Tags:

Use comma for seperation between tags

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

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

Usage example:

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

Style Definition: Heading 1
Style Definition: Heading 2
Style Definition: Heading 3
Style Definition: Heading 4
Style Definition: Heading 5
Style Definition: Heading 6
Style Definition: Title
Style Definition: Subtitle
Formatted: Font color: Red
Formatted Table

Deleted: DONE : Finished . ... [1]
Deleted: done

Deleted: For
Deleted: , use ### only

<STATUS>%%%TBC

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

Editor:

<ID>REF 0.1 - Overall comments on the paper

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

Referee Comment	The referees have raised a range of technical concerns on the analyses, including for the background mutation rate, the need to include statistical significance to support many of the claims, and the limitations of this data including cell lines used.
Author Response	We have tried to respond to extensively revise our manuscript in the new version. In summary, we have answered most of these comments. We felt many of them were good suggestions, so we expanded them in large while conserving the manuscript, particularly the suggestions related to - The overall value of this resource to cancer genomics - Network rewirings - Normal-tumor-stem cell comparisons - SVs statistics on networks - Discovery of SUB1 as a potential new oncogene One area that we wish to [illegible] 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 have shown how in certain contexts, the ENCODE3 date can help with existing driver discovery measures. We also want to emphasize that although some referees mentioned the limitation of cell line data used here, the usage of functional genomics data from tissue of origin is not necessarily a better option, as correctly pointed out by referee 4. The genomic and epigenomic heterogeneity in tumor cells, as well as heterogeneity in the tumor microenvironment are significant factors in tumor growth and development. We tried our best to validate, using external data set, the conclusions we draw from ENCODE call line data and found that our conclusions correlate well with the observations. We added more discussion in the revised manuscript about how technology advances, such as single cell sequencing, can help to provide further insights.

Formatted: Justified
Formatted Table
Deleted: We've
Deleted: our
Deleted: we've
Deleted: suggestion
Deleted: comparison to stem cell, SVs statistics on networks, and SUB1.
Deleted: we've
Deleted: ENCODE
Deleted: Excerpt From ... [2]

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

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

Referee Comment	The referees also find that the current manuscript provides limited context with prior studies using similar approaches for use of prior ENCODE and Epigenome Roadmap datasets in cancer genomics. They detail the need for clearer presentation in context of prior studies as well comparisons to demonstrate advance.
Author Response	We thank the referees for this comment. We want to note that many of the prior studies have been cited in our initial submission. Some papers, such as Martincorena et al 2017, came out after we submitted our paper in Aug 2017, so it is impossible us to cite in the initial submission. In the revised paper, we have clarified the unique aspects of our paper and provided clearer text with previous efforts.

Deleted: >%%%DONE
Formatted: Justified
Formatted Table
Deleted: and
Deleted: Excerpt From ... [4]
Deleted: Excerpt From ... [3]

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

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

Referee Comment	The referees also recommended that the current manuscript does not represent a distinct advance to the main ENCODE manuscript, as it does not report separate new datasets, methods, or clear novel findings. Some referees also recommended that this may be more suitable as Perspective in a specialized journal that further highlights the use on the current ENCODE datasets for cancer genomic studies.
Author Response	We 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

Deleted: >
Deleted: >%%%DONE
Formatted Table

	(2) 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. (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 in both proximal and distal networks. We feel that this is best exemplified in oncogenesis. **(2) Deep, integrative annotation – complementary to the Encyclopedia.** While the encyclopedia paper considers broad, "universal" annotations across cell-types (currently the centerpiece of ENCODE), it focuses on data common to most cell types (DHS, 2 histone marks and 2 TFs). It does not take advantage of the cell types richer in assays -- the other dimension of ENCODE (diagrammed in ENCODEC's first figure). The ENCODEC paper takes a complementary approach, constructing a more accurate annotation using a large battery of histone marks (>10), next generation assays such as STARR-seq and elements linked by ChIA-PET and Hi-C. **(3) Replication Timing.** Although a major feature of ENCODE is replication timing, none of the other papers feature it. Previous work on mutation burden calculation usually selects replication timing data from the HeLa cell line due to the limited data availability. The wealth of the ENCODE replication timing data greatly helps to parametrize somatic mutation rates.

Deleted: that

Deleted: .

	(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	

Referee #1 (Remarks to the Author):

<ID>REF1.0 – Preamble

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

We [redacted] appreciate the referee's feedback. Overall the reviewer mentioned that this is an interesting resource but the novelty of the paper is lacking. We thank the referee for his/her acknowledgement of the potential popularity of our resource for cancer genomics.

Regarding the novelty point, we think differently about the value of our paper. We want to make it clear that this 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 the deep annotations and network changes. We have listed some more details about the resource of this paper as below. Thus, where the referee asks for novelty in cancer gene discovery - we strongly feel that this is out of scope.

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	51 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 564 DNase-seq STARR-seq
	Enhancer-gene linkage	Annotation in BED format	2015 Histone ChIP-seq 329 RNA-seq

Deleted: Done
Deleted: ,,
Deleted: of
Deleted: his
Deleted: novelty
Deleted: ask
Formatted Table
Deleted: 135

	Extended gene	Annotation in BED format	1863 TF ChIP-seq 167 eCLIP Enhancer-gene linkage
SV and SNV callsets	Cancer cell lines	Variants in VCF format	WGS BioNano Hi-C 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 564 DNase-seq

Specifically for the BMR estimation part, the reviewer mentioned that there had 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. However, some of the references was out after our initial submission so we did not have a chance to add them. 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.

Deleted: have

Deleted: and in

Formatted: Font color: Red

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

Deleted: [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.] ... [5]

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

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

Deleted: DONE

Referee Comment	This manuscript describes how the ENCODE project data could be utilized to derive insights for cancer genome analysis. It has several examples to illustrate this point, e.g., how to

Formatted Table

	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.

Deleted: Excerpt From ... [6]

Deleted: Excerpt From ... [7]

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

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

Deleted: DONE

Deleted: [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]

Formatted Table

Referee Comment	Just to take the first application as an example, the problem of estimating background somatic mutation rate accurately in order to better identify cancer drivers has been studied extensively in the literature. One paper, "Mutational heterogeneity in cancer and the search for new cancer-associated genes" (Nature 2013), is cited in the current manuscript, but there are many others. For instance, Weinhold et al, 2014 (Genome-wide analysis of noncoding regulatory mutations in cancer, Nat Genetics), Araya et al, 2015 (Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations, Nat Genetics), and similar non-coding mutation identification papers all include steps to account for epigenetic features in their background rate calculation.
Author Response	We thank the reviewer for identifying these references. We recognize that genomic features have been previously been used to estimate BMR and improve driver mutation detection. Our aim here was neither claim a better BMR estimation model nor claim a novel discovery that "matched" features performs better. We made it

Deleted: epigenetic

Deleted: not to produce novel

Deleted: models, but rather

	clearer in our revised manuscript that our purpose is to showcase how ENCODE data can help BMR estimation in many models. With the wealth data available through ENCODE data, we had a much larger pool of features to choose from to potentially improve BMR estimation. There are thousands of histones modification marks that are released into a ready to use format (see details in table below). In addition, we have provided other data types, such as replication timing, that have been proven to be affect BMR but have not been widely by others. We believe that such data, when released into a ready to format, can help BMR estimation through many existing models.
Excerpt From Revised Manuscript	<table><tr><th>Cell Type</th><th># histone marks</th></tr><tr><td>tissue</td><td>818</td></tr><tr><td>primary-cell</td><td>521</td></tr><tr><td>cell-line</td><td>339</td></tr><tr><td>in-vitro-differentiated-cells</td><td>179</td></tr><tr><td>stem-cell</td><td>114</td></tr><tr><td>induced-pluripotent-stem-cell-line</td><td>46</td></tr></table>

Deleted: improve the performance of such

Deleted: 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.I 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).

Formatted: Justified

Deleted: 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. ... [8]

<ID>REF1.3 – BMR: lacking in novelty in the conclusion

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

50% MATCH

Deleted: of

Deleted: method

Deleted: 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.

Formatted Table

Deleted:

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 did cite this paper in our initial submission. It is worth mentioning that we are not trying to reproduce the discoveries 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. First, To select a perfect "matching" feature (no matter 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 types they investigated (Fig. 2a). The noticeably larger pool of functional characterization data from ENCODE3 can actually help to find a matching issue, especially for cancers types that cannot find an obvious "matching" feature from the Roadmap, such as prostate cancer. Second, 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 predicting BMR does not exclude that other "non-matched" features from being useful to improve 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 thousands of features in a ready-to-use format. Many of them are not mentioned in other literature, such as replication time from 51 tissue/cell lines. They have proven useful but are less frequently matched probably due to the lack of data incorporated into previous BMR models.
Excerpt From Revised Manuscript	

Deleted: also cited

Deleted: discovery

Deleted: not

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

Deleted:

Deleted: cancer

Deleted: way

Formatted: Underline

Formatted: Underline

Deleted: ENCODE

Formatted: Underline

Formatted: Underline

Deleted: on this

Formatted: Underline

Deleted: can not

Deleted: data

Deleted: The

Formatted: Underline

Deleted: prediction

Formatted: Underline

Deleted: mean

Formatted: Underline

Deleted: are not

Formatted: Underline

Deleted: improved

Deleted: 932 histone modification

Deleted: And also listed many

Deleted: features, especially the 51

Deleted: data, that

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

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

[JZ2DS: would you please add xx xx? Also add some text on the CTCF plot]

Deleted: DONE

Deleted:

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. 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 the comparison of cell line vs. tissue. We further investigated this comparison and extended this point more to the RNA-seq and ChIP-Seq data. We think slightly differently with the referee on this point. - On a large scale (up to mbp) • First, the Polak 2015 paper did not perform large-scale comparison across various cancer cell lines. As seen from the following figure, cell line data provides comparable, sometimes even better, correlation with mutation counts. We have added a new section in the supplementary file to discuss this. • As compared to cell line data, there are way less functional characterization data in tissues (such as prostate tissue). We have updated supplementary table 1 for a comparison of data richness in ENCODE3. • ENCODE is not just about cell lines, and there are many ENCODE tissue data for histones (339 cell line vs 818 tissue). We have added a supplementary table on this point. - On a small scale (less than kbp) Features, like expression levels and TF binding events, have been used widely to affect somatic mutation rates. As suggested by the referee, we systematically investigated the RNA-seq and TF ChIP-Seq data and found that many of the cancer transcriptome/TF binding landscape are quite similar to each other, as compared to the initial of primary cells. This has also been mentioned by previous reports, such as Lotem et al. 2005 and Hoadley et al. 2014. The fact that cancer cells lose diversity and showed a distinct pattern from the primary cells highlights

Formatted Table

Deleted: (see updated Figure 5).

Deleted:

Deleted: ... [9]

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

Deleted: , it is not always the case that cell-of-origin can be predicted perfectly using the epigenetic features (Fig. 4 b). ... [10]

Deleted:

Deleted: types. Here we used breast cancer as an example. 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).

Deleted: MCF-7

Deleted: (and

Deleted:).

Deleted: also found that histones from tissue and matched cell lines are actually quite correlated in

Deleted: larger scale (see heatmap below).

Deleted: ... [11]

Deleted: such data. On the contrary, the cell line

Deleted: has lots of advantage in terms of assay richness. For example, there is no data for

Deleted: from the roadmap, but

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

Deleted: like LNCap might further help under such condition

Deleted: 3. Some genomic features

Deleted: proven

Deleted: We

Deleted: scanned all

Deleted: cancerous

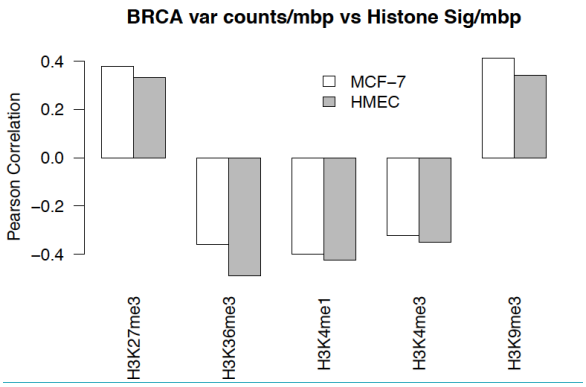
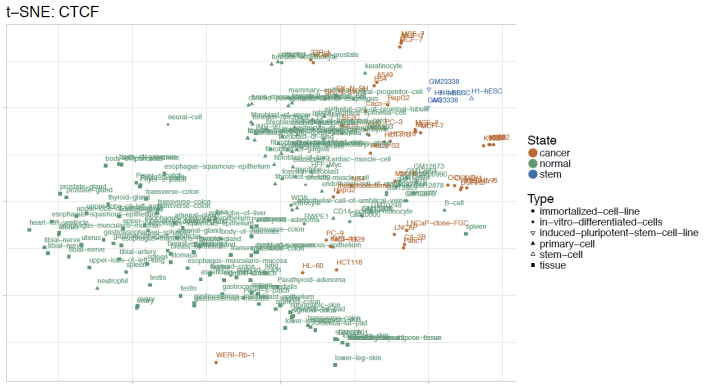
Deleted: non-cancerous cell types from ENCODE

Deleted: Our observation is consistent

Formatted: Font:Helvetica Neue

Deleted: For example, here is the projection of CTCF binding sites from all ChIP-Seq experiments.

Deleted: loose

	the values of cell line data. We have added this result into the main figure and supplementary files.
Excerpt From Revised Manuscript	1. Comparison of mutation rate vs features in tissue/cell lines. We provided the pearson correlation of the breast cancer mutations count per Mbp vs. various histone modification features in tissue and cell line. Cell line data provides comparable (and sometimes even better) correlation with mutation counts.  2. t-SNE plot of xxx CTCF ChIP-Seq data  3. Summary of ENCODE histone ChIP-seq data

Deleted: . [12]

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

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

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

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 others 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

Deleted: DONE
Deleted: ####Dictation ... [13]
Formatted: Justified
Formatted Table
Deleted: other people
Deleted:
Deleted: Seq

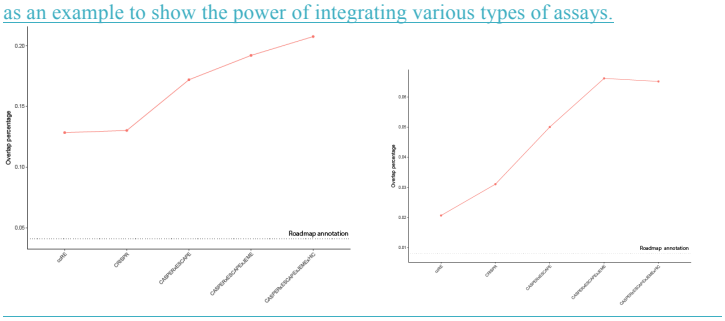
	as an example of the best we can do with this level of integration. The fact that we coupled this with quite successful validation that we believe points to the great value of the integrated incurred data.
Excerpt From Revised Manuscript	

<ID>REF1.6 – Novelty and presentation of the paper

<TYPE>$$$Presentation,$$$NoveltyPos,$$$NoveltyNeg,$$$Text
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%DONE
JZ2MTG: would u pls update the figure? The legend is too small to see and would you please change it to a barplot?

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 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. As for the enhancer part, with the ensemble method, for example, we can get more accurate annotation and pin-point to sequences where transcription factors would actually bind to. To estimate the false positive rate would not be very practical at this stage as there is no gold-standard experiment that could assert an predicted enhancer is definitely negative. Here we took the FANTOM enhancer data set and assess the overlap percentage of our enhancer annotation in each ensemble step. We show that each ensemble step indeed increases the percentage of overlap between our annotation and the FANTOM enhancer set. The overlap percentage for our annotation is much higher than that of the Roadmap annotation, and is also higher than the main encyclopedia enhancer annotation annotation (ccRE).

Deleted:
Deleted:
Deleted:
Formatted Table
Formatted: Justified
Deleted:

Excerpt From Revised Manuscript	We have performed QC of different types of enhancers in details in K562 and GM12878 as an example to show the power of integrating various types of assays. 

Referee #2 (Remarks to the Author):

<ID>REF2.0 – Preamble

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

Deleted: Done

We would like to appreciate the referee's feedback, especially about the positive comments on the value of our resource, the extended gene, and the network rewirings. Regarding the novelty point, Regarding the novelty of our work, this paper is unique in its highlighting of a number of ENCODE assays (e.g. replication timing, TF knockdowns, STARR-seq, ChIA-PET, and Hi-C), its deep, integrative annotations combining a wide variety of assays in specific cell types, and its analysis of networks. Note also that while we do NOT feel this is a cancer genomics paper, we 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 the novelty of this paper as below.

