# Tags:

Use comma for seperation between tags

| | |
|---|---|
| <ID> | REF 0.0 - title of the comment |
| <TYPE> | $$$BMR<br>$$$Power<br>$$$Presentation<br>$$$Annotation<br>$$$Network<br>$$$Hierarchy<br>$$$CellLine<br>$$$Stemness<br>$$$Validation<br>$$$NoveltyPos<br>$$$NoveltyNeg<br>$$$Minor<br>$$$Validation |
| <ASSIGN> | @@@ |
| <PLAN> | &&&AgreeFix - agree and fix<br>&&&DisagreeFix - disagree but we fix, obsequious, and we're safe<br>&&&OOS - out of scope<br>&&&Defer - help me |
| <STATUS> | %%%TBC: To Be Continued<br>%%%DONE : Finished<br>%%%MORE : Go above and beyond the scope of the question and indicates more analyses to be done |

PLEASE NOTE $$$ @@@ &&& %%% are reserved as above. For all other tags, use ### only.

Usage example:

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

**Style Definition:** Subtitle: Left, Line spacing: multiple 1.15 li

**Style Definition:** Comment Text: Font:(Default) Arial, (Asian) Arial, Left

**Style Definition:** Balloon Text: Font:(Asian) Arial, Left

**Deleted:** $$$BMR ... [1]

# Format:

Referee Comment: Courier New

Author Response: Helvetica Neue

Excerpt From Revised Manuscript: Times New Roman

# Referee expertise:

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

**Formatted:** Font:Times New Roman

**Deleted:**

**Deleted:**

**Deleted:**

**Deleted:**

**Deleted:**

**Deleted:**

**Deleted:**

**Deleted:**

**Formatted:** Normal, Space Before: 0 pt

# Editor:

## <ID>REF 0.1 - Overall comments on the paper

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

**Deleted:** -- Editor
**Deleted:** --
**Formatted:** Space Before: 18 pt
**Deleted:** –
**Deleted:** $$$
**Deleted:** @@@
**Deleted:** &&&
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**

| | |
|---|---|
| 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've tried to respond to extensively revise our manuscript in our new version. In summary, we've answered most of these comments. We felt many of them were good suggestions, so we expanded them in large conserving the manuscript, particularly the suggestion related to comparison to stem cell, SVs statistics on networks, and SUB1.<br>One area that we wish to push back a little on is asking us to compare our calculations to that for driver identification. The point of this paper is not to develop a novel method of driver discovery or to find new cancer drivers. The point is to highlight the use of ENCODE3 data in cancer genomics, particularly related to understanding the overall patterns of mutations, network rewiring, and variant prioritization. Obviously, the ENCODE data will be useful for people developing future driver discovery metrics but we believe that's out of scope for this paper. To respond to previous comments, we've shown how in certain contexts, the ENCODE date can help with existing driver discovery measures. |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single
**Formatted:** Justified, Line spacing: single
**Deleted:** these
**Deleted:** on
**Deleted:** ENCODE
**Deleted:** understand
**Formatted:** Font:Helvetica Neue
**Formatted:** Font:Times New Roman
**Formatted:** Line spacing: single

## <ID>REF0.2 – Overall comments on the paper

<TYPE>$$$Presentation

**Deleted:** -- Editor 0
**Deleted:** --
**Deleted:** $$$
**Deleted:** @@@

<ASSIGN>@@@MG @@@JZ
<PLAN>
<STATUS>%%%DONE

**Deleted:** @@@
**Deleted:** &&&compl)
**Formatted:** Normal

| | |
|---|---|
| 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 have clarified the unique aspects of our paper. |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Line spacing: single
**Formatted:** Font:Times New Roman
**Formatted:** Line spacing: single

## <ID>REF0.3 – Overall comments on the paper

<TYPE>$$$Presentation
<ASSIGN>@@@MG @@@JZ
<PLAN>
<STATUS>%%%DONE

**Deleted:** -- Editor 0
**Deleted:** --
**Deleted:** $$$
**Deleted:** @@@
**Deleted:** @@@
**Deleted:** &&&compl
**Formatted:** Normal

| | |
|---|---|
| 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 disagree with the reviewers on this point. We want to make it explicit that (1) this paper is to be considered as a "*resource*" paper, not a novel biology paper (2) that the current Encyclopedia *package is not meant to be structured like previous packages* (i.e. '12 ENCODE). The integrative analysis is meant to be spread over a number of papers and not centered on a single one. |

**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Justified
**Formatted:** Line spacing: single
**Formatted:** Font:Helvetica Neue, 11 pt
**Formatted:** Justified

(3) note that the ENCODE 3 "data" is not explicitly tied to any paper. Unlike previous roll-outs, ENCODE 3 does not associate particular data sets with specific papers and make use of these data contingent on that paper's publication (as codified in an agreement with NHGRI.)

Regarding the novelty of this paper, 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, and its analysis of networks.

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) 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. 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.

| | |
|---|---|
| | |
| Excerpt From Revised Manuscript | |

**Deleted:**

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

# Referee #1 (Remarks to the Author):

## <ID>REF1.0 – Preamble

Deleted: -- Ref 1.1 – Overall comments on the pap ... [2]

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

We ~~would like to~~ appreciate the referee's feedback. Overall the reviewer mentioned that this is an interesting resource but the novelty of the paper is lacking. Regarding the novelty point,, we think differently of the value of our paper. We want to make it clear that his paper is to be considered as a "resource" paper, not a novel biology 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. Thus, where the referee ask for novelty in cancer gene discovery - we strongly feel that this is out of scope.

OTHER IN LEVEL 2.

2.

| Contribution | Subtypes | Data types | ENCODE experiments |
|---|---|---|---|
| Processed raw signal tracks | Histone modification | Signal matrix in TSV format | 2015 Histone ChIP-seq |
| | DNase I hypersensitive site (DHS) | Signal matrix in TSV format | 564 DNase-seq |
| | Replication timing (RT) | Signal matrix in TSV format | 135 Repli-seq and Repli-ChIP |
| | TF hotspots | Signal track in bigWig format | 1863 TF ChIP-seq |
| Processed quantification matrix | Gene expression quantification | FPKM matrix in TSV format | 329 RNA-seq |
| | TF/RBP knockdowns and knockouts | FPKM matrix in TSV format | 661 RNAi KD + CRISPR-based KO |
| Integrative annotation | Enhancer | Annotation in BED format | 2015 Histone ChIP-seq<br>564 DNase-seq<br>STARR-seq |
| | Enhancer-gene linkage | Annotation in BED format | 2015 Histone ChIP-seq<br>329 RNA-seq |

| | | | |
|---|---|---|---|
| | Extended gene | Annotation in BED format | 1863 TF ChIP-seq<br>167 eCLIP<br>Enhancer-gene linkage |
| SV and SNV callsets | Cancer cell lines | Variants in VCF format | WGS<br>BioNano<br>Hi-C<br>Repli-seq |
| Network | RBP proximal network | Network in TSV format | 167 eCLIP |
| | Universal TF-gene proximal network | Network in TSV format | 1863 TF ChIP-seq |
| | Tissue-specific TF-gene proximal network | Network in TSV format | 1863 TF ChIP-seq |
| | Tissue-specific imputed TF-gene proximal network | Network in TSV format | 564 DNase-seq |
| | TF-enhancer-gene network level 1-3 | Network in TSV format | 2015 Histone ChIP-seq<br>564 DNase-seq |

Specifically for the BMR estimation part, the reviewer mentioned that there have been many existing references focusing on applications like cancer driver detection. First, we thank the referee for pointing out to a lot of related references. On the reference side, we have listed many of the papers as the referee suggested and compared them with our approach. We have acknowledged the efforts of many of these references and in the revised version we have further expanded our reference list for some the publications after our initial submission date.  We want to emphasize that the richness of the ENCODE data can actually help many of the methods used in these papers. With a larger pool of covariate selection, the estimation accuracy can be significantly improved.

DISC

[JZ2MG: I am a little bit confused, since this preamble actually contains some of the question. Then do we delete the questions that are mentioned here? I currently feel we should delete them, have some local version and can revert if this is not appropriate.]

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

## <ID>REF1.1 – Comments on the resource releases

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

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

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted:** Justified

**Formatted Table**

| | |
|---|---|
| | 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 the positive feedback. |
| Excerpt From Revised Manuscript | |

## <ID>REF1.2 –

## BMR: comparison with existing literature

<TYPE>$$$BMR $$$Text
<ASSIGN>@@@JZ,@@@WM @@@PDM
<PLAN>&&&OOS
<STATUS>%%%DONE
[JZ2MG: I feel there is some overlapping with the preample. It talks about reference, but I don't want to put it into the preamble since it is too long and no need to re-amphasize this point from our side]

| | |
|---|---|
| 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 |

**Formatted:** Line spacing: single

**Formatted:** Justified

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Deleted:** -- Ref 1

**Deleted:** Scope of the paper --
However, I find the current manuscript seriously lacking. The major problem is simply that most of these applications have already been in the literature for a while, often as high profile papers on their own. So the manuscript is not quite a review but does not seem to have any significant findings either. ... [3]

**Moved down [1]:** Excerpt From ... [4]

**Formatted:** Font:Times New Roman

**Formatted Table**

**Deleted:** -- Ref 1.3 –

**Deleted:** literatures --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** @@@

**Deleted:** @@@JZ @@@

**Deleted:** &&&compl

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

| | |
|---|---|
| | 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 reviewer for identifying these references. We recognize that epigenetic features have been previously been used to estimate BMR and improve driver mutation detection. Our aim was not to produce novel BMR estimation models, but rather to showcase how ENCODE data can help improve the performance of such models.<br>With the wealth data available through ENCODE data, we had a much larger pool of features to choose from to potentially improve BMR estimation. It is worth to mention that ENCODE data is not just cell line data, in fact XXX of this histone modification data is actually from real tissues.l Indeed, we found that application of some additional features from the this expansive set, especially the replication timing data, significantly improved BMR estimation in many cancer types (see Supplement Section S7).<br>For example, many prior efforts to model BMR have been limited by the availability of genomic assays, or by the availability of assays matched by cell-type. For example, Lawrence et al., 2013, used HeLa replication timing data and K562 chromatin state via Hi-C. Martincorena et al., 2017, included histone modification features, but not replication timing. The genomic signals we used from ENCODE have been processed uniformly and are provided in a ready-to-use format for the community.<br>We do not intend to claim it is a new discovery that using matched features are better, but rather to show that the breadth of ENCODE data allows for improved estimates of background mutation rate. We have further acknowledged prior efforts on this topic in our revised manuscript. |
| Excerpt From Revised Manuscript | |

**Deleted:** [[@@@@7mar: we have to say they don't have so much data here]] ... [5]

**Formatted:** Line spacing: single

**Formatted:** Justified

**Formatted:** Justified

**Deleted:** In addition, were able to use cell-type matched feature data across our BMR analysis. This includes more commonly used features for BMR modification, like the 932 histone modification features we used, but also many other features, especially the 51 replication time data, that have proven useful but are less frequently incorporated into BMR models. ... [6]

**Formatted:** Justified

**Deleted:** only

**Deleted:** ... [7]

**Moved (insertion) [1]**

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF1.3 – BMR: novelty of the method

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

| | |
|---|---|
| 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 pointing out the Polak 2015 paper. This is an important reference to relate various genomic features to cancer mutational landscape, and we also cited this paper in our initial submission.<br>It is worth mentioning that we are not trying to reproduce the discovery in that paper, but rather to show how the richness of ENCODE data can help BMR estimation. We also want to emphasize that two points here.<br>1. To select a perfect "matching" feature (not from matter tissue or cell line) is a non trivial problem due to the heterogeneity of cancer. Even in the Polak 2015 paper, H3K9me3 from Breast luminal epithelial cells is a significant feature in 5 out of cancer cancer types they investigated (Fig. 2a). The way larger pool of functional characterization data from ENCODE can help on this matching issue, especially for cancers types that can not find an obvious "matching" data from the Roadmap, such as prostate cancer.<br>2. The goal of the Polak 2015 paper is to predict the cell of origin, while we are aiming to improve the BMR estimation accuracy. The fact that "matched" cell type features performs better in prediction BMR does not mean that other "non-matched" features are not useful to improved the BMR prediction accuracy. 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 932 histone modification features in a ready-to-use format. And also listed many other features, especially the 51 replication time data, that have proven useful but are less frequently incorporated into previous BMR models. |

| | |
|---|---|
| Excerpt From Revised Manuscript | |

## <ID>REF1.4 – BMR: Tissues vs. Cell lines

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

| | |
|---|---|
| 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.<br>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 pointing out comparison of cell line vs tissue. We further investigated this comparison and extended this point more to the RNA-seq and ChIP-Seq data (see updated Figure 5). We think slightly differently with the referee on this point.<br>1. Regarding the cell line data, we still think they are quite useful to predict the mutation rates. Two points need to be noted here are:<br>(1.1) Even in the the Polak 2015 paper, it is not always the case that cell-of-origin can be predicted perfectly using the epigenetic features (Fig. 4 b).<br>(1.2) the Polak 2015 paper only compare among normal tissues from the Roadmap data and they did not perform large scale comparison across various cancer types. Here we used breast cancer as an example. We calculated the |

**Formatted:** Font:Times New Roman
**Formatted:** Line spacing: single
**Formatted Table**

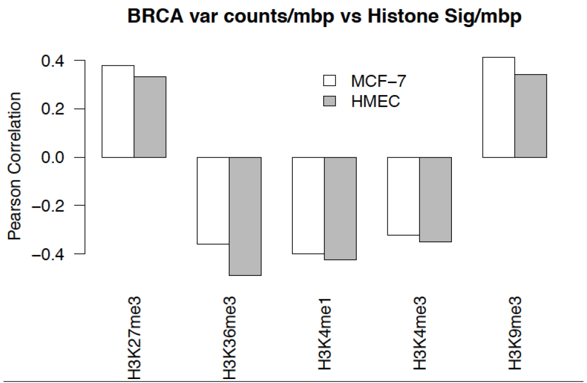
**Deleted:** -- Ref 1
**Deleted:** matching between
**Deleted:** &
**Deleted:** --
**Deleted:** $$$
**Deleted:** $$$Text @@@WM @@@
**Formatted:** Normal
**Deleted:** &&&compl
**Formatted:** Line spacing: single
**Formatted Table**
**Deleted:** Most large-scale cancer genome sequencing papers also have models at various levels sophistication, most of them including the issue of proper tissue-type matching.
**Moved (insertion) [2]**

**Deleted:** Thus, that "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.
**Formatted:** Line spacing: single
**Formatted:** Font:Helvetica Neue
**Moved (insertion) [3]**
**Formatted:** Font:Helvetica Neue
**Formatted:** Font:Helvetica Neue
**Deleted:** . This is an important reference to relate
**Formatted:** Justified
**Formatted:** Font:Helvetica Neue
**Moved (insertion) [4]**
**Formatted:** Font:Helvetica Neue

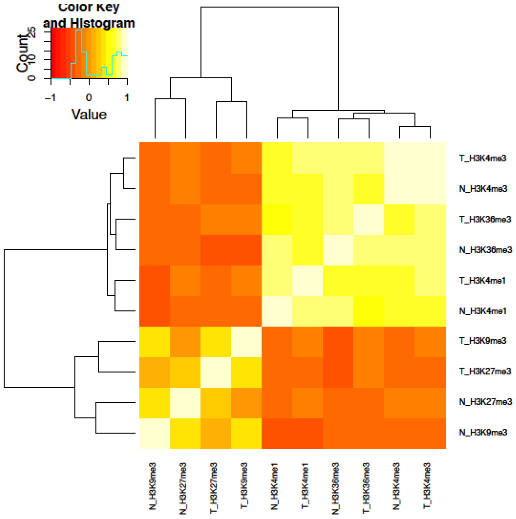
correlation of breast cancer mutation counts (from a patient cohort) per mbp with histone signals from both Breast tissue (the roadmap) and MCF-7 (an ENCODE cell line). As seen from the following figure, MCF-7 provides comparable (and sometimes even better correlation with mutation counts). We also found that histones from tissue and matched cell lines are actually quite correlated in a larger scale (see heatmap below).

**Moved (insertion) [5]**

**Formatted:** Font:Helvetica Neue



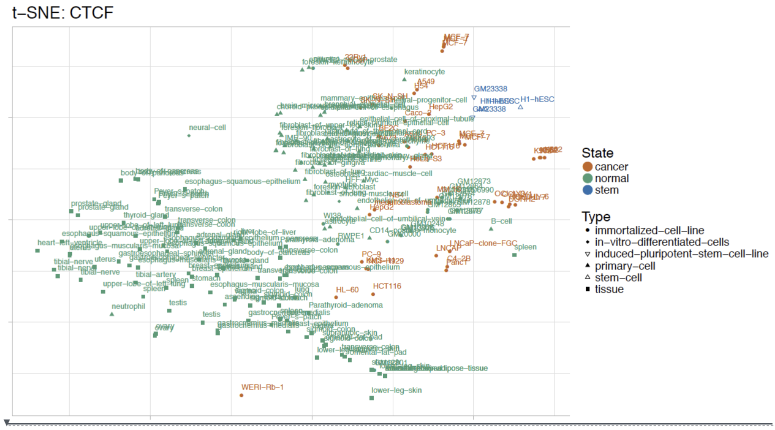
SUPP



2. In general, there are less such data. On the contrary, the cell line functional characterization data has lots of advantage in terms of assay richness. For

**Moved (insertion) [6]**

**Formatted:** Font:Helvetica Neue

| | |
|---|---|
| | example, there is no data for prostate tissue from the roadmap, but cell lines like LNCap might further help under such condition.<br>3. Some genomic features, like expression levels and TF binding events, have been proven to affect somatic mutation rates. We systematically scanned all the cancerous and non-cancerous cell types from ENCODE 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. Our observation is consistent by previous reports, such as Lotem et al. 2005 and Hoadley et al. 2014.  For example, here is the projection of CTCF binding sites from all ChIP-Seq experiments. The fact that cancer cells loose diversity and showed distinct pattern from the primary cells highlights the values of cell line data.<br> |
| Excerpt From Revised Manuscript | |

**Formatted:** Font:Helvetica Neue

**Deleted:** to cancer mutational

**Formatted:** Justified

**Formatted:** Font:Helvetica Neue

**Deleted:** , and we also cited this paper in our

**Formatted:** Font:Helvetica Neue

**Deleted:** submission

**Formatted:** Font:Helvetica Neue

**Deleted:** However, our point is not to provide a "novel" driver detection method, but rather to highlight the value of various types of ENCODE data, some of which are unique features, such as replication timing.

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF1.5 – Difference between ENCODEC and Prev. prioritization methods

<TYPE>$$$BMR,$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&DisagreeFix

**Deleted:** -- Ref 1

**Deleted:** BMR: matching

**Deleted:** Tissues & Cell lines --

**Deleted:** $$$BMR $$$Calc @@@JZ @@@JL&&&compl@@@@7mar: might consolidate with the previous one

<STATUS>%%%DONE
####Dictation

| | |
|---|---|
| Referee Comment | The rest of the sections (and their corresponding supplement sections) are variable in significance and quality. 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 other people have tried to prioritize non-coding elements before. This is definitely true and we are not claiming to be the first. However, 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. |
| Excerpt From Revised Manuscript | |

WE HAVE TRIED TO MAKE THIS CLEARER.

TUREN ISS

## <ID>REF1.6 –

## Novelty and presentation of the paper

<TYPE>$$$Presentation $$$NoveltyPos $$$NoveltyNeg $$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%DONE

| | |
|---|---|
| Referee Comment | Some newer assays such as STARR-seq are helpful, obviously, in better predicting enhancers, but, again, while the analysis done serves as illustrations how ENCODE data can |

**Formatted:** Line spacing: single

**Formatted Table**

**Moved up [2]:** Stepping back, it is not obvious to me that using the ENCODE c ... [8]

**Formatted:** Line spacing: single

**Deleted:** @@@@7mar: this is very difficult/differe ... [9]

**Moved up [4]:** . We calculated the correlation of t ... [11]

**Deleted:** For some specific cancer types, such pr ... [16]

**Formatted:** Font:Helvetica Neue

**Deleted:** As seen from the following figure, MCF-7 ... [12]

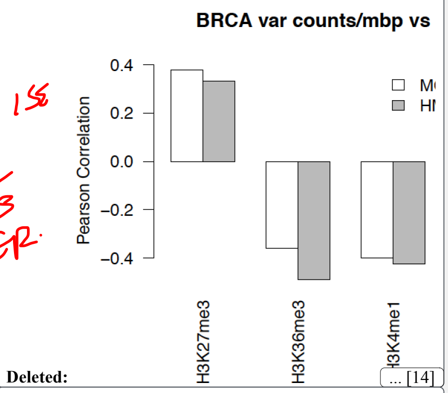
**Moved up [5]:** (and sometimes even better corre ... [13]

**Formatted:** Font:Helvetica Neue

**Moved up [6]:** In general, there are less such dat ... [15]

**Deleted:** note are: ... [10]

**Formatted:** Font:Helvetica Neue



**Deleted:** ... [14]

**Formatted:** Font:Helvetica Neue

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Deleted:** -- Ref 1

**Deleted:** Difference between ENCODEC and Fur ... [17]

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** @@@

**Deleted:** &&&compl

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

|  |  |
|---|---|
|  | be used, the supplement does not seem to give a convincing evidence of how the results found are novel. |
| Author Response | We thank the referee for praising the new STARR-seq assays and we have in fact tried to illustrate the value of novel assays such as STARR-Seq. We have modified both the main manuscript and the supplement to further highlight this. |
| Excerpt From Revised Manuscript |  |

# Referee #2 (Remarks to the Author):

## <ID>REF2.0 – Preamble

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

ORLI COVER

We would like to appreciate the referee's feedback, especially about the positive comments on the value of resource, extended gene, and network rewirings. Regarding the novelty point, Regarding the novelty of this paper, our paper 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, and its analysis of networks. Note also that while we do NOT feel this 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.

| Contribution | Subtypes | Data types | ENCODE experiments |
|---|---|---|---|
| Processed raw signal tracks | Histone modification | Signal matrix in TSV | 2015 Histone ChIP- |

**Formatted:** Font:Helvetica Neue

**Formatted:** Line spacing: single

**Formatted:** Justified

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Deleted:** -- Ref 1.8 – Novelty and presentation of the paper --
Personally, I wonder whether a review paper that gives an update to the ENCODE database and state the illustrative examples succinctly might be more appropriate than several studies, in which more work/descriptions are needed to show novelty, packaged together? ... [18]

**Moved down [7]:** Excerpt From ... [19]

**Formatted:** Font:Times New Roman

**Formatted Table**

**Deleted:** -- Ref 2.1 – Novelty of the paper -- ... [20]

| | | | |
|---|---|---|---|
| | | format | seq |
| | DNase I hypersensitive site (DHS) | Signal matrix in TSV format | 564 DNase-seq |
| | Replication timing (RT) | Signal matrix in TSV format | 135 Repli-seq and Repli-ChIP |
| | TF hotspots | Signal track in bigWig format | 1863 TF ChIP-seq |
| Processed quantification matrix | Gene expression quantification | FPKM matrix in TSV format | 329 RNA-seq |
| | TF/RBP knockdowns and knockouts | FPKM matrix in TSV format | 661 RNAi KD + CRISPR-based KO |
| Integrative annotation | Enhancer | Annotation in BED format | 2015 Histone ChIP-seq<br>564 DNase-seq<br>STARR-seq |
| | Enhancer-gene linkage | Annotation in BED format | 2015 Histone ChIP-seq<br>329 RNA-seq |
| | Extended gene | Annotation in BED format | 1863 TF ChIP-seq<br>167 eCLIP<br>Enhancer-gene linkage |
| SV and SNV callsets | Cancer cell lines | Variants in VCF format | WGS<br>BioNano<br>Hi-C<br>Repli-seq |
| Network | RBP proximal network | Network in TSV format | 167 eCLIP |
| | Universal TF-gene proximal network | Network in TSV format | 1863 TF ChIP-seq |
| | Tissue-specific TF-gene proximal network | Network in TSV format | 1863 TF ChIP-seq |
| | Tissue-specific imputed TF-gene proximal network | Network in TSV format | 564 DNase-seq |

**Deleted:** Referee ... [21]

**Inserted Cells**

**Inserted Cells**

**Merged Cells**

**Deleted:** The manuscript does not report new datasets in addition to the ENCODE release and offers limited conceptual novelty.

