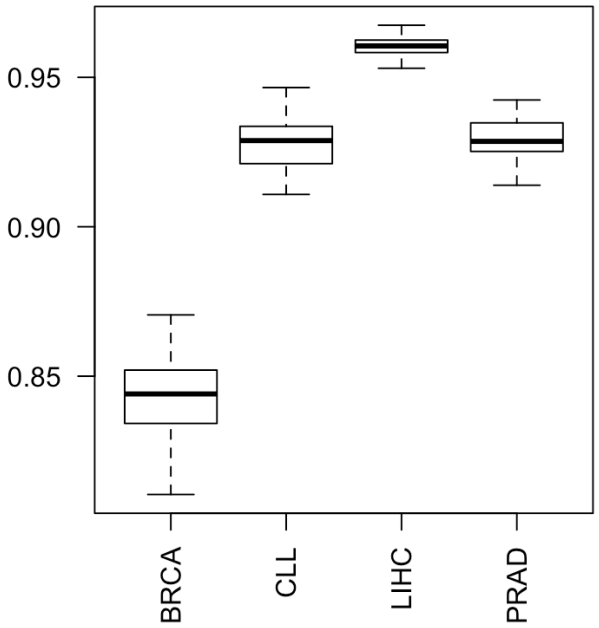
Cell Type	**# histone marks**
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

Excerpt 3 From Revised Supplementary file

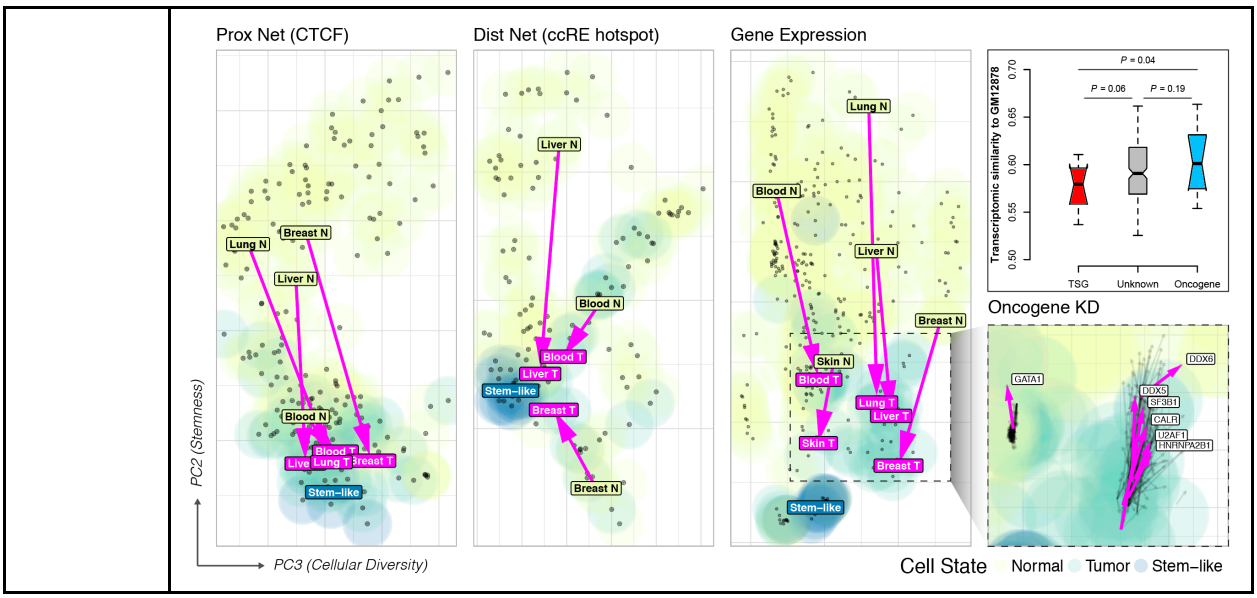
At 1mb bin resolution, we compared the performance of models using random features vs. computationally selecting best features sequential (forward selection). It has shown that by adding features appropriately from ENCODE3, we can noticeably improve the performance of BMR accuracy.



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



Excerpt 4 From Revised Supplementary file	3. 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.



Page 21: [90] Deleted **Author** **5/4/18 9:05:00 PM**

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 incurred data.

Also the network help tf & rbp prioritization

It is not just variant prioritization, but also regulators, that is not at all in any of the other papers. Figure 3 & 4

Page 22: [91] Deleted **Author** **5/4/18 9:05:00 PM**

Level	Annotation type	Example Applications
Element	- TF/RBP binding peaks & motifs - DHS peaks - Replication timing profiling - Enhancers level 1-3 - Hi-C TADs and ChIA-pet loops - SV and SNV in cell lines	- BMR estimation (Fig. 2) - Genome annotation (Fig. 6) - Variant prioritizations (Fig. 6)
Gene	- Extended genes definitions - RNA-Seq expressions - Expression changes after knockdowns	- Somatic & germline burdens (Fig.1) - Stemness analysis (Fig. 5) - Variant prioritizations (Fig.3 & Fig. 6)

Network	*Distal network:* - Enhancer-gene (computational) - Enhancer-gene (computational + Hi-C) - TF-Enhancer-gene *Proximal network:* Experimental based: - TF/RBP Universal networks (strong & weak) - TF/RBP tissue specific networks (binary & probabilistic) Imputed: - DHS imputed tissue specific TF networks	- TF/RBP Regulatory Activities (Fig.3) - Network rewiring (Fig. 4) - Network Hierarchies (Fig. 5) - TF binding disruptiveness (Fig. 5)

Page 23: [92] Deleted **Author** **5/4/18 9:05:00 PM**

We thank referee for the suggestion. The referee is pointing out that negative binomial regression has been used before. We also feel that the fact that other papers also used negative binomial regression bolsters the underlying technical validity of our argument. While we admit it does slightly undercut a claim of novelty in this regard, that is not central to our work. (reference all these papers)

Page 23: [93] Deleted **Author** **5/4/18 9:05:00 PM**

comparison to our 2067 features.

On our side, we think negative binomial regression is a standard statistical technique that has been used in many contexts. Also, ENCODE3 provides noticeably more covariate data, which is uniformly processed and less explored in the references mentioned by the referees. Some features, such as replication timing is well-known confounders but was not included in the

Page 23: [94] Deleted **Author** **5/4/18 9:05:00 PM**

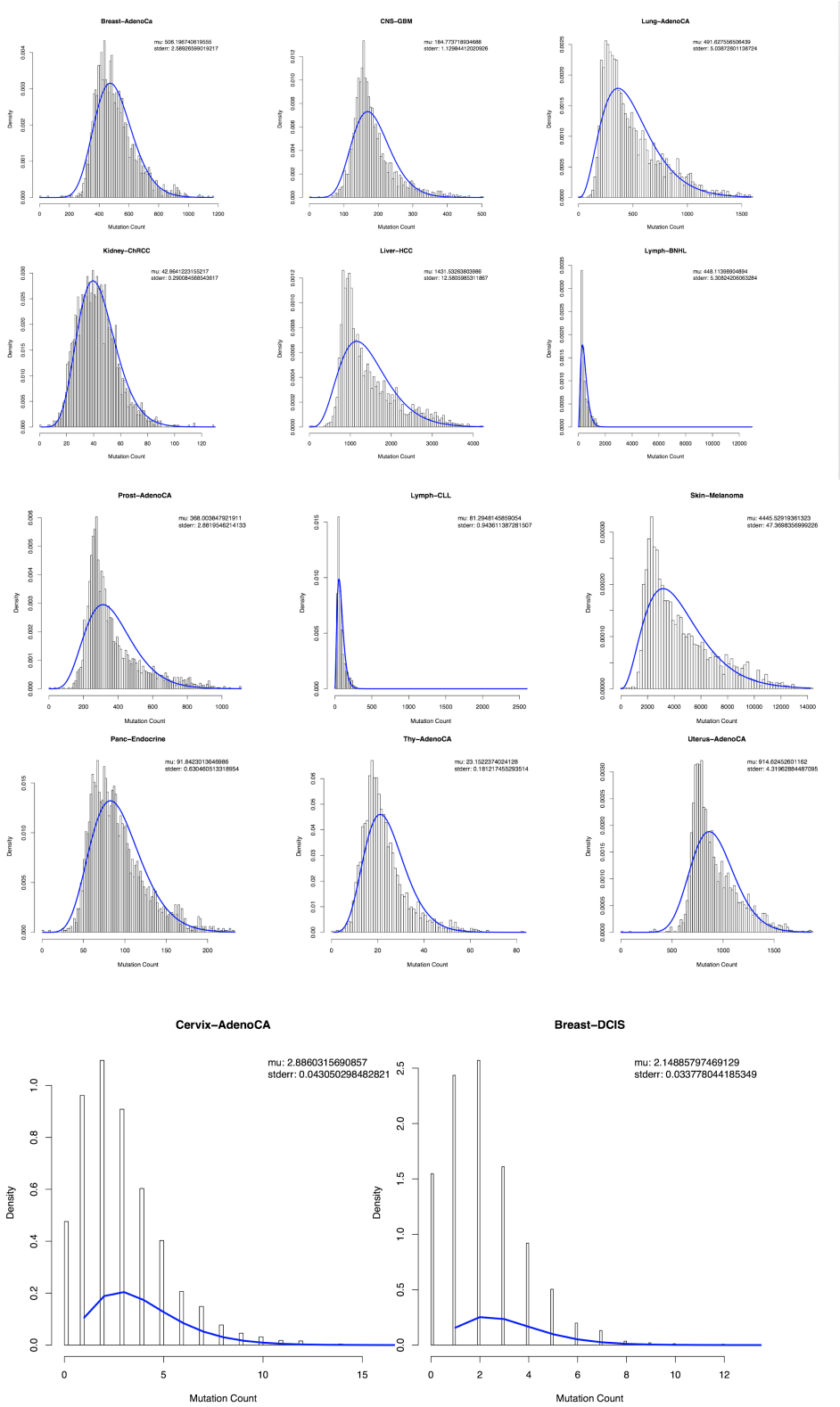
. paper. We are not aiming to make a new method for predicting background mutation rate, but rather to use a robust regression method that really takes into account

Page 23: [95] Deleted **Author** **5/4/18 9:05:00 PM**

very large amount of data and is able to leverage that to more successfully predict background mutation. Therefore, we did not directly use their approach.