Deleted: paper, our

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

Formatted Table

	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 564 DNase-seq STARR-seq
	Enhancer-gene linkage	Annotation in BED format	2015 Histone ChIP-seq 329 RNA-seq
	Extended gene	Annotation in BED format	1863 TF ChIP-seq 167 eCLIP Enhancer-gene linkage
SV and SNV callsets	Cancer cell lines	Variants in VCF format	WGS BioNano Hi-C 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 564 DNase-seq

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

<TYPE>$$$NoveltyPos
<ASSIGN>

<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

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

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

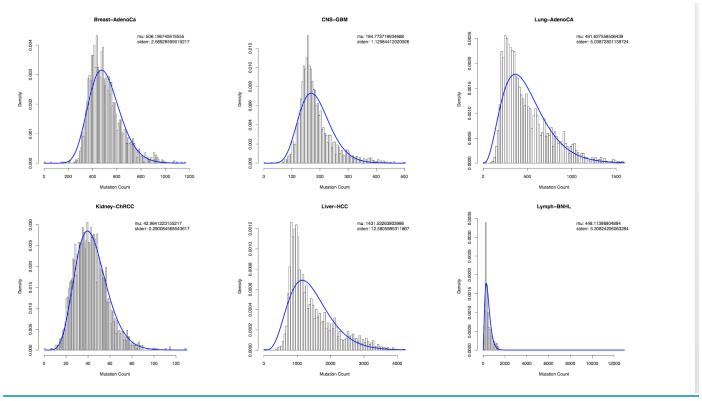
<TYPE>$$$BMR,$$$Text,$$$Cat
<ASSIGN>@@@JZ
<PLAN>&&&OOS
<STATUS>%%%100DONE

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	The referee is pointing out that negative binomial regression has been used before. This is a standard statistical technique that has been used in many contexts. 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. ENCODE3 provides noticeably more covariate data, which is uniformly processed and less explored in the references mentioned by the referees. There is new data type, such as replication timing, that is well-known confounders but not included in those papers. 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. Therefore, we did not directly use their approach.

Deleted: DONE
Formatted Table
Formatted: Justified
Deleted: Excerpt From ... [14]
Deleted: Excerpt From ... [15]
Deleted: DONE
Formatted Table
Deleted: In relation to the negative binomial regression, the
Deleted: the use of
Deleted: that's be
Moved (insertion) [1]
Moved (insertion) [2]
Deleted: The fact that it was earlier used in relation to background mutational rate shows that it
Deleted: an appropriate approach
Deleted: Excerpt From ... [16]
Moved up [1]: 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
Moved up [2]: While we admit it does slightly undercut a claim of novelty in this regard, that's not central to our work.
Deleted: Excerpt From ... [18]
Deleted: Excerpt From ... [17]

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

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

Referee Comment	Also, does Gamma-Poisson model fits data for most cancers well or is it just an approximation? One can use non-conjugate priors but this is probably beyond the scope of this work.
Author Response	We thank the referee for mentioning the goodness of fit of the Gamma-Poisson model. As suggested, we provided more figures in our supplementary file to investigate this. For most of the cancer types, the fitting of Gamma-Poisson is pretty good (as seen in the figures below). Also, we point out the fact that it has been used in other literature provides further technical support for this using. However, we agree that it is interesting to investigate other non-conjugate priors. As the referee mentioned, this is out of scope, but we have made a mention of this in the text.
Excerpt From Revised Manuscript	

Deleted: DONE

Formatted Table

Formatted: Justified

Deleted: pointing out

Deleted: problem and he/she is right that

Deleted: didn't' provide enough background. Following the referee's suggestion, we made new

Deleted: figures as requested. In

Deleted: .

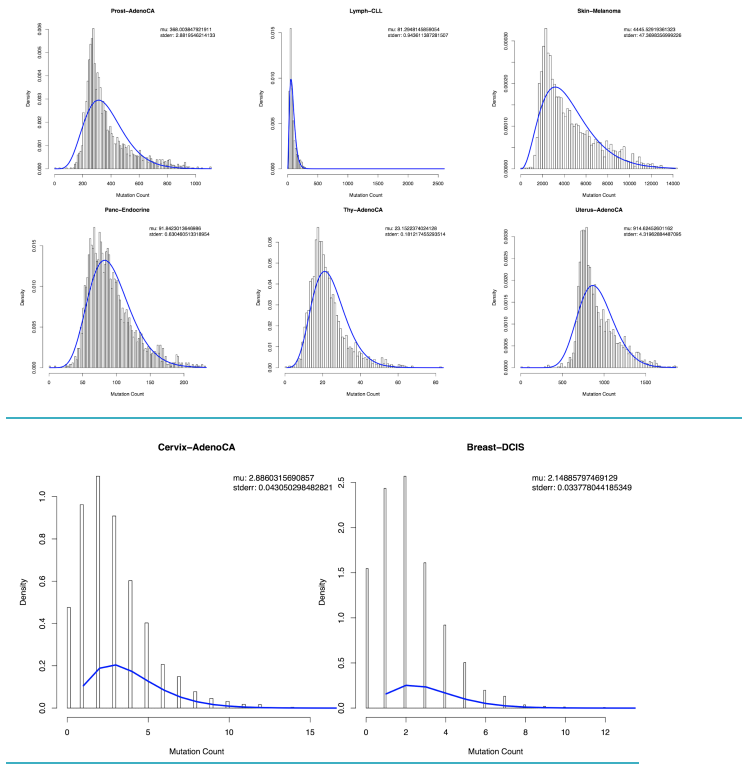
Deleted: that Inigo uses that and justifies andthis

Deleted: we choose Gamma-Poisson conjust it might

Deleted: .

Deleted: .

... [19]

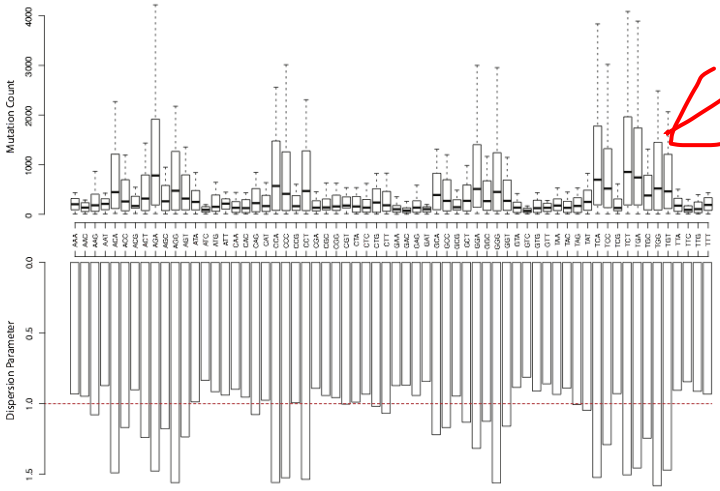
<ID>REF2.4 – Was the Poisson Model used for low mutation cancers

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

Deleted: 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?

Formatted Table

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 those 3mers with more variants, there is a tendency to introduce overdispersion and accept the Gamma-Poisson model. It could be either the power issue, or the level of heterogeneity among samples, or even both. We have 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. However, we think it is better to check how heterogeneous the count data is even after correcting enough covariate effects.
Excerpt From Revised Manuscript	

Deleted: higher number of

Deleted: of larger

Deleted: .

Deleted: might

Deleted: . A larger variation usually accepts the Negative binomial distribution. We've

Deleted: But

Deleted: ... [20]

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

<TYPE>$$$BMR,$$$Calc

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix
<STATUS>%%%50DONE

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 bring out this point and we don't see this as efficient either. The point of our approach is not to say that a few top components or a few features can predict a mutation rate accurately. Actually we want to show the opposite that the wealth of the ENCODE data is useful and that with additional data types, one gets a small but measurable continued improvement. 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 have replotted this figure and now simply show how one gets steady increase in predictions forms by just adding features one at a time. 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 ENCODE data provides a valuable resource for predicting BMR and we appreciated the referee helping us achieve clarity on this point. We put the main text figures into the supplementary files and made for the main.
Excerpt From Revised Manuscript	1. At 1mb bin resolution, we compared the performance of models using random features vs. computationally selecting best features sequential (forward selection). It has shown that by adding features appropriately from ENCODE3, we can noticeably improve the performance of BMR accuracy.

Deleted: DONE

Formatted Table

Deleted: component

Deleted: this out

Deleted: correctly

Deleted: actually

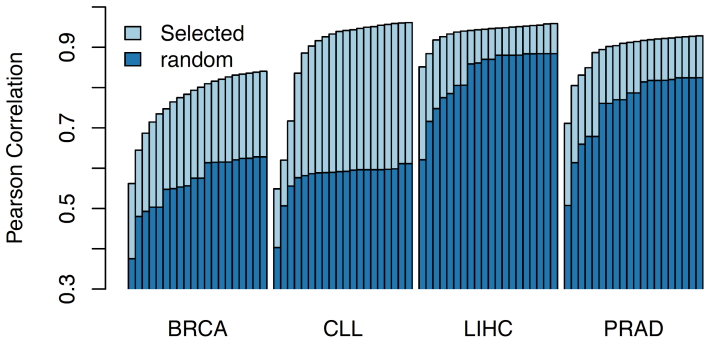
Deleted: in background mutation estimation. This may be because of the heterogeneity and the difficulty in matching samples, but may due to the correlated nature of the features themselves

Deleted: we've redone

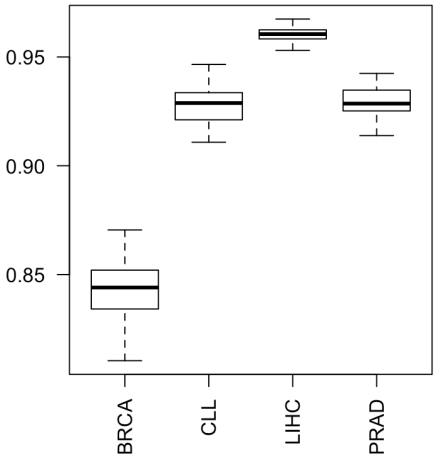
Deleted: end code

Deleted: corpus

Deleted: background mutation rate



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



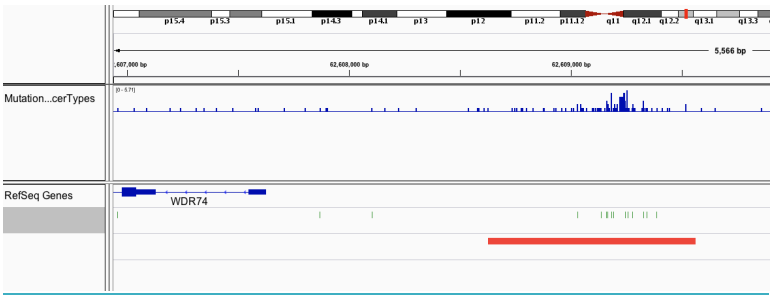
<ID>REF2.6 – Comments on the power analysis and compact annotations

<TYPE>$$$Power,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

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 calculation in our revised manuscript. In our initial submission, the assumption is that we were trimming off the nonfunctional sites while preserving the functional ones. Two examples can explain the motivation of this assumption. 1) Enhancers: Traditionally, enhancers were called as a 1kb peak regions, which admittedly introduced a lot of obviously nonfunctional sites. We believe we can get functional region more accurately by trimming the enhancers down using the exact shapes of many histone marks and further integration with STARR-seq and Hi-C data. 2) TFBS hotspots around the promoter region of WDR74. Instead of testing the conventional up to 2.5K promoter region, we can trim the test set to a core set of the promoter region where many TFs bind, which perfectly correlates with the mutation hotspots (red block) for this well-known driver site (blue line for pan-cancer and green line for liver cancer). Following the reviewer's suggestions, in our revised manuscript we show in a formal power analysis that the most important contribution to power comes from

Deleted: Done
Formatted: Font color: Red
Formatted: Justified
Formatted Table
Deleted: calculated
Deleted: not
Deleted: truly
Deleted: or important sites, but rather trimming unimportant sites. For instance, in the old way that we found enhancer sites by just calling
Deleted: 1KB region from a
Deleted: by almost any estimation included knots
Deleted: non
Deleted: sites. Trimming this
Deleted: a large battery
Deleted: the exact shape of the signal, we believe more accurately gets it the truly functional region, particularly when coupled
Deleted: accurate
Deleted: will hopefully increase power. Another case is the TF binding hotspot
Formatted: Underline
Deleted: without prior information
Deleted: TF binds to
Deleted:

Deleted:
Deleted: now in the supplement

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

Deleted: this

Deleted: However

Deleted: admittedly

Deleted: we're basically

Deleted: encode

Deleted: allows

Deleted: at

Deleted: functionally important

Deleted: It's

Deleted: really

Deleted: success

Deleted: It's

Deleted: I

Deleted: We've

<ID>REF2.7 – Q-Q plots

<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&Defer
<STATUS>%%%TBC
####Thinking
[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. We've done XXX & YYY now But we wish to make clear that the point of this paper is not driver detection

Formatted Table

	Our goal is BMR We show QQ w diff detection We actually show QQ plots with drivers Take some else's driver detection method, use our BMR model, show that it works better
Excerpt From Revised Manuscript	

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

<TYPE>$$$NoveltyPos
<ASSIGN>
<PLAN>&&&AgreeFix,&&&MORE
<STATUS>%%%75DONE
[JZ2JL: please add your figure here]

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 several new sections to highlight the value of extended genes, such as 1. We extensively expanded our power analysis part to include more extended gene analysis (as we pointed up in the response to <ID>REF2.6 – Comments on the power analysis and compact annotations) 2. We showed that by using the extended gene, we can better stratify the gene expressions
Excerpt From Revised Manuscript	

Deleted: DONE

Formatted Table

Deleted: a whole

Deleted: section of how

Deleted: could increase statistical power.

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

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

Deleted: DONE

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.

Formatted Table

Deleted: Excerpt From ... [21]
Deleted: Excerpt From ... [22]

<ID>REF2.11 – Confounding factors

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

Deleted: 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 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.

Formatted Table

Excerpt From Revised Manuscript	

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

<TYPE>$$$Power,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

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 expanded our power analysis in the revised manuscript.
Excerpt From Revised Manuscript	We showed in a formal power analysis that the most important contribution to power comes from including additional functional sites, which is of course by the extended gene concept 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. Admittedly, we are making assumptions and the referee is completely correct in pointing this out. We have tried to be more precise in the text that we are assuming that the large number of ENCODE assays, when integrated, allow us to more directly get the functional nucleotides, but this, of course, is an assumption. It is hard to tell to what degree one can succeed in finding the current events in cancer. It is hard to back this up with the gold standard, but we think that some of the points are self evidently obvious. We have tried to make this clear in text and thank the referee for pointing this out.

MERGE

Deleted: Done

Deleted: [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]

Formatted Table

Deleted: attempted to do this

Deleted: suppl figure XXXX

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

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

<STATUS>%%%75DONE

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

<ID>REF2.14 – Minor comment on terminology

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

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

Deleted: DONE
Formatted Table
Deleted: Excerpt From ... [23]
Deleted: Excerpt From ... [24]
Deleted: DONE
Formatted Table
Deleted: Excerpt From ... [25]
Deleted: Excerpt From ... [26]

Referee #3 (Remarks to the Author):

<ID>REF3.0 – Preamble

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

Deleted: 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

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

Referee Comment	It is difficult to understand the significant novel findings in this paper (compared to the main ENCODE paper). Perhaps, some of this is due to the data not being presented in a concise and clear manner. For example, I wonder whether the authors can add more details and straightforward directions when citing supplementary information. In the current main manuscript, the authors cited all supplementary information as (see suppl.). It might be hard for the reader to check where the authors refer to in the supplementary information. I think more direction, such as sup Fig1, sup Table 1, or section 7.2S etc, would be very helpful.