**Formatted Table**

**Deleted Cells**

**Formatted Table**

**Deleted:** @@@@mar7 earlier
More info ... [23]

**Moved down [8]:** Hi-C

**Deleted:** Author
@@@@mar7 earlier
More info ... [22]

**Deleted:** Universal and tissue-specific experimental based TF/RBP networks ... [24]

**Inserted Cells**

**Inserted Cells**

**Inserted Cells**

**Moved (insertion) [8]**

**Deleted:** Excerpt From

**Merged Cells**

**Deleted:**

|  |  |  |  |
|---|---|---|---|
|  | TF-enhancer-gene network level 1-3 | Network in TSV format | 2015 Histone ChIP-seq<br>564 DNase-seq |

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

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

|  |  |
|---|---|
| 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. |
| Excerpt From Revised Manuscript |  |

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

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

|  |  |
|---|---|
| Referee Comment | 1) The negative binomial regression (Gamma-Poisson mixture model) was introduced in Nik-Zainal et al. Nature 2016 and Marticorena et al., Cell 2017. Why was not this available method applied, and what is the benefit for the procedure used by the authors? |
| Author Response | In relation to the negative binomial regression, the referee is pointing out that the use of negative binomial regression has been used before. This is a standard statistical technique that's be used in many contexts. The fact that it was earlier |

**Formatted:** Normal, Space Before: 0 pt

**Deleted:** -- Ref 2.2

**Deleted:** --

**Deleted:** $$$

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Deleted:** -- Ref

**Deleted:** .3

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** @@@

**Deleted:** &&&compl

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

**Formatted:** Line spacing: single

**Deleted:** We thank the reviewer to point out these references and they are also good models. Actually one of mentioned paper (Marticorena, 2017) was published on Nov 2017, almost three months after our submission. It comes out three months after our initial submission so we did not cite in the last round. Admittedly, it decrease the novelty of our BMR estimation method, but it also proves that we are technically sound at this point. ... [25]

WE DID, IN FACT, APPLY THIS Z.

an established

| | |
|---|---|
| | used in relation to background mutational rate shows that it is an appropriate approach. Our paper is 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 the very large amount of data and is able to leverage that to more successfully predict background mutation.<br>The fact that the recent Martincorena et al 2017 paper uses this, we think only bolsters the underlying technical validity of our argument. While we admit it does slightly undercut a claim of novelty in this regard, that's not central to our work. |
| Excerpt From Revised Manuscript | |

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

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

| | |
|---|---|
| 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 pointing out the goodness of fit problem and he/she is right that we didn't' provide enough background. Following the referee's suggestion, we made new supplementary figures as requested. In most of the cancer types, the fitting of Gamma-Poisson is pretty good. Also we point out that Inigo uses that and justifies andthis provides further technical support for this |

**Deleted:** supplementary figures. We also added these references and clarified our point by proper acknowledgement. ... [26]

**Deleted:** aspect

**Deleted:** , loses novelty, but we are not claiming r ... [27]

**Moved (insertion) [7]**

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Formatted:** Heading 2, Space Before: 14 pt

**Moved down [9]:** Excerpt From ... [28]

**Formatted:** Font:Times New Roman

**Formatted Table**

**Deleted:**
$$$

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** (little) @@@

**Formatted:** Normal

**Deleted:** &&&compl

**Formatted:** Line spacing: single

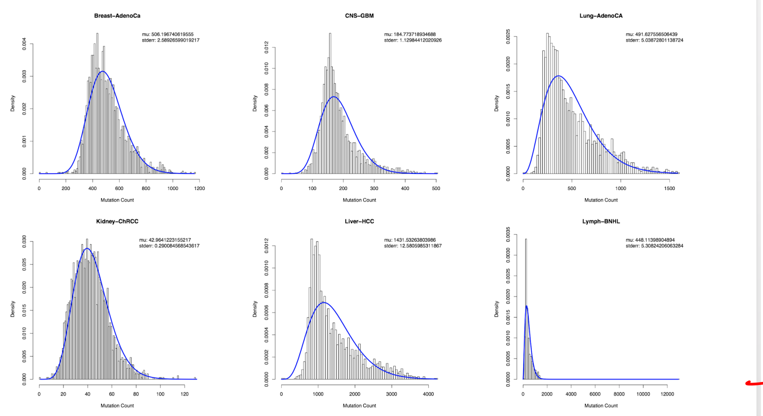
**Formatted Table**

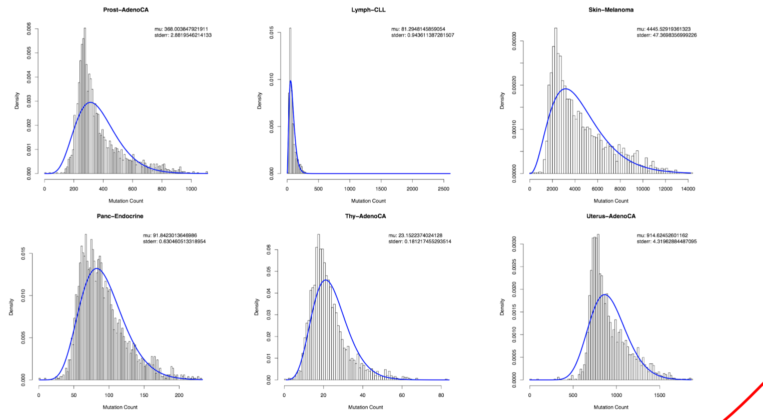
**Comment [1]:** Wait!

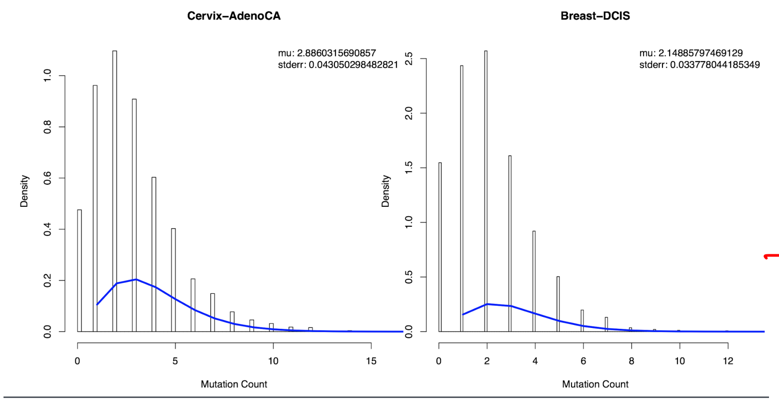
**Formatted:** Line spacing: single

**Deleted:** We have tried to use the Gamma-Poisson model to fit the variant counts per 1mb bins for many cancer types and the fitting are listed as below. We feels for most cancer types that have enough variants, it fits OK with the observed data. However, there might be some case, especially when somatic mutation count is relatively low, fitting is not that good. This is partially why we test the degree of overdispersion before we jump to the negative binomial model. But in our analysis of CLL, BRCA, and LIHC, we feel it is a good model. ... [29]

using. However, we agree that we choose Gamma-Poisson conjust it might interesting to investigate. As the referee this is out of scope but we have made a mention of this in the text.





| | |
|---|---|
| |  |
| Excerpt From Revised Manuscript | |

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

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

| | |
|---|---|
| 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 do feel this is a good point. To answer this question, 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 3mers with higher number of variants, there is a tendency of larger overdispersion. It might be the power |

SUPPL

**Moved down [10]:** Excerpt From ... [31]

**Moved down [10]:** Excerpt From ... [30]

**Formatted:** Font:Times New Roman

**Formatted:** Font:Times New Roman

**Formatted Table**

**Formatted Table**

**Deleted:** -- Ref 2.5

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** $$$Calc (little) @@@

**Deleted:** @@@

**Deleted:** &&&compl

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

**Formatted:** Line spacing: single

**Formatted:** Justified

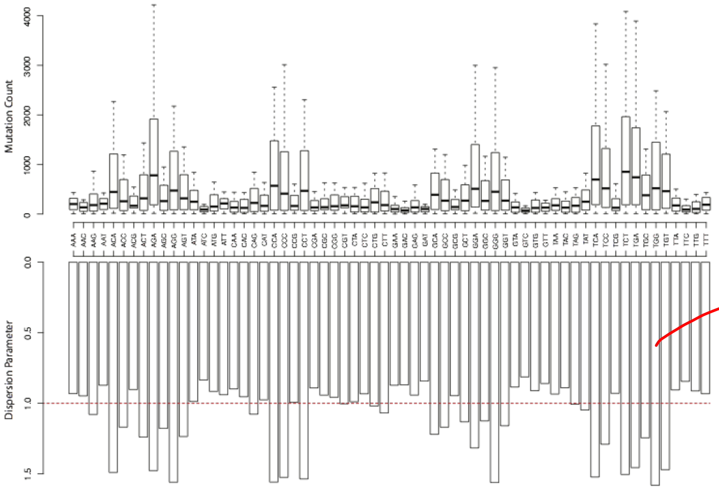
**Deleted:** pointing

**Deleted:** out

**Deleted:** great

**Deleted:** However, we also think it is due to variance-to-mean relationship.

issue, or the level of heterogeneity among samples. A larger variation usually accepts the Negative binomial distribution. We've put more in supplementary file. 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. But we think it is better to check how heterogeneous the count data is even after correcting enough covariate effects.



Excerpt From Revised Manuscript

## <ID>REF2.5 – Cross validation analysis to do model selection

<TYPE>$$$BMR $$$Calc

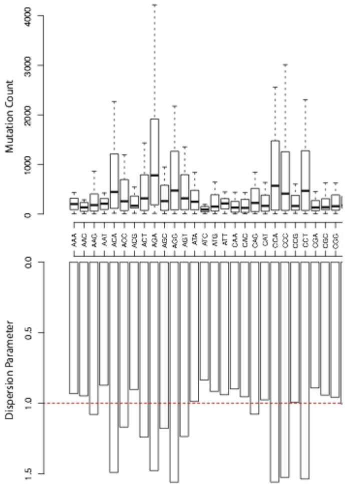
**Deleted:** We admit

**Deleted:** it is possible that with data sets with lower

**Deleted:** of variants Poisson model might be more likely to get rejected. However, it is

**Deleted:** to the heterogeneity of the data. ... [32]

**Deleted:** various



**Deleted:**

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Deleted:** -- Ref 2.6

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** @@@

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

| | |
|---|---|
| Referee Comment | 3) The approach with principal components used for the BMR estimation does not seem to work well. Starting with the second PC most components have roughly the same prediction power. One possibility is that higher principle components do not capture the additional signal and reflect noise in the data, and the correlation with mutation rate is due to an overfit of the NB regression (it is unclear whether it was analyzed with cross-validation). Another possibility is that the signal is spread over many components. In the latter case, this is not an optimal method choice. |
| Author Response | We thank the referee for pointing out the limited contribution from the higher order principal components. In fact, we actually wanted to point this out and we don't see this as efficient. The point of our approach is not to say that a few top components or a few features can predict a mutation rate correctly. Actually we want to show the opposite: that the wealth of the ENCODE data is ~~actually~~ useful and that with additional data types, one gets a small but measurable continued improvement in background mutation estimation. This may be because of the heterogeneity and the difficulty in matching samples, but may due to the correlated nature of the features themselves. We use principal components essentially as a way of doing ~~a principled unbiased feature selection but we realized that actually~~ didn't get across very clearly, so we've redone this figure now simply show how one gets steady increase in predictions forms by just adding features one at a time.<br>We hope this ~~gets the point across. The aim here is to not highlight a complicated mathematical~~ method but just simply to get across the idea that the very large end code data corpus provides a valuable resource for predicting background mutation rate and we appreciated the referee helping us achieve clarity on this point. |
| Excerpt From Revised Manuscript | |

INFORMAL

EDIT

RES IN SUPPL
MOD TEXT

**Deleted:** &&&compl

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

**Formatted:** Line spacing: single

**Deleted:** We thank the referees for understanding/agreeing with our point - a lot of data helps. Actually, PCs are not part of our model - it is just for the demonstration purpose. And we did not use it in our final BMR estimation. In the revised version, we actually used forward selection to show that adding more data will greatly help with the BMR prediction.

**Deleted:** Excerpt From ... [33]

**Formatted:** Line spacing: single

**Moved (insertion) [9]**

**Formatted:** Font:Times New Roman

**Formatted Table**

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

<TYPE>$$$Power $$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%Done

| | |
|---|---|
| Referee Comment | 4) I do not agree with the power analysis presented to support the idea of compact annotations. I understand that this is a toy analysis neglecting specific properties of mutation rate known for regulatory regions and also sequence context dependence of mutation rate. The larger issue is that the analysis assumes that ALL functional sites are within the compact annotation. In that case, power indeed would decrease with length. However, in case some of the functional sites are outside the compact annotation power would not decrease and is even likely to increase with the inclusion of additional sequence. Is there a justification for all functional sites to reside within compact annotations? Can this issue be explored? Some statistical tests incorporate weighting schemes. |
| Author Response | The referee is indeed correct and we expanded our power calculated in our revised manuscript. In our initial submission, we were not trimming truly functional or important sites, but rather trimming unimportant sites. For instance, in the old way that we found enhancer sites by just calling a 1KB region from a peak admittedly by almost any estimation included knots of obviously non functional sites. Trimming this down using a large battery of histone marks and the exact shape of the signal, we believe more accurately gets it the truly functional region, particularly when coupled with accurate STARR-seq and Hi-C data will hopefully increase power. Another case is the TF binding hotspot around the promoter region of WDR74. Instead of testing up to 2.5K promoter region without prior information, we can trim the test set to a core set of the promoter region where many TF binds to, 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). |

FIX

IN MAKING THESE POINTS.

**Deleted:** $$$

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** (from

**Deleted:** presentation) @@@JZ&&&TBC

**Formatted:** Normal

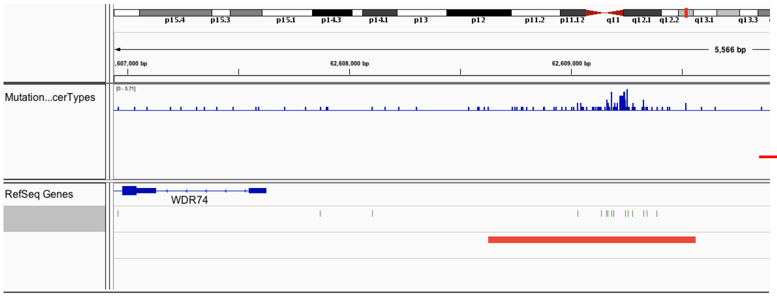
**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Underline

**Formatted:** Line spacing: single

**Deleted:** @@@@7mar - more & be positive - we've thoguth band we ... [34]

| | |
|---|---|
| | <br> p15.4<br>p15.3<br>p15.1<br>p14.3<br>p14.1<br>p13<br>p12<br>p11.2<br>p11.12<br>q11<br>q12.1<br>q12.2<br>q13.1<br>q13.3<br>5,566 bp<br>607,000 bp<br>62,608,000 bp<br>62,609,000 bp<br>Mutation...cerTypes<br>[0 - 5.71]<br>RefSeq Genes<br>WDR74<br><br>Following the reviewer's suggestions, we show in a formal power analysis now in the supplement that the most important contribution to power comes from including additional functional sites, this is of course by the extended gene concept and then 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.<br><br>However, we are admittedly making assumptions and the referee is completely correct in pointing this out. We have tried to be more precise in the text that we're basically assuming that the large number of encode assays when integrated allows us to more directly get at the functionally important nucleotides, but this of course is an assumption. It's hard to really tell to what degree one can success in finding the current events in cancer. It's hard to back this up with the gold standard, but I think that some of the points are self evidently obvious. We've tried to make this clear in text and thank the referee for pointing this out. |
| Excerpt From Revised Manuscript | |

SUPPL

REWRITE

## <ID>REF2.7 – Q-Q plots

<TYPE>$$$BMR $$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&Defer
<STATUS>%%%TBC
####Thinking

**Moved (insertion) [10]**

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Moved down [11]:** Excerpt From ... [35]

**Formatted:** Font:Times New Roman

**Formatted Table**

**Deleted:** -- Ref 2.8

**Deleted:** --

**Deleted:** $$$

**Formatted:** Not Highlight

**Deleted:** $$$

**Formatted:** Not Highlight

**Deleted:** $$$

**Deleted:** @@@JZ&&&TBC

**Formatted:** Normal

**Formatted:** Not Highlight

**Formatted:** Not Highlight

**Deleted:** Referee
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. ... [36]

[JZ2MG: not finished yet for this part]

| | |
|---|---|
| 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 | This is a good point.<br>We've done XXX & YYY now<br>But we wish to make clear that the point of this paper is not driver detection<br>Our goal is BMR<br>We show QQ w diff detection<br>We actually show QQ plots with drivers<br>Take some else's driver detection method, use our BMR model, show that it works better |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF2.8 – Value of the extended gene

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

**Deleted:** -- Ref 2.9 – Novelty

**Deleted:** paper --

**Deleted:** $$$

**Formatted:** Normal

| | |
|---|---|
| Referee Comment | 6) The idea of extended genes and the use of multiple information sources to construct them is a strength of the paper. |
| Author Response | We thank the reviewer for the positive remarks. We further highlighted this part in our revised manuscript and added a whole new section of how the extended gene could increase statistical power. VIZ |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

**Formatted:** Line spacing: single

**Formatted:** Justified

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF2.10 – BMR effect on local tri-nucleotide context

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

DOING2.

| | |
|---|---|
| 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 | 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. We made this point more clear in our revised manuscript. |
| Excerpt From Revised Manuscript | |

## <ID>REF2.11 – Confounding factors

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

| | |
|---|---|
| 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 incorporate such as TF binding, nucleosome occupancy, histone modification (which changes sharply in |

**Deleted:** -- Ref 2
**Deleted:** --
**Deleted:** $$$
**Deleted:** $$$
**Deleted:** @@@
**Deleted:** &&&compl
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Justified
**Formatted:** Line spacing: single
**Formatted:** Justified
**Moved (insertion) [11]**
**Formatted:** Font:Times New Roman
**Formatted:** Line spacing: single
**Deleted:** Excerpt From
@@@@7mar: the ref to point this out ... we actually did have some in the submission... was buried in the supp. we 've tried to make ... [37]
**Deleted:** --
**Deleted:** $$$
**Deleted:** &&&compl
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Justified
**Formatted:** Line spacing: single
**Formatted:** Justified

| | |
|---|---|
| | 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 | |

+ SUPPL

Formatted: Font:Times New Roman

Formatted: Line spacing: single

## <ID>REF2.12 – Power analysis of extended genes

<TYPE>$$$Power,$$$Calc,
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%Done

DOING

[JZ2MG: as discussed we are only supposed to put text here but the real analysis into the supplementary file. However, this could be very inconvenient for the referee since if he wants to check the part, he needs to go to the supp with >100 pages. Please suggest here]

Deleted: -- Ref 2

Deleted: --

Deleted: $$$

Deleted: $$$

Deleted: (

Deleted: presentation) @@@JZ&&&TBC

Formatted: Normal

| | |
|---|---|
| Referee Comment | It would be great to see a formal analysis about how extended genes increase power of cancer driver discovery. |
| Author Response | We thank the referee for this comment and encouraging us to do a formal analysis. We have attempted to do this in suppl figure XXXX. |
| Excerpt From Revised Manuscript | |

Formatted: Line spacing: single

Formatted Table

Formatted: Justified

Formatted: Line spacing: single

Formatted: Justified

Formatted: Font:Times New Roman

Formatted: Line spacing: single

## <ID>REF2.13 – Minor, comment on burden test

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

Deleted: -- Ref 2

Deleted: Comment: Burden --

Deleted: $$$

Deleted: $$$Minor $$$

Deleted: @

Formatted: Normal

| | |
|---|---|
| Referee Comment | 1) I would not use the term "burden test". This usage is slightly confusing because this term is commonly used in human genetics where it refers to a case-control test. |
| Author Response | We thank the referee to point out this. We have changed our terminology in our revised manuscript. |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

**Formatted:** Line spacing: single

**Formatted:** Justified

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF2.14 – Minor comment on terminology

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

**Deleted:** -- Ref 2

**Deleted:** Comment: Terminology --

**Deleted:** $$$

**Deleted:** $$$Minor $$$

**Deleted:** &&&compl

**Formatted:** Normal

| | |
|---|---|
| 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. |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted:** Justified

**Formatted Table**

**Comment [3]:** I don't think this is correct!

**Formatted:** Line spacing: single

**Deleted:** to xxx

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

# Referee #3 (Remarks to the Author):

## <ID>REF3.0 – Preamble

**Deleted:** -- Ref 3

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

In relation to the supplement and genomics, the referee points out that it's 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 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 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. We've tried to revise it to make these clearer and also to move more appointives into the main text, though we think given the current main text limitations of a typical paper nature and the scale of the results in the data in this paper, it's simply 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

**Deleted:** --

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

**Deleted:** $$$

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

| | |
|---|---|
| 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 information. I think more direction, such as sup Fig1, sup Table 1, or section 7.2S etc, would be very helpful. |

| | |
|---|---|
| Author Response | We tried the new way of citing supplementary info. |
| Excerpt From Revised Manuscript | |

Formatted: Line spacing: single

Formatted: Justified

Formatted: Font:Times New Roman

Formatted: Line spacing: single

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

Deleted: -- Ref 3.2 – BMR -- ... [38]

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

Formatted: Normal

| | |
|---|---|
| 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 | |
| Excerpt From Revised Manuscript | |

Formatted: Line spacing: single

Formatted Table

Formatted: Line spacing: single

Formatted: Font:Times New Roman

Formatted: Line spacing: single

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

Deleted: -- Ref 3

Deleted: --

Deleted: $$$

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

Formatted: Normal

| | |
|---|---|
| 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 | |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF3.4 – Regarding enhancer detection algorithm

**Deleted:** -- Ref 3.4 – Untitled -- ... [39]

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

**Formatted:** Normal

| | |
|---|---|
| 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 | |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF3.5 – Regression coefficients of BMR

**Deleted:** -- Ref 3

**Deleted:** Untitled -- ... [40]

<TYPE>$$$BMR
<ASSIGN>
<PLAN>&&&AgreeFix

<STATUS>%%%TBC

Formatted: Normal

| | |
|---|---|
| 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 | |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF3.6 – Validation of extended gene

**Deleted:** -- Ref 3.6 – Untitled -- ... [41]

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

**Formatted:** Normal

| | |
|---|---|
| Referee Comment | For Figure 2C, more explanation is needed on how to form an extended gene. 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? Are there any novel oncogenes detected by the method? |
| Author Response | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF3.7 – Logic gates

**Deleted:** -- Ref 3

**Deleted:** Untitled --

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

**Deleted:** $$$

**Formatted:** Normal

| | |
|---|---|
| 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 | |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF3.8 – Network hierarchy

**Deleted:** -- Ref 3

**Deleted:** Hieratchy --

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

**Deleted:** $$$

**Formatted:** Normal

| | |
|---|---|
| Referee Comment | For Figure 4, what does the star symbol (*) mean in the legend? Did the authors use a different grey color to show |

**Formatted:** Line spacing: single

**Formatted Table**

|  |  |
|---|---|
|  | the connection between TFs? I'm not able to read the grey gradient for the edges. |
| Author Response | We thank referee for point out this issue. We have updated the figure 4 to show the significance testing of network hierarchy analysis. 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 (***). |
| Excerpt From Revised Manuscript |  |

## <ID>REF3.9 – Network rewiring

<TYPE>$$$Network

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

|  |  |
|---|---|
| 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 |  |
| Excerpt From Revised Manuscript |  |

**Formatted:** Line spacing: single

**Moved down [12]:** Excerpt From ... [43]

**Moved down [12]:** Excerpt From ... [42]

**Deleted:** -- Ref 3

**Formatted:** Font:Times New Roman

**Formatted:** Font:Times New Roman

**Formatted Table**

**Formatted Table**

**Formatted:** Font:Italic

**Formatted:** Font:Italic

**Deleted:** Untitled -- ... [44]

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

# Referee #4 (Remarks to the Author):

## <ID>REF4.0 – Preamble

Deleted: -- Ref 4

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

[JZ2MG: do we need a preamble here? I don't feel strongly]
We would like to appreciate the referee's feedback and positive comments about our resource. We found that many of the suggestions, such as further power analysis, stemness & rewiring, comparison of cell line vs tissue, cross validation using primary cancer data, are quite valuable. As suggested, we have significantly expanded them while preserving our original goal in our revised manuscript.