We have been misunderstood. Does not make a big deal of their paper

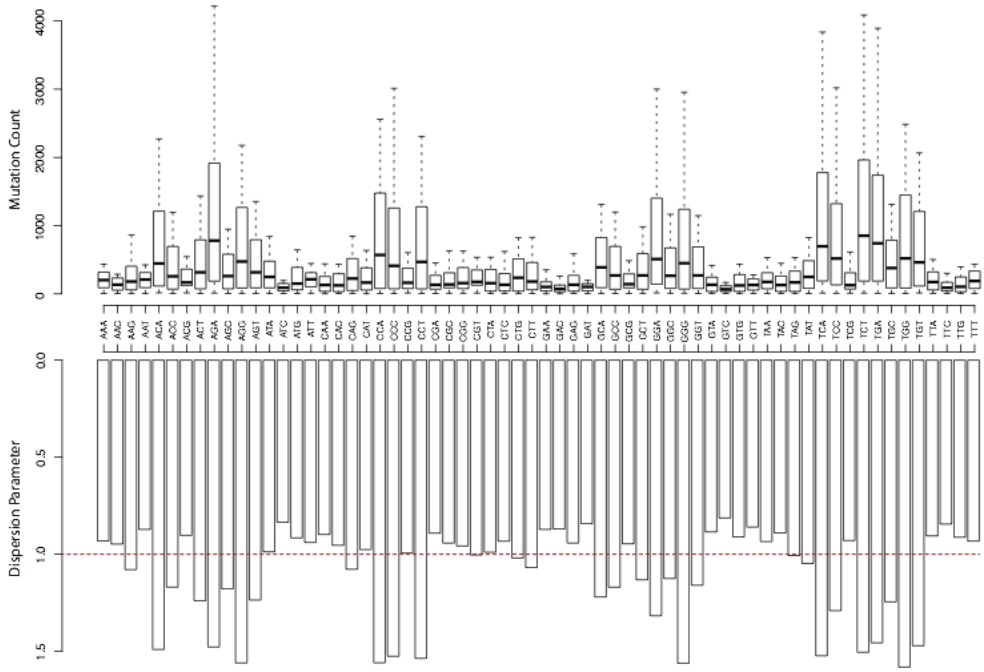
Page 25: [96] Deleted **Author** **5/4/18 9:05:00 PM**



Page 27: [97] Deleted	Author	5/4/18 9:05:00 PM

We also want to point out that the overdispersion problem on count data is also confounded by omitting related covariates. That is the main reason why we want to introduce more feature

candidates from ENCODE and at the same time avoid overfitting. Many other methods (such as Marticorena, 2017) directly use Negative Binomial regression without checking whether it is necessary. It is simpler to not introduce additional parameters. However, we think it is better to check how heterogeneous the count data is even after correcting enough covariate effects.



Page 28: [98] Deleted **Author** **5/4/18 9:05:00 PM**

Page 28: [98] Deleted **Author** **5/4/18 9:05:00 PM**

Page 28: [98] Deleted **Author** **5/4/18 9:05:00 PM**

Page 28: [98] Deleted **Author** **5/4/18 9:05:00 PM**

Page 28: [98] Deleted **Author** **5/4/18 9:05:00 PM**

Page 28: [98] Deleted **Author** **5/4/18 9:05:00 PM**

Page 28: [99] Deleted **Author** **5/4/18 9:05:00 PM**

, but

Page 28: [99] Deleted **Author** **5/4/18 9:05:00 PM**

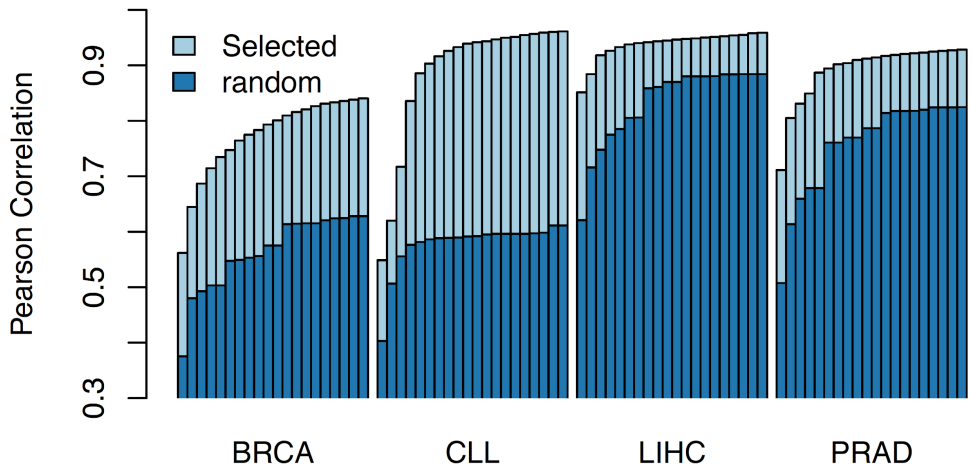
, but

Page 28: [100] Deleted	Author	5/4/18 9:05:00 PM

this gets the point across. The aim here is not to highlight a complicated mathematical method but just simply to get across the idea that the extensive ENCODE data provides a valuable resource for predicting BMR and we appreciated the referee helping us achieve clarity on this point. We put the main text figures into the supplementary files and made for the main.

PCA : Moving it to the supplement

Page 28: [101] Deleted	Author	5/4/18 9:05:00 PM



2.

Page 28: [102] Formatted	Unknown

Font:(Default) Times New Roman, (Asian) Times New Roman

Page 28: [102] Formatted	Unknown

Font:(Default) Times New Roman, (Asian) Times New Roman

Page 30: [103] Deleted	Author	5/4/18 9:05:00 PM

Following the reviewer's suggestions, in our revised manuscript we show in a formal power analysis that the most important contribution to power comes from including additional functional sites, which is of course by the extended gene concept. secondarily, from removing non-functional sites, but to a lesser extent. The assumption in our compacting annotations is that we can accurately distinguish the more important functional nucleotides from the less important ones through the guidance of many functional characterization assays.

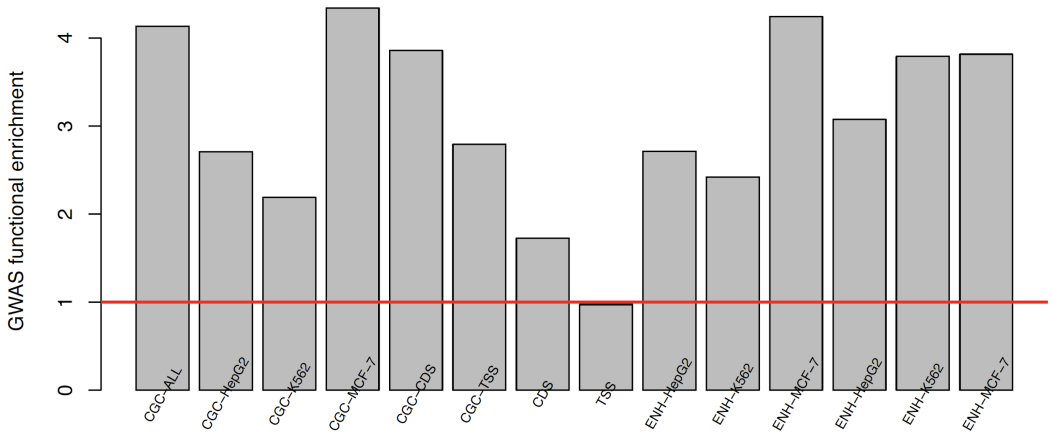
Agree but not too weak, add the math

Admittedly, we are making assumptions and the referee is completely correct in pointing this out. We have tried to be more precise in the text that we are assuming that the large number of ENCODE assays, when integrated, allow us to more directly get the functional nucleotides, but this, of course, is an assumption. It is hard to tell to what degree one can succeed in finding the current events in cancer. It is hard to back this up with the gold standard, but we think that some

of the points are self evidently obvious. We have tried to make this clear in text and thank the referee for pointin[5]g this out.

Page 34: [104] Deleted	Author	5/4/18 9:05:00 PM

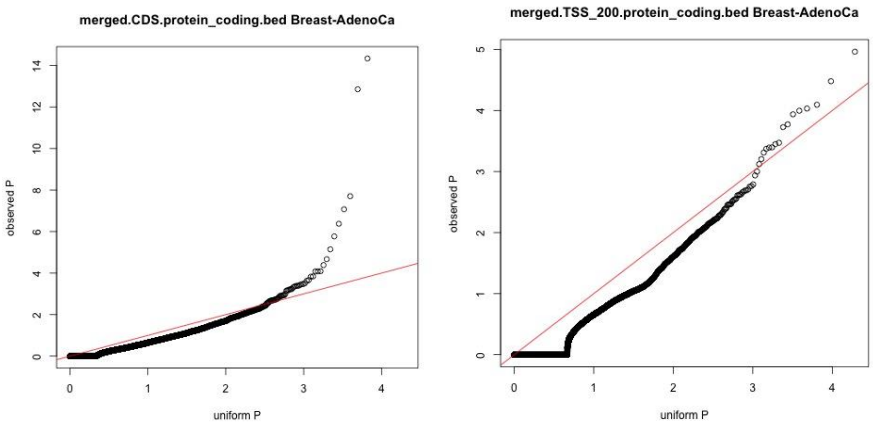
Expression stratification example?[6] we apply a method in biorxiv, we extended their method..