Formatted Table

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

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

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

Referee Comment	In the second paragraph of page 3, it says 'using matched replication timing data in multiple cancer types significantly outperforms an approach in a which one restricts the analysis to replication timing data from the unmatched HeLa-S3 cell line.' This statement is confusing and does Figure 2A or 2B supported it?
Author Response	
Excerpt From Revised Manuscript	

Formatted Table

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

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

Referee Comment	In Figure 1, "top tier" should point to cell types that is mentioned in the content. However, we also see SNV, SV, Mutation, etc.
Author Response	
Excerpt From Revised Manuscript	

Formatted Table

<ID>REF3.4 – Regarding enhancer detection algorithm

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

Referee Comment	What is a single shape algorithm? The authors point to Supplementary data, but there is no definition there either. Do the authors mean the complete graphs or connected components?
Author Response	
Excerpt From Revised Manuscript	

Formatted Table

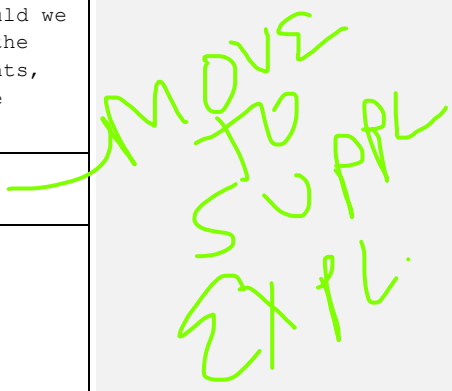
<ID>REF3.5 – Regression coefficients of BMR

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

<STATUS>%%%TBC

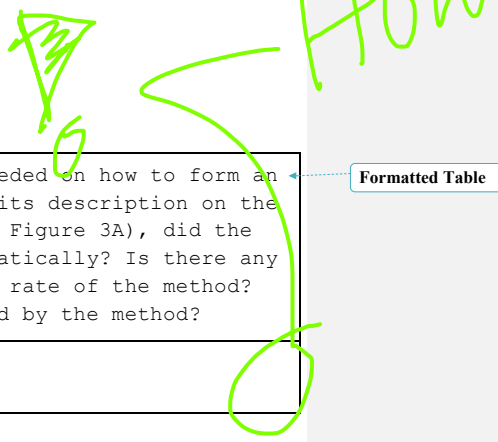
Referee Comment	For Figure 2B, what does 'regression coefficients of remaining features' mean? Does that means beta_0 or the remaining regression noise? From Figure 2B, the coefficient to regression is rounded to -0.001 and 0.001. How should we understand these values? If the coefficients are for the main features, we would be expecting higher coefficients, wouldn't we? In this case, does it means the lower the better?
Author Response	
Excerpt From Revised Manuscript	

Formatted Table



<ID>REF3.6 – Validation of extended gene

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



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 Table

Excerpt From Revised Manuscript	

<ID>REF3.7 – Logic gates

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

Referee Comment	Are circuit gates necessary for Fig 3B? There are OR, AND and NOT gates used. For Figure 3C(i), what is the meaning of the values between the green and yellow dots (MYC and *)? The figure legends are not explaining the figure very well and many details are omitted.
Author Response	
Excerpt From Revised Manuscript	

Formatted Table

WE HAVE REDRAWN FIG.

<ID>REF3.8 – Network hierarchy

<TYPE>$$$Hierarchy
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%50DONE

Deleted: TBC

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 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 Table

Referee #4 (Remarks to the Author):

<ID>REF4.0 – Preamble

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

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 and 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.

<ID>REF4.1 – Strengths of the Paper

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

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.

Deleted: Done
Deleted: ... [27]
Formatted: Font color: Black
Deleted: &
Deleted: Done
Formatted Table
Deleted: Excerpt From ... [28]
Deleted: Excerpt From ... [29]

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

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

Referee Comment	Yet, efforts to make the manuscript more readable will be quite important. For instance, I could understand several sections of the manuscript after reading carefully the not so short supplementary part. The strategy of sample selection was easier to understand after seeing the first figure of the supplementary information, as well as fig S1-3 regarding the number of normal vs cancer cell lines. I'm not sure what the space limitation for this manuscript will be, but clarity should be an important component of a Nature paper.
Author Response	We thank the referee for pointing out that it is sometimes hard to see the full documentation of our methods in the main part and one has to look at the extensive supplements. We are well aware of this fact. The very large scale of the supplement is typical for large genomic paper. We, in fact, have been actively discussing with Nature Publishing and other companions about the supplement with regard to the main text. We have attempted to put important contents in the supplement and to structure it very carefully. We admit that maybe this construction is not that intuitive. We are prepared to work very hard to make the structure of the supplement understandable. We have tried to revise it to make these clearer and also to move more into the main text, though we think given the current main text limitations of a typical paper in Nature and the scale of the results in the data in this paper, it is not easy to put everything into the main text. We are preparing to work constructively with the referees and the others to make this clear.
Excerpt From Revised Manuscript	

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

<TYPE>$$$Text,$$$Presentation

Deleted: Done
Formatted Table
Deleted: In relation to the supplement and genomics,
Deleted: points
Deleted: it's
Deleted: things
Deleted: We've
Deleted: appointives
Deleted: nature
Deleted: it's simply impossible

<ASSIGN>@@@DC,@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

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

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

<TYPE>$$$CellLine,$$$Validation
<ASSIGN>@@@JZ,@@@DL,@@@Peng
<PLAN>&&&MORE
<STATUS>%%%50DONE

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 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).
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

Deleted: Done
Formatted Table
Deleted: we've
Deleted: Excerpt From ... [31]
Deleted: Excerpt From ... [30]
Deleted: >
Deleted: Done
Formatted Table
Deleted: ... [32]
Formatted: Font:Helvetica Neue
Formatted: Font:Helvetica Neue
Formatted: Justified

extensively explored the chromatin landscape and expression patterns across all of available ENCODE primary cells and tissues, and compared them 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 examples 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.

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

####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

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)

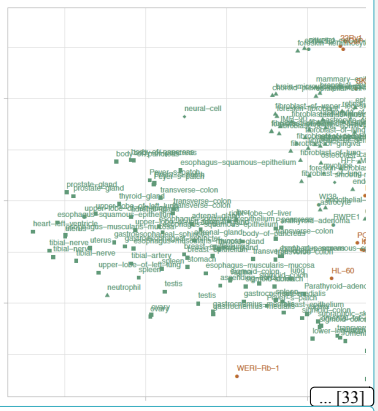
Human cancers overexpress genes that are specific to a variety of normal human tissues
http://www.pnas.org/content/102/51/18556
"The results indicate that many genes that are overexpressed in human cancer cells are specific to a variety of normal tissues, including normal tissues other than those from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

Formatted: Font:Helvetica Neue

Deleted: an example

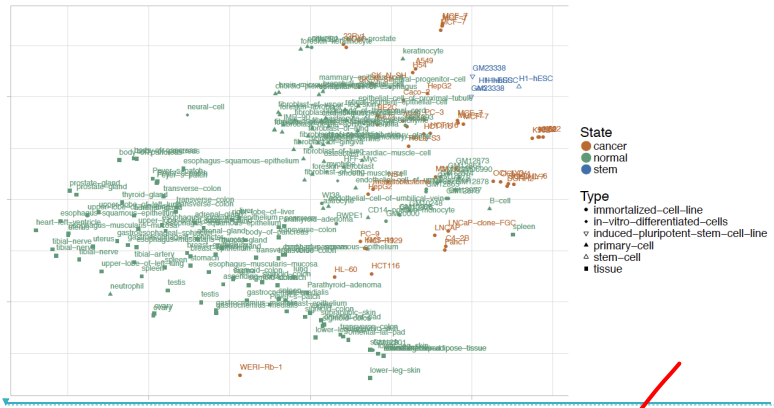
Formatted: Font:Helvetica Neue

t–SNE: CTCF



Deleted: ... [33]

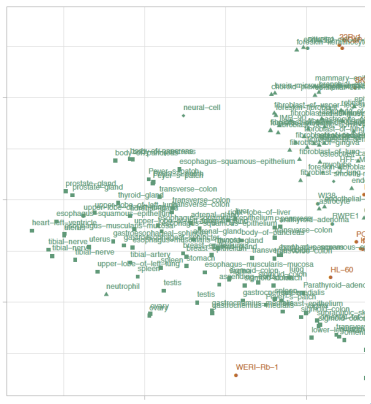
Formatted: Font:Helvetica Neue

	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) ###
Excerpt From Revised Manuscript	t–SNE: CTCF  <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).>

Formatted: Line spacing: single

Deleted: (DL maybe) . ... [34]

t–SNE: CTCF



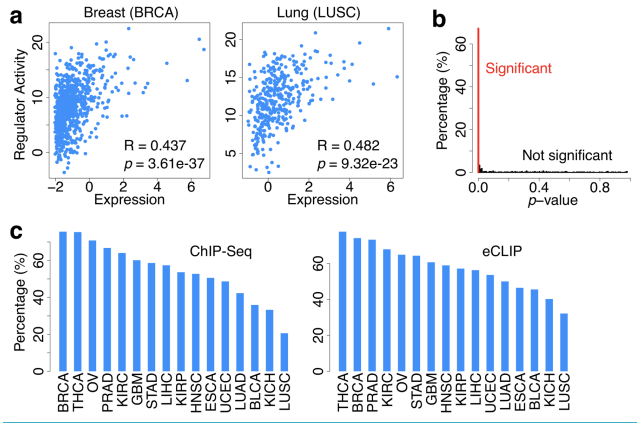
Deleted:

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

<TYPE>$$$CellLine,$$$Validation
<ASSIGN>@@@JZ,@@@DL,@@@Peng,@@@DC
<PLAN>
<STATUS>%%%75DONE
[JZ2PE: use the cristina leslie ATAC-Seq data set]

Deleted: TBC

Deleted: . ... [35]

Referee Comment	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 take the referee's comment to heart and we agree with the reviewer that it is important to verify the discoveries from cell lines from primary cancers. In the revision, we compared the concordance level of our conclusions made from ENCODE cell line data to observations from patients with primary cancers. And we clarified that although ENCODE data are profiled in cell culture models, the regulatory targets are still representative of the gene regulations in human cancers. We have added a new section in the revised supplementary file for more discussions.
Excerpt From Revised Manuscript	We predicted the regulatory activities of transcription factor (TF) MYC using a ChIP-Seq profile in MCF-7 cells. We found that the MYC regulatory activity is highly correlated with the MYC expression across TCGA breast tumors (Supplementary Figure Xa). For most TFs, their regulatory activities predicted using ENCODE ChIP-Seq profile in cell lines are significantly correlated with their expression levels across breast tumors (Supplementary Figure Xb). Moreover, using the same MCF-7 ChIP-Seq profile, the MYC regulatory activity predicted for lung tumors is also significantly correlated with MYC expression level in TCGA lung cancer (Supplementary Figure Xa). These results indicate that the ChIP-Seq profiles from a particular cell line can capture regulatory targets in human tumors from diverse cancer types. To select ChIP-Seq or eCLIP profiles that are representative of the regulatory targets in human cancers, we only reported the results of TFs or RBPs whose regulatory activities are significantly correlated with their gene expression level in each TCGA cohort (Supplementary Figure Xc). 

Formatted Table

Deleted: thank

Deleted: referee for this

Deleted: human clinical relevance of cell line data.

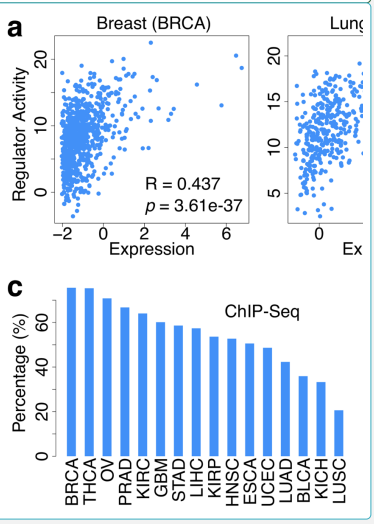
Deleted: For example, we predicted the regulatory activities of transcription factor (TF) MYC using

Deleted: ChIP-Seq profile

Deleted: MCF7 cells. The MYC regulatory activity is highly correlated with

Deleted: 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.

Moved down [3]: 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).



Deleted:

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

Deleted: MCF7

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

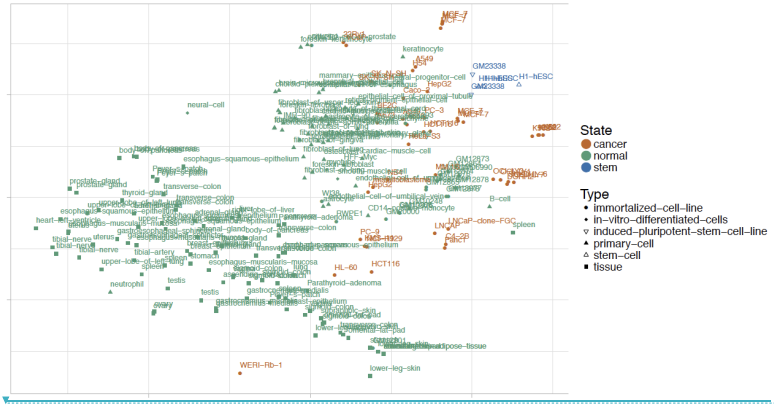
<TYPE>$$$Stemness$$$Calc
<ASSIGN>@@@DL,@@@PE,@@@DC
<PLAN>&&&AgreeFix,&&&MORE
<STATUS>%%%75DONE

Deleted: 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	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

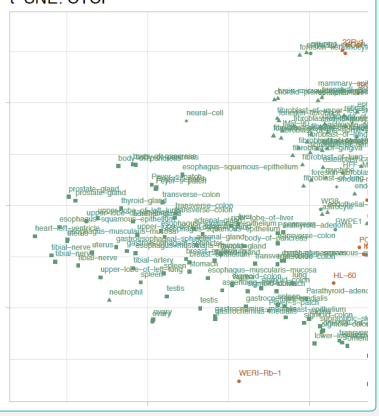
Formatted Table

Deleted: > PE's imputed network stuff ... [36]

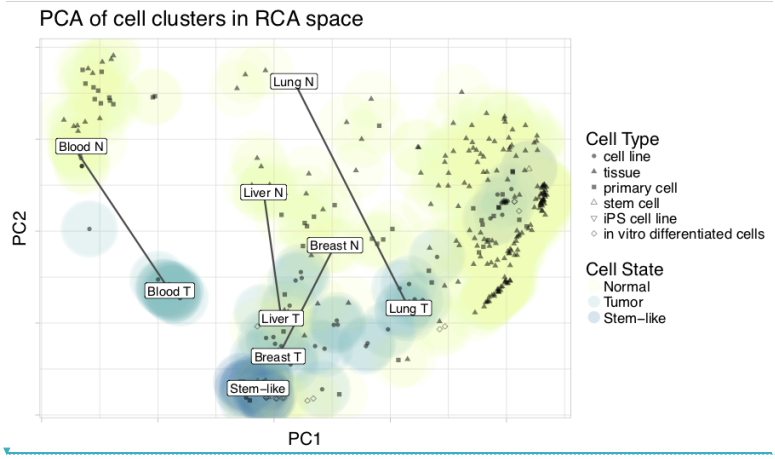
	coverage and has the most amount of other assays in ENCODE. In our revised manuscript, we have expanded our analysis to other stem cells. We have compared other available stem-related cell types, as suggested by the referee, to H1-hESC to show that H1-hESC is not very different from other stem cells 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	t–SNE: CTCF  <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).>

Deleted:

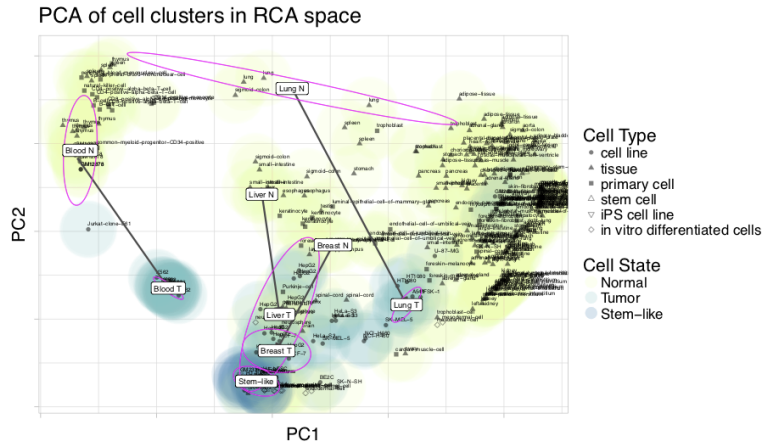
t–SNE: CTCF



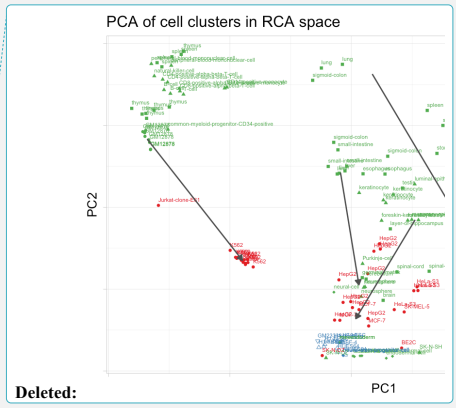
Deleted:



<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.>



<Shadow figure of RCA>



Deleted:

<ID>REF4.7 – Fixes for Figure 1

<TYPE>$$$Presentation,$$$Later

<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE

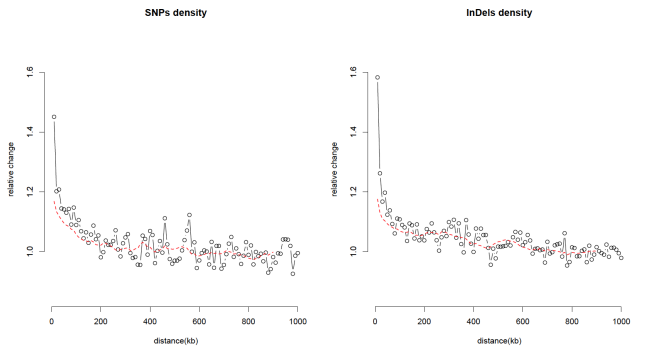
Referee Comment	4) I have difficulties to fully understand Fig.1, in particular the patient cohort (PC) at the bottom of the "depth approach" (just above the green box of cell -specific analysis). The two rows are at the bottom of the columns report mutation and expression, but they belong to the columns of the cell lines (K562, HepG2, etc). I just simply do not understand that part of the figure, in particular the relation between cell lines and the patient cohort (the figure legend does not help, and also supplementary material did not help).
Author Response	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,&&&MORE
<STATUS>%%%30DONE
[JZ2MG: to disc next week]
[JZ2DL, XM, TG, STL: woiuld you please help to fill in the stuff?]