2

## <ID>REF4.1 – Strengths of the Paper

Deleted: --

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

Deleted: $$$
Deleted: &&&compl

Formatted: Normal
Formatted: Line spacing: single
Formatted: Justified
Formatted Table

| | |
|---|---|
| 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: Line spacing: single
Formatted: Justified

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

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

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

**Deleted:** -- Ref 4

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$Text @@@

**Deleted:** @@@DC&&&TBC

**Formatted:** Normal

| | |
|---|---|
| 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 | In relation to the supplement and genomics, the referee points out that it's 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 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 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. We've tried to revise it to make these clearer and also to move more appointives into the main text, though we think given the current main text limitations of a typical paper nature and the scale of the results in the data in this paper, it's simply impossible to put everything into the main text. We are preparing to work constructively with the referees and the others to make this clear. |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

**Formatted:** Line spacing: single

**Deleted:** We thanks the referee for these comments. We've tried to fix the presentation of the Supplement by a) shortening certain sections, and b) improving the overall organization and readability by more clearly labeling sub-sections.

**Deleted:** Excerpt From
*(See Supplement)* ... [45]

**Formatted:** Line spacing: single

| | |
|---|---|
| Excerpt From Revised Manuscript | |

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

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

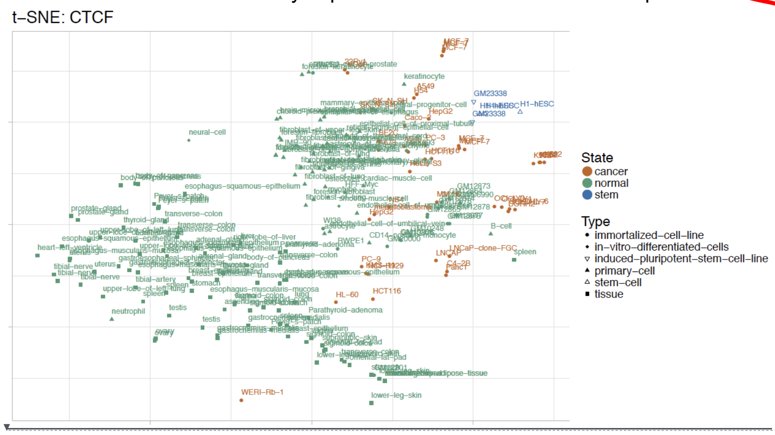
| | |
|---|---|
| Referee Comment | 1) The manuscript is quite complex and efforts are needed to improve clarity. Some of the text can seem to be somehow redundant or not needed (for instance, general comments about the ENCODE project; or the Step-Wise prioritization scheme (page7; other parts at page 7, for instance). |
| Author Response | We thank the referee for his/her suggestions on our presentations. As requested, we've trimmed and edited these sections in our revised manuscript. |
| Excerpt From Revised Manuscript | |

## <ID>REF4.4 –Comparison of tissues to cell lines

<TYPE>$$$CellLine,$$$Validation
<ASSIGN>@@@JZ,@@@DL,@@@Peng
<PLAN>
<STATUS>%%%Done

| | |
|---|---|
| Referee Comment | 2) 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 |

**Moved (insertion) [12]**
**Formatted:** Font:Times New Roman
**Formatted:** Font:Italic
**Formatted Table**
**Deleted:** $$$
**Deleted:** (wait after updated version of manuscript)--
**Deleted:** $$$
**Deleted:** $$$Text $$$Later
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Justified
**Formatted:** Line spacing: single
**Deleted:** &
**Formatted:** Font:Times New Roman
**Formatted:** Line spacing: single
**Formatted:** Normal, Space Before: 0 pt
**Deleted:** ... [46]
**Formatted:** Font color: Black
**Deleted:** &&&More $$$Calc @@@PE @@@
**Formatted:** Font color: Black
**Deleted:** @@@JZ @@@
**Formatted:** Font color: Black
**Deleted:** &&&TBC
**Formatted:** Font color: Black
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Justified

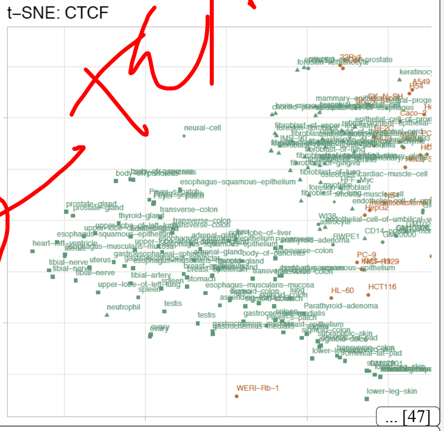
| | |
|---|---|
| | further selected for cell growth very extensively. Many of the cell lines may have/have accumulated further mutation and rearrangements, if compared to what cancer cells are at the moment that they leave the human body. The authors accurately acknowledge, in the discussion, stating 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 (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).<br>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 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 we made a new figure in our revised version. 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 with existing immortalized cell lines with deep annotations. We have chosen CTCF ChIP-seq and RNA-seq, which has the most abundant number of cell types in ENCODE, as an example to highlight this point. We looked at differential binding patterns of CTCF at promoter regions across cell types. The t-SNE plot of CTCF network shows that most of normal cell lines form a cluster together with healthy primary cells, and cancer cell lines can be linearly separable from their normal counterparts.<br><br> |

SEE COMM-ENT

**Formatted:** Line spacing: single

**Deleted:** great

**Deleted:**



... [47]

**Deleted:**

... [48]

####7mar - Thx you for this comment... you are right... we've made we new fig. Bc it in fact does show ...

####7mar - get pe to do this timputed on the leslie data & also some transcriptome analysis

Deleted: @@@@

more

####7mar either for imputed network OR for the transcription, we take the referee's comment to heart & try to do they we .... as the the ref suggested
Take one TF from the imputed network
Ask PE on tumor data ATAC-seq paper

Deleted: @@@@

Try to use some of the imputed stuff on roadmap tissue to show similar results
Let peng to use PE's network, compare results?
To use the imputed network in tissue and used the KD data in cell line as a validation
KD in tissue external data
**** we've really made better use of the encode knockdown data and highlight &&&&& & knockdowns

### PDM references ###
A pathology atlas of the human cancer transcriptome
http://science.sciencemag.org/content/357/6352/eaan2507
"analyses revealed that gene expression of individual tumors within a particular cancer varied considerably and could exceed the variation observed between distinct cancer types." (RNA-seq, Uhlen et al. 2017)

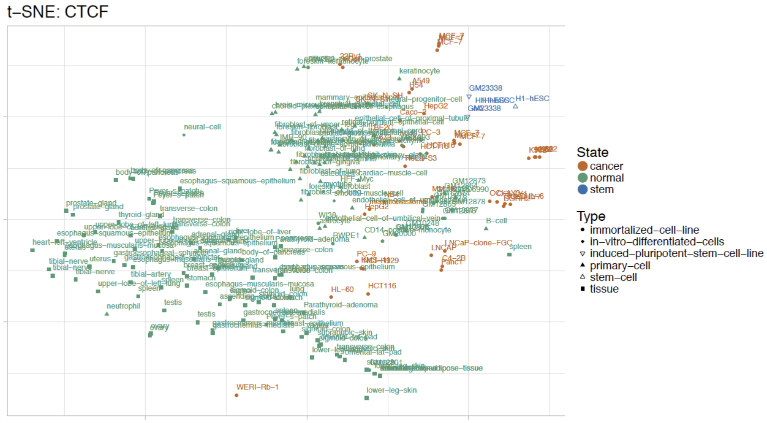
GOOD

Human cancers overexpress genes that are specific to a variety of normal human tissues
http://www.pnas.org/content/102/51/18556
"The results indicate that many genes that are overexpressed in human cancer cells are specific to a variety of normal tissues, including normal tissues other than those from which the cancer originated." (microarray, Lotem et al. 2005)

Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin.
https://www.ncbi.nlm.nih.gov/pubmed/25109877
"Five subtypes were nearly identical to their tissue-of-origin counterparts, but several distinct cancer types were found to converge into common subtypes."
(5 genome-wide platforms, incl. RNA-seq, 1 proteomic platform, Hoadley et al. 2014)
###

(DL maybe)

Formatted: Line spacing: single

| | |
|---|---|
| | |
| Excerpt From Revised Manuscript |  <br><Figure update candidate: CTCF regulatory networks based on all available ENCODE ChIP-seq shows clustering of stem-like state cell types (Blue). All cancer cell lines (Red) were clustered closer to stem-like cell types than normal cell types (Green).> |

## <ID>REF4.5 – Validate the cell line results using tissue data

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

[JZ2MG: almost done, but need to gather figures from multiple persons here]
[JZ2MG: If we have Peng's result, do we need to have PE's imputed network comparison from the Leslie lab?]

| | |
|---|---|
| Referee Comment | It would be appropriate to (computationally) verify at least a small part of the data in other systems, taking from |

**Moved down [13]:** Excerpt From ... [50]

**Moved down [13]:** Excerpt From ... [49]

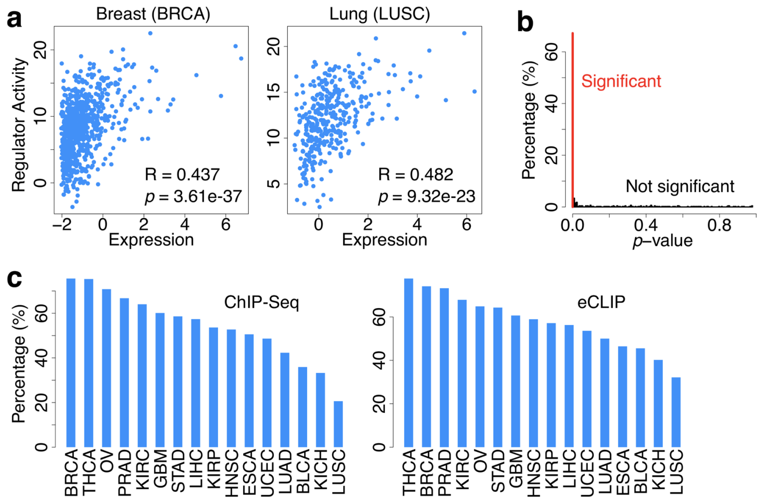
**Formatted:** Font:Times New Roman

**Formatted:** Font:Times New Roman

**Formatted Table**

**Formatted Table**

**Deleted:** -- Ref 4

| | |
|---|---|
| | published studies including normal cells control and primary cancers. |
| Author Response | We thank the referee for this comment and we agree with the reviewer that it is important to verify the human clinical relevance of cell line data. In the revision, 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. For example, we predicted the regulatory activities of transcription factor (TF) MYC using a ChIP-Seq profile in MCF7 cells. 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 MCF7 ChIP-Seq profile, the MYC regulatory activity predicted for lung tumors is also significantly correlated with MYC expression level in TCGA lung cancer (Supplementary Figure Xa). These results indicate that the ChIP-Seq profiles from a particular cell line can capture n 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). |
| Excerpt From Revised Manuscript | <br> <br>**Supplementary Figure X. The clinical relevance of ENCODE cell line data in human primary tumors.** |

2

SUPP

?

| | |
|---|---|
| | **(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 MCF7 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.<br>**(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.<br>The fraction of regulators with statistically significant correlations in different cancer types for ChIP-Seq and eCLIP networks. In each TCGA cancer type, we computed the correlations between regulator expression levels and regulatory activities across tumors for all regulators (TFs, or RBPs). We selected regulators with statistically significant correlations through two-sided student t test (FDR < 0.05). |

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

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

| | |
|---|---|
| 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 | > PE's imputed network stuff<br>> histones DHS<br>&&&&&& explicit imputed network<br>Expand the resource - |

**Deleted:** --

**Deleted:** $$$

**Deleted:**

**Deleted:** &&&More @@@

**Formatted:** Font color: Black

**Formatted:** Font color: Black

**Deleted:** @@@DL&&&

**Formatted:** Font color: Black

**Formatted:** Font color: Black

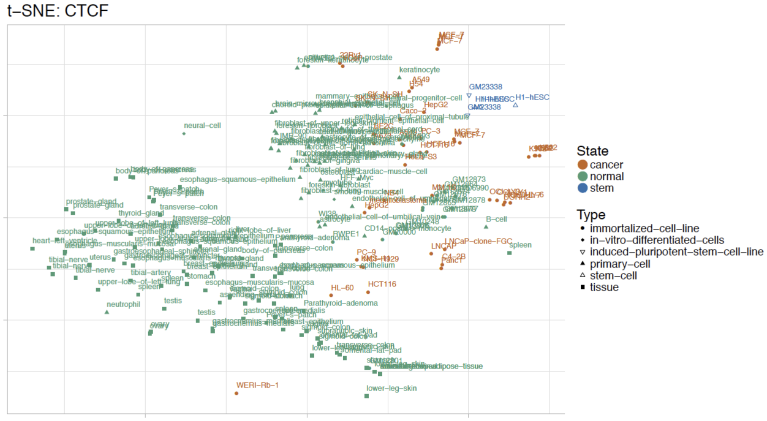
**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

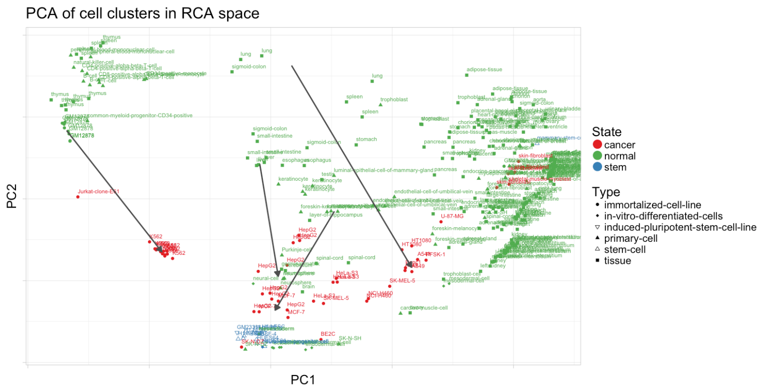
**Formatted:** Line spacing: single

| | |
|---|---|
| | === <br>We thank the referees for bringing this point out and we have done what they suggested. 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. 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 . Another analysis we added was to look at gene expression profiles of all available ENCODE cell types. In agreement with the previous analysis, gene expression profiles of stem-like cell types were very similar to each other and formed a cluster when projected onto 2D RCA space. |
| Excerpt From Revised Manuscript |  <br>&lt;Figure update candidate: CTCF regulatory networks based on all available ENCODE ChIP-seq shows clustering of stem-like state cell types (Blue). Promoter network of CTCF was projected onto 2D space using t-SNE. All cancer cell lines (Red) were clustered closer to stem-like cell types than normal cell types (Green).&gt; |

**Formatted:** Justified

**Deleted:** ... [51]

**Deleted:** However, we

| | |
|---|---|
| | <br> <br><Figure update candidate: Gene expression profiles of all available ENCODE RNA-seq experiments show that all stem-like cell types form a cluster (Blue). Gene expression quantifications were projected onto 2D space using reference component analysis.> |

## <ID>REF4.7 – Fixes for Figure 1

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

| | |
|---|---|
| Referee Comment | 4) I have difficulties to fully understand Fig.1, in particular the patient cohort (PC) at the bottom of the "depth approach" (just above the green box of cell -specific analysis). The two rows are at the bottom of the columns report mutation and expression, but they belong to the columns of the cell lines (K562, HepG2, etc). I just simply do not understand that part of the figure, in particular the relation between cell lines and the patient cohort (the figure legend does not help, and also supplementary material did not help). |
| Author Response | DL - think about how we can change the figure<br>(We fixed the figure. Less data, more on overview schematic) |

**Moved down [14]:** Excerpt From ... [52]
**Moved down [14]:** Excerpt From ... [53]
**Deleted:** -- Ref 4.6
**Formatted:** Font:Times New Roman
**Formatted:** Font:Times New Roman
**Formatted Table**
**Formatted Table**
**Deleted:** --
**Deleted:** $$$
**Deleted:** $$$
**Deleted:** @@@
**Deleted:** &&&
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted:** Justified
**Formatted Table**
**Formatted:** Line spacing: single
**Formatted:** Font:Helvetica Neue
**Deleted:**
**Formatted:** Font:Helvetica Neue

|  |  |
|---|---|
|  | We thank referee for the suggestion. In the revision we have extensively revised the figure 1. We understand that numbers at the mutation and expression rows can be misleading, so we have separated cohort-based data matrix out of cell-type data matrix. In addition, more emphasis was put into the overview schematic to highlight the value of ENCODEC as a resource. |
| Excerpt From Revised Manuscript |  |

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

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

|  |  |
|---|---|
| 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 extremely good suggestions and we wished we had taken that more in this mission.<br>REWRITE<br>In the revision, we're definitely taking this comments to heart and have added in main text figures that look at the degree to which structural variants, or SVs, mature background mutational rate, and they also affected a network wiring. 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 |

**Deleted:** ... [54]

**Moved (insertion) [13]**

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Moved down [15]:** Excerpt From ... [55]

**Formatted:** Font:Times New Roman

**Formatted Table**

**Deleted:**
$$$

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$NETWORK $$$

**Deleted:** &&&More @@@

**Deleted:** (rewire) @@@

**Deleted:** +

**Deleted:** (expression & elements vs SV) @@@

**Deleted:** (mechanism) &&&

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

**Formatted:** Line spacing: single

**Deleted:** &&&& SVs ... [56]

|  |  |
|---|---|
|  | of new mains figures that address this and we quite heartly thank the referee for pointing this out. |
| Excerpt From Revised Manuscript |  |

Deleted: Excerpt From ... [57]

Formatted: Line spacing: single

Moved (insertion) [14]

Formatted: Font:Times New Roman

Formatted Table

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

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

IN 4, 4 REF

USE AGAINST MATCH IN #1 WD.

|  |  |
|---|---|
| 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 | ###JZ: strength of cell line, no heterogeneity, emphasize this, co-expression network<br>### Can mention something related to single cells<br>### Some clinically significant changes will occur in<br><br>####7mar - high level is how to connect the reference cell line to annotation to patient .... key pt of the paper ... peng's figure<br>Individualize the network a little bit<br><br>###WUM text###<br>The referee is correct that tissue heterogeneity represents a source of complexity not directly modeled in our resource, a limitation which we now discuss with greater emphasis. |

Deleted: $$$

Deleted: liens --

Deleted: $$$

Deleted: $$$

Deleted: @@@

Deleted: @@@

Deleted: @@@

Formatted: Not Highlight

Deleted: @@@PDM&&&compl

Formatted: Normal

Formatted: Line spacing: single

Formatted Table

Formatted: Justified

Formatted: Line spacing: single

Deleted: @@@@

Deleted: Nonetheless, some of our analyses *indirectly* model tissue heterogeneity, and some other of our analyses that do not model tissue heterogeneity should be particularly robust to this potential source of variance.

Nonetheless, some of our analyses are should be particularly robust to the presence and activities of stromal and infiltrating cells. For example, our BMR calculations should not largely be affected by stromal tissue epigenetics, because clonally-amplified mutations detected by bulk sequencing will tend to accrue to a much greater extent in cells descendant from the cell-of-origin of the cancer cell much more so than associated normal tissue.

More generally, in the coming years, we might be able to better model this complexity making use of new single-cell epigenetic data, which is just beginning to emerge. https://www.nature.com/articles/s41467-018-03149-4
Another possibility for future improvements that we mention in our updated discussion section is the potential to model regulatory networks and the BMR separately for each major subclone present in a patient cancer sample, whose differential mutations can be approximately inferred using existing computational tools.
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003665

###PDM text###
As the reviewer correctly states, 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. Nonetheless, we feel there remains value in single-cell comparisons between tumor and normal cells.

Among the strengths of cell-line comparisons is the ability to perform well-controlled analyses of cancel cell function in a way that is not possible with whole tumor specimens. For example, the detailed gene co-expression network analyses we highlight in our manuscript (see section XXX), were made possible by a homogenous cancer-cell population with robust and uniform expression signal. Such an analysis in whole-tumor specimens would be challenging due to the need for deconvolution of expression signals originating from various cell types present in tumors.

Apart from the advantage of single-cell analyses of enabling examination of complex cancer cell biology, there is, moreover, reason to believe that single-cell analyses may capture important tumor biology present *in vivo*. Cancers that result from a single progenitor cell, or homogenous progenitor population, provide a justification for the use of single-cell analyses and comparisons. There is evidence that a number of cancers may develop according to the cancer stem-cell model, which posits that it is only a small population of stem-like cells that are responsible for tumor development and observed intratumoral heterogeneity

NEED TO WRITE

**Deleted:** One way in which we indirectly model tissue heterogeneity is by incorporating the patient's tumor's transcriptome when constructing patient-specific regulatory networks. Paracrine signalling by stromal tissue can trigger a signalling cascade that results in altered TF expression and therefore potentially global gene regulation in a patient sample. We empirically take such consequences into account by adding or removing regulatory network edges from patient-specific regulatory networks based on patient-specific TF expression levels, which implicitly takes into account the role

**Deleted:** normal cell signalling on those TF levels in cancer cells.
In

**Deleted:** is

**Deleted:** BMR calculations.

**Deleted:** In

**Comment [4]:** Is this what we want to say?

Could be useful to say something of the format -- "XYZ analysis would not be possible, except by using a cell-line". Not sure this is the best example.