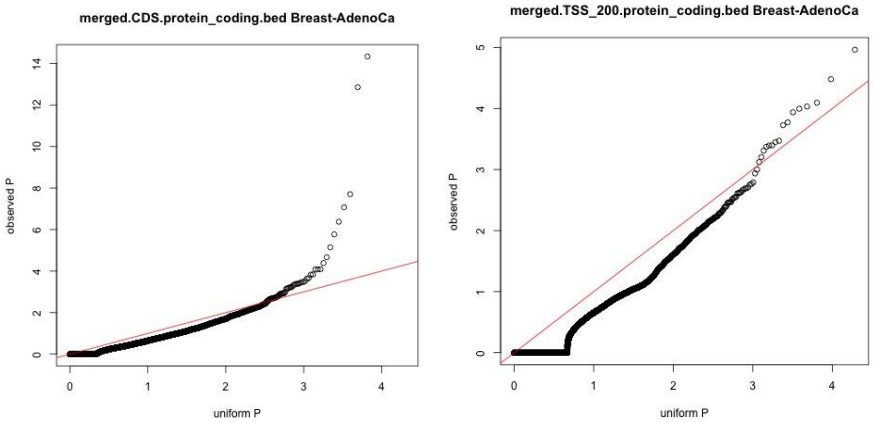
Page 34: [105] Moved to page 37 (Move #9)	Author	5/4/18 9:05:00 PM

Referee Comment	5) 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.

Page 34: [106] Deleted	Author	5/4/18 9:05:00 PM

Author Response	We thank the referees for this comment. We have updated the QQ-plots in our revised manuscript and they look fine.
Excerpt From Revised Manuscript (in supplement)	

Page 34: [107] Deleted	Author	5/4/18 9:05:00 PM

Author Response	We thank the referees for this comment. We have updated the QQ-plots in our revised manuscript and they look fine.
Excerpt From Revised Manuscript (in supplement)	

Page 35: [108] Deleted	Author	5/4/18 9:05:00 PM

extended genes, such as

1. We extensively expanded our power analysis part to include more

Page 35: [109] Deleted	Author	5/4/18 9:05:00 PM

2. We showed that by using the extended gene, we can better stratify the gene expressions and regulations
3. We explored the cancer related GWAS SNPs and showed that extended genes in matched cell types showed noticeable improvement. (See details in Excerpt 2 to REF 2.6 above)

One point we want to make clear is that the application of the extended gene is more than driver discovery hence the revisions have tried to highlight other areas, such as GWAS, gene expression and/or regulations stratification mentioned above, where the extended gene is useful in cancer.

Page 40: [110] Deleted	Author	5/4/18 9:05:00 PM

.

This is of

Page 42: [111] Deleted	Author	5/4/18 9:05:00 PM

Excerpt From Revised Manuscript	

Page 42: [112] Deleted	Author	5/4/18 9:05:00 PM

Excerpt From Revised Manuscript	

Page 42: [113] Deleted **Author** **5/4/18 9:05:00 PM**

We have changed figure
Agree and fix....

Page 44: [114] Deleted **Author** **5/4/18 9:05:00 PM**

be descbie
See

Page 48: [115] Deleted **Author** **5/4/18 9:05:00 PM**

Page 48: [116] Deleted **Author** **5/4/18 9:05:00 PM**

***Are there any [7]novel oncogenes detected by the method?

Page 48: [117] Deleted **Author** **5/4/18 9:05:00 PM**

Think about how we should responsed
Break this out

Page 54: [118] Deleted **Author** **5/4/18 9:05:00 PM**

Author
Response

Page 55: [119] Moved to page 48 (Move #12) **Author** **5/4/18 9:05:00 PM**

Excerpt From Revised Manuscript	

Page 58: [120] Deleted **Author** **5/4/18 9:05:00 PM**

Excerpt From Revised Manuscript	*[JZ2MG: is there an excerpt here?]*

Page 58: [121] Deleted **Author** **5/4/18 9:05:00 PM**

Excerpt From Revised Manuscript	*[JZ2MG: is there an excerpt here?]*

Page 59: [122] Deleted **Author** **5/4/18 9:05:00 PM**

[JZ2MG: ongoing]

Page 59: [123] Moved to page 20 (Move #5) **Author** **5/4/18 9:05:00 PM**

Ask

Page 59: [124] Deleted **Author** **5/4/18 9:05:00 PM**

Author Response	Mention that theres a lot of tissue in ENCODE We take the referee's comment to heart and we agree with the reviewer agree that it is important to verify the discoveries from cell lines from in primary cancers. There are lots of tissue

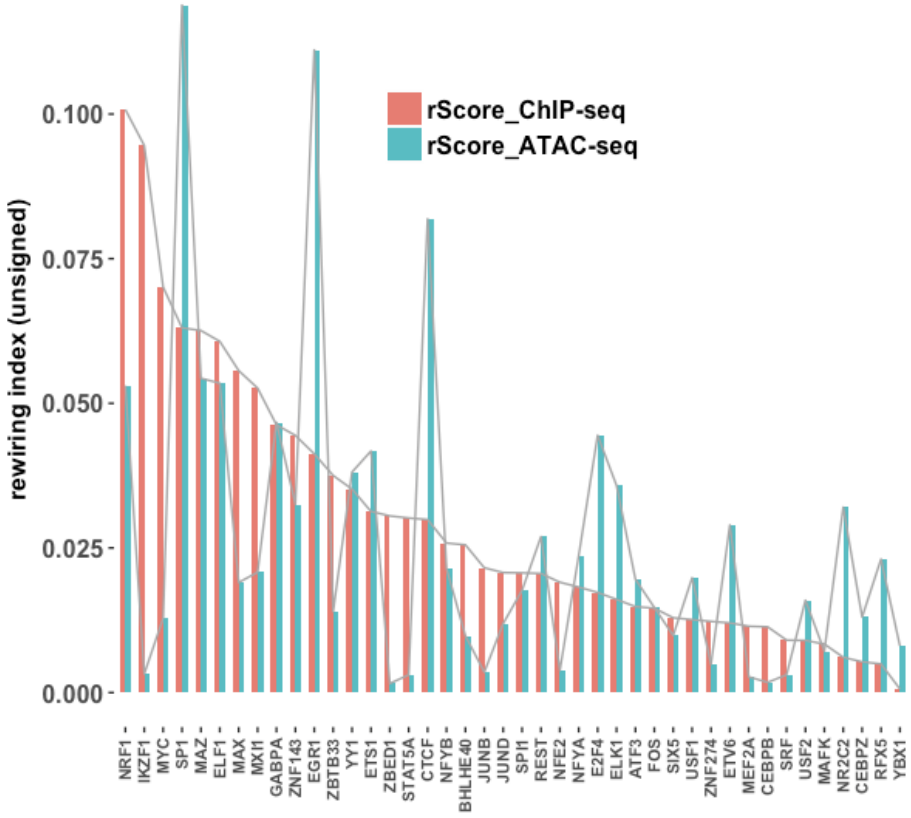
Page 59: [125] Deleted **Author** **5/4/18 9:05:00 PM**

In addition, we built an imputed network from a published dataset outside ENCODE and evaluated the rewiring of regulatory network. We used ATAC-seq dataset from the paper {\cite: Philip, Mary, et al. "Chromatin states define tumour-specific T cell dysfunction and reprogramming." Nature 545.7655 (2017): 452.} and show that the rewiring from ChIP-seq based network can be recapitulated using T cell ATAC-seq data.

{result doesn't look good, we may end up not using ATAC-seq dataset here.}

Try

Page 59: [126] Deleted **Author** **5/4/18 9:05:00 PM**



[[to add ATAC-seq from Christina Leslie lab tissue rewiring using imputed]]

Page 61: [127] Deleted	Author	5/4/18 9:05:00 PM

Author Response	We thank referee for bringing this point and we feel it is a good comment. Actually, the referee is correct many of the cancer transcriptome is similar to each other and In relation to this & other points we made a new figure in our revised version. Which is shown in the response to point 4.7

Excerpt 1 From	One of the strengths of ENCODE release 3 is massive expansion of functional genomic data into various primary cells and tissue types. In this revision, we have extensively

Revised Manuscript Author Response	explored the chromatin landscape and expression patterns across all of available ENCODE primary cells and tissues, and compared themWe agree with existing immortalized cell lines with deep annotations. We have chosen CTCF ChIP-seq, distal enhancers, and RNA-seq, which has the most abundant number of cell types in ENCODE, as examplesthe referee that many cancer transcriptomes de-differentiate and lose diversity during tumorigenesis. We aimed to highlight this point. We performed RCA/PCA using deep integration of the ENCODE resources. In relation to this and other points, we have expanded our analysis on RNA-Seq, shRNA RNA-Seq, and TF ChIP-seq data and found that cancer cells tend to cluster together and stay away from their normal counterparts. Please refer the updated main figure on stemness in the revised manuscript and made a new figure, which is shown in the response to the point REF4.6.

Page 61: [128] Deleted **Author** **5/4/18 9:05:00 PM**

existing immortalized cell lines with deep annotations.

We have chosen CTCF ChIP-seq, distal enhancers, and RNA-seq, which has the most abundant number of cell types in ENCODE, as examples

Page 61: [129] Deleted **Author** **5/4/18 9:05:00 PM**

RNA-Seq, shRNA RNA-Seq, and TF ChIP-seq data and found that cancer cells tend to cluster together and stay away from their normal counterparts.

Please refer the updated main figure on

Page 62: [130] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 62: [131] Formatted Table **Author** **5/4/18 9:05:00 PM**

Formatted Table

Page 62: [132] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 62: [133] Deleted **Author** **5/4/18 9:05:00 PM**

We thank the referees for bringing this point out and we have done what they suggested.

- Regarding "H1 may not necessarily be the best cells to compare with tumor phenotype"

We have chosen H1-hESC because it offers the broadest ChIP-seq coverage and has the most amount of other assays in ENCODE. In our revised manuscript, we have expanded our analysis to other stem cells.

- Regarding "other stem cells (like tissutal stem cells)"
We have compared other available stem-related cell types, as suggested by the referee, to H1-hESC to show that H1-hESC is not very different from other stem cells from tissues. We have evaluated regulatory activity of all ENCODE biosamples and across all available stem-like cells in ENCODE and measured the distance between stem-like cells. We show that H1-hESC is not far distinct from other stem-like cells. As shown earlier, one analysis we have added is to look at regulatory networks of CTCF, one of the most widely assayed TF in ENCODE. As expected, all of 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 .