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

Deleted: TBC
Formatted Table
Deleted: DL - think about how we can change the figure ... [37]
Deleted:
Deleted: TBC,%%%MORE
Deleted:
Formatted Table

Author Response	The referee asked us to comment on the relationship of structural variants, BMR, and network wiring. We think these are very good suggestions and we wished we had taken that more in this mission. In the revision, we have definitely taken this comments to heart and have added in main text figures that look at the degree to which structural variants, or SVs, mature background mutational rate, and they also affected the network rewiring. We think this is an ideal illustration of the ENCODE data since, in addition to mapping a lot about the function of the genome, some of the new incurred data sets actually give rise to structural variants meaning that structural variants are an integral output of the product. Relating them to network wiring and background mutation rate is an ideal illustration of the value of the data and the project. We have constructed a number of new main figures that address this and we quite heartly thank the referee for pointing this out. To summarize our conclusion, First, we did observe an elevated SNV/indel rate around the breakpoints. Second, we explored the SV introduced enhancer gain/loss events and relate them to gene expression changes. Third, we studied the relationship of SNVs to network rewirings
Excerpt 1 From Revised Manuscript	Regarding the relationship of SNV to SV 

Deleted: extremely

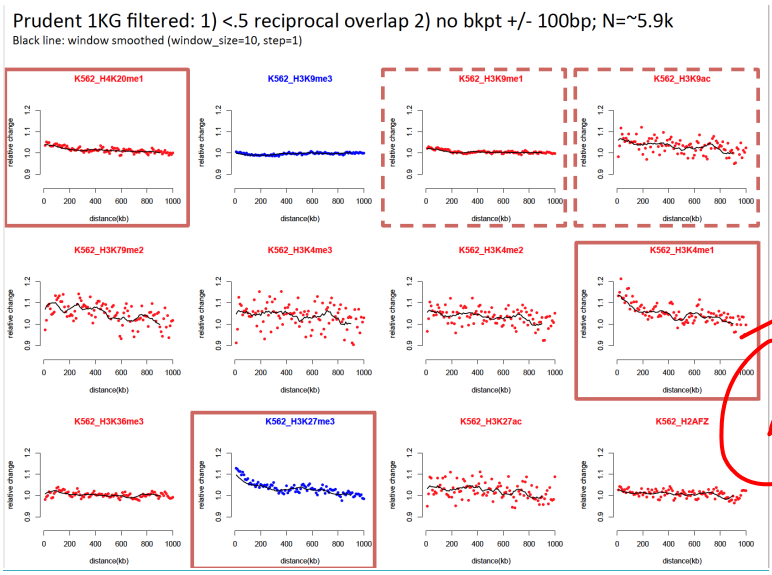
Deleted: we're

Deleted: taking

Deleted: a

Deleted: wiring

Deleted: mains



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

<TYPE>$$$CellLine,$$$Text
<ASSIGN>@@@WM,@@@JZ,@@@MRS
<PLAN>&&&AgreeFix
<STATUS>%%%50DONE
[JZ2MG: special attention. To disc next week]
make a response for Orli using 4.9, the other referee thinks matching doesn't make sense

Referee Comment	6) Most cancers are not necessarily represented by a single cell type used to obtain genomics data in this study, but contains numerous types of cells with different mutations, as well as normal cells, infiltrating cells, all in a three dimensional structure, often producing metastatic colonizing other organs. However, this study focuses only on comparisons between cells. These limitations should be better discussed, also to put in perspective future studies on single cells.
Author Response	We thank the referee for bringing this up and we totally agree with the referee that genomic and epigenomic heterogeneity in tumor cells, as well as heterogeneity in

Deleted: TBC

Formatted Table

Moved down [4]: 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. ... [41]

Deleted: ###JZ: strength of cell line, no heterogeneity, emphasize this, co-expression ~~network~~ ... [38]

Deleted: reference cell line to annotation to patient .... key pt of the paper ... peng's figure ... [39]

Deleted: greater emphasis. ... [40]

Formatted: Justified

Deleted: is correct that tissue heterogeneity represents a source of complexity not directly modeled in our resource, a limitation which

Deleted: now discuss

Deleted: 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 ... [42]

	the tumor microenvironment (e.g., immune cell infiltrates, hormonal factors, normal cell populations, etc.) are significant factors in tumor growth and development. This is a limitation of the current technique, which we now discuss with greater emphasis. 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 (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. 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. In addition, 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). As we have shown in the revised manuscript that, the conclusions we made from the cell lines correlate well with the observations from primary cancer patients.
Excerpt From Revised Manuscript	We predicted the regulatory activities of transcription factor (TF) MYC using a ChIP-Seq profile in MCF-7 cells. We found that the MYC regulatory activity is highly correlated with the MYC expression across TCGA breast tumors (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,

Deleted: Nonetheless, we feel there remains value in single-cell comparisons between tumor and normal cells.

Moved (insertion) [4]

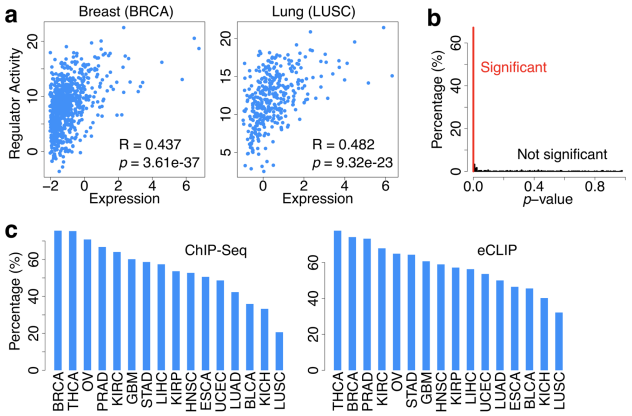
Deleted: Even

Formatted: Justified

Formatted: Justified

using the same MCF-7 ChIP-Seq profile, the MYC regulatory activity predicted for lung tumors is also significantly correlated with MYC expression level in TCGA lung cancer (Supplementary Figure Xa). These results indicate that the ChIP-Seq profiles from a particular cell line can capture regulatory targets in human tumors from diverse cancer types. To select ChIP-Seq or eCLIP profiles that are representative of the regulatory targets in human cancers, we only reported the results of TFs or RBPs whose regulatory activities are significantly correlated with their gene expression level in each TCGA cohort (Supplementary Figure Xc).

Moved (insertion) [3]



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

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

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

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

	with statistically significant correlations through two-sided student t test (FDR < 0.05).

<ID>REF4.10 – lncRNAs and BMR

<TYPE>$$$BMR,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%50DONE

Deleted: 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.
Excerpt From Revised Manuscript	

Formatted Table
Formatted: Justified

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

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

Deleted: Done

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

Formatted Table
Formatted: Justified

Author Response	We thank the referee to point out this and we have fixed in our revised manuscript
Excerpt From Revised Manuscript	

<ID>REF4.12 – (Minor) Figure legends

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

Deleted: Done

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

Formatted Table

Referee #5 (Remarks to the Author):

<ID>REF5.0 – Preamble

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

We would like to appreciate the referee's feedback. We found that many of the suggestions, such as further power analysis, the false positive rate of rewiring, comparison with other networks, cross-validation using external data, are quite valuable and we significantly 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 this 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 the 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 564 DNase-seq STARR-seq
	Enhancer-gene linkage	Annotation in BED format	2015 Histone ChIP-seq 329 RNA-seq

Deleted: Done

Deleted:

Deleted: his

Formatted Table

	Extended gene	Annotation in BED format	1863 TF ChIP-seq 167 eCLIP Enhancer-gene linkage
SV and SNV callsets	Cancer cell lines	Variants in VCF format	WGS BioNano Hi-C 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 564 DNase-seq

Specifically for the BMR estimation part, the reviewer mentioned that there had 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 help many of the methods used in these papers. With a larger pool of covariate selection, the estimation accuracy can be significantly improved.

Deleted: have

Formatted: Underline

Deleted: actually

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

Deleted:

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

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

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

Deleted: 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

Formatted Table

Author Response	We thank the referee for the positive comment.

Deleted: Excerpt From ... [44]

Deleted: Excerpt From ... [43]

<ID>REF5.2 – BMR

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

Deleted: Done

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 bringing out these references. We did notice that epigenetic features have been used to estimate BMR and improve driver mutation detection. 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. It is worth to mention that we have released way more genomic features in a ready-to-use format and have shown that it would noticeably improve BMR estimate accuracy if appropriately used. We want to further emphasize two points here. 1. ENCODE3 uniformly processed 2017 histone modification data, which makes a much larger pool of features to choose from to potentially improve BMR estimation. Also, the majority of them are actually from real tissues and primary cells (1339 out of 2017). 2. ENCODE3 provides way more replication timing data. Previously, researchers either use no or only HeLa replication timing for all cancer types (Martincorena et

Formatted Table

Formatted: Justified

Formatted: Font:Helvetica Neue

Deleted: identifying

Formatted: Font:Helvetica Neue

Deleted: recognize

Formatted: Font:Helvetica Neue

Deleted: previously been

Formatted: Font:Helvetica Neue

Deleted: 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. ... [45]

Formatted: Font:Helvetica Neue

Moved (insertion) [5]

Formatted: Font:Helvetica Neue, 11 pt

Formatted: Justified

Formatted: Font:Helvetica Neue

Moved (insertion) [6]

Formatted: Font:Helvetica Neue

Formatted: Line spacing: multiple 1.15 li

	al., 2017, Lawrence et al., 2013), or any of the 16 repli-Seq data from previous ENCODE release. We largely extended this number to 51 cell types (12 cell lines).
Excerpt From Revised Manuscript	Table S1. Summary of ENCODE3 histone ChIP-Seq data

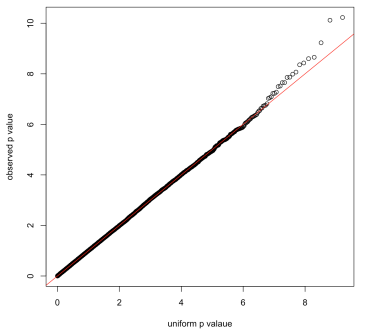
Cell Type	# histone marks
tissue	818
primary-cell	521
cell-line	339
in-vitro-differentiated-cells	179
stem-cell	114
induced-pluripotent-stem-cell-line	46

<ID>REF5.3 – TCGA benchmark

<TYPE>$$$BMR $$$Calc
<ASSIGN>@@@JZ,@@@WM
<PLAN>&&&MORE
<STATUS>%%%75DONE
[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 thank the referee for bringing out this point. We agree that it is important to benchmark the mutation rate estimation. However, we are part of the PCAWG noncoding driver detection group for the joint analysis of TCGA and ICGC data. From our experience in this group, we did not find a gold standard for the whole genome mutation rate estimation. Alternatively, we evaluated the BMR estimation to the commonly used permutation set, which random select a new position within a 50kb window of each somatic variant while preserving the local context.

Deleted: on the gene level
Deleted: $$$
Deleted: >%%%
Deleted: TBC
Formatted Table
Deleted: . ... [46]

	1. We applied our mutation driver detection method on the CDS regions of ~20k protein coding regions on the permuted dataset for breast cancer, and found no driver there. QQ plot was added into the supplementary site. 2. We down sampled the simutated dataset and xxxx (WM to fill in)
Excerpt From Revised Manuscript	1. QQ plot of the observed vs. uniform p value from Breast cancer permuted data set. Red line is the diagonal line. 

<ID>REF5.4 – Improvements of the BMR

<TYPE>$$$BMR $$$Calc
<ASSIGN>@@@JZ@@@WM
<PLAN>&&&MORE &&&DisagreeFix &&&OOS
<STATUS>%%%TBC
[JZ2MG: need more advice here? Does it look good?]

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 ####7mar:fight-outofscope ####7mar - comparisons w/ other methods

Deleted: $$$

Deleted: >%%%

Deleted: [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]

Formatted Table

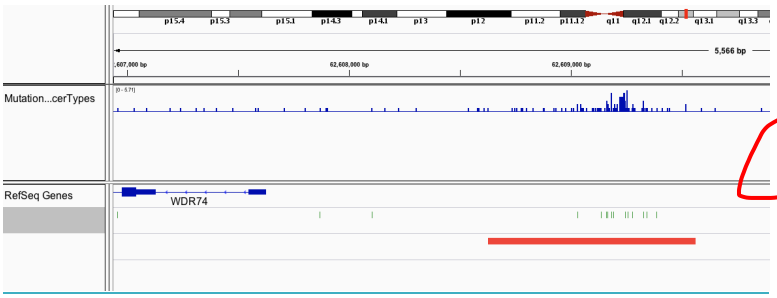
	###21mar - Inigo's paper is not about BMR/driver discovery ### in response doc, praise referee, do analysis to compare Inigo's method
Excerpt From Revised Manuscript	

<ID>REF5.6 – Power analysis

<TYPE>$$$BMR $$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&MORE
<STATUS>%%%75DONE
[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 genes should be cited (Lawrence et al, 2014; Martincorena, 2017).
Author Response	We thank the referee for his/her positive comment on the value of selecting genomic element and suggestion on the power analysis. In our revised manuscript, we expanded our power calculation extensively (see details below). In terms of reference, we cited the Lawrence et al, 2014 paper (and the paper before this one in the same group) in our initial submission and added the Martincorena, 2017, which is published after our submission in Aug 2017. In our initial submission, the assumption is that we were trimming off the nonfunctional sites while preserving the functional ones. Two examples can explain the motivation of this assumption. 1) Enhancers: Traditionally, enhancers were called as a 1kb peak regions, which admittedly introduced a lot of obviously nonfunctional sites. We believe we can get functional region more accurately by trimming the enhancers down using the exact shapes of many histone marks and further integration with STARR-seq and Hi-C data.