**Comment [5]:** Might also be useful to say, even with 3-D data, from multiple cell types (e.g., stromal cells, immune cells, etc.), some of analyses, like BMR prediction, may not change much -- see WUM text.

| | |
|---|---|
| | (PMID: 24607403). Understanding the biology of a single cells in the progenitor population may be sufficient to gain perspective on the tumor landscape as a whole.<br>Even when there is genomic heterogeneity observed across tumor clones and subclones, the main driver mutations and phenotypic traits may be widely shared among cells (PMID: 3944607, 21376230). For example, in a single-cell sequencing analysis of colon cancer, the primary drivers TP53 and APC were present in the majority of cells across clones, with other mutations showing greater heterogeneity. (PMID: 24699064) Furthermore, even when there is substantial initial genomic and phenotypic heterogeneity, tumors may tend to converge to a genomic and phenotypic equilibrium (e.g, to a stem-like state) as has been shown in a number of studies on breast cancer tumor evolution (PMID: 21854987, 21498687, 22472879). |
| Excerpt From Revised Manuscript | |

## <ID>REF4.10 – lncRNAs and BMR

<TYPE>$$$BMR $$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%Done

| | |
|---|---|
| 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. We have added the analysis of lncRNA by comparing BMRs in genes and lncRNAs. |

2:

2IDW

**Deleted:** Excerpt From ... [58]

**Formatted:** Line spacing: single

**Moved (insertion) [15]**

**Formatted:** Font:Times New Roman

**Formatted Table**

**Formatted:** Heading 2, Space Before: 14 pt

**Deleted:** $$$

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** @@@

**Formatted:** Normal

**Deleted:** &&&TBC

**Deleted:** Referee
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? ... [59]

**Formatted:** Line spacing: single

**Formatted Table**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

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

**Deleted:** -- Ref 4.10

**Deleted:** suppl. --

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

**Deleted:** $$$

**Deleted:** $$$Minor &&&TBC

**Formatted:** Normal

| | |
|---|---|
| Referee Comment | In the supplementary material, there is room to improve figures (some numbers are too small). |
| Author Response | We thank the referee to point out this and we have fixed in our revised manuscript |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Justified

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF4.12 – (Minor) Figure legends

**Deleted:** -- Ref 4.11

**Deleted:** --

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

**Deleted:** $$$

**Deleted:** $$$Minor&&&TBC

**Formatted:** Normal

| | |
|---|---|
| Referee Comment | Figure legends. Figure legends are essential but I struggled to understand the figures based on the legends only. |
| Author | We thank the referee to point out this and we have fixed in our revised manuscript |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

**Formatted:** Line spacing: single

**Formatted:** Justified

|  |  |
|---|---|
| Response |  |
| Excerpt From Revised Manuscript |  |

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing:  single

# Referee #5 (Remarks to the Author):

## <ID>REF5.0 – Preamble

**Deleted:** -- Ref 5

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

We would like to appreciate the referee's feedback. We found that many of the suggestions, such as further power analysis, false positive rate of rewiring, comparison with other networks, cross validation using external data, are quite valuable and we expanded them in our revised manuscript as suggested. The referee mentioned that but the novelty of the paper is lacking. We also thank the referee to point out his/her confusion about whether this is prospective or biology paper. We want to make it clear that his paper is to be considered as a "resource" paper, not a novel biology 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.

| **Contribution** | **Subtypes** | **Data types** | **ENCODE experiments** |
|---|---|---|---|
| Processed raw signal tracks | Histone modification | Signal matrix in TSV format | 2015 Histone ChIP-seq |
| | DNase I hypersensitive site (DHS) | Signal matrix in TSV format | 564 DNase-seq |
| | Replication timing (RT) | Signal matrix in TSV format | 135 Repli-seq and Repli-ChIP |
| | TF hotspots | Signal track in bigWig format | 1863 TF ChIP-seq |
| Processed quantification matrix | Gene expression quantification | FPKM matrix in TSV format | 329 RNA-seq |
| | TF/RBP knockdowns and knockouts | FPKM matrix in TSV format | 661 RNAi KD + CRISPR-based KO |
| Integrative annotation | Enhancer | Annotation in BED format | 2015 Histone ChIP-seq<br>564 DNase-seq<br>STARR-seq |
| | Enhancer-gene linkage | Annotation in BED format | 2015 Histone ChIP-seq<br>329 RNA-seq |

| | | | |
|---|---|---|---|
| | Extended gene | Annotation in BED format | 1863 TF ChIP-seq<br>167 eCLIP<br>Enhancer-gene linkage |
| SV and SNV callsets | Cancer cell lines | Variants in VCF format | WGS<br>BioNano<br>Hi-C<br>Repli-seq |
| Network | RBP proximal network | Network in TSV format | 167 eCLIP |
| | Universal TF-gene proximal network | Network in TSV format | 1863 TF ChIP-seq |
| | Tissue-specific TF-gene proximal network | Network in TSV format | 1863 TF ChIP-seq |
| | Tissue-specific imputed TF-gene proximal network | Network in TSV format | 564 DNase-seq |
| | TF-enhancer-gene network level 1-3 | Network in TSV format | 2015 Histone ChIP-seq<br>564 DNase-seq |

Specifically for the BMR estimation part, the reviewer mentioned that there have been many existing references focusing on applications like cancer driver detection. First, we thank the referee for pointing out to a lot of related references. On the reference side, we have listed many of the papers as the referee suggested and compared them with our approach. We have acknowledged the efforts of many of these references and in the revised version we have further expanded our reference list for some the publications after our initial submission date. We want to emphasize that the richness of the ENCODE data can actually help many of the methods used in these papers. With a larger pool of covariate selection, the estimation accuracy can be significantly improved.

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

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

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

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

**Deleted:** --

**Deleted:** $$$NoveltyPos

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

| | |
|---|---|
| Author Response | We thank the referee for the positive comment. |
| Excerpt From Revised Manuscript | |

## <ID>REF5.2 – BMR

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

HOW MANY IMPROVES

| | |
|---|---|
| Referee Comment | 1. The manuscript does not clearly state innovation and novelty over previously published data and methods. Several published studies have used epigenomic data types, including replication time and histone modifications from ENCODE and other sources, to model background mutational background density and define genomic elements of interest. The use of the Negative Binomial/gamma-Poisson distributions to model mutational background in cancer has also been published (Imielinski et al 2016; Martincorena et al, 2017). |
| Author Response | We thank the reviewer for identifying these references. We recognize that epigenetic features have been previously been used to estimate BMR and improve driver mutation detection. Our aim was not to produce novel BMR estimation models, but rather to showcase how ENCODE data can help improve the performance of such models.<br>With the wealth data available through ENCODE data, we had a much larger pool of features to choose from to potentially improve BMR estimation. It is worth to mention that ENCODE data is not just cell line data, in fact XXX of this histone modification data is actually from real tissues. Indeed, we found that application of some additional features from the this expansive set, especially the replication |

**Formatted:** Line spacing: single

**Formatted:** Justified

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Formatted:** Heading 2, Space Before: 14 pt

**Deleted:** -- Ref 5.2 – Untitled @@@@7mar: change title--
it is not clear what are the main findings in the paper and their statistical and biological significance. The manuscript seems to be somewhat confused between a perspective piece or a guide to ENCODE data for the cancer community (which should be published in a more specialized journal), and a genomics study with clear findings. ... [60]

**Moved down [16]:** Excerpt From ... [61]

**Formatted:** Font:Times New Roman

**Formatted:** Normal

**Formatted Table**

**Deleted:** -- Ref 5.3 – Novelty of the paper --
As it is, the manuscript falls short of the novelty characteristic of publications in Nature. The main concepts presented in this manuscript have been explored extensively before; albeit not with the same amount of ENCODE data specifically (e.g. Martincorena et al (2017); Lawrence et al (2013); Polak et al (2015); Imielinski (2017); Roadmap Epigenomics). The cancer genome community has been using ENCODE and Roadmap data in various ways, including in papers such as Tomokova et al. (2017), Schuster-Böckler and Lehner (2012), Frigola et al. (2017), Sabarinathan et al. (2016), Morganella et al. (2016), Supek and Lehner (2015). There is no clear comparison to prior work and no demonstration of improved results compared to those in the literature. ... [62]

**Formatted:** Not Highlight

**Formatted:** Line spacing: single

**Formatted Table**

**Deleted:** Similar to comment to referee 2

**Formatted:** Line spacing: single

| | |
|---|---|
| | timing data, significantly improved BMR estimation in many cancer types (see Supplement Section S7).<br>For example, many prior efforts to model BMR have been limited by the availability of genomic assays, or by the availability of assays matched by cell-type. For example, Lawrence et al., 2013, used HeLa replication timing data and K562 chromatin state via Hi-C. Martincorena et al., 2017, included histone modification features, but not replication timing. The genomic signals we used from ENCODE have been processed uniformly and are provided in a ready-to-use format for the community.<br>We do not intend to claim it is a new discovery that using matched features are better, but rather to show that the breadth of ENCODE data allows for improved estimates of background mutation rate. We have further acknowledged prior efforts on this topic in our revised manuscript. |
| Excerpt From Revised Manuscript | |

?

## <ID>REF5.3 – TCGA benchmark on the gene level

<TYPE>$$$BMR$$$Calc
<ASSIGN>@@@JZ,@@@WM
<PLAN>%%%MORE
<STATUS>%%%TBC
[JZ2WM: can you please help to paste your stuff here?]

?

| | |
|---|---|
| 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're part of pcawg ... there's no benchmark,<br>There's a driver comparison but this is different |

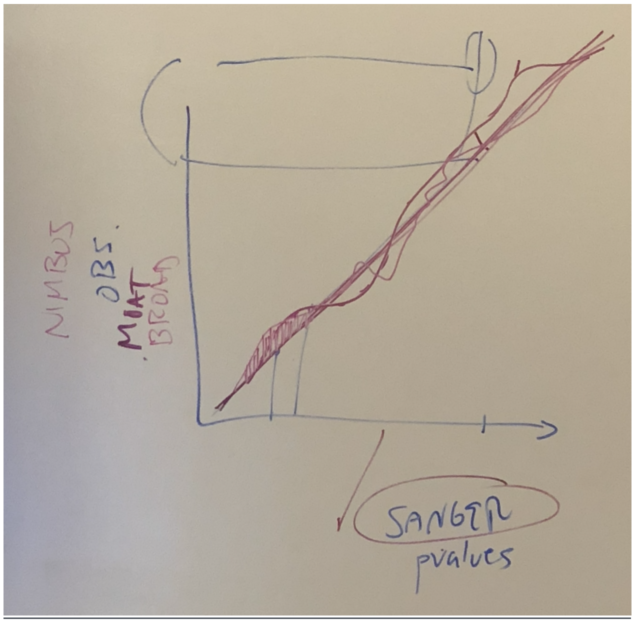
**Moved down [17]:** Excerpt From ... [64]
**Moved down [17]:** Excerpt From ... [63]
**Deleted:** -- Ref 5.5
**Formatted:** Font:Times New Roman
**Formatted:** Font:Times New Roman
**Formatted Table**
**Formatted Table**
**Deleted:** --
**Deleted:** $$$
**Deleted:**
**Deleted:** $$$Later @@@
**Deleted:** @@@JZ (hard comparison) &&&
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted:** Justified
**Formatted Table**
**Formatted:** Line spacing: single
**Deleted:** Non driver TCGA gene (remove cancer ... [65]

Best we find is tcga pancan but this is genes
We tried this we got...

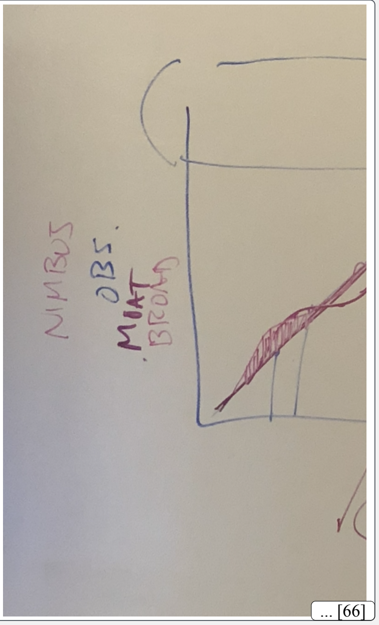
####7mar - WM & esther // running est. program on our data set // could use the sanger randomized or the broad model to compare against nimbus but not do a q-q for driver detection

Deleted: @@@@

WM 3/13: [Esther can’t help us - MutSigNC doesn’t store, allegedly, the BMRs, only the p-values. New idea: Derive implicit BMR from PCAWG Sanger sims using downsampling. For each patient in (a subset of) PCAWG We will probably win since Sanger overfits]



####7mar - compare the sanger rand v us (nimbus) in a qq

Deleted:


... [66]

Excerpt From Revised Manuscript

Formatted: Font:Times New Roman

Formatted: Line spacing:  single

## <ID>REF5.4 – Improvements of the BMR

<TYPE>$$$BMR$$$Calc
<ASSIGN>@@@JZ@@@WM
<PLAN>%%%MORE
<STATUS>%%%TBC

[JZ2MG: more discussion next week. To say BMR more accurate is OK, but to say we are more sensitive in driver detection is not OK. Not sure it is OK to say this is out of scope]

| | |
|---|---|
| Referee Comment | 3. 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 | Part of the previous<br>####7mar:fight-outofscope<br>####7mar - comparisons w/ other methods |
| Excerpt From Revised Manuscript | |

## <ID>REF5.6 – Power analysis

<TYPE>$$$BMR$$$Calc
<ASSIGN>@@@JZ
<PLAN>%%%MORE
<STATUS>%%%TBC

[JZ2MG: seems that this referee need to see results not just math equations]

| | |
|---|---|
| Referee Comment | 4. The power considerations for selecting genomic elements are valuable. Again, sensitivity/specificity analyses of driver discovery with large sets, or long vs. reduced element size need to be added. Prior efforts to address this problem with restricted hypothesis testing for cancer |

**Deleted:** -- Ref 5.6 – Addressing improvements
**Deleted:** --
**Deleted:** $$$
**Formatted:** Not Highlight
**Deleted:**
**Formatted:** Not Highlight
**Deleted:** @@@
**Formatted:** Not Highlight
**Deleted:** &&&
**Formatted:** Normal
**Formatted:** Not Highlight
**Formatted:** Line spacing: single
**Formatted Table**
**Comment [9]:** Wait
**Formatted:** Line spacing: single
**Deleted:** @@@@
**Deleted:** @@@@
**Deleted:** Excerpt From ... [67]
**Formatted:** Line spacing: single
**Moved (insertion) [16]**
**Formatted:** Font:Times New Roman
**Formatted Table**
**Formatted:** Normal, Space Before: 0 pt
**Deleted:** $$$Power
**Deleted:** --
**Deleted:** $$$Power
**Formatted:** Not Highlight
**Deleted:** @@@
**Formatted:** Not Highlight
**Deleted:** &&&
**Formatted:** Not Highlight
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**

| | |
|---|---|
| | genes should be cited (Lawrence et al, 2014; Martincorena, 2017). |
| Author Response | |
| Excerpt From Revised Manuscript | |

END LETTER

IS THIS APPLIC. UECITE.

## <ID>REF5.7 – Comparing power analysis to other work

<TYPE>$$$Power$$$Text
<ASSIGN>@@@JZ
<PLAN>%%%MORE
<STATUS>%%%TBC

| | |
|---|---|
| Referee Comment | 5. "Increased" power of the combined strategy is suggested in the manuscript, yet comparison to prior work is missing. |
| Author Response | Following the reviewer's suggestions, we show in a formal power analysis now in the supplement that the most important contribution to power comes from including additional functional sites, this is of course by the extended gene concept and then 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.<br>However, we are admittedly making assumptions and the referee is completely correct in pointing this out. We have tried to be more precise in the text that we're basically assuming that the large number of encode assays when integrated allows us to more directly get at the functionally important nucleotides, but this of course is an assumption. It's hard to really tell to what degree one can success in finding the current events in cancer. It's hard to back this up with the gold standard, but I think that some of the points are self evidently obvious. We've tried to make this clear in text and thank the referee for pointing this out. |

DOESN'T ADDR

**Deleted:** JZ's presentation ... [68]

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Deleted:** -- Ref 5.8

**Deleted:** --

**Deleted:** $$$

**Deleted:**

**Deleted:** @@@

**Deleted:** &&&

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Deleted:** We thank for the referee to point this out. In our revised manuscript, we have added a whole new section in the supplementary file to discuss this problem. In summary, previous power calculations was based on the assumption that all functional sites are within the test region, hence it is better to have short and accurate annotations. However, we found that this assumption is pretty strong and is not realistic for some cases. ... [69]

|  |  |
|---|---|
| Excerpt From Revised Manuscript |  |

## <ID>REF5.8 – Calculation of power

<TYPE>$$$Power$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%Done

MTG ?

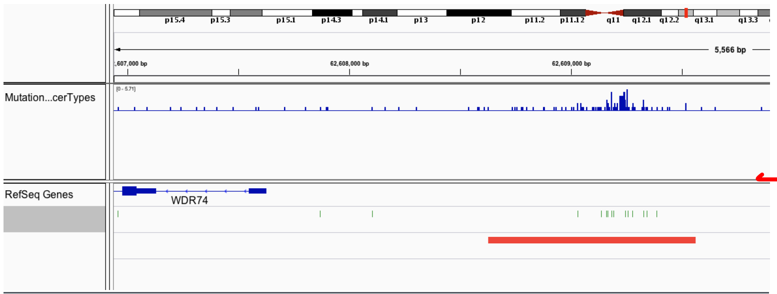
|  |  |
|---|---|
| Referee Comment | 6. The authors claim that reduction of functional elements increases power to discover recurrently mutated elements. This point needs quantitative support in the main manuscript (some analysis is given in the supplemental). For example, in the enhancer list derived from the ensemble method, what fraction of enhancers are estimated to be false positives? |
| Author Response | We thank the referee for pointing out the importance of power calculations. As suggested we have added more in both main manuscript and supplementary file. In our initial submission, we were not trimming truly functional or important sites, but rather trimming unimportant sites. For instance, in the old way that we found enhancer sites by just calling a 1KB region from a peak admittedly by almost any estimation included knots of obviously non functional sites. Trimming this down using a large battery of histone marks and the exact shape of the signal, we believe more accurately gets it the truly functional region, particularly when coupled with accurate STARR-seq and Hi-C data will hopefully increase power. Another case is the TF binding hotspot around the promoter region of WDR74. Instead of testing up to 2.5K promoter region without prior information, we can trim the test set to a core set of the promoter region where many TF binds to, 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). |

**Moved (insertion) [17]**
**Formatted:** Font:Times New Roman
**Formatted:** Line spacing: single

**Moved down [18]:** Excerpt From ... [70]
**Formatted:** Font:Times New Roman
**Formatted Table**
**Deleted:** -- Ref 5.9
**Deleted:** --
**Deleted:** $$$Annotation $$$Calc @@@JZ &&&TBC

**Formatted:** Line spacing: single
**Formatted Table**

**Formatted:** Line spacing: single
**Deleted:** (JZ's presentation) ... [71]

| | |
|---|---|
| |  p15.4<br>p15.3<br>p15.1<br>p14.3<br>p14.1<br>p13<br>p12<br>p11.2<br>p11.12<br>q11<br>q12.1<br>q12.2<br>q13.1<br>q13.3<br>5,566 bp<br>,607,000 bp<br>62,608,000 bp<br>62,609,000 bp<br>Mutation...cerTypes<br>[0 - 5.71]<br>RefSeq Genes<br>WDR74 |
| Excerpt From Revised Manuscript | |

SUPPL.

## <ID>REF5.9 – Assessing quality of enhancer gene linkage annotation

<TYPE>$$$Annotation$$$Text
<ASSIGN>@@@KevinYip @@@SKL
<PLAN>%%%MORE
<STATUS>%%%TBC
[JZ2MG: next week will check the status of KevinYip, SKL stuff added]
[JZ2XK: can you please update this figure and check this text?]

| | |
|---|---|
| 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 by data supporting its accuracy and better performance compared to existing approaches. |
| Author Response | We thank the referee for the comments. We have done as suggested. We have added a few sentences to the main text better discuss the enhancer definition and gene linkage prediction. We have created suppl. Section XXX that shows the performance of JEME + Hi-C. |

**Moved down [19]:** Excerpt From ... [72]

**Moved down [19]:** Excerpt From ... [73]

**Deleted:** -- Ref 5.10

**Formatted:** Font:Times New Roman

**Formatted:** Font:Times New Roman

**Formatted Table**

**Formatted Table**

**Deleted:** --

**Deleted:** $$$

**Deleted:**

**Deleted:** @@@

**Deleted:** @@@

**Deleted:** &&&

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Justified

**Deleted:** :

**Deleted:** desc. The methods

**Deleted:** we

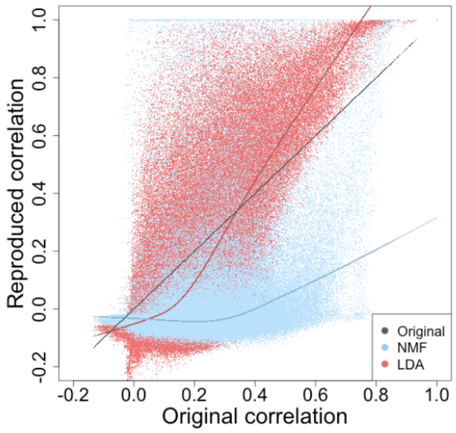
Also we have compared the gene community model with other methods. Mix membership model is a hierarchical Bayesian topic model framework and can help to uncover the underlining semantic structure of a document collection. The core of topic models is Latent Dirichlet Allocation(LDA), which cast the mixed-membership (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.