MARK'S DICTATION 4/22/2018
Comment on stem cells.

Page 62: [134] Formatted **Author** **5/4/18 9:05:00 PM**
Justified

Page 62: [135] Formatted **Author** **5/4/18 9:05:00 PM**
Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [135] Formatted **Author** **5/4/18 9:05:00 PM**
Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [136] Deleted **Author** **5/4/18 9:05:00 PM**
think was very good. We initially focused on H1 because of course, that's the main, the stem cell with the most data and end code. However, the referees comment really thought us to think about this as a

Page 62: [137] Formatted **Author** **5/4/18 9:05:00 PM**
Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [138] Formatted **Author** **5/4/18 9:05:00 PM**
Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [139] Formatted **Author** **5/4/18 9:05:00 PM**
Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [140] Formatted **Author** **5/4/18 9:05:00 PM**
Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [141] Deleted **Author** **5/4/18 9:05:00 PM**

Page 62: [142] Formatted **Author** **5/4/18 9:05:00 PM**
Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [143] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [144] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [145] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [146] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [147] Deleted **Author** **5/4/18 9:05:00 PM**

with an end code. This makes for a very nice picture which we now include as

Page 62: [148] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [149] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [149] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [150] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [151] Formatted **Author** **5/4/18 9:05:00 PM**

Justified

Page 62: [152] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [152] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [153] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [153] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [154] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [155] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [155] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [155] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [155] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Helvetica Neue, 12 pt, Pattern: Clear

Page 62: [156] Deleted **Author** **5/4/18 9:05:00 PM**

We also want to highlight here that there are many tissue types avaialble from encodee

Page 62: [157] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 62: [157] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 62: [158] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 62: [159] Commented **Patrick McGillivray** **5/4/18 10:31:00 PM**

Peng made a useful comment here, that it is unusual to examine PC2 and PC3, and that choosing these components while rejecting PC1 due to potential 'batch effect' is a bit challenging to justify.

Page 62: [160] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Times New Roman, 10 pt

Page 62: [161] Formatted **Author** **5/4/18 9:05:00 PM**

Line spacing: single

Page 62: [162] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Times New Roman, 10 pt

Page 62: [162] Formatted **Author** **5/4/18 9:05:00 PM**

Font:Times New Roman, 10 pt

Page 64: [163] Deleted **Author** **5/4/18 9:05:00 PM**

In the revision, we have definitely taken this comments to heart and have added in main text figures that look at the degree to which structural variants, or SVs, measure background mutational rate, and they also affected the network rewiring. We think this is an ideal illustration of the ENCODE data since, in addition to mapping a lot about the function of the genome, some of the new incurred data sets actually give rise to structural variants meaning that structural variants are an integral output of the product. Relating them to network wiring and background mutation rate is an ideal illustration of the value of the data and the project. We have constructed a number of new main figures that address this and we quite heartly thank the referee for pointing this out. To summarize our conclusion,

- 1. we did observe an elevated SNV/indel rate around the breakpoints and found an elevated mutation rate around the breakpoints (Excerpt 1) we made a supplement figures
- 2. we explored

Page 65: [164] Deleted **Author** **5/4/18 9:05:00 PM**

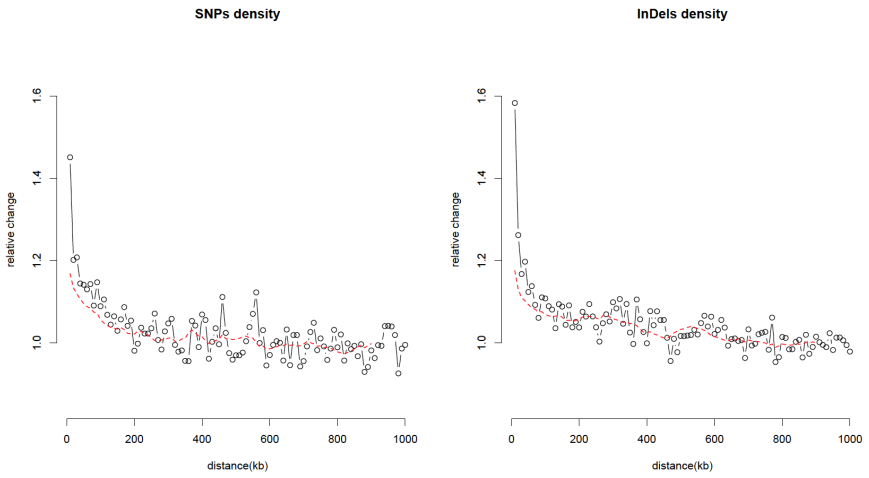
.(Excerpt 2)

-

•

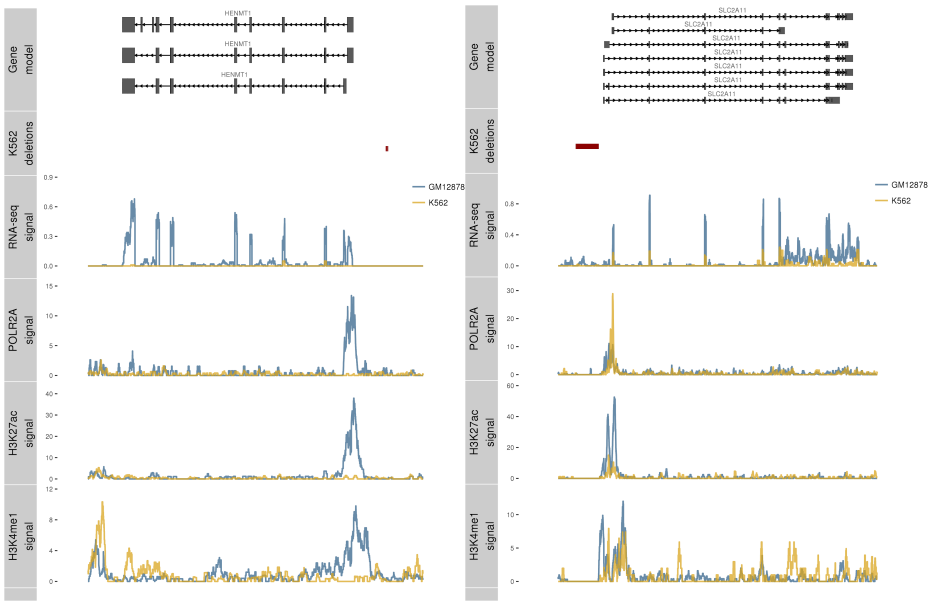
Page 65: [165] Deleted **Author** **5/4/18 9:05:00 PM**

Regarding the relationship of SNV to SV



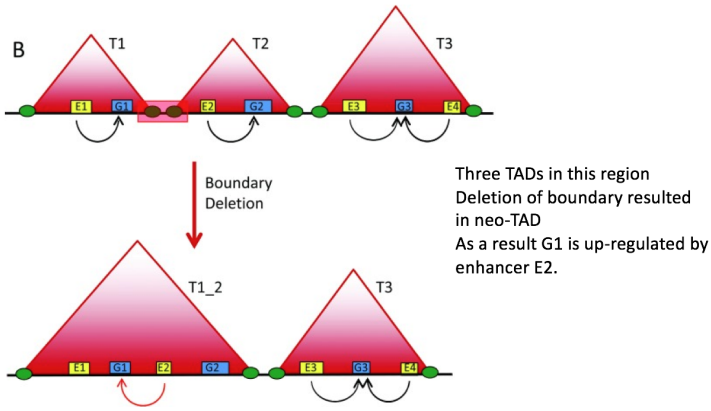
Page 66: [166] Deleted **Author** **5/4/18 9:05:00 PM**

Promoter and SV examples:

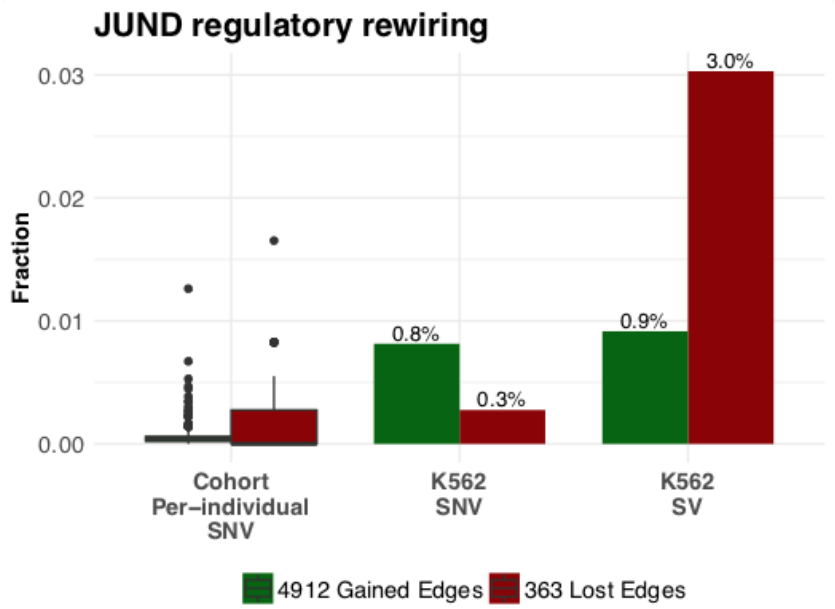


Page 67: [167] Deleted **Author** **5/4/18 9:05:00 PM**

Excerpt 4 From Revised Manuscript	Feng's figure



Page 67: [168] Deleted **Author** **5/4/18 9:05:00 PM**



Page 70: [169] Deleted **Author** **5/4/18 9:05:00 PM**

Page 71: [170] Deleted **Author** **5/4/18 9:05:00 PM**

Excerpt From Revised Manuscript	

Page 71: [171] Deleted **Author** **5/4/18 9:05:00 PM**

Excerpt From Revised Manuscript	

Page 71: [172] Deleted **Author** **5/4/18 9:05:00 PM**

Excerpt From Revised Manuscript	

Page 73: [173] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [174] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [175] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [176] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [177] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [178] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [179] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [180] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [181] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [182] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 73: [183] Commented **Patrick McGillivray** **5/4/18 9:23:00 PM**

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

Perhaps 'main results' in substitution.