Deleted: $$$
Deleted: >%%%
Deleted: TBC
Formatted Table
Formatted: Justified
Deleted: . ... [47]

	2) TFBS hotspots around the promoter region of WDR74. Instead of testing the conventional up to 2.5K promoter region, we can trim the test set to a core set of the promoter region where many TFs bind, which perfectly correlates with the mutation hotspots (red block) for this well-known driver site (blue line for pan-cancer and green line for liver cancer). Following the reviewer's suggestions, in our revised manuscript we show in a formal power analysis that the most important contribution to power comes from including additional functional sites, which is of course by the extended gene concept 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. Admittedly, we are making assumptions and the referee is completely correct in pointing this out. We have tried to be more precise in the text that we are assuming that the large number of ENCODE assays, when integrated, allow us to more directly get the functional nucleotides, but this, of course, is an assumption. It is hard to tell to what degree one can succeed in finding the current events in cancer. It is hard to back this up with the gold standard, but we think that some of the points are self evidently obvious. We have tried to make this clear in text and thank the referee for pointing this out.
Excerpt From Revised Manuscript	

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

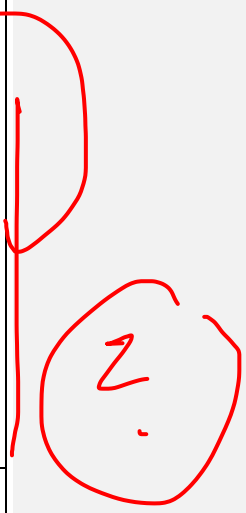
<TYPE>$$$Power $$$Text
<ASSIGN>@@@JZ
<PLAN>&&&MORE
<STATUS>%%%TBC
[JZ2MG: can we say this is out of scope here? Please advise]

Deleted: $$$

Deleted: >%%%

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. 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	

Formatted Table



<ID>REF5.8 – false positive rates of enhancers

Deleted: Calculation of power

<TYPE>$$$Power $$$Text

Deleted: $$$

<ASSIGN>@@@JZ @@@MTG

<PLAN>&&&AgreeFix

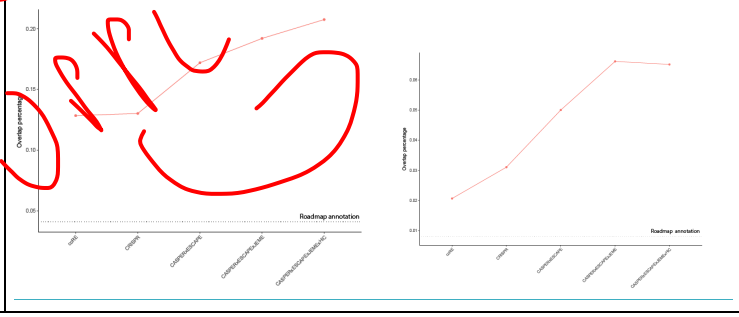
<STATUS>%%%DONE

Deleted: Done

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?

Formatted Table

Formatted: Justified

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. As for the enhancer part, with the ensemble method, for example, we can get more accurate annotation and pin-point to sequences where transcription factors would actually bind to. To estimate the false positive rate would not be very practical at this stage as there is no gold-standard experiment that could assert an predicted enhancer is definitely negative. Here we took the FANTOM enhancer data set and assess the overlap percentage of our enhancer annotation in each ensemble step. We show that each ensemble step indeed increases the percentage of overlap between our annotation and the FANTOM enhancer set. The overlap percentage for our annotation is much higher than that of the Roadmap annotation, and is also higher than the main encyclopedia enhancer annotation annotation (ccRE).
Excerpt From Revised Manuscript	

Deleted: In our initial submission, we were not trimming truly functional or important sites, but rather trimming unimportant sites. For instance, in

Deleted: old way that we found

Deleted: 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

Deleted: exact shape of the signal, we believe

Deleted: accurately gets it the truly functional region, particularly when coupled with

Deleted: STARR-seq and Hi-C data will hopefully increase power. Another case is the TF binding hotspot around

Deleted: promoter region of WDR74. Instead of testing up to 2.5K promoter region without prior information, we can trim

Deleted: test set to a core

Deleted: promoter region where many TF binds to, which perfectly correlates with

Deleted: mutation hotspots (red block) for this well known driver site (blue line for pan-cancer and green line for liver cancer). ... [48]

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

<TYPE>$$$Annotation $$$Text
<ASSIGN>@@@KevinYip,@@@SKL
<PLAN>&&&MORE
<STATUS>%%%50DONE

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. In the revised supplementary file, we have added two sections to discuss these points.

Deleted: $$$

Deleted: >%%%

Deleted: TBC

Deleted: [JZ2MG: next week will check the status of KevinYip, SKL stuff added] ... [49]

Formatted: Justified

Formatted Table

Formatted: Font:Helvetica Neue

Deleted: We have done as suggested. We

Formatted: Font:Helvetica Neue

Deleted: a few sentences

Formatted: Font:Helvetica Neue

Deleted: the main text better

Formatted: Font:Helvetica Neue

	1. Regarding the gene-enhancer linkages *2. Regarding the gene community methods* We have compared the gene community model with other methods like NMF by extending our analysis from 122 GM12878 and K526 dataset to all the 862 TF ChIP-Seq assays included in ENCODE data portal. Analysis showed that our method can better preserve the data structure after dimension reduction.
Excerpt From Revised Manuscript	Mix membership model is a hierarchical Bayesian topic model framework and can help to uncover the underlying semantic structure of a document collection. The core of topic models is Latent Dirichlet Allocation(LDA), which cast the mixed-membership (topics) problem into a hidden variable model of documents. The LDA model has been widely used to analyze a wide variety of data types, including but not limited to text and document data, genotype data, survey and voting data. The advantage of LDA over other algorithms (like SVD, PLSI) used in semantic analysis has been described in Blei 2003. 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. To explore how well the LDA mixed-membership analysis on TF regulatory network, we extend our dataset from 122 GM and K526 samples to all the 862 TF ChIP-Seq assays included in ENCODE data portal. In order to get a reliable correlation, we also increase the number of topic to 50 as the number of TF sample increases. The non-negative matrix factorization (NMF) and Kmeans clustering are used for comparison because the nature of regulatory network requires a non-negative decomposition. The same target dimension K =50 was used to NMF and target number of clusters K=50 for Kmeans. The Euclidean distance between each data the centroidds are used to calculated the correlation. As shown in the figure, the x-axis is original correlation of two TF regulatory target, y-axis is reproduced correlation from LDA document to topic distribution and NMF decomposed matrix. The solid line is the 'loess' smoothing curve for the scattered dots. We can see the LDA method can reproduce the original correlation better than either NMF or Kmeans. Overall correlation between the reproduced pairwise correlation and the original correlation were 0.123 in Kmeans, 0.404 in NMF and 0.788 in LDA.

Deleted: definition and gene linkage prediction. We have created suppl. Section XXX that shows the performance of JEME + Hi-C.

Formatted: Font:Helvetica Neue, Italic, Underline

Formatted: Font:Helvetica Neue, Italic, Underline

Formatted: Font:Helvetica Neue, Italic, Underline

Formatted: Font:Helvetica Neue

Formatted: Justified

Deleted: Also we

Formatted: Font:Helvetica Neue

Deleted: . Mix membership model is a hierarchical Bayesian topic model framework and can help to ... [50]

Moved down [7]: The core of topic models is Latent Dirichlet Allocation(LDA), which cast the mixed- ... [51]

Formatted: Font:Helvetica Neue

Deleted: SVD, PLSI) used in semantic

Formatted: Font:Helvetica Neue

Deleted: has been described in Blei 2003. ... [52]

Deleted: GM

Formatted: Font:Helvetica Neue

Deleted: samples

Formatted: Font:Helvetica Neue

Formatted: Font:Helvetica Neue

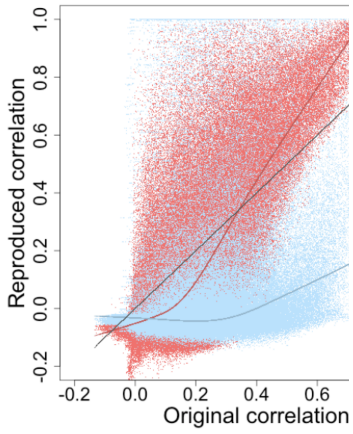
Deleted: In order to get a reliable correlation, we also increase the number of topic to 50 as the numbe ... [53]

Moved down [8]: . As shown in the figure, the x-axis is original correlation of two TF regulatory target, y- ... [54]

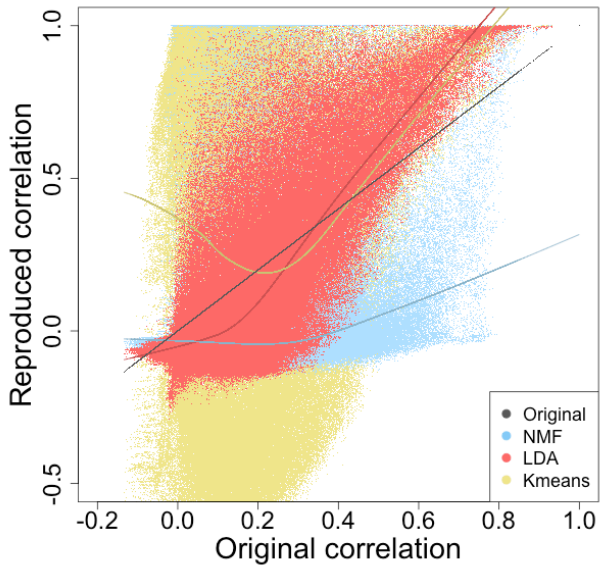
Deleted: the NMF.

Moved (insertion) [7]

Moved (insertion) [8]



Deleted:

<ID>REF5.10 – What data sets are used

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

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.

Deleted: Done

Deleted: [JZ2MG: to disc next week can we use other public data?]

Formatted: Justified

Formatted Table

Excerpt From Revised Manuscript	

<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	

Formatted Table

Deleted: . ... [55]

<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

Formatted Table

	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: Excerpt From ... [56]

Deleted: Excerpt From ... [57]

<ID>REF5.13 – Sequence coverage

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

Deleted: 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	

Formatted Table

Deleted:

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

<TYPE>$$$Power,$$$Calc
<ASSIGN>@@@JZ
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE
[JZ2MG: feel the three power related questions can be combined]

Deleted: Done

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. In our initial submission, the assumption is that we were trimming off the nonfunctional sites while preserving the functional ones. Two examples can explain the motivation of this assumption.

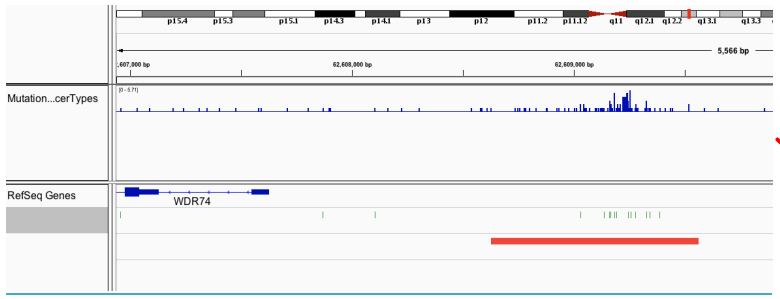
Formatted Table

Deleted: When all

Deleted: sites are within the test region, a shorter or "compact" annotation

Deleted: 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

Deleted: old way that we found enhancer sites by just calling a 1KB region from a

	1) Enhancers: Traditionally, enhancers were called as a 1kb peak regions, which admittedly introduced a lot of obviously nonfunctional sites. We believe we can get functional region more accurately by trimming the enhancers down using the exact shapes of many histone marks and further integration with STARR-seq and Hi-C data. 2) TFBS hotspots around the promoter region of WDR74. Instead of testing the conventional up to 2.5K promoter region, we can trim the test set to a core set of the promoter region where many TFs bind, which perfectly correlates with the mutation hotspots (red block) for this well-known driver site (blue line for pan-cancer and green line for liver cancer).
Excerpt From Revised Manuscript	

Deleted: by almost any estimation included knots

Deleted: non functional sites. Trimming this down using a large battery of histone marks and the exact shape of the signal, we

Deleted: gets it the truly functional region, particularly when coupled

Deleted: accurate

Deleted: will hopefully increase power. Another case is the TF binding hotspot

Formatted: Underline

Deleted: without prior information

Deleted: TF binds to

Deleted:

Deleted: ... [58]

<ID>REF5.15 – BCL6 Questions

<TYPE>$$$Annotation,$$$Calc
<ASSIGN>@@@XK,@@@TG
<PLAN>&&&AgreeFix
<STATUS>%%%20Done
[JZ2MG: checking the SV status now, to report next week]

Deleted: TBC

Deleted: to be added to

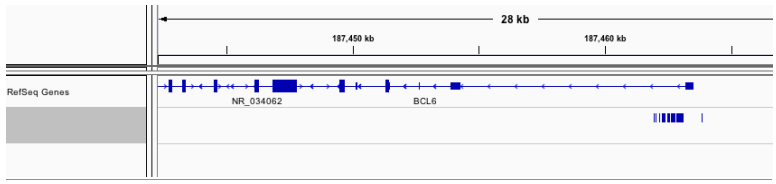
Deleted: disc agenda

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	We thank the referee for this comment. As suggested, we found that the there is a mutation hotspot near the first intron of BCL6.

Formatted Table

Formatted: Justified

Deleted: BCL6 mutations were found in promoter region. ... [59]

Excerpt From Revised Manuscript	

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

<TYPE>$$$Network,$$$Calc
<ASSIGN>@@@Peng,@@@JZ
<PLAN> &&&AgreeFix
<STATUS>%%%75DONE

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 this point and we also feel that it is important to make comparison with other existing networks with statistical evaluation. We made the following revisions in the updated manuscript. **1. Regarding the proximal regulatory element network:** *1.1 Comparison with Biogrid and String experimental interactions.* We showed that the ENCODE ChIP-seq/eCLIP based networks can capture a higher fraction of standard interactions (from manually curated networks from TTRUST) than protein physical networks, including Biogrid and String experimental interactions (see details below). *1.2 Comparison with DHS-based imputed networks* *1.3 False positive rate estimation of the ChIP-Seq based networks* The ENCODE consortium has always enforced a strict data quality standards for all ENCODE produced transcription factor ChIP-seq experiments, which allow us to rigorously control the false positives.

Deleted: Done
Formatted Table
Formatted: Justified
Formatted: Font:Helvetica Neue, 11 pt
Formatted: Font:Helvetica Neue, 11 pt
Formatted: Font:Helvetica Neue, 11 pt
Moved up [5]: ... [60]
Deleted: .
Formatted: Font:Helvetica Neue, 11 pt
Formatted: Font:Helvetica Neue
Deleted: To make
Formatted: Font:Helvetica Neue
Deleted: 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 networks: 1, transcriptional regulations between TFs and target genes; 2, post-transcriptional regulations between RBPs and target genes.
Moved up [6]: ... [61]
Deleted: To evaluate the quality of ENCODE transcriptional regulatory networks, we utilized
Formatted: Font:Helvetica Neue
Deleted: 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
Formatted: Font:Helvetica Neue
Deleted: recapitulate. The ENCODE network can
Formatted: Font:Helvetica Neue
Formatted: Font:Helvetica Neue
Formatted: Font:Helvetica Neue
Deleted: Supplementary Figure X). Moreover,

	2. Regarding the distal regulatory element network: With the ChIP-Seq, DHS, STARR-Seq, ChIA-PET, and Hi-C experiment, ENCODE has a distal TF-enhancer-gene network of high quality, which is less discussed and investigated previously. We feel this is one of the unique aspect of our resource. *2.1 High quality of enhancer definitions after integrating many histone ChIP-seq and DHS, and STARR-Seq data* Here we took the FANTOM enhancer data set and assess the overlap percentage of our enhancer annotation in each ensemble step. We show that each ensemble step indeed increases the percentage of overlap between our annotation and the FANTOM enhancer set. The overlap percentage for our annotation is much higher than that of the Roadmap annotation, and is also higher than the main encyclopedia enhancer annotation annotation (ccRE). *2.2 High quality of enhancer-gene linkages*
Excerpt 1 From Revised Manuscript	*Regarding Comparison with Biogrid and String experimental interactions.* To evaluate the quality of ENCODE transcriptional regulatory networks, we utilized the TRRUST database, which manually curated transcriptional regulations from Pubmed articles (Han et al., 2018). We defined the TRRUST interactions as the standard and tested the fraction of standard interactions that other networks can recapitulate. The ENCODE network can capture a higher fraction of standard interactions than protein physical networks, including Biogrid and String experimental interactions (Supplementary Figure X). Moreover, the fraction of standard networks that ENCODE network recapitulated is consistently higher than random. These results supported the higher relevance of ENCODE networks on transcriptional regulation compared to other networks. We also constructed another post-transcriptional network between RBPs and target genes through linking the RBP binding sites on gene 3'UTR regions. To the best of our knowledge, the current study is the first one to study RBP-gene interactions systematically; thus we are not aware of any previous resources that can provide gold standard regulations for comparison.

Formatted: Font:Helvetica Neue

Deleted: fraction of standard networks that ENCODE

Deleted: recapitulated

Deleted: consistently higher than random. These results supported the higher relevance of ENCODE networks on transcriptional regulation compared to other networks.