Deleted: @@@@7mar - DC write to KY ^& CC everyone

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. 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.

surp

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) are used for comparison because the nature of regulatory network requires a non-negative decomposition. The same target dimension K =50 are used. As shown in the figure, the x-axis is original correlation of two TF regulatory target, y-axis is reproduced correlation from LDA document to topic distribution and NMF decomposed matrix. The solid line is the 'loess' smoothing curve for the scattered dots.  We can see the LDA method can reproduce the original correlation better than the NMF.

| | |
|---|---|
| Excerpt From Revised Manuscript |  |

## <ID>REF5.10 – What data sets are used

<TYPE>$$$BMR
<ASSIGN>@@@JZ
<PLAN>&&&Defer
<STATUS>%%%Done
[JZ2MG: to disc next week can we use other public data?]

| | |
|---|---|
| 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. |
| Excerpt From Revised Manuscript | |

**Moved down [20]:** Excerpt From ... [75]
**Moved down [20]:** Excerpt From ... [74]
**Formatted:** Font:Times New Roman
**Formatted:** Font:Times New Roman
**Formatted Table**
**Formatted Table**
**Deleted:** -- Ref 5.11
**Deleted:** --
**Deleted:** $$$
**Deleted:** $$$Punt @@@
**Deleted:** &&&TBC
**Formatted:** Font:Bold
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**

**Formatted:** Line spacing: single
**Deleted:** JZ: ... [76]
**Deleted:** you, we
**Deleted:** & in summzraized
**Deleted:** maintext
**Moved (insertion) [18]**
**Formatted:** Font:Times New Roman
**Formatted:** Line spacing: single

**Moved down [21]:** Excerpt From ... [77]
**Formatted:** Font:Times New Roman
**Formatted Table**

## <ID>REF5.11 – Mutational signatures

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

| | |
|---|---|
| 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 revised manuscript. |
| Excerpt From Revised Manuscript | |

**Deleted:** -- Ref 5.12 – Signature & Mut. rate -- ... [78]

**Deleted:** $$$

**Deleted:** @@@

**Formatted:** Normal

**Deleted:** &&&compl

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Justified

**Deleted:** @@@@7mar - this is a good point. We try to discuss this in the disc. Not however, that no one yet takes this into account but this is the direction to go

**Deleted:** Excerpt From ... [79]

**Formatted:** Line spacing: single

**Moved (insertion) [19]**

**Formatted:** Font:Times New Roman

**Formatted Table**

## <ID>REF5.12 – Additional QQ plots

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

| | |
|---|---|
| Referee Comment | 10. The significance analysis of cancer cohorts (Figure 2) should highlight known cancer genes versus those newly found in this study. 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. Yes, we have provided the QQ plot in the supplementary file in our initial submission. |

**Deleted:**
$$$

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** &&&compl

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Justified

**Formatted:** Line spacing: single

**Formatted:** Justified

**Deleted:** @@@@7mar - @@@@praise

| | |
|---|---|
| Excerpt From Revised Manuscript | |

## <ID>REF5.13 – Sequence coverage

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

| | |
|---|---|
| Referee Comment | Do the authors include sequence coverage in their method? |
| Author Response | Thanks for pointing this out. We did not consider coverage but this is a good point. We included in the discussion in our revised manuscript. |
| Excerpt From Revised Manuscript | |

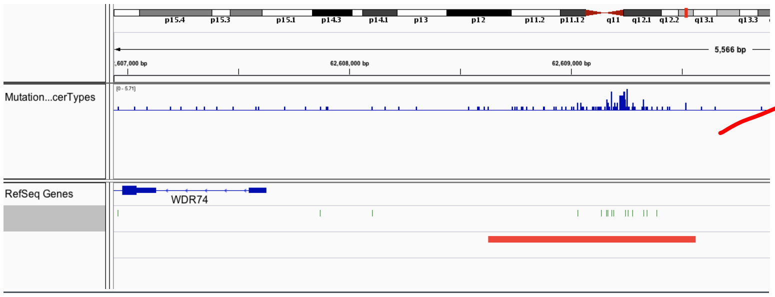
## <ID>REF5.14 – Power analysis for compact annotation

<TYPE>$$$Power $$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%Done
[JZ2MG: feel the three power related questions can be combined]

| | |
|---|---|
| Referee Comment | How do the new “compact annotations” lead to improved results over traditional annotations? |
| Author Response | We thank the referee for pointing this out. We have made it more clear in our supplementary file. When all the functional sites are within the test region, a shorter |

**Formatted:** Font:Times New Roman
**Formatted:** Line spacing: single
**Deleted:** -- Ref 5.14
**Deleted:** --
**Deleted:** $$$
**Deleted:** $$$
**Deleted:** &&&compl
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Line spacing: single
**Formatted:** Justified
**Deleted:** @@@@7mar - for driver discov , we are not driver discovery paper
**Deleted:** Excerpt From ... [80]
**Formatted:** Line spacing: single
**Moved (insertion) [20]**
**Formatted:** Font:Times New Roman
**Formatted Table**
**Formatted:** Normal, Space Before: 0 pt
**Deleted:** $$$
**Deleted:** Compact Annotation --
**Deleted:** $$$
**Deleted:** $$$
**Deleted:** @@@
**Deleted:** &&&TBC
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Line spacing: single
**Deleted:** We demonstrate through power analysis

RSF

| | |
|---|---|
| | or "compact" annotation can significantly reduce noise level and increase statistical power. For example, if we were not trimming truly functional or important sites, but rather trimming unimportant sites, the test power will increase. For instance, in the old way that we found enhancer sites by just calling a 1KB region from a peak admittedly by almost any estimation included knots of obviously non functional sites. Trimming this down using a large battery of histone marks and the exact shape of the signal, we believe more accurately gets it the truly functional region, particularly when coupled with accurate STARR-seq and Hi-C data will hopefully increase power. Another case is the TF binding hotspot around the promoter region of WDR74. Instead of testing up to 2.5K promoter region without prior information, we can trim the test set to a core set of the promoter region where many TF binds to, 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).<br><br> p15.4<br>p15.3<br>p15.1<br>p14.3<br>p14.1<br>p13<br>p12<br>p11.2<br>p11.12<br>q11<br>q12.1<br>q12.2<br>q13.1<br>q13.3<br>5,566 bp<br>,607,000 bp<br>62,608,000 bp<br>62,609,000 bp<br>Mutation...cerTypes<br>[0 - 5.71]<br>RefSeq Genes<br>WDR74 |
| Excerpt From Revised Manuscript | |

## <ID>REF5.15 – BCL6 Questions

<TYPE>$$$Annotation $$$Calc
<ASSIGN>@@@XK @@@TG
<PLAN>&&&AgreeFix
<STATUS>%%%TBC
[JZ2MG: to be added to the disc agenda next week]

**Deleted:** Excerpt From
@@@@7mar power analysis - please realize our goal is not to do driver discover. ... [81]

**Formatted:** Line spacing: single

**Moved (insertion) [21]**

**Formatted:** Font:Times New Roman

**Formatted Table**

**Deleted:**
$$$

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Formatted:** Not Highlight

**Deleted:** @@@

**Formatted:** Not Highlight

**Deleted:** @@@

**Formatted:** Not Highlight

**Deleted:** &&&

**Formatted:** Not Highlight

**Formatted:** Normal

| | |
|---|---|
| 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 | BCL6 mutations were found in promoter region.<br><br>XK, TG<br>@@@7mar - yuck!<br>Are any SVs associated with BCL6? |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF5.16 – ChIP-seq vs other computational based networks

<TYPE>$$$Network $$$Calc
<ASSIGN>@@@Peng @@@JZ
<PLAN> &&&AgreeFix
<STATUS>%%%Done

**Deleted:** -- Ref 5.17

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** @@@

**Deleted:** @@@

**Deleted:** &&&TBC

**Formatted:** Normal

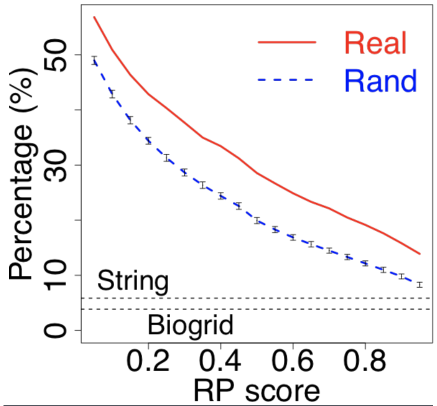
**Formatted:** Line spacing: single

**Formatted Table**

| | |
|---|---|
| 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 up and we also feel that it is important to make comparison with other networks with statistical evaluation. We made the following revisions.<br>1. To make the statement more accurate, we changed our previous sentence from “more accurate and experimentally based regulatory linkages” to “ENCODE TF and RBP networks provide experimentally based linkages that are more relevant to gene expression regulation that other network types.” As stated, we constructed two ENCODE regulatory |

**Formatted:** Line spacing: single

**Deleted:** @@@@7mar - Compare to J stam & pe

| | |
|---|---|
| | networks: 1, transcriptional regulations between TFs and target genes; 2, post-transcriptional regulations between RBPs and target genes.<br>2. 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. |
| Excerpt From Revised Manuscript |  <br>**Supplementary Figure X. ENCODE networks captured a higher fraction of curated regulations than other networks.** The TRRUST database manually curated 8,412 transcriptional regulatory interactions from Pubmed articles (Han et al., 2018). We computed the fractions of TTRUST interactions that other networks can recapitulate. Since each ENCODE ChIP-Seq interaction has a regulatory potential (RP) score, we showed the fractions with different RP thresholds. The random fraction for ENCODE network was estimated through 100 perturbed TTRUST networks using the stub-rewiring method that preserved the gene network degrees (Milo et al., 2002). |

WHAT HAPPENED TO STAM CMP (2.)

**Moved down [22]:** Excerpt From ... [83]

**Moved down [22]:** Excerpt From ... [82]

**Formatted:** Font:Times New Roman

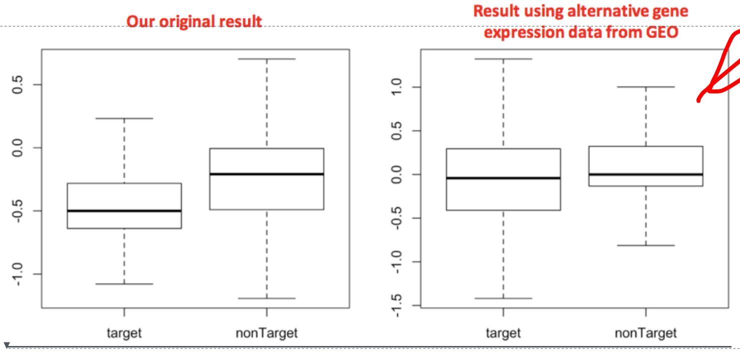
**Formatted:** Font:Times New Roman

**Formatted Table**

**Formatted Table**

<ID>REF5.17 – MYC KD

<TYPE>$$$Network $$$Text
<ASSIGN>@@@DC
<PLAN>&&&AgreeFix
<STATUS>%%%DONE

| | |
|---|---|
| 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. 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.<br>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 MCF7 cell line). These comparable results in an alternative cell line suggests that these results are robust.<br>We also found another array based MYC knockdown data the results correlate well with our discoveries. |
| Excerpt From Revised Manuscript | <br>Our original result<br>Result using alternative gene expression data from GEO<br> |

SUPP

Deleted: $$$

Deleted: --

Deleted: $$$

Deleted: $$$

Deleted: @@@

Deleted: &&&compl

Formatted: Normal

Formatted: Line spacing: single

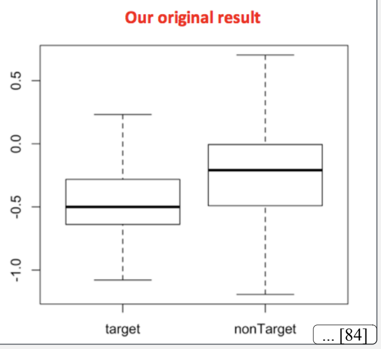
Formatted Table

Formatted: Line spacing: single

Formatted: Justified

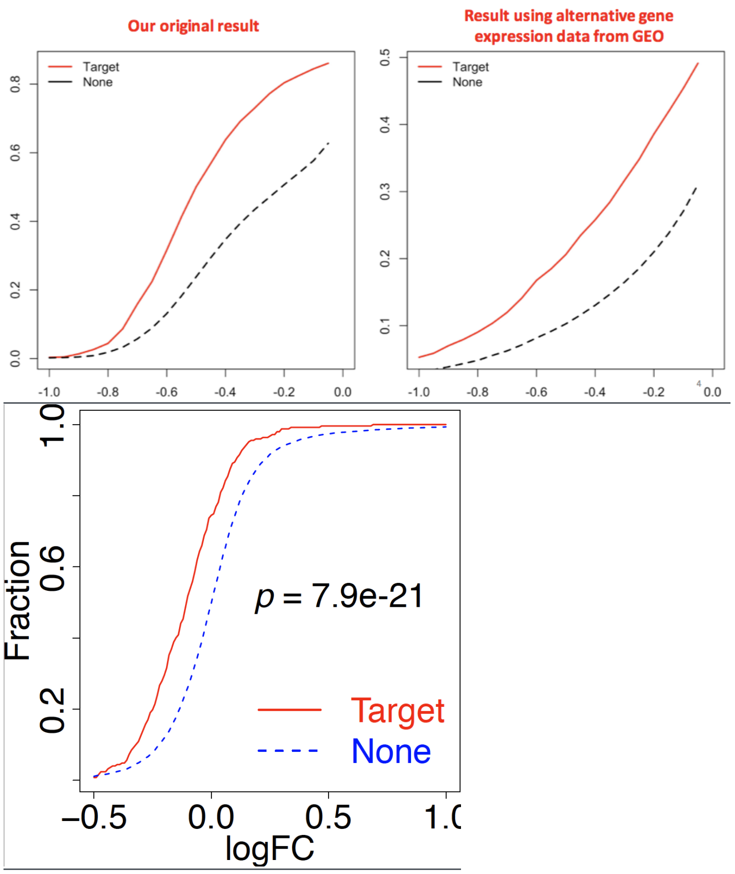
Formatted: Font:Times New Roman

Formatted: Line spacing: single



Our original result

Deleted: ... [84]

| | |
|---|---|
| |  |

## <ID>REF5.18 – SUB1 analysis

<TYPE>$$$NoveltyPos $$$Calc
<ASSIGN>@@@MRS @@@JL @@@YY
<PLAN>&&&MORE
<STATUS>%%%TBC
[JZ2YY: would you please add your stuff here?]
[JZ2Peng: write something about sub1 decay rate]

| | |
|---|---|
| Referee Comment | 14. SUB1 is a potentially interesting new cancer gene. The authors should further explore the biology of this gene. |

**Deleted:** -- Ref 5.19

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** &&&More @@@

**Deleted:** @@@

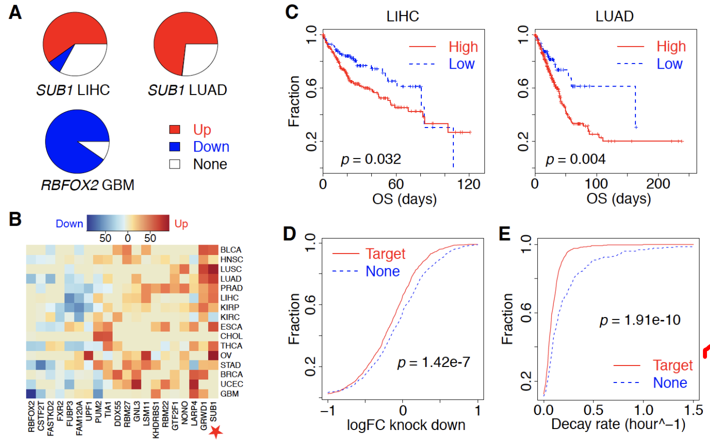
**Deleted:** @@@

**Deleted:** &&&

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

| | |
|---|---|
| Author Response | We thank the referees for the positive comments. We did follow up with SUB1 in this round of revision.<br>1. We checked SUB1 regulation potential in different cancer types and found that they are consistent as below. |
| Excerpt From Revised Manuscript | <br> <br>**Inference of RNA binding proteins that drive tumor specific expression patterns**. Based on ENCODE eCLIP data, we applied RABIT framework to identify RNA binding proteins (RBP), whose target genes are differentially regulated in diverse TCGA cancer types. (A) For each RBP, the percentage of patients with target genes significantly up regulated (red), down regulated (blue) or not regulated (white) is shown for each cancer type. (B) Hierarchically clustered heatmap was used to show the percentage of patients in each cancer type with RBP target significantly up regulated (red) or down regulated (blue). (C) All TCGA Liver Hepatocellular Carcinoma (LIHC) lung adenocarcinoma (LUAD) patients are divided to two groups according to the *SUB1* activity predicted by RABIT. The overall survival was shown in each group by KM plot. The association between RABIT regulatory activity and overall survival was tested CoxPH regression. (D) The cumulative distributions of gene expression after *SUB1* knock down in HepG2 cell are shown for predicted target genes and none-target genes. The comparison between two categories of expression changes is done through Wilcoxon rank-sum test. (E) The mRNA decay rates are compared between predicted *SUB1* targets and none-target genes as part D. |

2

## <ID>REF5.19 – Significance of regulatory network hierarchy

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

**Formatted:** Line spacing: single

**Deleted:** &&&& we've done more with the network

**Moved down [23]:** Excerpt From ... [85]

**Moved down [23]:** Excerpt From ... [86]

**Formatted:** Font:Times New Roman

**Formatted:** Font:Times New Roman

**Formatted Table**

**Formatted Table**

**Deleted:** -- Ref 5.20

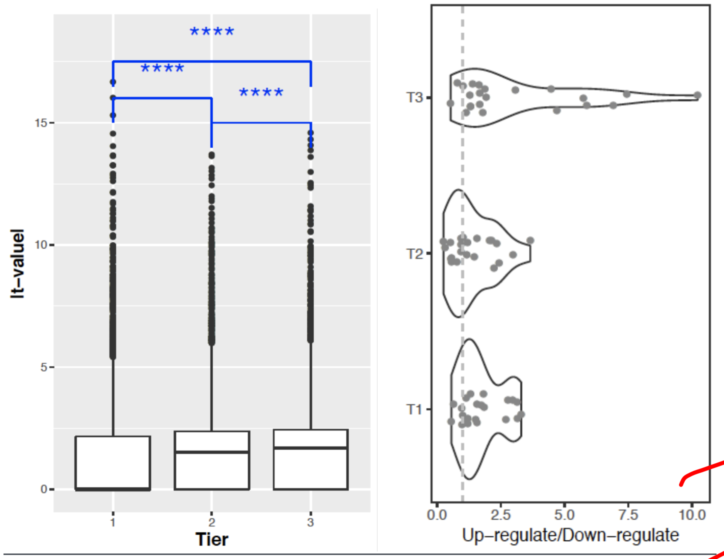
**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$Text @@@

**Deleted:** &&&compl

<table>
<tr><td>Referee Comment</td><td>15. The manuscript claims that transcription factors placed at the top level of the network hierarchy are enriched in cancer-associated genes and drive expression changes. Both claims need to be supported with statistical tests.</td></tr>
<tr><td>Author Response</td><td>We thank the referees for the positive comments. We've done a statistical significance test as requested. 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).



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

MANUSCRIPT

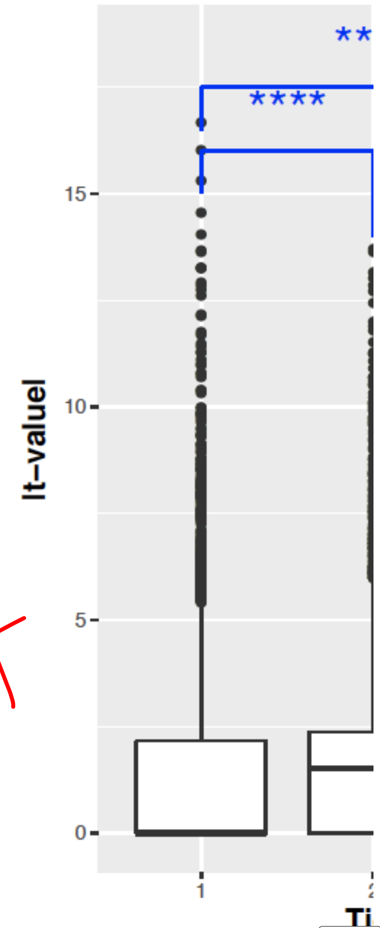
**Formatted:** Normal

**Formatted:** Line spacing: single

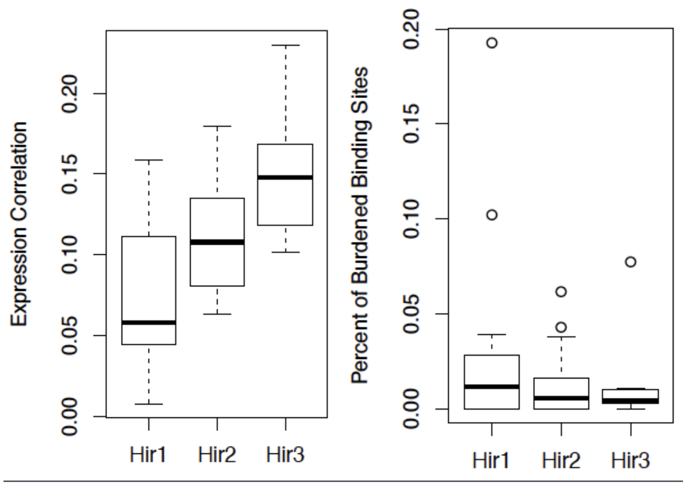
**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Justified



**Deleted:** ... [87]

| | |
|---|---|
| |  |
| Excerpt From Revised Manuscript | |

## <ID>REF5.20 – Rewiring of regulatory network

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

| | |
|---|---|
| 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? |

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Deleted:** -- Ref 5.21

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** @@@

**Deleted:** &&&TBC

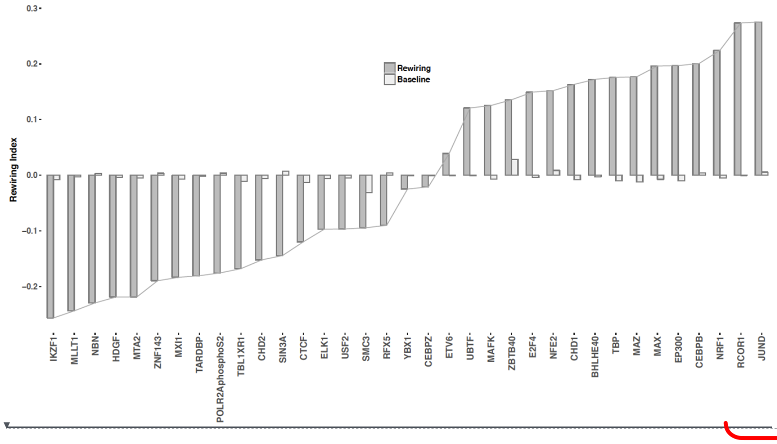
**Formatted:** Normal

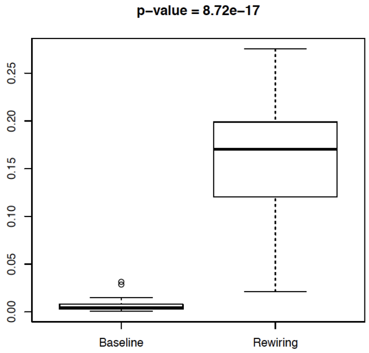
**Formatted:** Line spacing: single

**Formatted Table**

Author Response

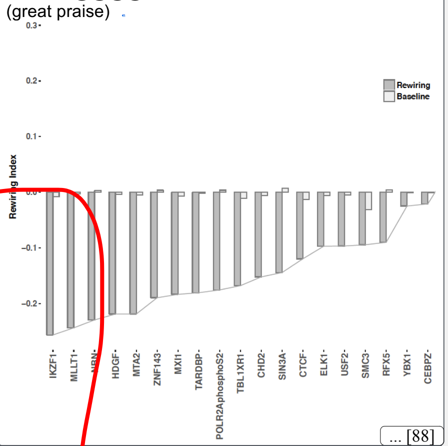
We would like to truly thank referee for pointing out this issue. We agree with the referee that we need to be more clear about the rewiring of regulatory network in the revised manuscript. We would like to clarify that the rewiring index is based on the fraction of regulatory edge changes between two cellular contexts. The rewiring index is also 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 performed the same rewiring analysis between isogenic replicates of the same cellular context. The edge changes between two networks will be simply a noise from ChIP-seq experiments.