Page 73: [184] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [185] Deleted	Author	5/4/18 9:05:00 PM

The referee mentioned his/her confusion about whether this is a prospective or a biology paper. We thank the referee for this point

Page 73: [186] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt, Bold

Page 73: [187] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [188] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [189] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [190] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [191] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [192] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [193] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [194] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [195] Commented	Patrick McGillivray	5/4/18 9:23:00 PM

Just flagging that numbers need completion here.

Perhaps also some work on wording/grammar.

E.g. "2017 histone ChIP-Seq data."

What are the units here? 'Data' could mean a number of different things.

Page 73: [196] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [197] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 73: [198] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [199] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 74: [200] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 74: [201] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 74: [202] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 74: [203] Commented **Patrick McGillivray** **5/4/18 5:25:00 PM**

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.

Page 74: [204] Formatted **Author** **5/4/18 9:05:00 PM**

Justified

Page 74: [205] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 74: [206] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 74: [207] Deleted **Author** **5/4/18 9:05:00 PM**

Level	Annotation type	Example Applications
Element	- Enhancers level 1-3 - SV and SNV in cell lines - Extended genes definitions	- BMR estimation (Fig. 2) - Genome annotation (Fig. 6) - Variant prioritizations (Fig. 6) - Somatic & germline burdens (Fig.1)
Network	*Distal network:* - Enhancer-gene (computational) - Enhancer-gene (computational + Hi-C) - TF-Enhancer-gene *Proximal network:* Experimental based: - TF/RBP Universal networks (strong & weak) - TF/RBP tissue specific networks (binary & probabilistic) Imputed: - DHS imputed tissue specific TF networks	- TF/RBP Regulatory Activities (Fig.3) - Network rewiring (Fig. 4) - Network Hierarchies (Fig. 5) - Stemness analysis (Fig. 5) - TF binding disruptiveness (Fig. 5)

Page 74: [208] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [209] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [210] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [211] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [212] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [213] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [214] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [214] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [214] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [215] Commented	Patrick McGillivray	5/4/18 6:15:00 PM

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.

Page 74: [216] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [217] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [218] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [219] Commented	Patrick McGillivray	5/4/18 5:40:00 PM

Again, not sure about the word goal in this context.

Suggest perhaps 'main result' instead.

Page 74: [220] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [220] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [221] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [222] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [223] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [224] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [225] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [225] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [226] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [226] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [227] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 74: [228] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 77: [229] Deleted	Author	5/4/18 9:05:00 PM

We thank the reviewer for bringing out these references. In our revised manuscript, we tried to make it clear that we are not claiming to have developed negative binomial regression or to be the first to apply it to cancer genomics. We want to point out that negative binomial regression is a very standard statistical technique that has been used in many contexts in genomics. In fact, some of the references, such as Martincorena et al. 2017, came out after our initial submission in Aug 2017, and some of them have diverse focuses such as positive selection patterns instead of BMR estimation in noncoding regions. We have tried to give a better context of existing work in our revised manuscript.

We want to further clarify that the main points in our paper are

Page 77: [230] Formatted	Author	5/4/18 9:05:00 PM

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)

Page 77: [231] Deleted	Author	5/4/18 9:05:00 PM

. The ENCODE3 rollout dramatically expands the number of available features to use for negative binomial regression to 2069 (as compared to 8 in

Page 77: [232] Deleted	Author	5/4/18 9:05:00 PM

They are released in a ready to use format

There are 2017 histone modification data from xxx histone modification types and xxx cell types

The majority (1,339) of the histone data are from real tisse or primary cells

We expanded replication timing data from simply HeLa (cite MutsigCV) or several cell lines to 52 datasets, including xx tissues.

2. The above expansion can noticeably

Page 77: [233] Deleted	Author	5/4/18 9:05:00 PM

accuracy either using the features directly or principal components.

Page 77: [234] Deleted	Author	5/4/18 9:05:00 PM

. While it's valuable matching a cancer cell to its cell of origin, tumors, *as also mentioned by multiple referees*

Page 77: [235] Commented	Patrick McGillivray	5/4/18 5:57:00 PM

Having 'a goal of demonstrating the value of the data' sounds relatively biased, and a bit unlike a scientific study.

Perhaps this whole section could be reworded:

e.g., 'Our main result related to BMR estimation is a more accurate model enabled by the expanded number of features available through ENCODE data...'

Page 77: [236] Deleted	Author	5/4/18 9:05:00 PM

sets provide the best overall fit to estimate background mutation rate.

Page 79: [237] Deleted	Author	5/4/18 9:05:00 PM

We know about this stuff

Page 80: [238] Deleted	Author	5/4/18 9:05:00 PM

Page 80: [239] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 80: [239] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 80: [240] Deleted	Author	5/4/18 9:05:00 PM

1. We

Page 80: [241] Deleted	Author	5/4/18 9:05:00 PM

on the permuted dataset for breast cancer and

Page 80: [242] Deleted	Author	5/4/18 9:05:00 PM

the

Page 80: [242] Deleted	Author	5/4/18 9:05:00 PM

the

Page 80: [242] Deleted	Author	5/4/18 9:05:00 PM

the

Page 80: [242] Deleted	Author	5/4/18 9:05:00 PM

the

Page 80: [242] Deleted	Author	5/4/18 9:05:00 PM

the

Page 80: [243] Formatted	Unknown

Font:(Default) Times New Roman, (Asian) Times New Roman

Page 80: [243] Formatted	Unknown

Font:(Default) Times New Roman, (Asian) Times New Roman

Page 80: [244] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 80: [244] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 80: [244] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 80: [245] Deleted	Author	5/4/18 9:05:00 PM

Results show that we have comparable performance with the permutations dataset.

Page 80: [246] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 80: [246] Formatted	Author	5/4/18 9:05:00 PM

Font:12 pt

Page 82: [247] Deleted	Author	5/4/18 9:05:00 PM

Admittedly, we are making assumptions and the referee is completely correct in pointing this out. We have tried to be more precise in the text that we are assuming that the large number of ENCODE assays, when integrated, allow us to more directly get the functional nucleotides, but this, of course, is an assumption. It is hard to tell to what degree one can succeed in finding the current events in cancer. It is hard to back this up with the gold standard, but we think that some of the points are self evidently obvious. We have tried to make this clear in text and thank the referee for pointing this out.

Page 84: [248] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 84: [249] Formatted Table **Author** **5/4/18 9:05:00 PM**

Formatted Table

Page 84: [250] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [251] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [252] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [252] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [253] Formatted **Author** **5/4/18 9:05:00 PM**

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

Page 84: [254] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [255] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [256] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [257] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [258] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [259] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [260] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [261] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [261] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [262] Deleted **Author** **5/4/18 9:05:00 PM**

discoveries. However, we did labeled the known driver genes in our calculations with supporting pubmed IDs. We further compared our results with the PCAWG reports (unpublished data).

Page 84: [263] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt, Highlight

Page 84: [264] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [265] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [266] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [267] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [268] Deleted **Author** **5/4/18 9:05:00 PM**

added the following two aspects here.

1. Germline SNV analysis

We have extracted cancer GWAS sites from GWAS Catalog and calculated the GWAS SNP enrichment in different annotation categories. We found that the

Page 84: [269] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [270] Deleted **Author** **5/4/18 9:05:00 PM**

gave us much better cancer associated SNP enrichment (details see excerpt 1 below).

Page 84: [271] Formatted **Author** **5/4/18 9:05:00 PM**

Font:12 pt

Page 84: [272] Deleted **Author** **5/4/18 9:05:00 PM**

2. Expression stratification analysis using extended genes

We used the mutation status separate the patients into groups with or without mutations depending on different types of annotations. The extended gene annotation, which include coding sequence, proximal and distal regulatory elements, demonstrated the best largest expression changes among the mutated and non-mutated patient groups (details see excerpt 2 below).

Page 84: [273] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 84: [274] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 84: [275] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 84: [276] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 84: [277] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 84: [278] Formatted **Author** **5/4/18 9:05:00 PM**

Font:10 pt

Page 84: [279] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 84: [280] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 84: [281] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 84: [281] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 84: [282] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 84: [283] Formatted	Author	5/4/18 9:05:00 PM

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"

Page 84: [284] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 84: [284] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 84: [285] Formatted	Author	5/4/18 9:05:00 PM