Formatted: Font:Helvetica Neue

Formatted: Font:Helvetica Neue

Deleted: 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

Formatted: Font:Helvetica Neue

Formatted: Font:Arial

Deleted: knowledge,

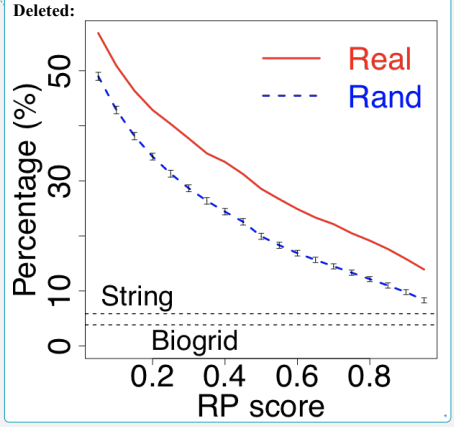
Formatted: Font:Arial

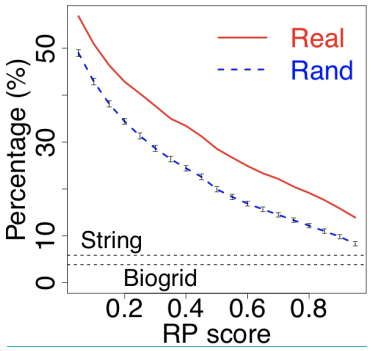
Deleted: current study is

Formatted: Font:Arial

Deleted: 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.

Deleted:





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

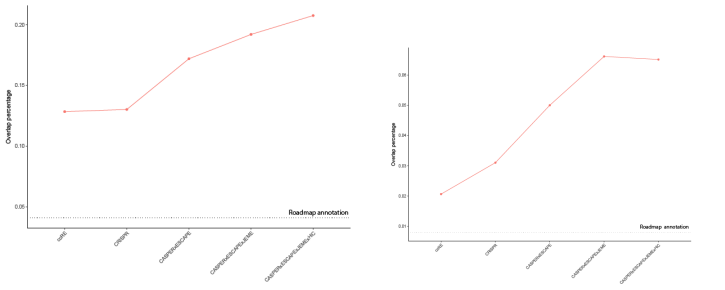
Excerpt 2 From Revised Manuscript

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

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

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

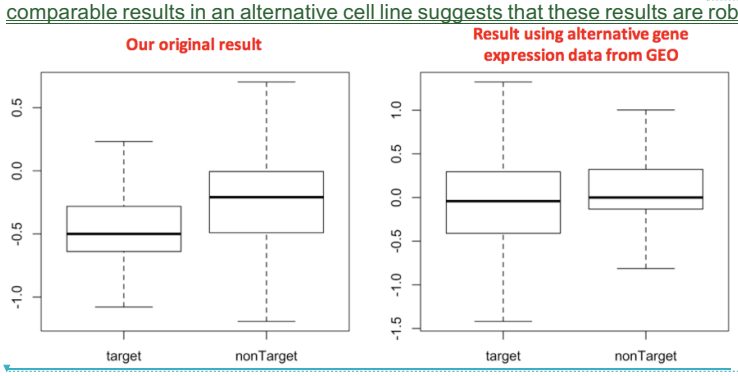
Self-consistency Ratio	Rescue Ratio	Resulting Data Status	Flag colors
Less than 2	Less than 2	Ideal	None
Less than 2	Greater than 2	Acceptable	Yellow
Greater than 2	Less than 2	Acceptable	Yellow
Greater than 2	Greater than 2	Concerning	Orange

	After extensive quality controls for the concordance between replicates, peaks are called using macs2 {"Zhang et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* (2008) vol. 9 (9) pp. R137"} with p-value cutoff of 0.01.
Excerpt 3 From Revised Manuscript	*Regarding quality of enhancers* As for the enhancer part, with the ensemble method, for example, we can get more accurate annotation and pin-point to sequences where transcription factors would actually bind to. To estimate the false positive rate would not be very practical at this stage as there is no gold-standard experiment that could assert an predicted enhancer is definitely negative. Here we took the FANTOM enhancer data set and assess the overlap percentage of our enhancer annotation in each ensemble step. We show that each ensemble step indeed increases the percentage of overlap between our annotation and the FANTOM enhancer set. The overlap percentage for our annotation is much higher than that of the Roadmap annotation, and is also higher than the main encyclopedia enhancer annotation annotation (ccRE). 
Excerpt 4 From Revised Manuscript	

<ID>REF5.17 – MYC KD

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

<STATUS>%%%100DONE

Referee Comment	13. MYC is known to have profound effects on gene networks. Have the authors considered comparing the results from their MCF7 knockdown experiment to existing data from similar MYC knockdowns to validate the behavior of the network?
Author Response	We thank the referee for this suggestion and we feel this is a good comment.As suggested we searched for external dataset from multiple platform and cell types and used them to compare with our discoveries. Both datasets confirmed our claims.
Excerpt From Revised Manuscript	1. We carried out these analyses after first identifying an alternative dataset. Specifically, we identified a dataset of gene expression for both MYC knockdowns (as well as a corresponding control) in Gene Expression Omnibus (GEO accession number GSE86504). For these alternative data, gene expression was measured by RNA-seq in the HT1080 cell line. We note that, even though these alternative analyses were conducted on a different cell line, the results we obtain (shown below in the right panels, and now made available in the supplementary materials) validate the behavior of the network, and they are consistent with our previous results (in which gene expression was measured in the MCF-7 cell line). These comparable results in an alternative cell line suggests that these results are robust. 

Deleted: DONE

Formatted Table

Formatted: Justified

Deleted: . We carried out these analyses after first identifying an alternative

Deleted: . 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

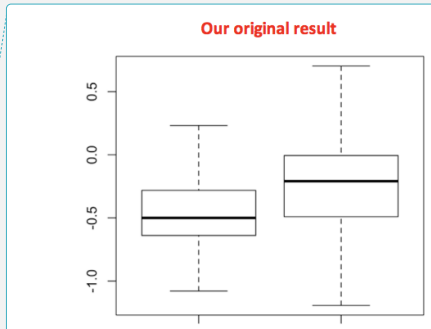
Deleted: line. ... [62]

Moved down [9]: These comparable results in an alternative cell line suggests that these results are robust.

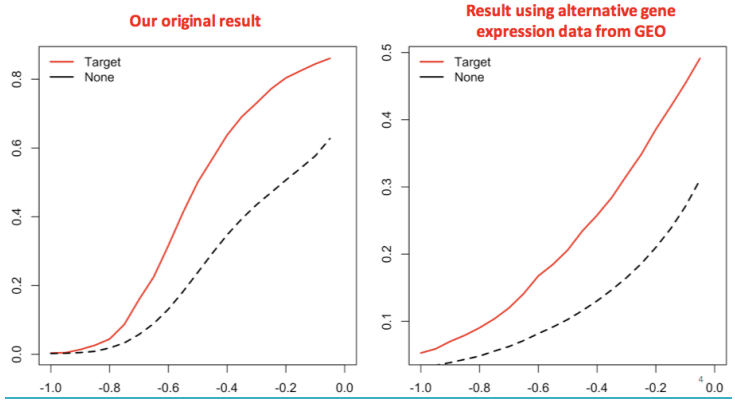
Deleted: ... [63]

Formatted: Line spacing: single

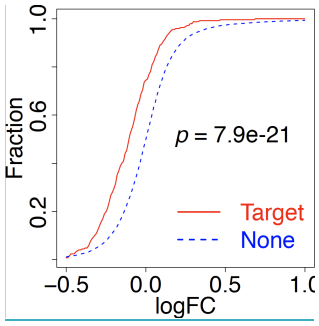
Moved (insertion) [9]



Deleted: ... [64]



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



<ID>REF5.18 – SUB1 analysis

<TYPE>$$$NoveltyPos,$$$Calc
<ASSIGN>@@@MRS,@@@JL,@@@YY
<PLAN>&&&MORE
<STATUS>%%%85DONE
[JZ2Peng: write something about sub1 decay rate]

Deleted: TBC

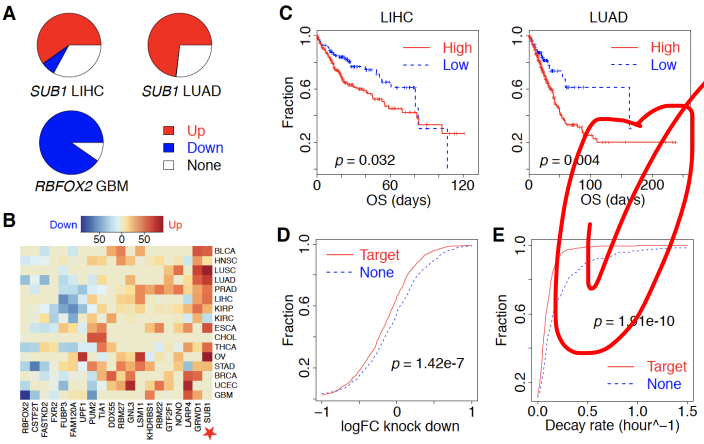
Deleted: [JZ2YY: would you please add your stuff here?]

Referee Comment	14. SUB1 is a potentially interesting new cancer gene. The authors should further explore the biology of this gene.
Author Response	We thank the referees for the positive comments. We did follow up with SUB1 in this round of revision.

Formatted Table

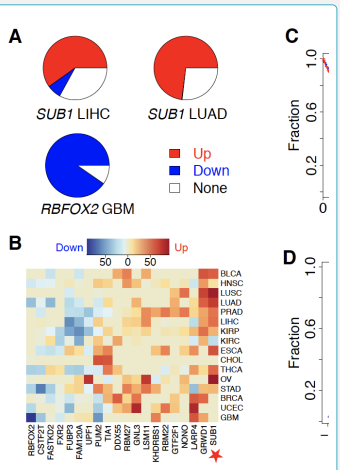
1. We checked SUB1 regulation potential in different cancer types and found that they are consistent as below. We also found that SUB1 tends to bind to the 3UTRs to stabilize its target mRNA. The decay rate of SUB1 is slower than non-targets (p value=1.91e-10).
2. We checked the 3' UTR expression level of SUB1 target genes and found that the target genes are significantly down-regulated upon SUB1 KD. In addition, we found enrichment of SUB1 target genes for CGC (Cancer Gene Census) genes.

Excerpt 1 From Revised Manuscript



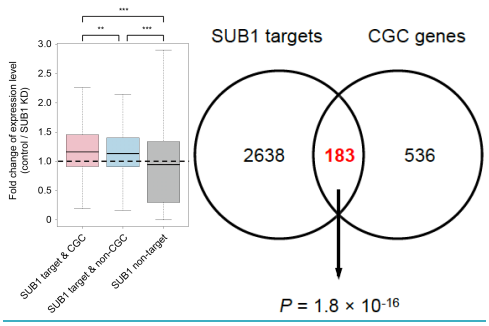
Inference of RNA binding proteins that drive tumor specific expression patterns. Based on ENCODE eCLIP data, we applied RABIT framework to identify RNA binding proteins (RBP), whose target genes are differentially regulated in diverse TCGA cancer types. (A) For each RBP, the percentage of patients with 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.

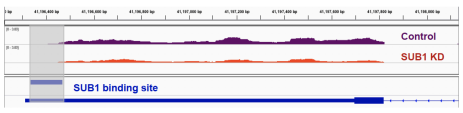
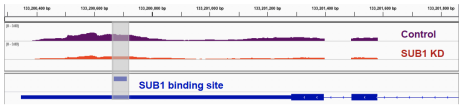
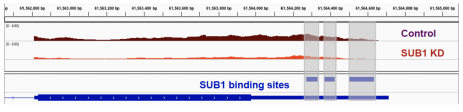
Formatted: Justified, Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, ... + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"



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

Deleted:

Excerpt 2 From Revised Manuscript	Comparison  Here we show some IGV examples together with SUB1 binding sites on the 3' UTRs.

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

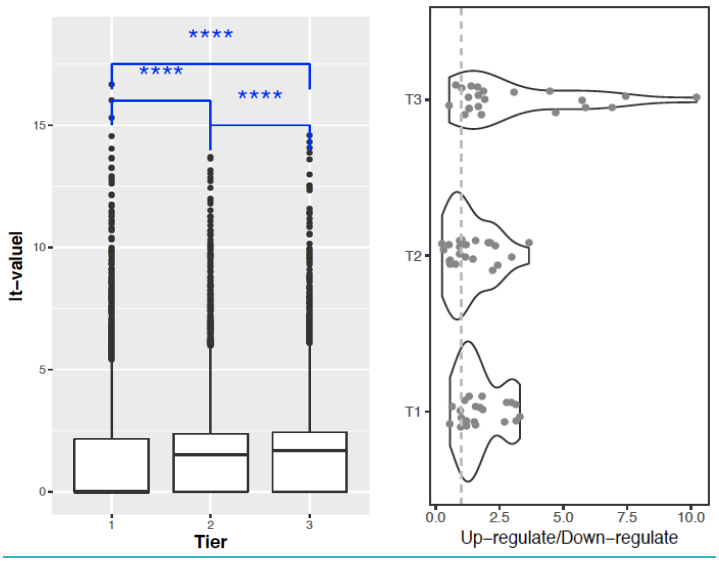
<ID>REF5.19 – Significance of regulatory network hierarchy

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

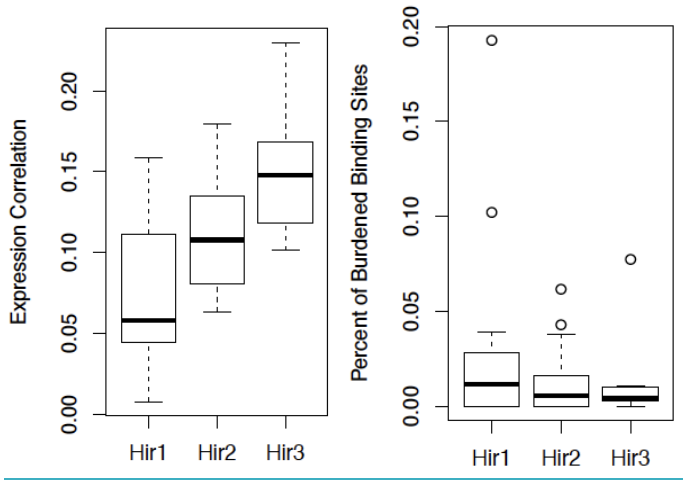
Deleted: DONE

Referee Comment	15. The manuscript claims that transcription factors placed at the top level of the network hierarchy are enriched in cancer-associated genes and drive expression changes. Both claims need to be supported with statistical tests.

Formatted Table

Author Response	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).
Excerpt From Revised Manuscript	Supplementary Figure X. 