**Formatted:** Line spacing: single

**Deleted:** @@@@7mar we TRULY thank the referee (great praise)



... [88]

**Formatted:** Justified

**Deleted:**

... [89]

SUPPL

| | |
|---|---|
| | 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. |
| Excerpt From Revised Manuscript | |

## <ID>REF5.21 – Rewiring analysis in the stem cells

<TYPE>$$$Stemness $$$Calc
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

| | |
|---|---|
| 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 | ####7mar we truly thank referee. Took referee’s comment to heart, made hugh improvement<br>We truly thank referee for the pointing this out. We took referee’s suggestion to heart and we now have added a statistical significance testing for H1 stem cell model in the revised manuscript.<br>#### to do - same as 16<br>#### False positive rate analysis<br>#### Think about test of significance (have some more analysis) DL/JZ disc. |

**Formatted:** Justified

**Moved (insertion) [22]**

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Moved down [24]:** Excerpt From ... [90]

**Formatted:** Font:Times New Roman

**Formatted Table**

**Deleted:** -- Ref 5.22

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** @@@

**Deleted:** &&&

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Deleted:** @@@@

**Deleted:** Statistical

**Deleted:** DL :

**Deleted:**

|  |  |
|---|---|
| Excerpt From Revised Manuscript |  |

## <ID>REF5.22 – Selection of regions for validation testing

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

|  |  |
|---|---|
| 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 pointing this out. We had some of the details in the supplementary but they weren't that well spelled out . We've redone supplementary section 6 and to answer this question.<br>The eight regions were selected from our integrative promoter and enhancer regions in MCF-7 cell lines. We prioritized these regulatory regions based on motif breaking power as described in section 6.1 S. 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 3 replicates for each mutant and control experiments.<br>Error bar is representing 95% confidence interval across 3 replicates. |

SUPPL

**Moved down [25]:** Excerpt From ... [92]

**Moved down [25]:** Excerpt From ... [91]

**Formatted:** Font:Times New Roman

**Formatted:** Font:Times New Roman

**Formatted Table**

**Formatted Table**

**Deleted:** -- Ref 5.23

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** @@@

**Deleted:** @@@

**Deleted:** &&&compl

**Formatted:** Normal

**Formatted:** Line spacing: single

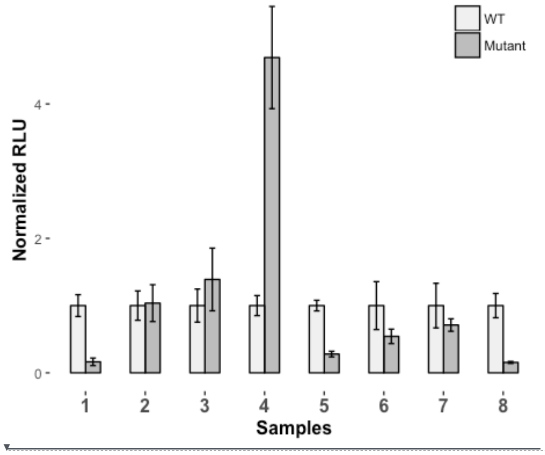
**Formatted Table**

**Formatted:** Line spacing: single

**Deleted:** JZ, DL: we can answer ... [93]

**Formatted:** Justified

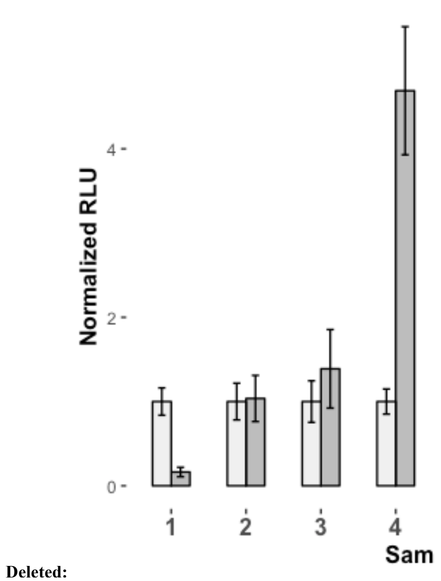
**Formatted:** Justified

| | |
|---|---|
| |  |
| Excerpt From Revised Manuscript | |

## <ID>REF5.23 – Presentation and revision to manuscript

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

| | |
|---|---|
| Referee Comment | 19. The authors should consider moving the general overview diagrams that constitute much of the main figures to the supplement, and in turn present data-rich figures from there with the main manuscript. |
| Author Response | We thank for the referee for this comments. |




**Deleted:**

**Moved (insertion) [23]**

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Moved down [26]:** Excerpt From ... [94]

**Deleted:** -- Ref 5.24 – (Minor)

**Formatted:** Font:Times New Roman

**Formatted Table**

**Deleted:** --

**Deleted:** $$$

**Deleted:** $$$

**Deleted:** &&&

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

| | |
|---|---|
| | We have tried to revise the figures as requested<br>We have fixed figure XX & YY. |
| Excerpt From Revised Manuscript | |

Formatted: Font:Times New Roman

Formatted: Line spacing: single

## <ID>REF5.24 – Difference between ENCODEC and existing prioritization methods

Deleted: -- Ref 5.25 – (Minor) How related to Fun ... [95]

<TYPE>$$$Validation $$$Text
<ASSIGN>&&&AgreeFix
<PLAN>&&&AgreeFix
<STATUS>%%%TBC

Deleted: $$$

Deleted: &&&

Formatted: Normal

| | |
|---|---|
| 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. |
| Excerpt From Revised Manuscript | |

Formatted: Line spacing: single

Formatted Table

Formatted: Line spacing: single

Deleted: How are we diff funseq ... [96]

Deleted: Excerpt From ... [97]

Formatted: Line spacing: single

Moved (insertion) [24]

Formatted: Font:Times New Roman

Formatted Table

## <ID>REF5.25 – BMR

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

| | |
|---|---|
| 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 | |

## <ID>REF5.26 – Citation of previous work

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

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

**Formatted:** Heading 2, Space Before: 14 pt
**Deleted:** ... [98]
**Deleted:** ) BMR -- ... [99]
**Deleted:** &&&
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Line spacing: single
**Deleted:** Excerpt From ... [100]
**Formatted:** Line spacing: single
**Moved (insertion) [25]**
**Formatted:** Font:Times New Roman
**Formatted Table**
**Formatted:** Heading 2, Space Before: 14 pt
**Deleted:** ... [101]
**Deleted:** ) Untitled -- ... [102]
**Deleted:** &&&
**Formatted:** Normal
**Formatted:** Line spacing: single
**Formatted Table**
**Formatted:** Line spacing: single
**Deleted:** Excerpt From ... [103]
**Formatted:** Line spacing: single

|  |  |
|---|---|
| Excerpt From Revised Manuscript |  |

**Moved (insertion) [26]**

**Formatted:** Font:Times New Roman

**Formatted Table**

## <ID>REF5.27 – Tumor normal comparison and composite model

<TYPE>$$$Minor,$$$CellLine
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%TBC
[JZ2MG, JZ2DL: to disc next week]

MORE DISCUSS

**Deleted:** ... [104]

**Deleted:** ) Untitled -- ... [105]

**Deleted:** &&&

**Formatted:** Normal

|  |  |
|---|---|
| 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 | JZ: I assume that we used Roadmap normal? There is no ChIP-Seq data there! But we did use the DHS data for the imputed network! |
| Excerpt From Revised Manuscript |  |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF5.28 – Use of H1 for stemness calculation

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

**Deleted:** -- Ref 5.29

**Deleted:** --

**Deleted:** $$$

**Deleted:** &&&

$$$Stemness %%%TBC

| | |
|---|---|
| Referee Comment | 24. The authors use the H1 embryonic stem  cell line as model for "stemness" in cancer. Tumor "stemness" often resembles tissue progenitors, not embryonic stem cells. In the absence of reliable data for such progenitors the authors should note this caveat with their analysis. |
| Author Response | |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF5.29 – Validation of prioritized element

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

**Deleted:** -- Ref 5.30 – Untitled -- ... [106]

**Deleted:** &&&

**Formatted:** Normal

| | |
|---|---|
| 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 | |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF5.30 – SYCP2 and beyond

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

| | |
|---|---|
| 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 | |
| Excerpt From Revised Manuscript | |

**Deleted:** -- Ref 5.31 – Untitled -- ... [107]

**Deleted:** &&&

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF5.31 – Utility of ENCODEC

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

| | |
|---|---|
| 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 | |

**Deleted:** -- Ref 5.32 – Untitled -- ... [108]

**Deleted:** &&&

**Formatted:** Normal

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

| | |
|---|---|
| Excerpt From Revised Manuscript | |

Formatted: Font:Times New Roman

Formatted: Line spacing: single

## <ID>REF5.32 – P-value of survival analysis

Deleted: -- Ref 5.33 – Untitled -- ... [109]

<TYPE>$$$Minor,$$$Presentation

Deleted: &&&

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

Formatted: Normal

| | |
|---|---|
| Referee Comment | 28. In Figure 2e, a p-value should be given with the analysis. |
| Author Response | |
| Excerpt From Revised Manuscript | |

Formatted: Line spacing: single

Formatted Table

Formatted: Line spacing: single

Formatted: Font:Times New Roman

Formatted: Line spacing: single

## <ID>REF5.33 – Q-value of extended gene analysis

Deleted: -- Ref 5.34 – Untitled -- ... [110]

<TYPE>$$$Minor,$$$Presentation

Deleted: &&&

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

Formatted: Normal

| | |
|---|---|
| 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- |

Formatted: Line spacing: single

Formatted Table

Formatted: Line spacing: single

|  |  |
|---|---|
|  | value to the figure would be important, so we have updated the figure in the revised manuscript. |
| Excerpt From Revised Manuscript |  |

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF5.34 – Presentation issue with network hierarchy

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

**Deleted:** -- Ref 5.35

**Deleted:** --

**Deleted:** $$$

**Deleted:** &&&TBC

**Formatted:** Normal

|  |  |
|---|---|
| Referee Comment | 30. Figure 4 would benefit from labeling of the network tiers. |
| Author Response | We thank reviewer for the comment. We fixed the labeling of the network tiers in the revised manuscript. |
| Excerpt From Revised Manuscript |  |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF5.35 – Presentation

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

**Deleted:** -- Ref 5.36

**Deleted:** --

**Deleted:** $$$

**Deleted:** &&&

**Formatted:** Normal

|  |  |
|---|---|
| Referee Comment | 31. In Figure 6b, it should be clarified whether “samples” refers to genomic locations, patients, or cell lines. The |

**Formatted:** Line spacing: single

**Formatted Table**

| | |
|---|---|
| | 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. We have updated the figure in the revised manuscript. |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

## <ID>REF5.36 – Supplementary document

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

**Deleted:** -- Ref 5.37

**Deleted:** --

**Deleted:** $$$

**Deleted:** &&&TBC

**Formatted:** Normal

| | |
|---|---|
| Referee Comment | 32. The supplement contains multiple reference errors. |
| Author Response | We’ve made numerous improvements to the supplementary document. |
| Excerpt From Revised Manuscript | |

**Formatted:** Line spacing: single

**Formatted Table**

**Formatted:** Line spacing: single

**Formatted:** Font:Times New Roman

**Formatted:** Line spacing: single

**Page 1: [1] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

$$$BMR
$$$Power
$$$Presentation
$$$Annotation
$$$Network
$$$Hierarchy
$$$CellLine
$$$Stemness
$$$Validation
$$$NoveltyPos
$$$NoveltyNeg

@@@ : assignment

&&&TBC: To Be Continued
&&&compl: Completed
&&&More : go above and beyond the scope of the question and indicates more analyses to be done

---

**Page 7: [2] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

## -- Ref 1.1 – Overall comments on the paper --

## $$$

**Page 10: [3] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

## Scope of the paper --

## $$$NoveltyNeg $$$Text @@@JZ(@@@WM @@@MRS ) &&&compl

| | |
|---|---|
| Referee Comment | However, I find the current manuscript seriously lacking. The major problem is simply that most of these applications have already been in the literature for a while, often as high profile papers on their own. So the manuscript is not quite a review but does not seem to have any significant findings either. |

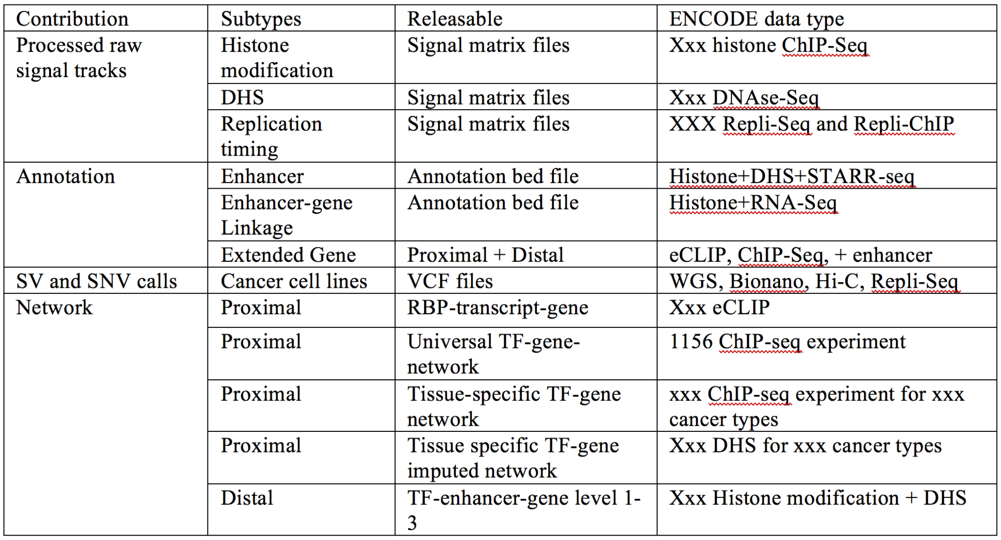
Author Response

We thank the reviewer for pointing out the existence of other literature that relates to the significant problems we address. We have summarized various references mentioned by the referees and made comparisons as below.

| 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, since this paper is out 3 months after our submission | Cited | Use 169 epigenetic features to predict gene level BMR | No replication timing data is used |
| Imielinski (2017) | No | Yes | | |
| 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 | NOT specifically for BMR |

Regarding the novelty of this paper, we disagree with the reviewer. We want to make it clear that his paper is to be considered as a "*resource*"

|  |  |
|---|---|
|  | paper, not a novel biology 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. |



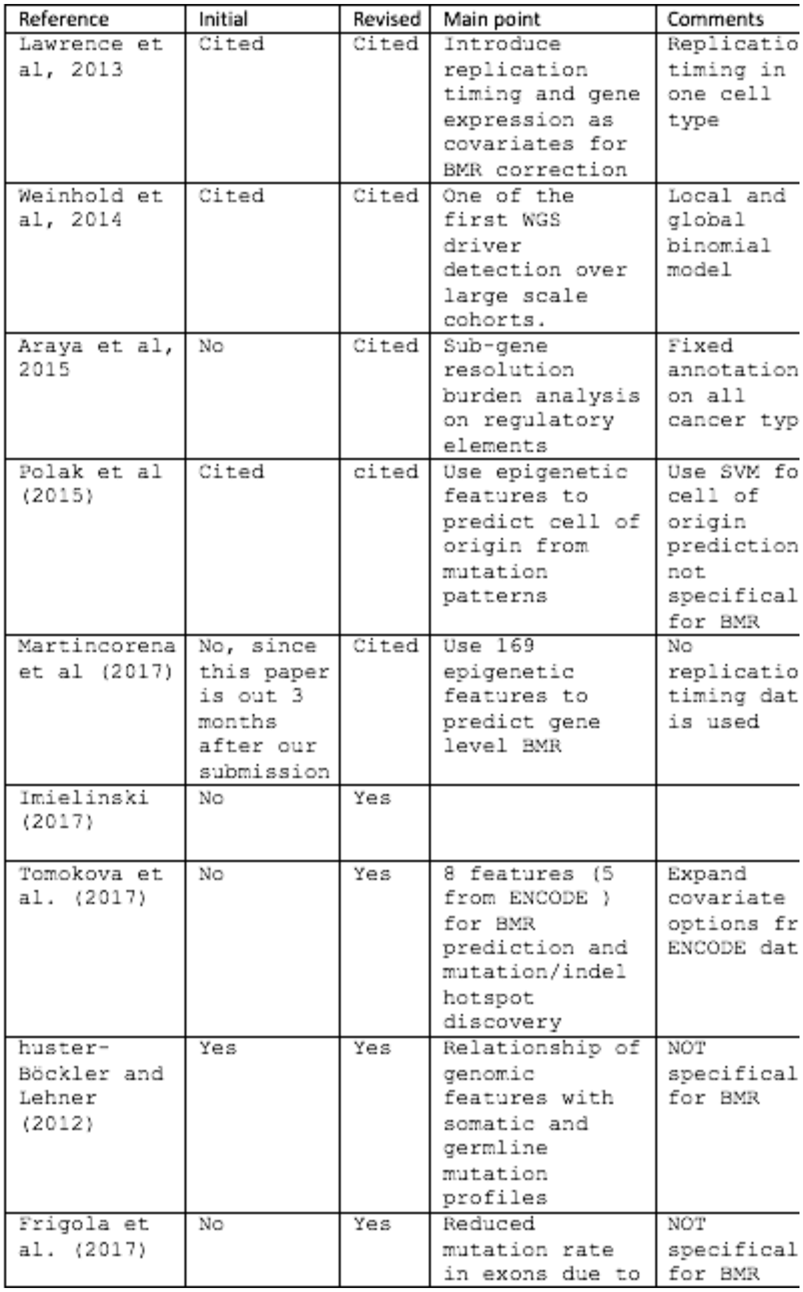
| Contribution | Subtypes | Releasable | ENCODE data type |
|---|---|---|---|
| Processed raw signal tracks | Histone modification | Signal matrix files | Xxx histone ChIP-Seq |
|  | DHS | Signal matrix files | Xxx DNAse-Seq |
|  | Replication timing | Signal matrix files | XXX Repli-Seq and Repli-ChIP |
| Annotation | Enhancer | Annotation bed file | Histone+DHS+STARR-seq |
|  | Enhancer-gene Linkage | Annotation bed file | Histone+RNA-Seq |
|  | Extended Gene | Proximal + Distal | eCLIP, ChIP-Seq, + enhancer |
| SV and SNV calls | Cancer cell lines | VCF files | WGS, Bionano, Hi-C, Repli-Seq |
| Network | Proximal | RBP-transcript-gene | Xxx eCLIP |
|  | Proximal | Universal TF-gene-network | 1156 ChIP-seq experiment |
|  | Proximal | Tissue-specific TF-gene network | xxx ChIP-seq experiment for xxx cancer types |
|  | Proximal | Tissue specific TF-gene imputed network | Xxx DHS for xxx cancer types |
|  | Distal | TF-enhancer-gene level 1-3 | Xxx Histone modification + DHS |

**Page 10: [4] Moved to page 11 (Move #1)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

|  |  |
|---|---|
| Excerpt From Revised Manuscript |  |

**Page 11: [5] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

[[@@@@7mar: we have to say they don’t have so much data here]]
*** We have to say why these papers don't do as accurate estimation



| Reference | Initial | Revised | Main point | Comments |
|---|---|---|---|---|
| Lawrence et al, 2013 | Cited | Cited | Introduce replication timing and gene expression as covariates for BMR correction | Replicatio 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 typ |
| Polak et al (2015) | Cited | cited | Use epigenetic features to predict cell of origin from mutation patterns | Use SVM fo cell of origin prediction not specifical for BMR |
| Martincorena et al (2017) | No, since this paper is out 3 months after our submission | Cited | Use 169 epigenetic features to predict gene level BMR | No replicatio timing dat is used |
| Imielinski (2017) | No | Yes | | |
| Tomokova et al. (2017) | No | Yes | 8 features (5 from ENCODE ) for BMR prediction and mutation/indel hotspot discovery | Expand covariate options fr ENCODE dat |
| huster-Böckler and Lehner (2012) | Yes | Yes | Relationship of genomic features with somatic and germline mutation profiles | NOT specifical for BMR |
| Frigola et al. (2017) | No | Yes | Reduced mutation rate in exons due to | NOT specifical for BMR |

**Page 11: [6] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

In addition, were able to use cell-type matched feature data across our BMR analysis. This includes more commonly used features for BMR modification, like the 932 histone modification features we used, but also many other features, especially the 51 replication time data, that have proven useful but are less frequently incorporated into BMR models.

**Page 11: [7] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

**Page 16: [8] Moved to page 13 (Move #2)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 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 16: [9] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

@@@@7mar: this is very difficult/different problem, more data is good, polak there is no cell line data invovled

Thanks for pointing out the Polak 2015 paper. (Note we did cite this in our manuscript.)
1. First we want to emphasize some specific type of data from ENCODE such as Hi-C and replication timing. By pointing out that using data from a matched cell line is better, is not used as a novel conclusion (as we also cited the Polak 2015 paper), but rather to emphasize the value of our data. Take replication timing as an example, a lot previous work (lawrence et al. 2013) actually use replication timing data from HeLa cell line due to the limited choice. In our revised manuscript, we described that there are 51 high quality replication timing data, which is quite valuable for cancer genomics.
2

**Page 16: [10] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

note are:
(A) Even in the the Polak 2015 paper, it is not always the case that cell-of-origin can be predicted perfectly using the epigenetic features (Fig. 4 b).
(B) the Polak 2015 paper only compare among normal tissues from the Roadmap data and they did not compare cell line data at all.
Here we used breast cancer as an example to show the importance of cell line data

**Page 16: [11] Moved to page 13 (Move #4)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

. We calculated the correlation of breast cancer mutation counts (from a patient cohort) per mbp with histone signals from both Breast tissue (the roadmap) and MCF-7 (an ENCODE cell line).

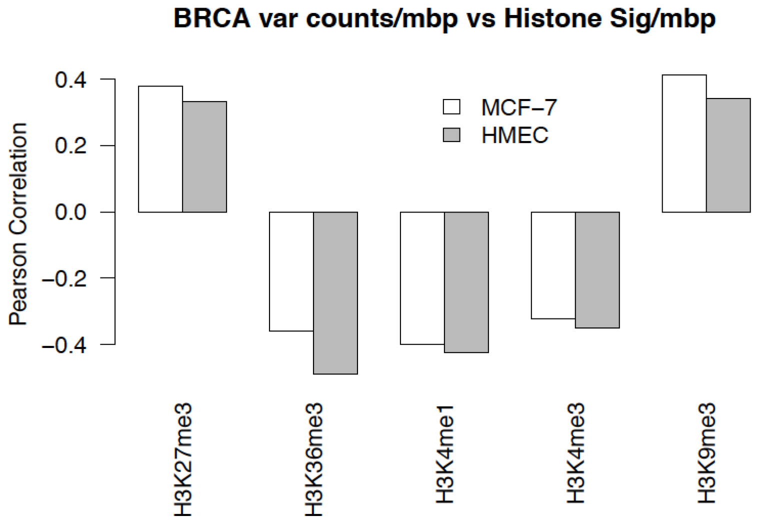
**Page 16: [12] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

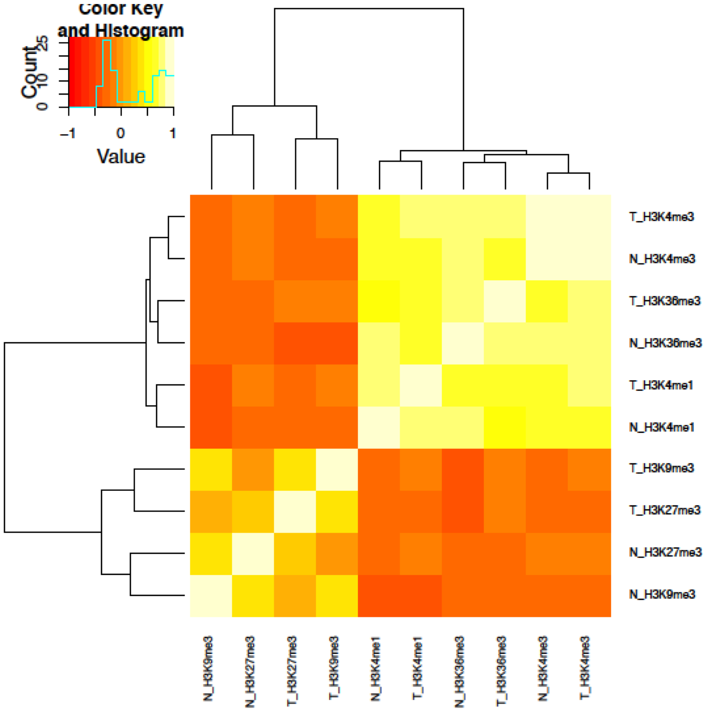
As seen from the following figure, MCF-7 provides similar

**Page 16: [13] Moved to page 14 (Move #5)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

(and sometimes even better correlation with mutation counts). We also found that histones from tissue and matched cell lines are actually quite correlated in a larger scale (see heatmap below).