Font:10 pt

Page 84: [286] Formatted	Author	5/4/18 9:05:00 PM

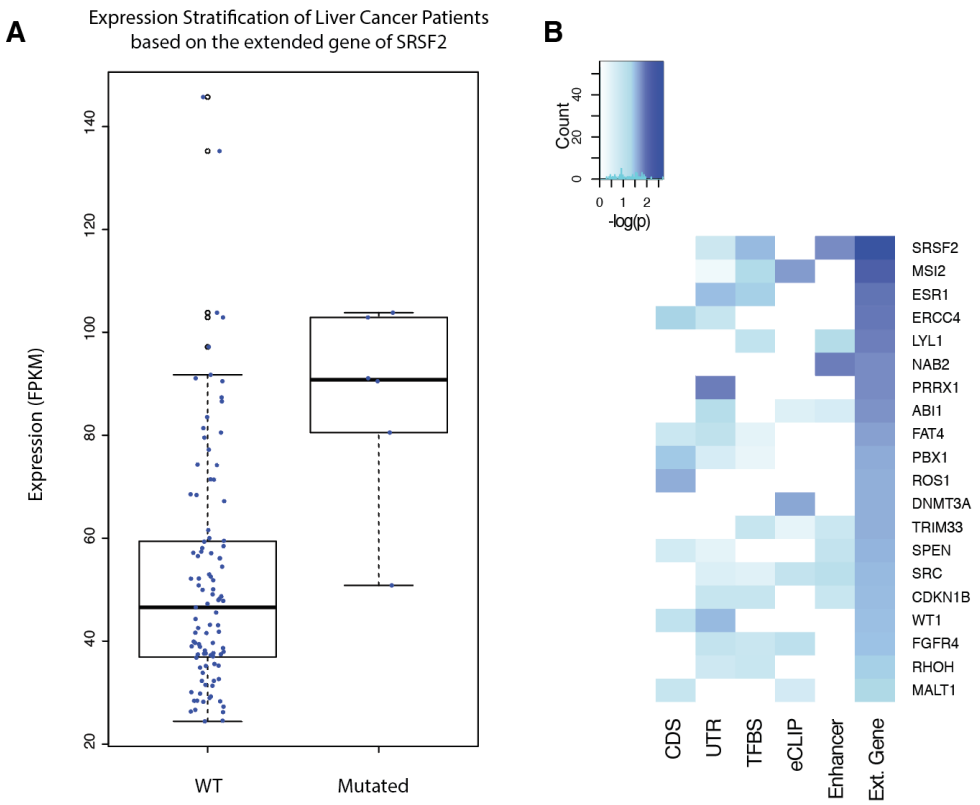
Font:10 pt

Page 84: [287] Formatted	Author	5/4/18 9:05:00 PM

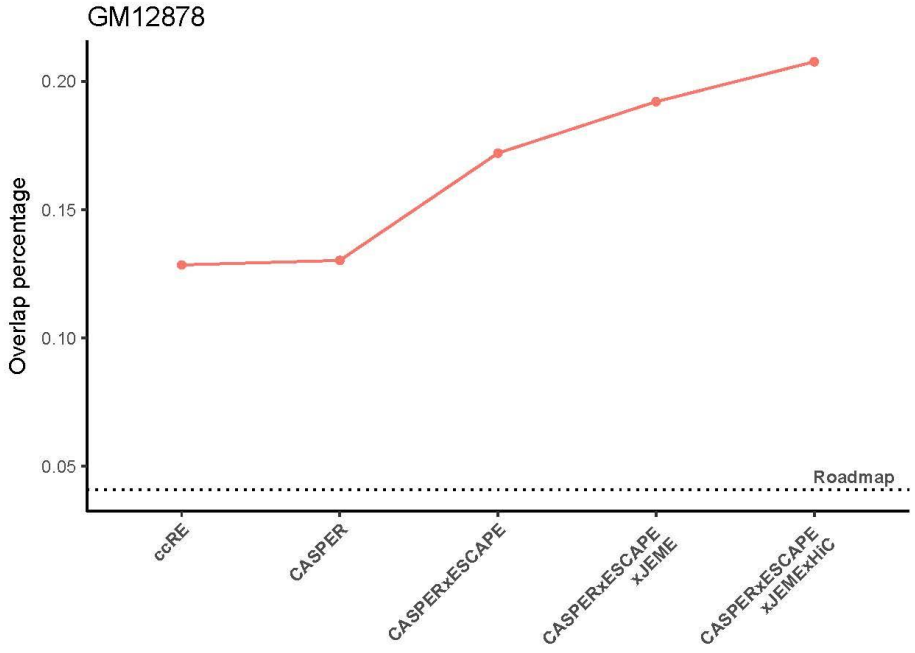
Font:10 pt

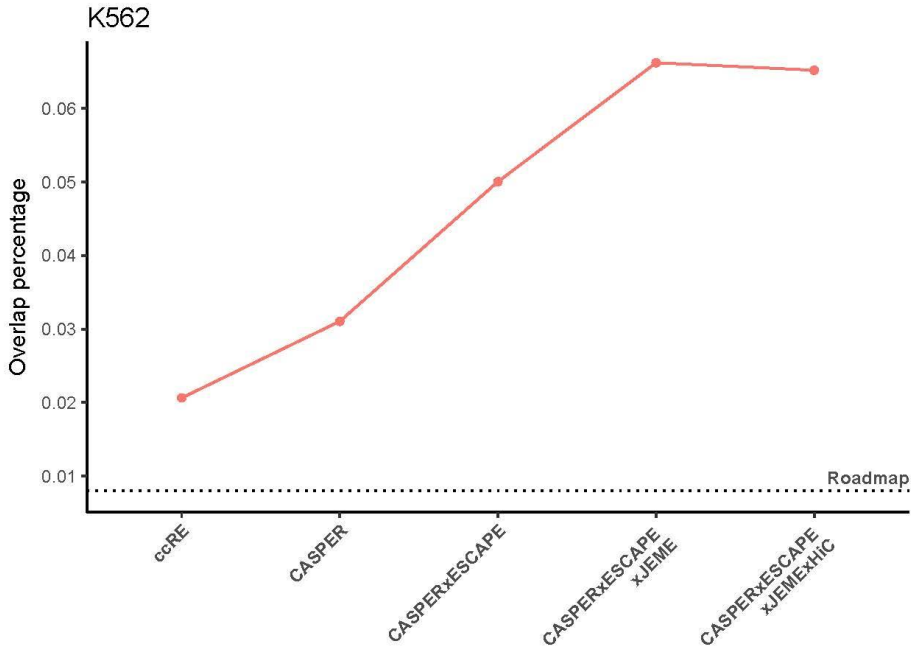
Page 85: [288] Deleted	Author	5/4/18 9:05:00 PM

For a given test region, we consider the expression (FPKM) of patients with a mutation or no mutation in that region to be separate distributions. By using a wilcoxon two-sided test, we test to see whether the expression of mutated patients versus non-mutated patients is different. The test regions we consider are the CDS, UTR, TFBS, eCLIP, Enhancer, and Extended Gene Definition. We find that in many genes, the p-value associated with expression stratification between the two groups is much more significant when using the extended gene than any of its individual parts, suggesting an advantage of the extended gene. Furthermore, when performing this analysis on liver cancer patients using the HepG2 annotations, we find that mutations in the extended gene of SRSF2 give the strongest p-value for stratifying expression of that gene. SRSF2 is a well known splicing factor involved heavily in driving hepatocellular carcinoma development. \cite{28082404}. The specific case of SRSF2 is shown in Panel A. Mutated samples in the extended gene definition are more likely to have higher expression of SRSF2 when compared to WT. Panel B below shows the -log p-value of stratifying expression of mutated and non-mutated patients in different genes using different test regions.



Page 87: [289] Deleted	Author	5/4/18 9:05:00 PM





Page 88: [290] Deleted	Author	5/4/18 9:05:00 PM

incorporates HiC data to our previous computational approach JEME (cite) to better capture the physical interactions between enhancers and target genes. We benchmark

Page 88: [291] Deleted	Author	5/4/18 9:05:00 PM

against published chromHMM enhancer-gene linkages using the GTEx whole blood eQTLs in GM12878. We showed that our JEME+HiC gene-enhancer linkages showed noticeably highly enrichment in whole blood eQTLs. (Excerpt 1)

Page 88: [292] Deleted	Author	5/4/18 9:05:00 PM

methods

We have compared the gene community model

Page 88: [293] Deleted	Author	5/4/18 9:05:00 PM

by extending our analysis from 122 GM12878 and K526 dataset to 862 TF ChIP-Seq assays included in ENCODE data portal. Analysis showed that our method can better preserve the

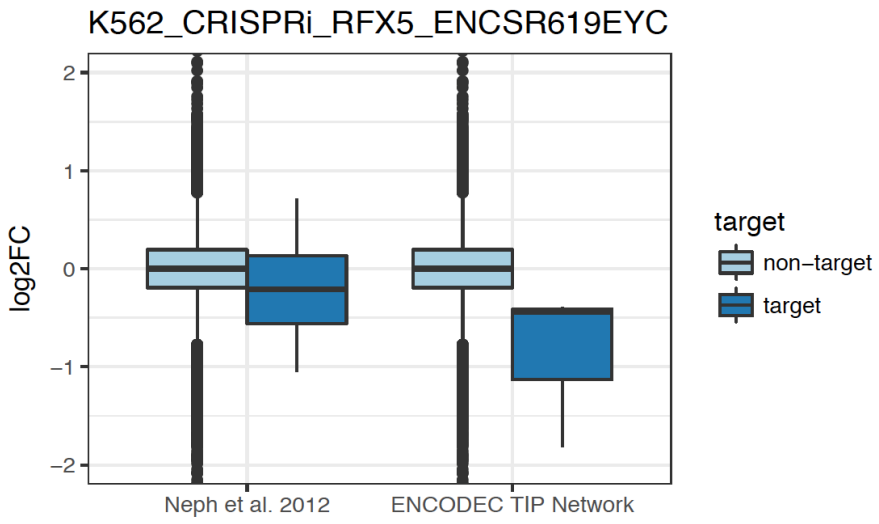
Page 88: [294] Formatted	Author	5/4/18 9:05:00 PM

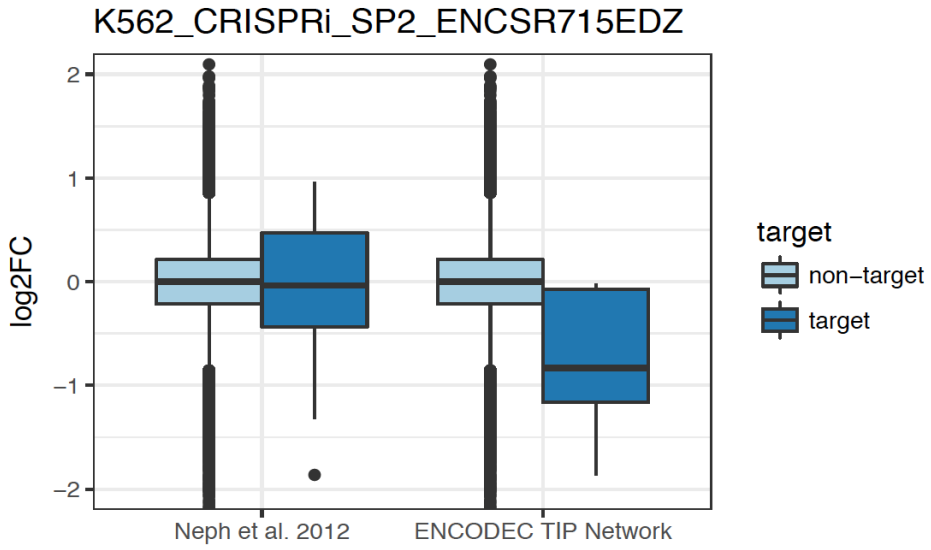
Justified, Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