Deleted: ... [65]



<ID>REF5.20 – Rewiring of regulatory network

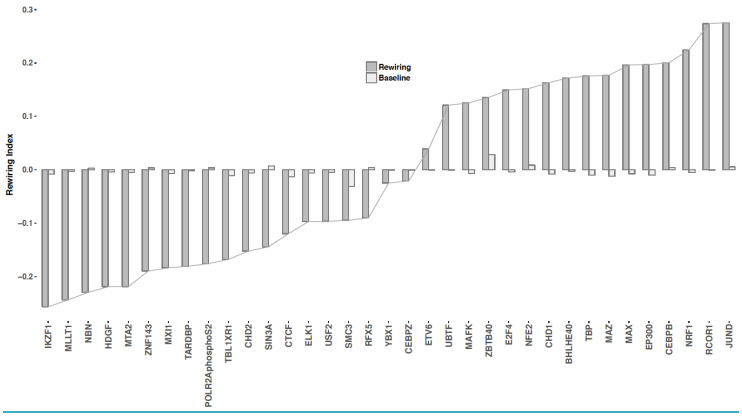
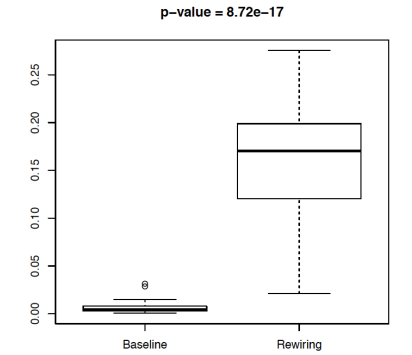
<TYPE>$$$Network,$$$Calc
<ASSIGN>@@@DL
<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Deleted: DONE

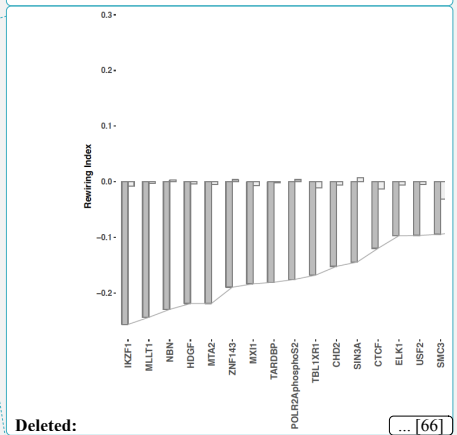
Referee Comment	16. In the tumor-normal network comparison, is the fraction of edge changes related to the total number of edges for a given TF? This analysis should further clearly state its null hypothesis (what changes are expected?). What happens when edges are randomly permuted?
Author Response	We 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

Formatted Table
Formatted: Justified

Deleted: We would like to truly

	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. 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	 

Formatted: Justified



Deleted: ... [66]

Deleted: ... [67]

<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	We 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.
Excerpt From Revised Manuscript	

Formatted: Justified
Formatted Table
Deleted: ####7mar we truly thank referee. Took referee's comment to heart, made hugh improvement
... [68]
Deleted: truly
Deleted:
... [69]

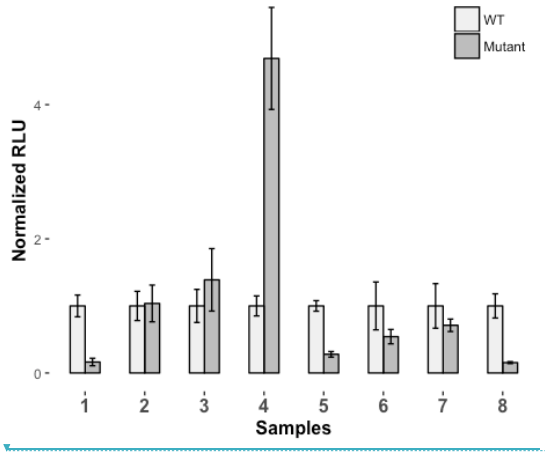
<ID>REF5.22 – Selection of regions for validation testing

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

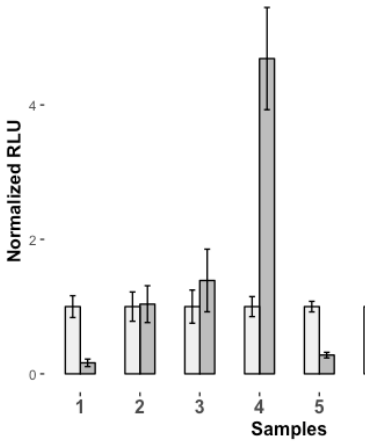
Deleted: DONE

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. 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

Formatted: Justified
Formatted Table

	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. Error bar is representing 95% confidence interval across 3 replicates. 

Deleted:



Deleted: Excerpt From ... [70]

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

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

Deleted: >&&&AgreeFix

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	We thank for the referee for this comments.

Formatted Table

Response	We have tried to revise the figures as requested We have fixed figure XX & YY.
Excerpt From Revised Manuscript	

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

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

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

<ID>REF5.25 – BMR

<TYPE>$$$Minor,$$$BMR
<ASSIGN>@@@JZ

Deleted: >&&&AgreeFix

Deleted: TBC

Formatted: Justified

Formatted Table

Deleted: Excerpt From ... [71]

<PLAN>&&&AgreeFix
<STATUS>%%%100DONE

Referee Comment	21. When the authors describe recurrent events, are these significant? If so, please provide p-values (and q-values, when applicable).
Author Response	We thank the referee to point this out. We have the values and q-values all deposited into our online resource and supplementary files. We have made this clearer in our revised manuscript.

<ID>REF5.26 – Citation of previous work

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

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

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

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

Referee Comment	23. The use of a "composite normal" is not optimal for tissue or tumor-type specific analyses that the authors advocate. Although the described data resource (ENCODE) may not provide normal control data, normal tissue data from the Roadmap Epigenomics could be included instead (or in addition) to improve the quality of the tumor-normal comparisons.

Deleted: TBC
Formatted Table
Formatted: Justified
Deleted: Excerpt From ... [72]
Deleted: TBC
Formatted Table
Formatted: Justified
Deleted: Excerpt From ... [73]
Deleted: TBC
Deleted: [JZ2MG, JZ2DL: to disc next week]
Formatted Table
Formatted: Justified

Author Response	We thank the referee for bringing this out. We did noticed the Roadmap data. Actually, in the new release, ENCODE3 reprocess the complete set of roadmap data and we did include that in our data tables (Figure 1 and supplementary table xxx).

Deleted: JZ: I assume that we used Roadmap normal? There is no ChIP-Seq data there!
Excerpt From ... [74]

Deleted: Excerpt From ... [75]

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

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

Deleted: TBC

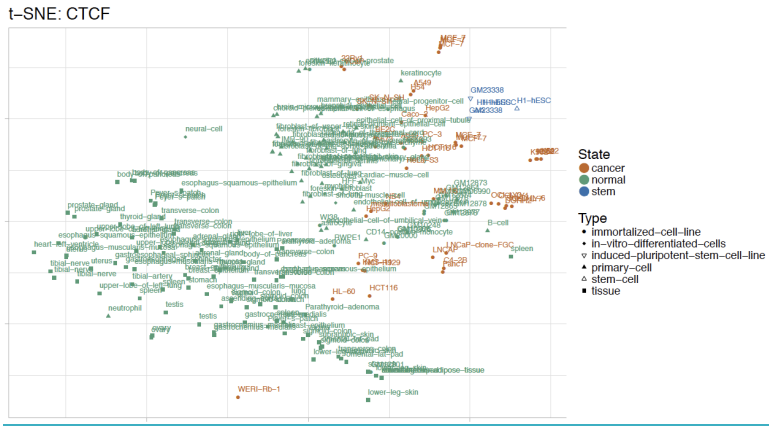
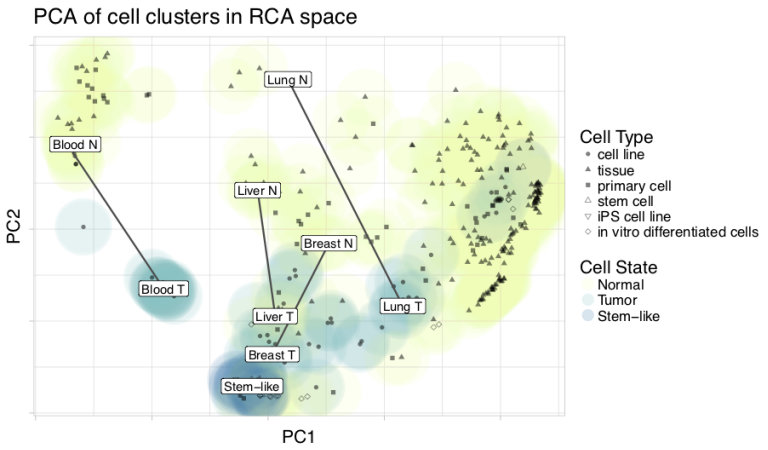
Deleted: $$$Stemness %%%TBC

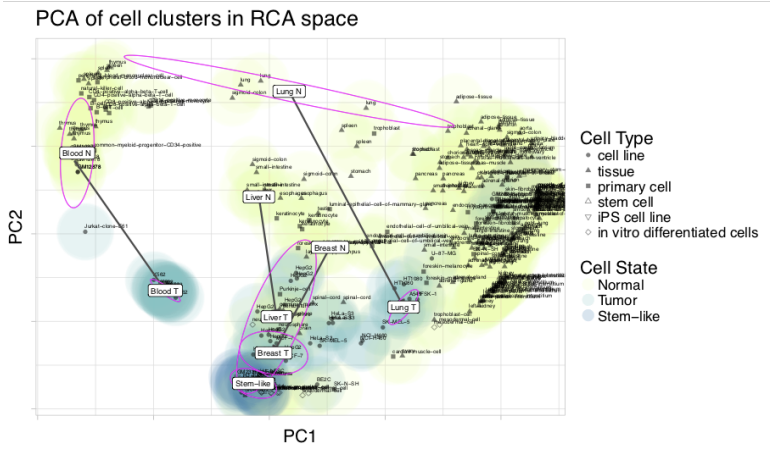
Formatted: Normal

Formatted Table

Formatted: Justified

Referee Comment	24. The authors use the H1 embryonic stem cell line as model for "stemness" in cancer. Tumor "stemness" often resembles tissue progenitors, not embryonic stem cells. In the absence of reliable data for such progenitors the authors should note this caveat with their analysis.
Author Response	We thank the referees for bringing this point out. We agree with the referee that the use of H1 embryonic stem cell line for measuring "stemness" should be further discussed. We, therefore, have revised the manuscript with two additional analysis to show that use of H1-hESC maybe a suitable substitute for a such analysis, especially in the absence of the proper progenitor cell data. We agree with the referee that tissue progenitors of matching cell type would be the ideal pairing to look at "stemness" in cancer. However, as the referee has noted, we mainly have chosen H1-hESC because it offers the broadest TF ChIP-seq coverage and also one of the top-tier cell lines with most variety of experimental assays in ENCODE. We first aimed to evaluate regulatory networks of all ENCODE biosamples including many available stem-like cells and profile their differences. We show that H1-hESC is not far distinct from other stem-like cells, and it is a good representation of stem-like state. We used a regulatory networks of CTCF, one of the most widely assayed TF in ENCODE, to examine their regulatory patterns across different cell types. 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. Second 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 (reference component analysis) space.
Excerpt From Revised Manuscript	 <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).>  <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.>  <Shadow figure of RCA>

<ID>REF5.29 – Validation of prioritized element

<TYPE>$$$Minor,$$$Validation
<ASSIGN>
<PLAN>&&&AgreeFix
<STATUS>%%%75DONE
[JZ2DL: could you please help to add the tracks? Reason for this?]

Deleted: TBC

Referee Comment	25. P-values should be given in Figure 6B for the luciferase reporter assay. The authors may also want to explain why candidate 5, rather than candidate 4 with a much larger expression fold difference was chosen for follow-up.
Author Response	We thank the referee for this comment. We added all the details of regions we tested into the revised supplementary file. The reason we selected candidate 4 is that it is the highest scored variants in our analysis. We made this more clear in our new version.

Formatted Table
Formatted: Justified

Excerpt From Revised Manuscript	

<ID>REF5.30 – SYCP2 and beyond

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

Deleted: ? 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	

Formatted Table

<ID>REF5.31 – Utility of ENCODEC

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

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

Formatted: Justified

Formatted Table

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

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

Deleted: >

Deleted: TBC

Referee Comment	28. In Figure 2e, a p-value should be given with the analysis.
Author Response	We thank referee for the comment. We now have updated figure 2e with p-value.
Excerpt From Revised Manuscript	

Formatted Table

Formatted: Justified

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

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

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

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

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

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

Deleted: TBC

Formatted Table

Formatted: Justified

Deleted: Done

Formatted Table

Formatted: Justified

<ID>REF5.35 – Presentation

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

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

<ID>REF5.36 – Supplementary document

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

Referee Comment	32. The supplement contains multiple reference errors.
Author Response	We thank the referee on this comment and we have made numerous improvements to the supplementary document.

Deleted: TBC
Formatted Table
Formatted: Justified
Formatted: Justified
Deleted: Done
Formatted Table
Deleted: We've

Excerpt From Revised Manuscript	

Page 1: [1] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

DONE : Finished

%%%MORE : Go above and beyond the scope of the question and indicates more analyses

Page 3: [2] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 4: [3] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 4: [4] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 9: [5] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

[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

Page 10: [6] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 10: [7] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 11: [8] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

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.

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.

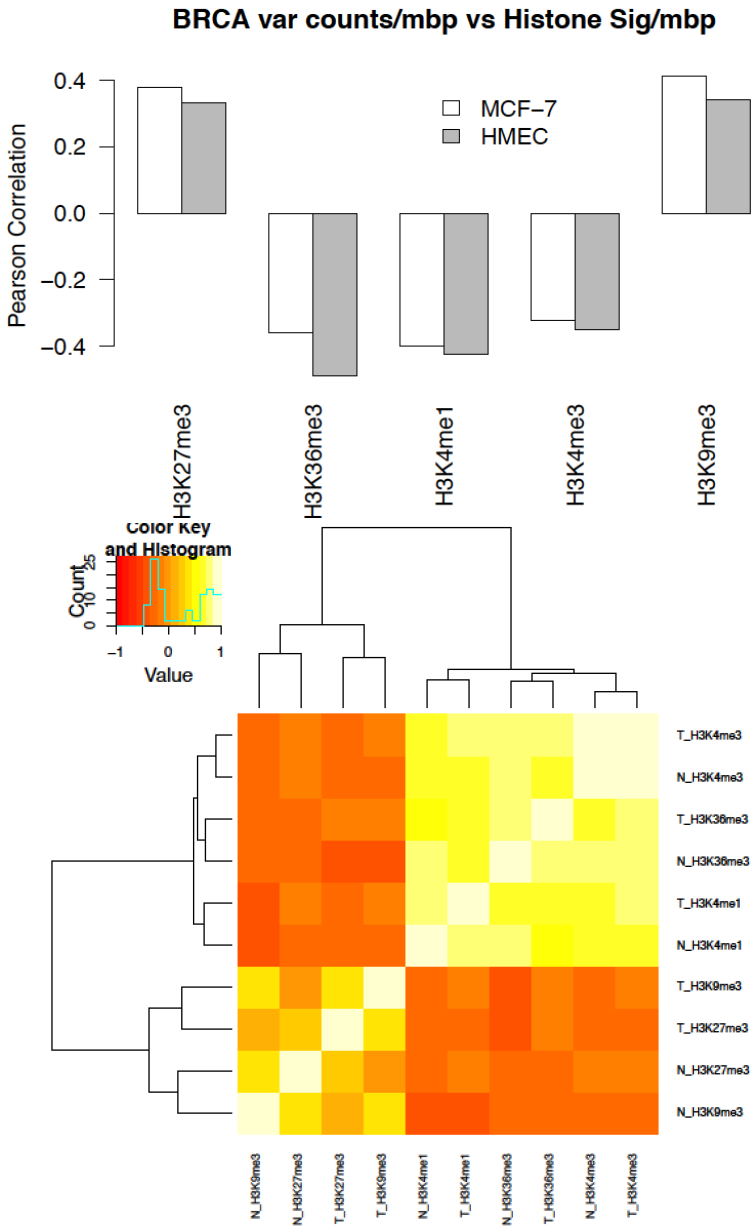
Page 13: [9] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

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:
(1.1) Even in the

Page 13: [10] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

, it is not always the case that cell-of-origin can be predicted perfectly using the epigenetic features (Fig. 4 b).
(1.2) the Polak 2015 paper only compare among normal tissues from the Roadmap data and they

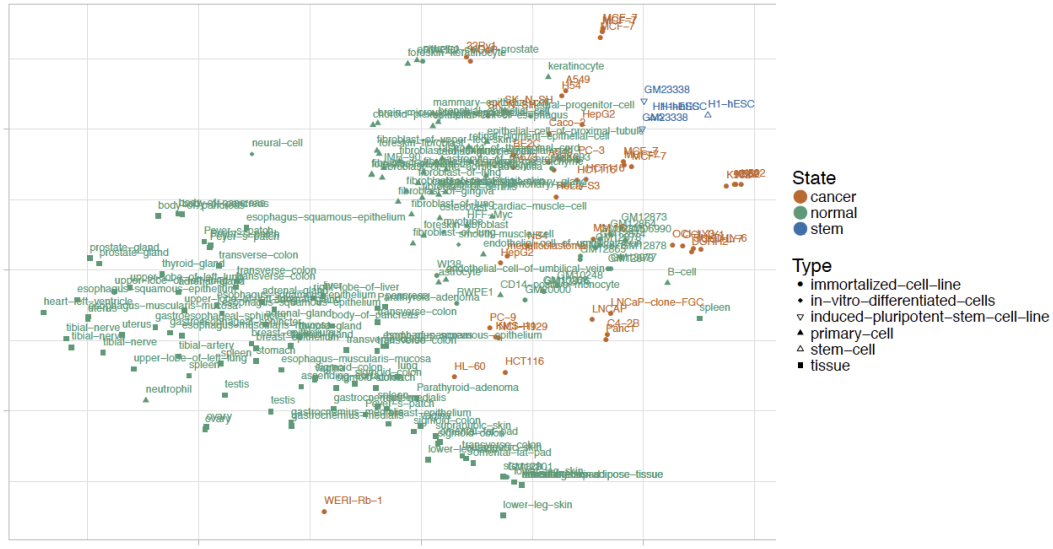
Page 13: [11] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**