**Page 16: [14] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**





3.

**Page 16: [15] Moved to page 14 (Move #6) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

In general, there are less such data. On the contrary, the cell line functional characterization data has lots of advantage in terms of assay richness.

**Page 16: [16] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

For some specific cancer types, such prostate cancer, cell lines like LNCap might further help.

**Page 16: [17] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

## Difference between ENCODEC and FunSeq --

## $$$BMR $$$Text @@@JZ&&&compl

| | |
|---|---|
| Referee Comment | The rest of the sections (and their corresponding supplement sections) are variable in significance and quality. That ENCODE data helps in prioritiz<br>, and so the value of the described analysis less clear. |
| Author Response | Variant/regulator prioritization is one of the most important applications of the ENCODEC resource. We want to clarify that our current approach is completely different from our previous approach (as shown in Fig 6 in the initial submission). ENCODE3 largely expanded its data richness in several top-tier cell types. With the increased number and novel types of assays, our current prioritization scheme now follows a tissue specific manner. It adopts a top-down scheme: 1) first combine cohort level expression level to prioritize key regulators; 2) then combine patient expression profiles and epigenomic features to prioritize key regulatory elements; 3) the pinpoint the SNVs after incorporating final scale features like motif breaking, conservation, and etc. Non of the tissue-specific features, network perturbations, and integration of external expression/mutation features are included before.<br>Also, it is worth mentioning that we did not claim this is a novel noncoding variant prioritization method, but rather an application about how the new release of ENCODEs data can help us to better interpret variants. |
| Excerpt From Revised Manuscript | |

## -- Ref 1.7 –

| Page 17: [18] Deleted | jingzhang.wti.bupt@gmail.com | 3/17/18 5:29:00 PM |
|---|---|---|

## -- Ref 1.8 – Novelty and presentation of the paper --

$$$Presentation $$$NoveltyPos $$$NoveltyNeg $$$Text @@@JZ&&&compl

| | |
|---|---|
| Referee Comment | Personally, I wonder whether a review paper that gives an update to the ENCODE database and state the illustrative examples succinctly might be more appropriate than several studies, in which more work/descriptions are needed to show novelty, packaged together? |
| Author Response | Note 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.<br>In our revised manuscript, we added new data types, like several whole genome sequencing of the cell lines and further incorporated more TF/RBP knockdown data to validate our prioritized known and novel key regulators, such as TP53, ESR1, ZNF687, and SUB1. |

**Page 17: [19] Moved to page 12288 (Move #7)jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 17: [20] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

## -- Ref 2.1 – Novelty of the paper --

$$$NoveltyNeg $$$Text @@@JZ @@@DC&&&compl

**Page 18: [21] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

Referee
Comment

**Page 18: [22] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

Author
Response

**Page 18: [23] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

@@@@mar7 earlier
In addition, we wish to point out that unlike previous roll-outs, ENCODE 3 does not associate specific data sets with specific papers. In addition, there are no dependencies between any of the papers in this package. All the ENCODE data is open to the public and is not associated with, for instance, the encyclopedia paper or a particular companion paper.

We thank the referee for pointing out the dataset problem, which we believe is more of a presentation issue. In addition to the massive traditional assays such as ChIP-seq, DNAse-Seq, and RNA-seq, we incorporated a list of new data types as summarized below.

| Assay | More info |
|---|---|
| STARR-seq | K562, MCF-7, LNCaP, HepG2 |
| Hi-C | K562, MCF-7, LNCaP, HepG2, etc ... |
| Replication timing | Xxx cell lines |
| CRISPERi based knockout | 77 |
| shRNA based knowck down | 533 |
| SV/SNV call set | Xxx cell lines |
| Bionano | Xxx cell lines |
| WGS | Xxx cell lines |

We thank the reviewers about the comments on presentation of this manuscript. We want to emphasize that the main goal of ENCODEC is about ENCODE resource for cancer community, instead of novel scientific cancer discoveries. By integrating the novel types of assays with massive traditional assays, we provide the following list of resources.

Ready to use signal files that can help BMR estimation, including the ones that are quite limited in existing methods such as replication timing and Hi-C

Accurate and compact annotation for assay rich cell lines for somatic mutation hotspot detection.

Accurate enhancer gene linkage supported by multiple type of advanced assays like STARR-seq and

**Page 18: [24] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

Universal and tissue-specific experimental based TF/RBP networks
Imputed TF networks for more than 20 cancer types
Paired tumor to normal networks to investigate network perturbations
High quality SNV and SV calls from WGS and other types of assays

**Page 19: [25] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

We thank the reviewer to point out these references and they are also good models. Actually one of mentioned paper (Marticorena, 2017) was published on Nov 2017, almost three months after our submission. It comes out three months after our initial submission so we did not cite in the last round. Admittedly, it decrease the novelty of our BMR estimation method, but it also proves that we are technically sound at this point.

We want to emphasize that the goal of this paper is not to propose novel cancer driver detection method, but rather than highlight that ENCODE data can help BMR estimation, also in those model.

In our revised manuscript, we tuned down this part by moving two sub-panels (A & B) in Figure 2 to

**Page 20: [26] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

supplementary figures. We also added these references and clarified our point by proper acknowledgement.

@@@@7mar: the fact that it is published,

**Page 20: [27] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

, loses novelty, but we are not claiming novelty

@@@@7mar this isn't so neg. It bolsters good

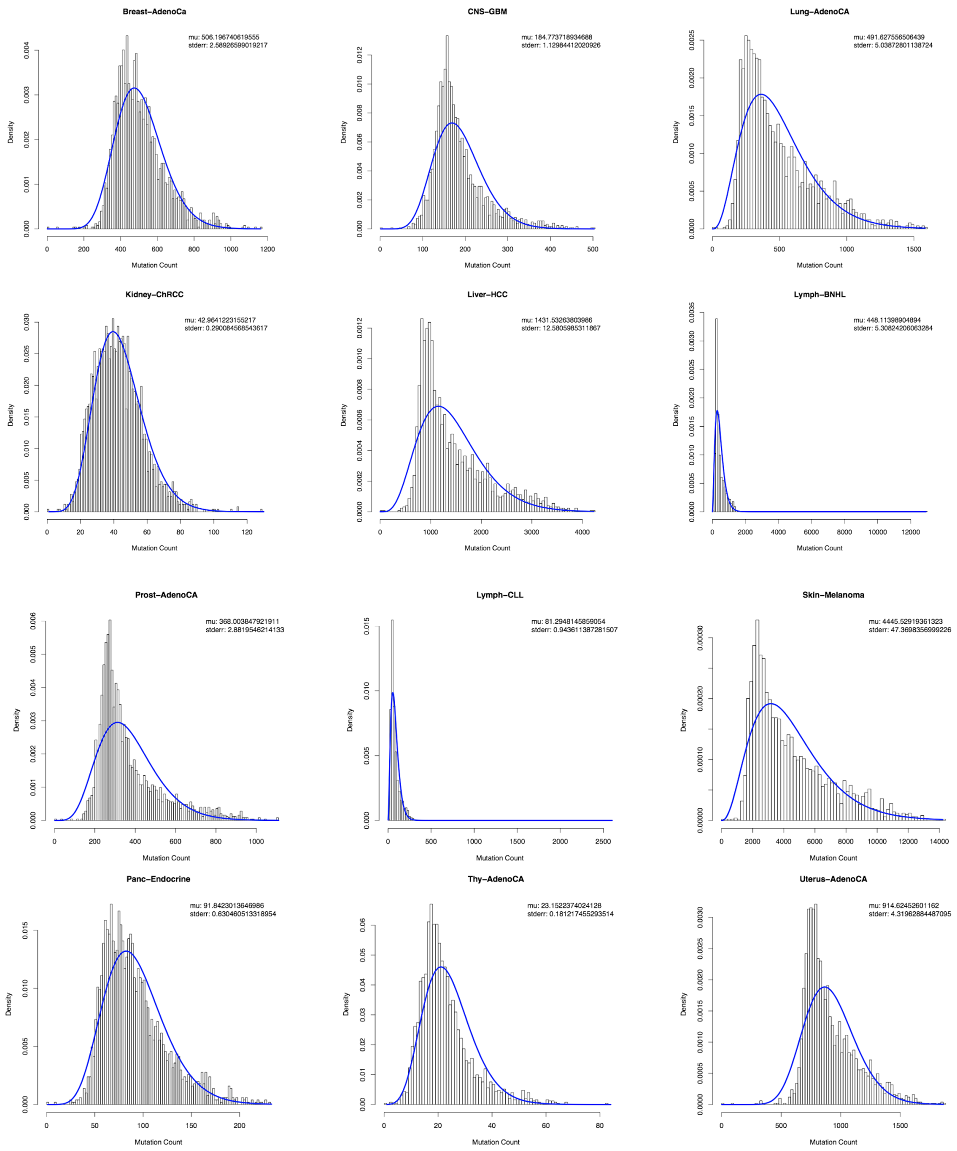
**Page 20: [28] Moved to page 24 (Move #9) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

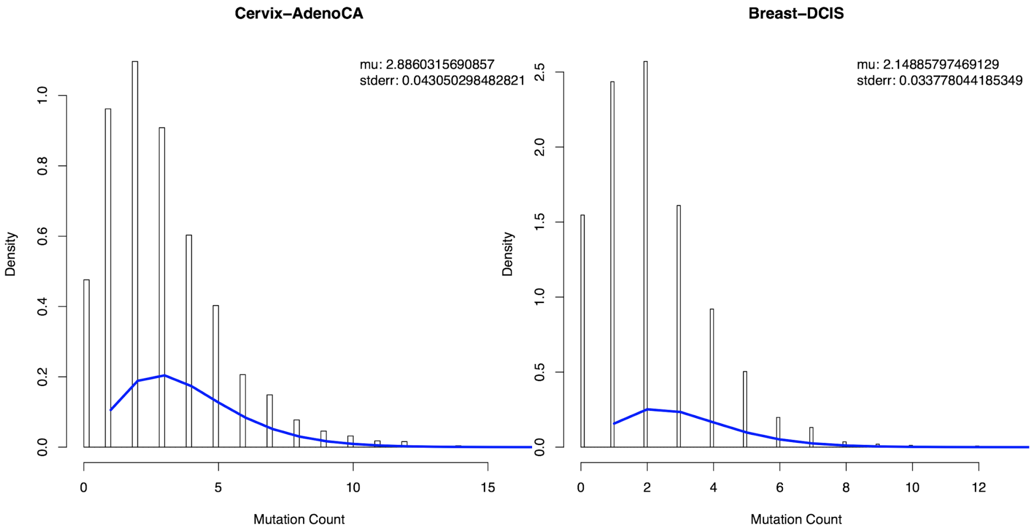
| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 20: [29] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

We have tried to use the Gamma-Poisson model to fit the variant counts per 1mb bins for many cancer types and the fitting are listed as below. We feels for most cancer types that have enough variants, it fits OK with the observed data. However, there might be some case, especially when somatic mutation count is relatively low, fitting is not that good. This is partially why we test the degree of overdispersion before we jump to the  negative binomial model. But in our analysis of CLL, BRCA, and LIHC, we feel it is a good model.

@@@@7mar: more positive comment - we've a suppl. That show goodness of fit.. When we have a enough counts it work.





**Page 22: [30] Moved to page 26 (Move #10)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 22: [31] Moved to page 26 (Move #10)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 23: [32] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

to the heterogeneity of the data.

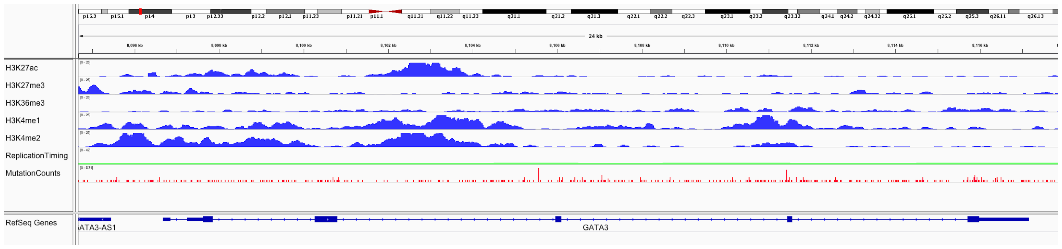
**Page 24: [33] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscrip t | |

**Page 25: [34] Deleted** jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM

@@@@7mar - more & be positive - we've thoguth band we

We thank the referee for accurately pointing out this problem. The current power analysis, which is also mentioned in previous literatures [[cite. Jz2add]] assumes that all the functional sites are within the test regions, is a fairly strong assumption and usually far away from the truth. In some cases, we do feel that the true functional sites might be allocated across various coding/noncoding elements. One example is the GATA3 case, there might be some mutational hotspots outside of the coding regions only. Some kind of joint test might increase detection power.



Actually this is the reason why we are proposing testing the extended gene regions. To illustrate this concept, we added a whole section of new power analysis in our supplementary file to discuss cases when and how power can be increased by joint testing.

@@@@7mar: write more and be more positive

**Page 26: [35] Moved to page 28 (Move #11)** jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 26: [36] Deleted** jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM

| | |
|---|---|
| Referee Comment | 5) Some of the QQ-plots in supplementary figures look problematic. Also, for some tumors with low count[1] |

| | |
|---|---|
| | `statistics QQ-plots are expected to always be deflated, so the interpretation of QQ-plots may be non-trivial.` |

**Page 28: [37] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | @@@@7mar: the ref to point this out ... we actually did have some in the submission... was buried in the supp. we 've tried to make |

-- Ref 2

**Page 32: [38] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

-- Ref 3.2 – BMR --

$$$

**Page 33: [39] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

-- Ref 3.4 – Untitled --

$$$

**Page 33: [40] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

Untitled --

$$$

**Page 34: [41] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

-- Ref 3.6 – Untitled --

$$$

**Page 36: [42] Moved to page 39 (Move #12)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 36: [43] Moved to page 39 (Move #12) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 36: [44] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

Untitled --

$$$

**Page 38: [45] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | *(See Supplement)* |

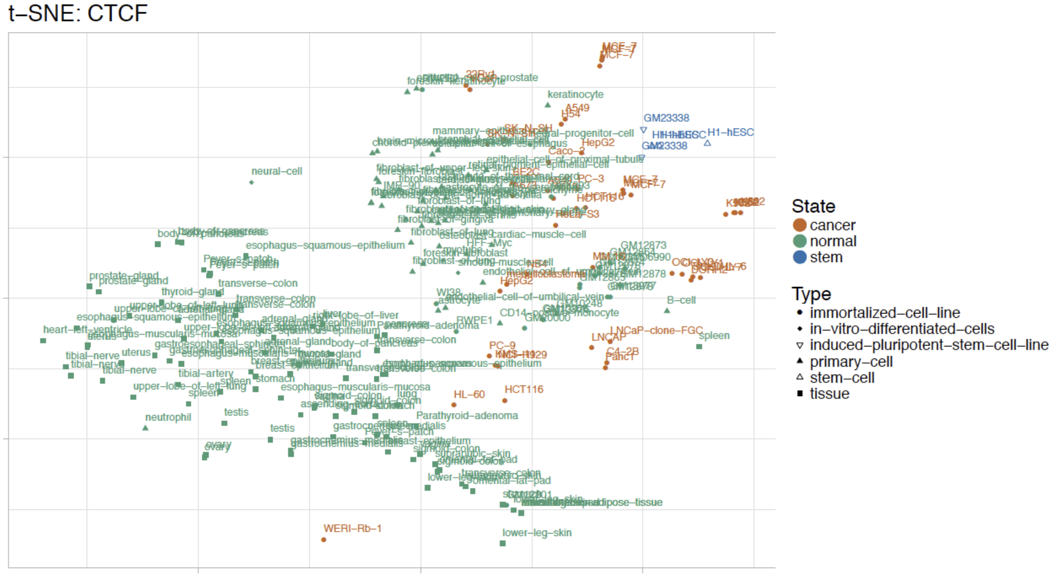
**Page 39: [46] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

-- Ref 4.4 – Can you validate the cell line results using tissue data --

$$$

**Page 40: [47] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

**Page 40: [48] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**



@@@@

**Page 42: [49] Moved to page 47 (Move #13) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 42: [50] Moved to page 47 (Move #13) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 45: [51] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

We admit that H1-hESC may not be the most ideal stem cells to compare with tumor phenotype.

**Page 46: [52] Moved to page 48 (Move #14) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 46: [53] Moved to page 48 (Move #14) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 47: [54] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

**Page 47: [55] Moved to page 50 (Move #15) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 47: [56] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

&&&& SVs

BMR

@@@@7mar - great suggestion [[[@@@@thxu @@@@gr8 @@@@fight

**Page 48: [57] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 50: [58] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 50: [59] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

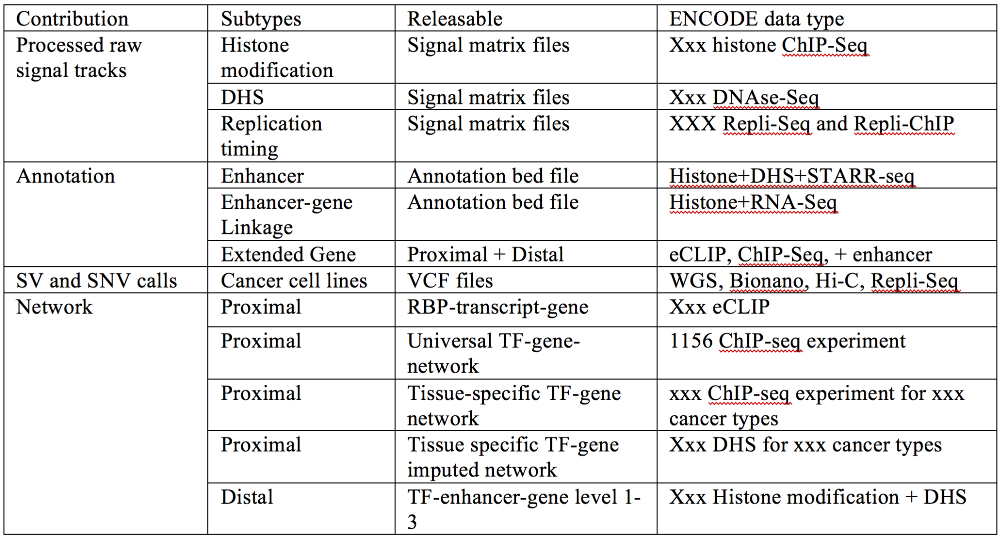
| | |
|---|---|
| Referee Comment | 7) When analyzing the BMR in cancer, did the author estimate the mutation rate in the lncRNAs? Is there any other interesting lesson from the analysis[2] of the non-coding regions and their mutations rate? |

**Page 56: [60] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

## -- Ref 5.2 – Untitled @@@@7mar: change title--

### $$$Presentation $$$Text @@@WM @@@JZ @@@PDM&&&compl

| | |
|---|---|
| Referee Comment | it is not clear what are the main findings in the paper and their statistical and biological significance. The manuscript seems to be somewhat confused between a perspective piece or a guide to ENCODE data for the cancer community (which should be published in a more specialized journal), and a genomics study with clear findings. |
| Author Response | We thank the referee for pointing this out. We want to make it clear that his paper is to be considered as a "resource" paper, not a novel biology 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. |



| Contribution | Subtypes | Releasable | ENCODE data type |
|---|---|---|---|
| Processed raw signal tracks | Histone modification | Signal matrix files | Xxx histone ChIP-Seq |
| | DHS | Signal matrix files | Xxx DNAse-Seq |
| | Replication timing | Signal matrix files | XXX Repli-Seq and Repli-ChIP |
| Annotation | Enhancer | Annotation bed file | Histone+DHS+STARR-seq |
| | Enhancer-gene Linkage | Annotation bed file | Histone+RNA-Seq |
| | Extended Gene | Proximal + Distal | eCLIP, ChIP-Seq, + enhancer |
| SV and SNV calls | Cancer cell lines | VCF files | WGS, Bionano, Hi-C, Repli-Seq |
| Network | Proximal | RBP-transcript-gene | Xxx eCLIP |
| | Proximal | Universal TF-gene-network | 1156 ChIP-seq experiment |
| | Proximal | Tissue-specific TF-gene network | xxx ChIP-seq experiment for xxx cancer types |
| | Proximal | Tissue specific TF-gene imputed network | Xxx DHS for xxx cancer types |
| | Distal | TF-enhancer-gene level 1-3 | Xxx Histone modification + DHS |

@@@@7mar: lump these together

Our goal is to integrate a number of assays (e.g. replication timing, STARR-seq and Hi-C), to provide deep, integrative annotations and various networks across many cell types.

**Page 56: [61] Moved to page 59 (Move #16) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 56: [62] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

## -- Ref 5.3 – Novelty of the paper --

## $$$NoveltyNeg $$$Text @@@WM @@@PDM @@@JZ &&&compl

| | |
|---|---|
| Referee Comment | As it is, the manuscript falls short of the novelty characteristic of publications in Nature. The main concepts presented in this manuscript have been explored extensively before; albeit not with the same amount of ENCODE data specifically (e.g. Martincorena et al (2017); Lawrence et |

| | |
|---|---|
| | al (2013); Polak et al (2015); Imielinski (2017); Roadmap Epigenomics). The cancer genome community has been using ENCODE and Roadmap data in various ways, including in papers such as Tomokova et al. (2017), Schuster-Böckler and Lehner (2012), Frigola et al. (2017), Sabarinathan et al. (2016), Morganella et al. (2016), Supek and Lehner (2015). There is no clear comparison to prior work and no demonstration of improved results compared to those in the literature. |
| Author Response | @@@@7mar: data matrix of publication, table showing importance<br>@@@@7mar:<br>Fight back: There is no clear comparison to prior work and no demonstration of improved results compared to those in the literature.<br><br>We thank the referee to point out many related references. We tried to cite some of the in our manuscript. But note that some important reference, such as Martincorena 2017, came out after our submission in Aug 2017. As a summary, we listed the papers above into the following paper for comparison. |

| 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, since this paper is out 3 months after our submission | Cited | Use 169 epigenetic features to predict gene level BMR | No replication timing data is used |
| Imielinski (2017) | No | Yes | | |
| 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 | NOT specifically for BMR |

We agree with the reviewer that the concept of using genomics features can help to estimate BMR. However, our goal in this manuscript is to demonstrate that ENCODE data is quite useful for a variety of models[3], rather than to develop a novel cancer driver detection method. The BMR part takes only two sub-panels of Fig. 2, and we do have many other aspects in the manuscript to go beyond this point. For example,

| | |
|---|---|
| | 1. We provided accurate noncoding annotation by integrating multiple novel assays such as Hi-C and STARR-seq, which may increase power in somatic mutation burden test.<br>2. We integrated more than 1000 ChIP-seq/eCLIP experiments to provide detailed TF/RBP networks. By combining cohort RNA-seq data, we identified both known (TP53 and ESR1) and novel (SUB1) cancer-associated regulators<br>3. Through whole genome sequencing data, we provided high-quality SV calls in top cancer cell lines, and investigate their effects on enhancers and networks.<br>4. For the first time, we have incorporated thousands of ChIP-seq experiments to directly observe the tumor-to-normal network perturbations and quantify it such changing events |
| Excerpt From Revised Manuscript | @@@@7mar: copy the same block, keep the referees separated |

## -- Ref 5.4 – BMR --

$$$BMR $$$Text @@@JZ @@@PDM @@@WM&&&TBC

**Page 57: [63] Moved to page 61 (Move #17) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

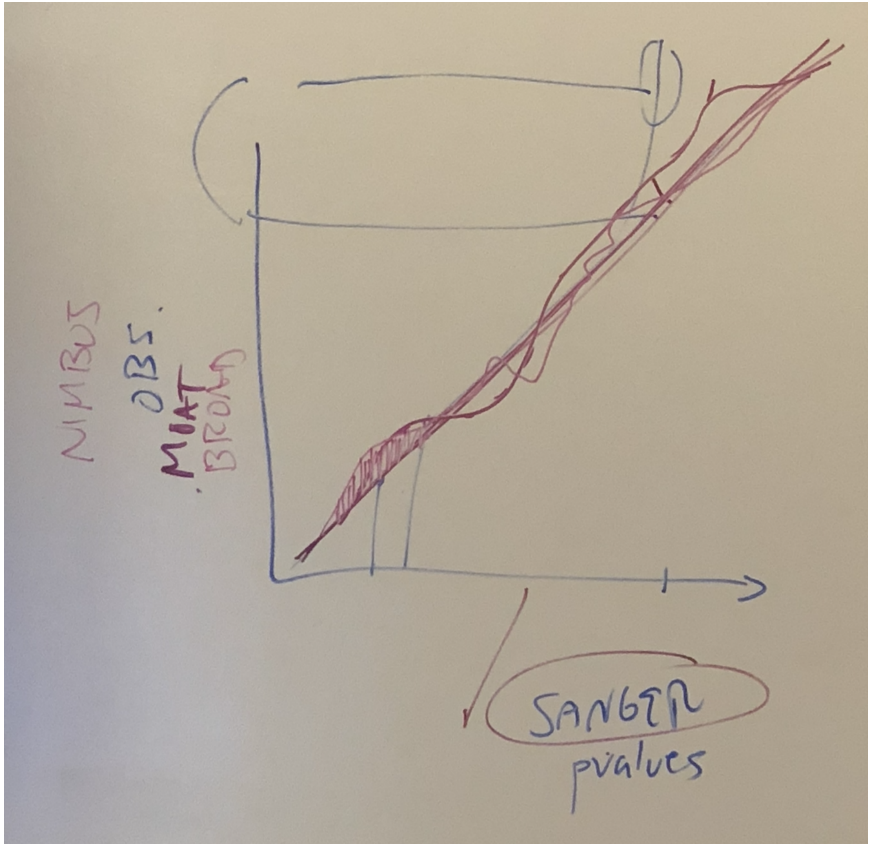
**Page 57: [64] Moved to page 61 (Move #17) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 57: [65] Deleted** jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM

Non driver TCGA gene (remove cancer genes)
Calc bmr and compare with benchmark?