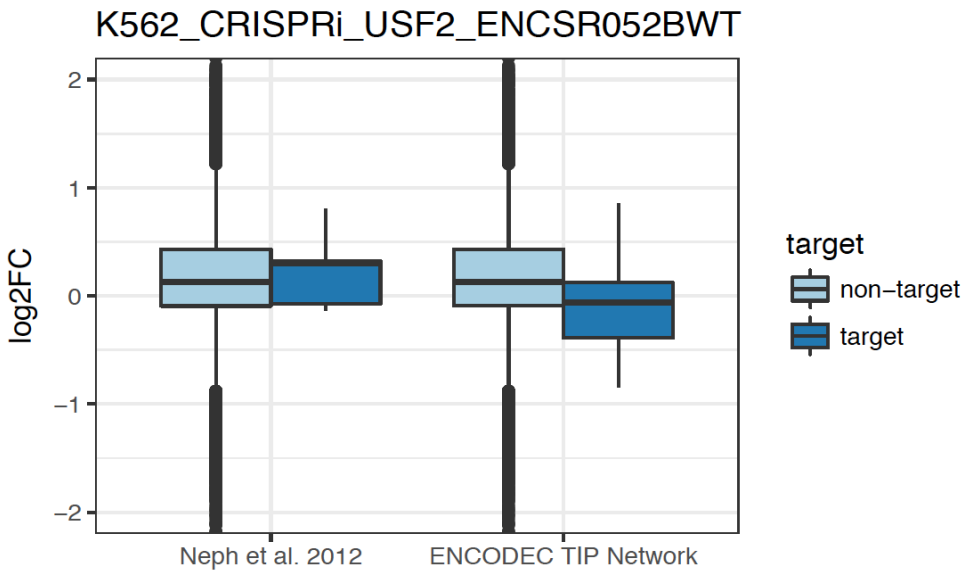
Page 88: [295] Deleted	Author	5/4/18 9:05:00 PM

Manuscript
(in

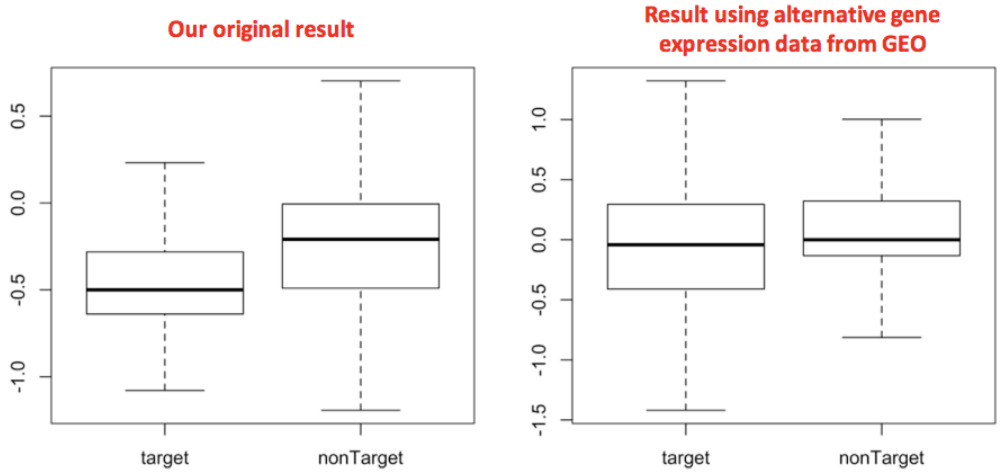
Page 96: [296] Deleted	Author	5/4/18 9:05:00 PM

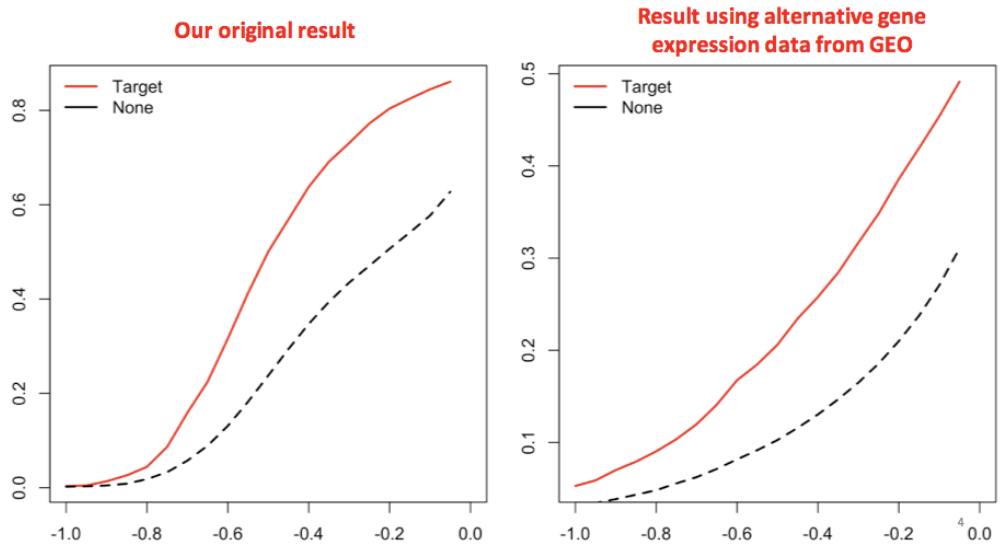

K562_CRISPRi_RFX5_ENCSR619EYC
log2FC
2
1
0
−1
−2
Neph et al. 2012
ENCODEC TIP Network
target
non−target
target


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



Page 99: [297] Deleted	Author	5/4/18 9:05:00 PM

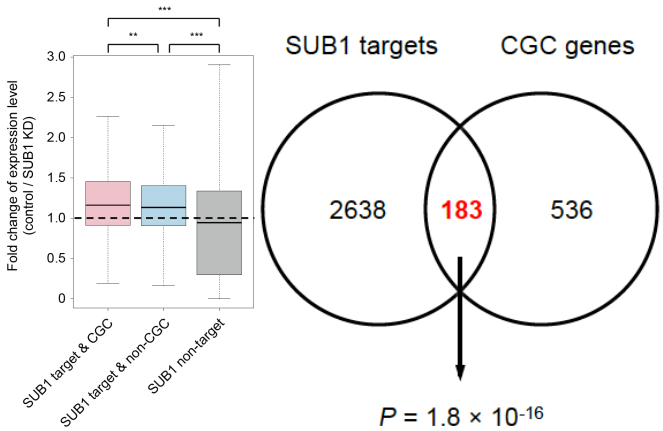


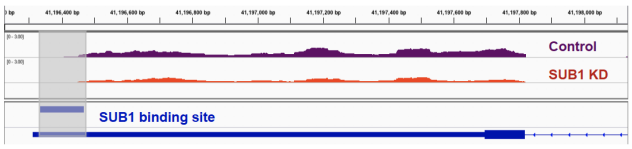
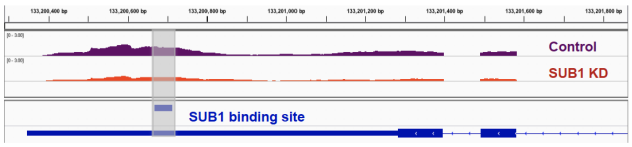
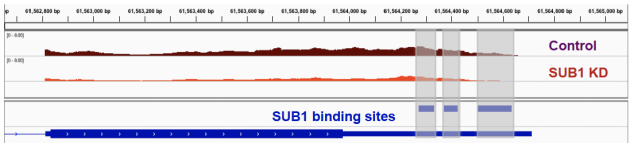


Page 100: [298] Deleted **Author** **5/4/18 9:05:00 PM**

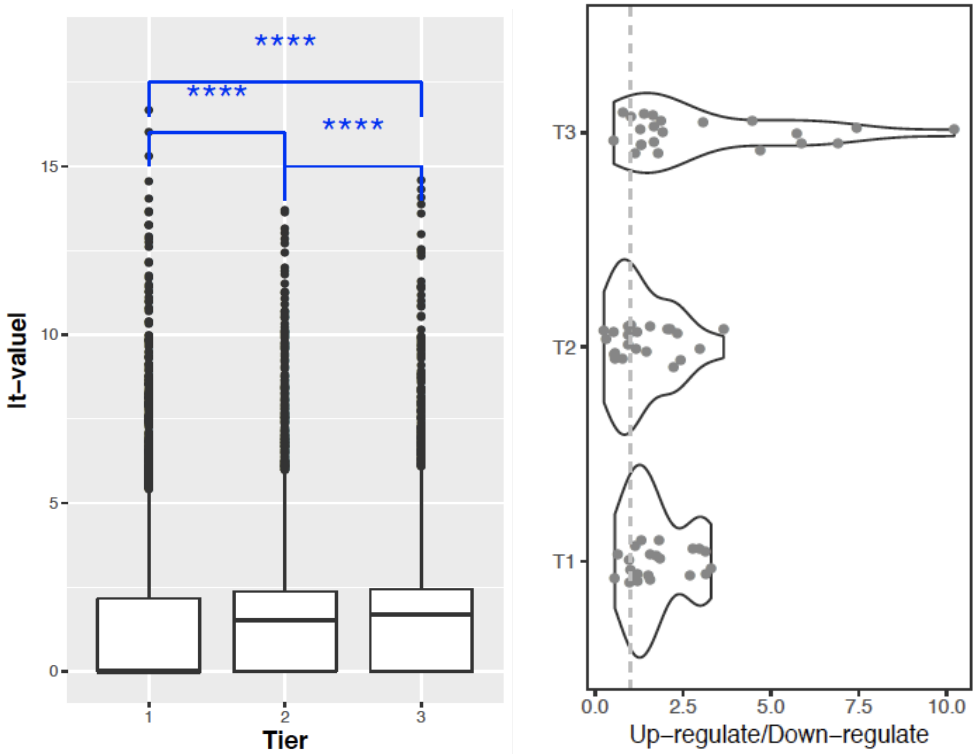
, and actually, we were able to elaborate on this considerably.