**Page 58: [66] Deleted** jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM



@@@@

**Page 59: [67] Deleted** jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 60: [68] Deleted** jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM

JZ's presentation
@@@@7mar outofscope + combine w/ 5.8

**Page 60: [69] Deleted** jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM

We thank for the referee to point this out. In our revised manuscript, we have added a whole new section in the supplementary file to discuss this problem. In summary, previous power calculations was based on the assumption that all functional sites are within the test region, hence it is better to have short and accurate annotations. However, we found that this assumption is pretty strong and is not realistic for some cases.

Instead, we added a whole section where some functional sites are allocated across multiple regions and then a combined strategy is better.

**Page 61: [70] Moved to page 64 (Move #18) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 61: [71] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

(JZ's presentation)

@@@@7mar - we this is reasonable... we agree... we'll calc some fdr
We need to write 1 para in the main text that summarizes this with some numbers

**Page 62: [72] Moved to page 65 (Move #19) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 62: [73] Moved to page 65 (Move #19) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 64: [74] Moved to page 66 (Move #20) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 64: [75] Moved to page 66 (Move #20) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 64: [76] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

JZ:

@@@@7mar -

**Page 64: [77] Moved to page 67 (Move #21) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 65: [78] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

-- Ref 5.12 – Signature & Mut. rate --

$$$

**Page 65: [79] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 66: [80] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 67: [81] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | @@@@7mar power analysis - please realize our goal is not to do driver discover. |

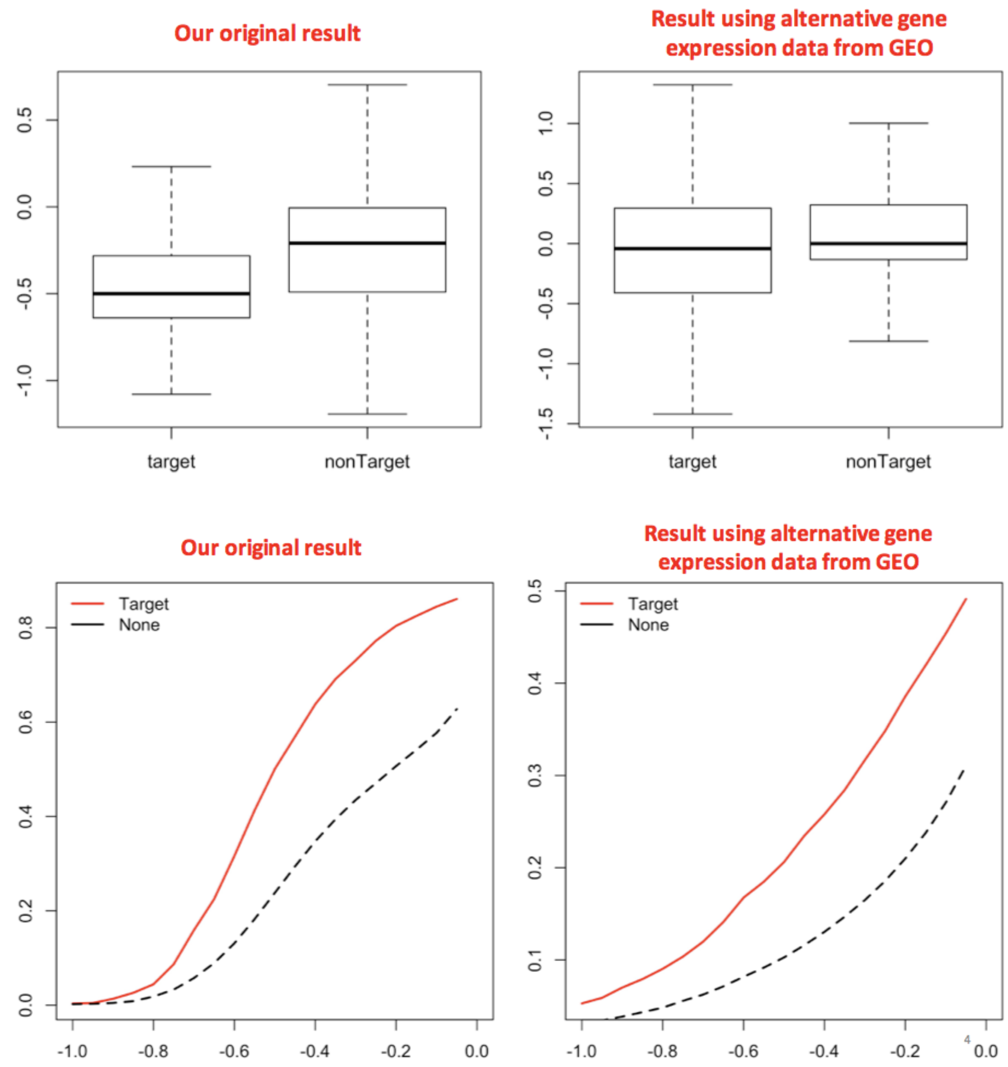
**Page 69: [82] Moved to page 76 (Move #22) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 69: [83] Moved to page 76 (Move #22) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 70: [84] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**



Our original result

Result using alternative gene expression data from GEO

Our original result

Result using alternative gene expression data from GEO

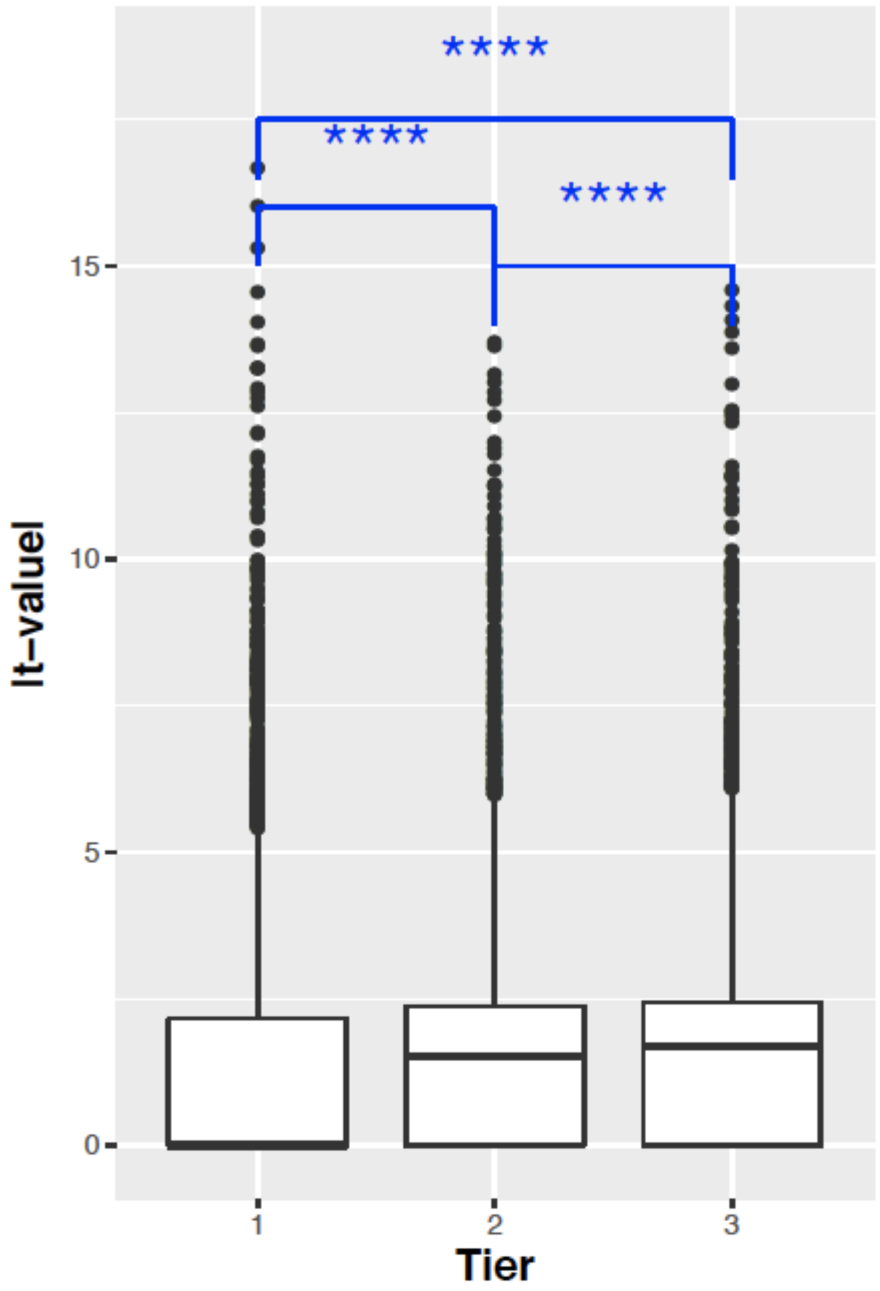
**Page 72: [85] Moved to page 78 (Move #23)  jingzhang.wti.bupt@gmail.com  3/17/18 5:29:00 PM**

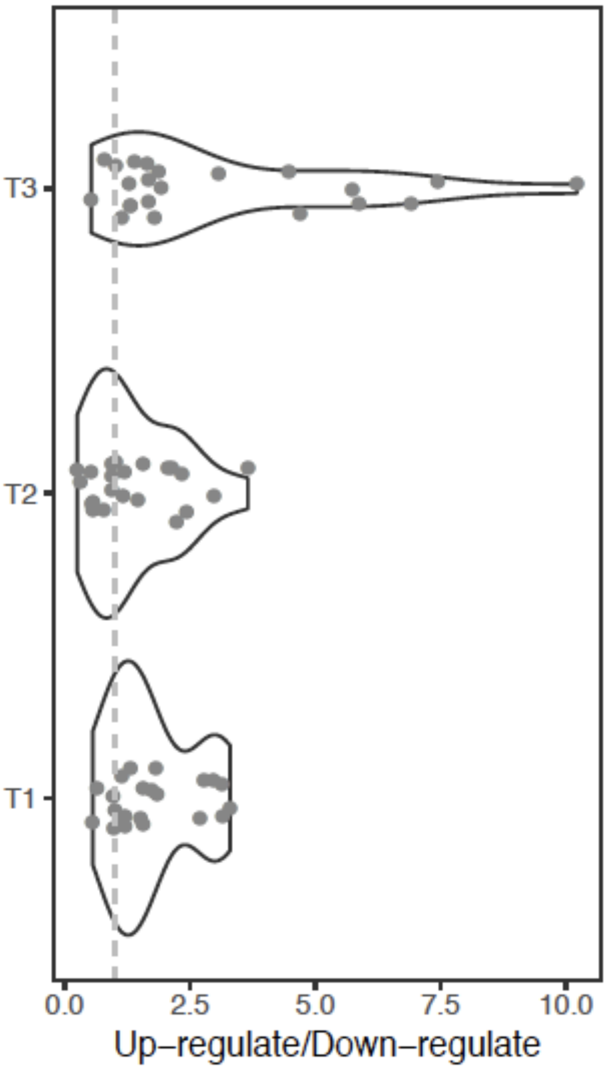
| | |
|---|---|
| Excerpt From Revised Manuscript | |

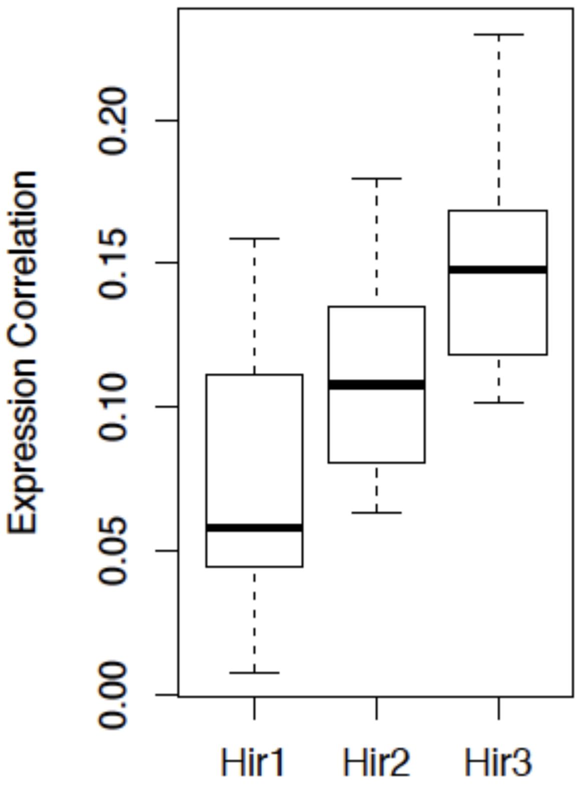
**Page 72: [86] Moved to page 78 (Move #23)  jingzhang.wti.bupt@gmail.com  3/17/18 5:29:00 PM**

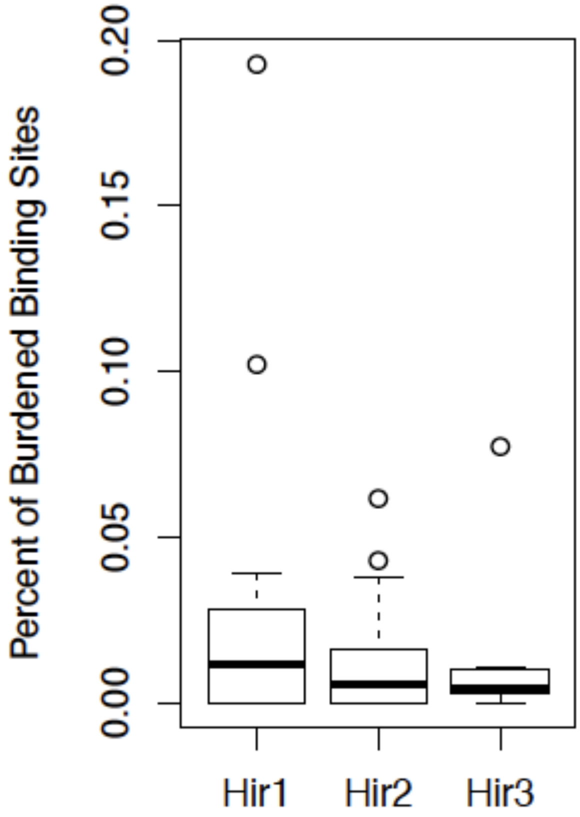
| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 73: [87] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**




T3
T2
T1
0.0
2.5
5.0
7.5
10.0
Up−regulate/Down−regulate

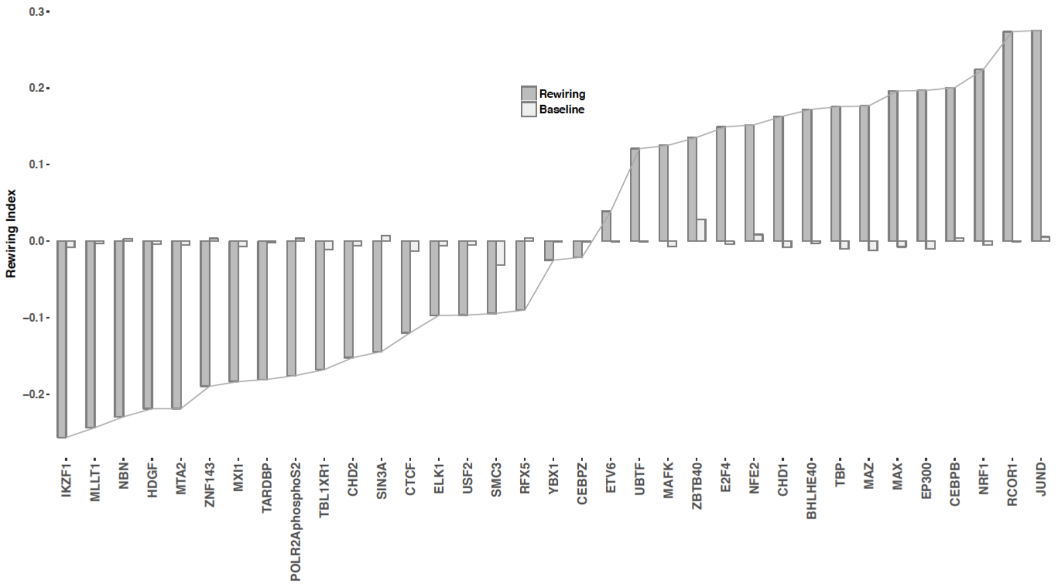

Expression Correlation
0.00
0.05
0.10
0.15
0.20
Hir1
Hir2
Hir3

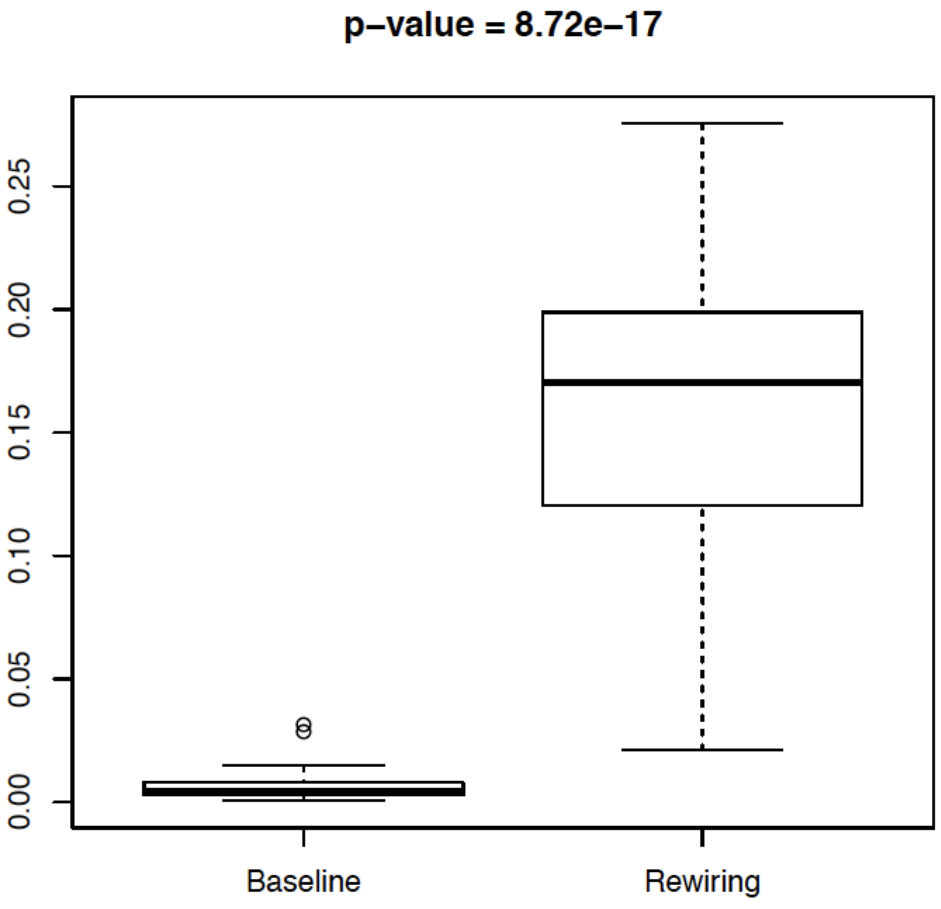


| Page 75: [88] Deleted | jingzhang.wti.bupt@gmail.com | 3/17/18 5:29:00 PM |
|---|---|---|

@@@@7mar we TRULY thank the referee (great praise)

| Page 75: [89] Deleted | jingzhang.wti.bupt@gmail.com | 3/17/18 5:29:00 PM |
|---|---|---|





Page 76: [90] Moved to page 79 (Move #24) jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 77: [91] Moved to page 80 (Move #25)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 77: [92] Moved to page 80 (Move #25)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 77: [93] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

JZ, DL: we can answer

**Page 78: [94] Moved to page 81 (Move #26)** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 79: [95] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

## -- Ref 5.25 – (Minor) How related to FunSeq --

## $$$

**Page 79: [96] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

How are we diff funseq
BMR
Rewiring
Tissue specific

**Page 79: [97] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 80: [98] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

## -- Ref 5.26 – (

**Page 80: [99] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

## ) BMR --

$$$

**Page 80: [100] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 80: [101] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

-- Ref 5.27 – (

**Page 80: [102] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

) Untitled --

$$$

**Page 80: [103] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

| | |
|---|---|
| Excerpt From Revised Manuscript | |

**Page 81: [104] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

-- Ref 5.28 – (

**Page 81: [105] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

) Untitled --

$$$

**Page 82: [106] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

-- Ref 5.30 – Untitled --

$$$

**Page 83: [107] Deleted jingzhang.wti.bupt@gmail.com 3/17/18 5:29:00 PM**

-- Ref 5.31 – Untitled --

$$$

**Page 83: [108] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

-- Ref 5.32 – Untitled --

$$$

**Page 84: [109] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

-- Ref 5.33 – Untitled --

$$$

**Page 84: [110] Deleted** **jingzhang.wti.bupt@gmail.com** **3/17/18 5:29:00 PM**

-- Ref 5.34 – Untitled --

$$$