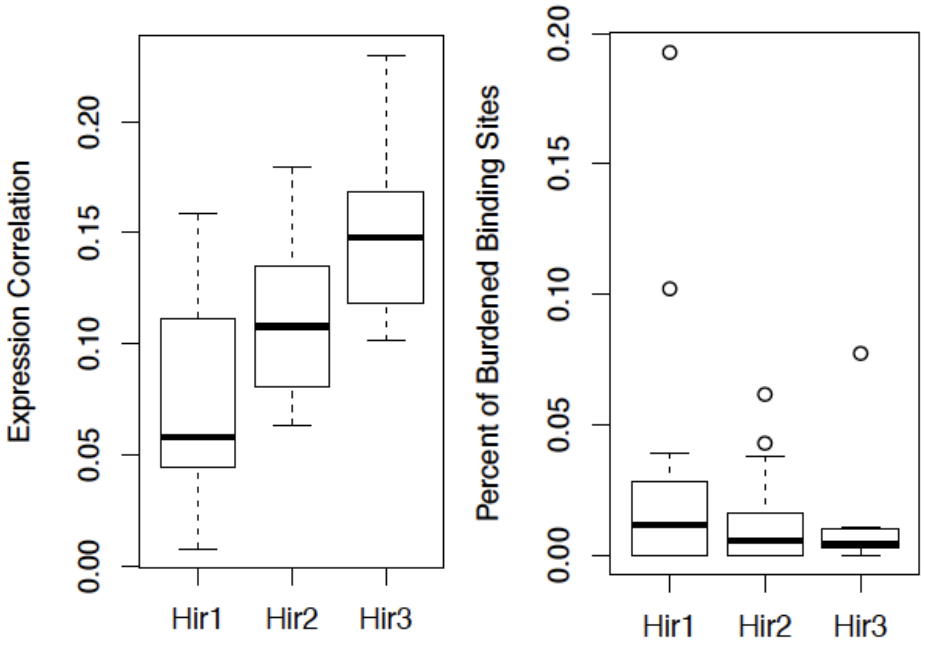
Page 101: [299] Deleted **Author** **5/4/18 9:05:00 PM**



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 103: [300] Deleted **Author** **5/4/18 9:05:00 PM**





Page 104: [301] Deleted	Author	5/4/18 9:05:00 PM

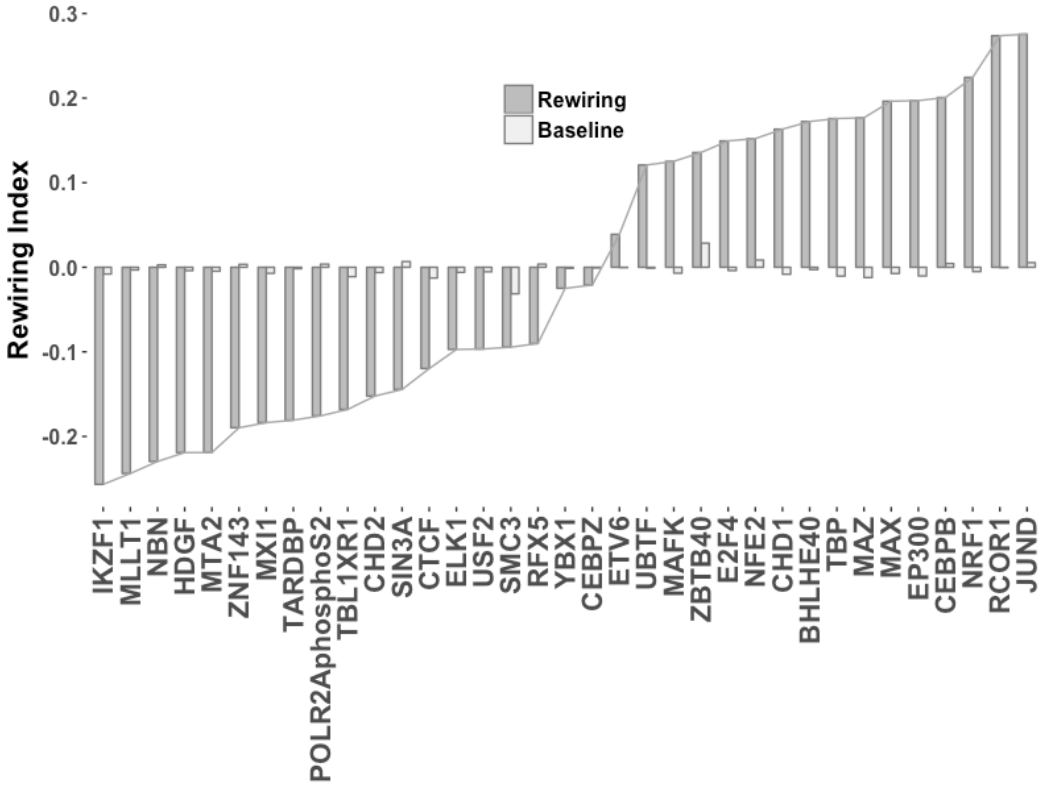
We

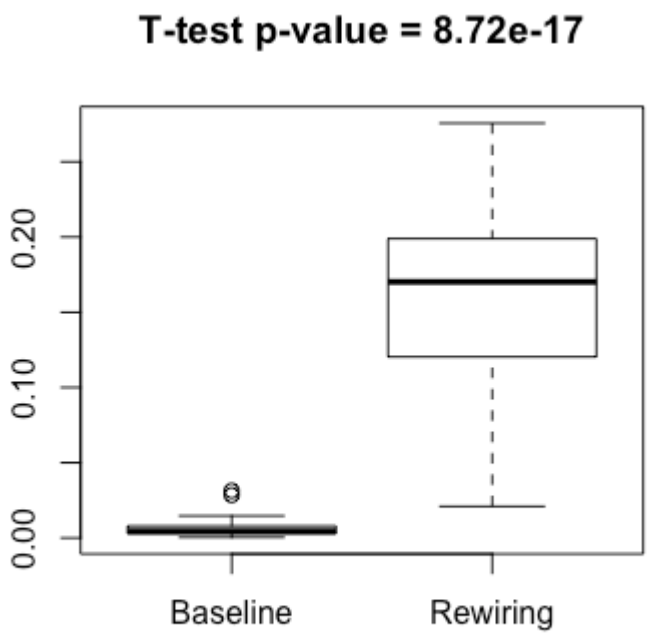
Page 104: [302] Deleted	Author	5/4/18 9:05:00 PM

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.

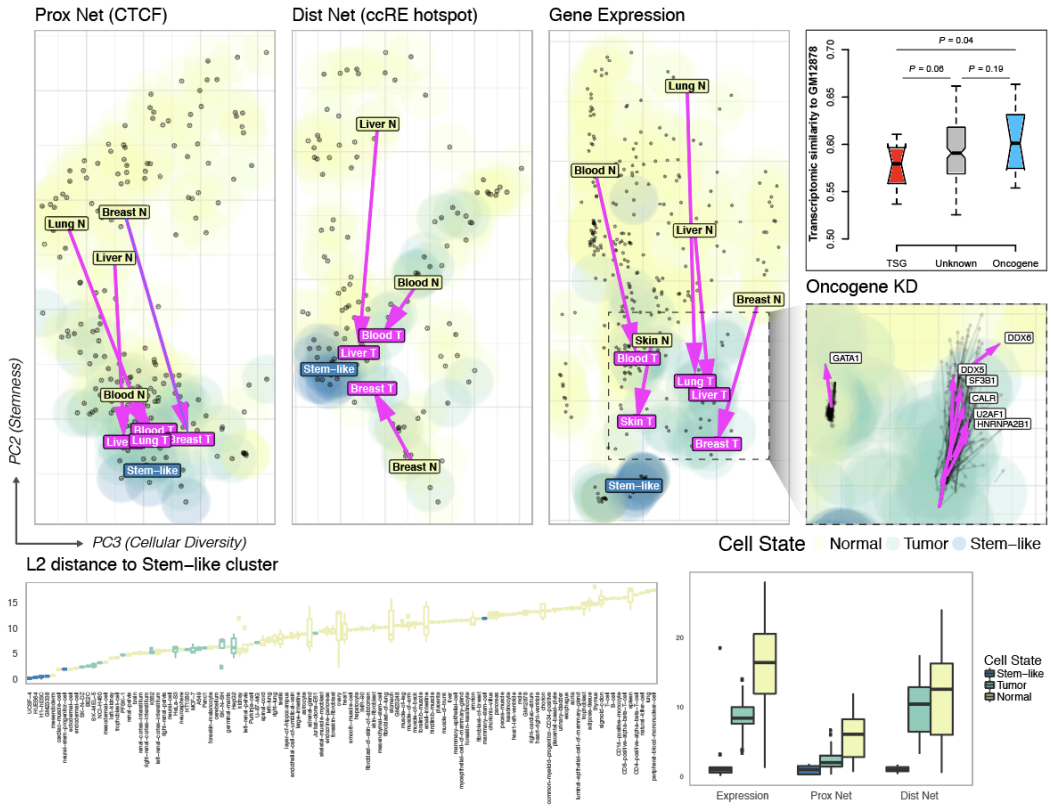
As expected, when two cellular context are similar, as shown in “baseline”, minimal number of edges do change targets.

Page 105: [303] Deleted	Author	5/4/18 9:05:00 PM





Page 110: [304] Deleted	Author	5/4/18 9:05:00 PM



Page 112: [305] Deleted	Author	5/4/18 9:05:00 PM