2. In general

Page 14: [12] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

t–SNE: CTCF



Page 15: [13] Deleted	**jingzhang.wti.bupt@gmail.com**	**3/23/18 4:23:00 PM**

####Dictation

Page 19: [14] Deleted	**jingzhang.wti.bupt@gmail.com**	**3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 19: [15] Deleted	**jingzhang.wti.bupt@gmail.com**	**3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 19: [16] Deleted	**jingzhang.wti.bupt@gmail.com**	**3/23/18 4:23:00 PM**

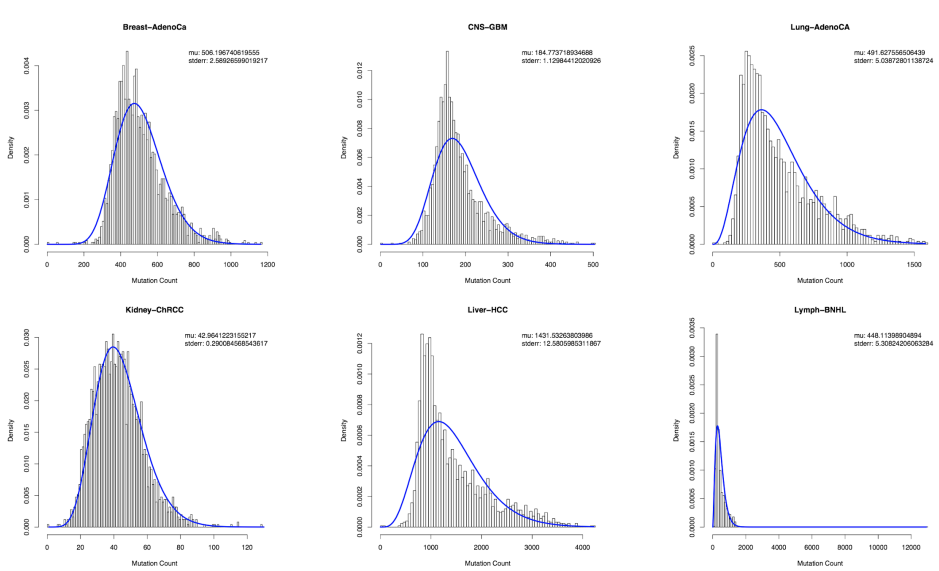
Page 19: [17] Deleted	**jingzhang.wti.bupt@gmail.com**	**3/23/18 4:23:00 PM**

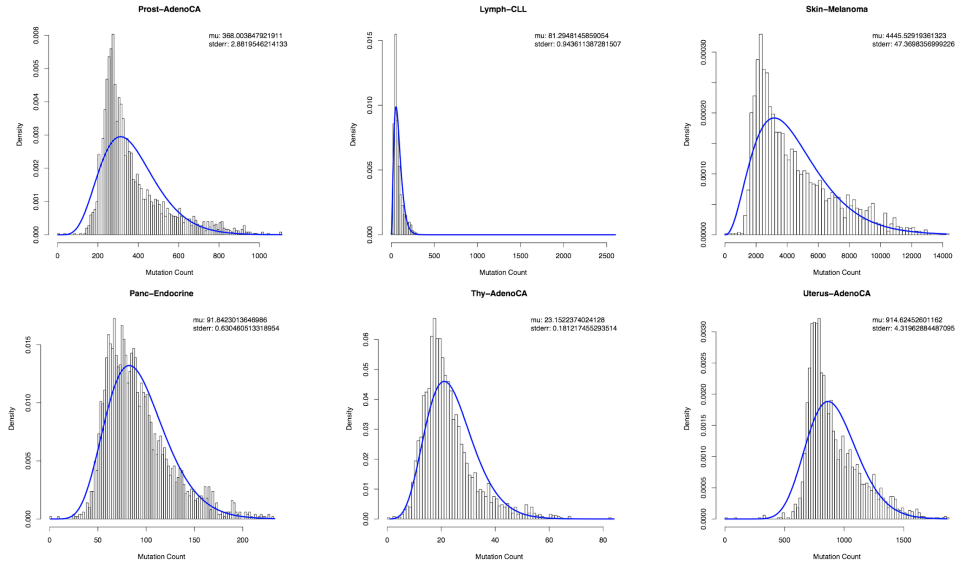
Excerpt From Revised Manuscript	

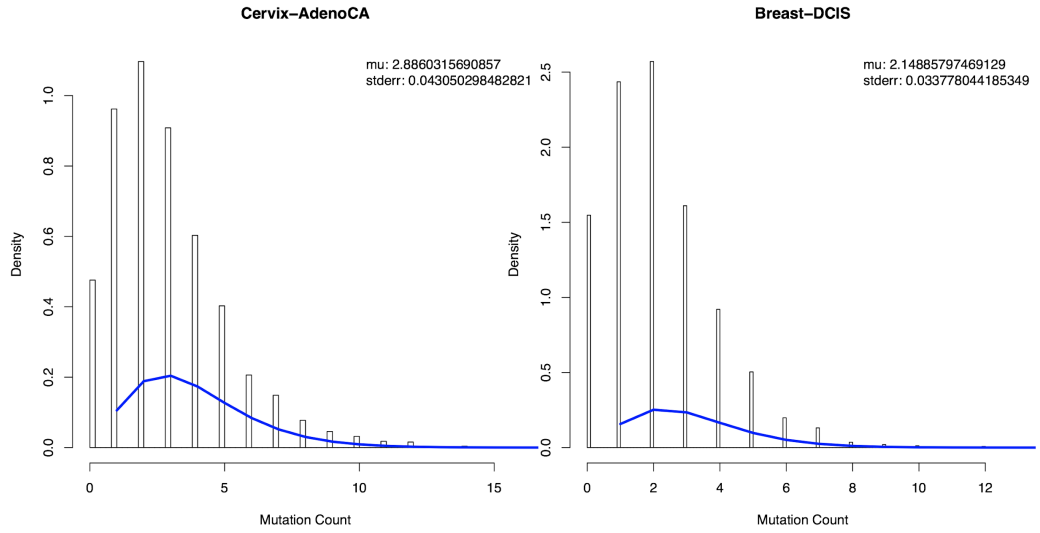
Page 19: [18] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

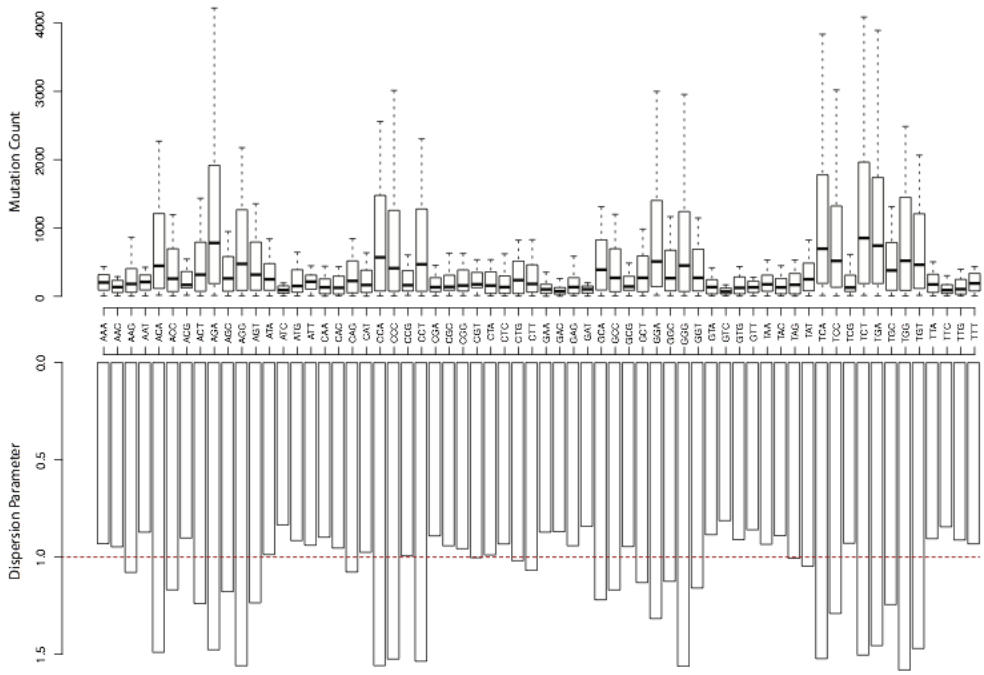
Page 20: [19] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**







Page 22: [20] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM



Page 28: [21] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

Excerpt From Revised Manuscript	

Page 28: [22] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

Excerpt From Revised Manuscript	

Page 30: [23] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

Excerpt From Revised Manuscript	

Page 30: [24] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 30: [25] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 30: [26] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 37: [27] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

[JZ2MG: do we need a preamble here? I don't feel strongly]

Page 37: [28] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 37: [29] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 39: [30] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 39: [31] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

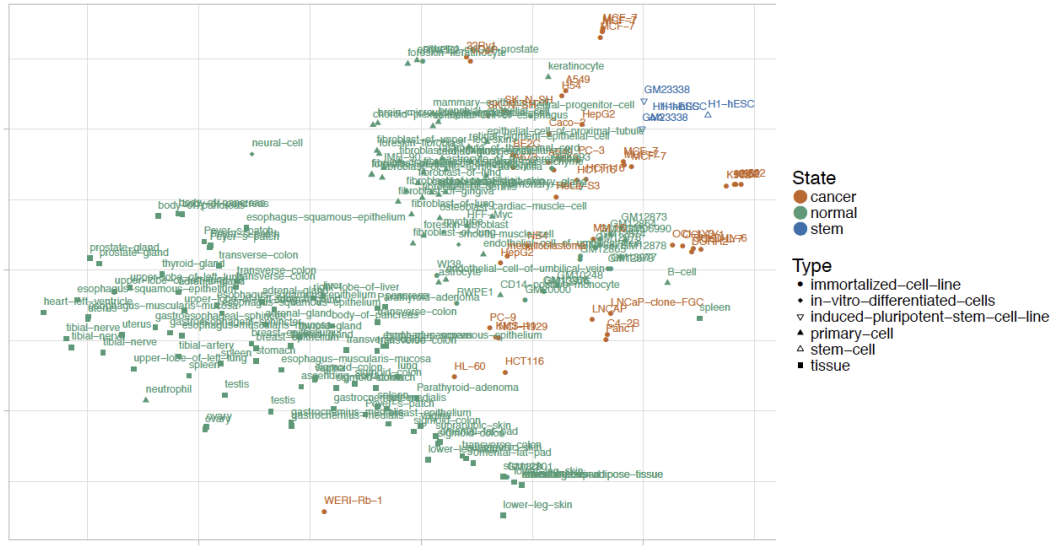
Excerpt From Revised Manuscript	

Page 39: [32] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

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.

Page 40: [33] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

t–SNE: CTCF



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

Page 41: [34] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

(DL maybe)

Page 41: [35] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

[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?]

Page 43: [36] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

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

===

Page 46: [37] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

DL - think about how we can change the figure

(We fixed the figure, Less data, more on overview schematic)

Page 48: [38] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

###JZ: strength of cell line, no heterogeneity, emphasize this, co-expression network
Can mention something related to single cells
Some clinically significant changes will occur in

####7mar - high level is how to connect

Page 48: [39] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

reference cell line to annotation to patient .... key pt of the paper ... peng's figure
Individualize the network a little bit

###WUM text###
The

Page 48: [40] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

greater emphasis.

Page 48: [41] Moved to page 63432 (Move #4)jingzhang.wti.bupt@gmail.com **3/23/18 4:23:00 PM**

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.

Page 48: [42] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

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,

Page 56: [43] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 56: [44] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 56: [45] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

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.

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).

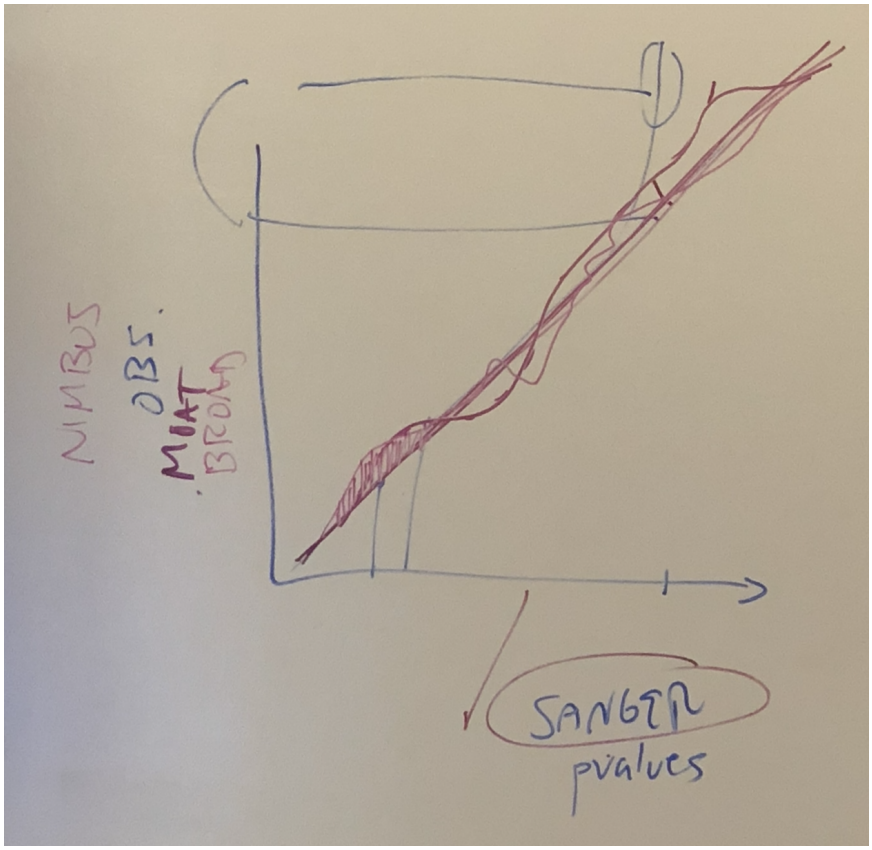
For example, many prior efforts to model BMR have been limited by the availability of genomic assays, or by the availability of assays matched by cell-type. For example, Lawrence et al., 2013, used HeLa replication timing data and K562 chromatin state via Hi-C. Martincorena et al., 2017, included histone 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.

Page 57: [46] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

* we're part of pcawg ... there's no benchmark,
There's a driver comparison but this is different
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

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]

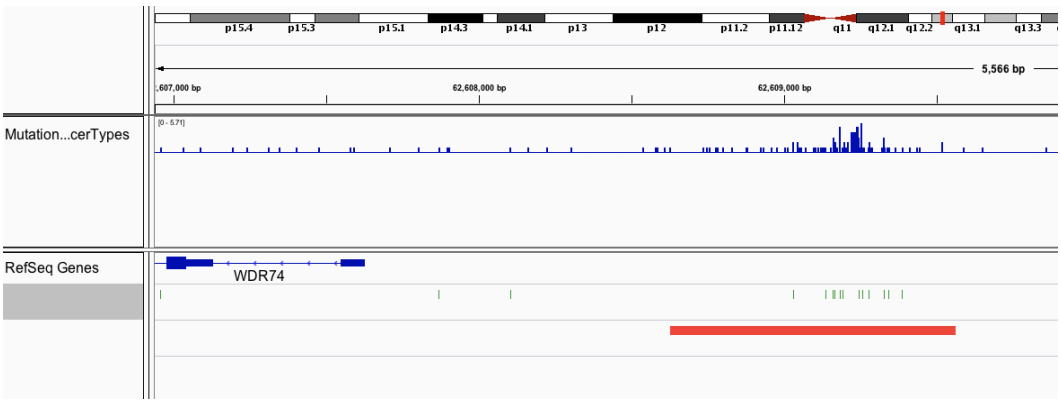


####7mar - compare the sanger rand v us (nimbus) in a qq

Page 59: [47] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

Page 62: [48] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

mutation hotspots (red block) for this well known driver site (blue line for pan-cancer and green line for liver cancer).



Page 62: [49] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

[JZ2MG: next week will check the status of KevinYip, SKL stuff added]
[JZ2XK: can you please update this figure and check this text?]

Page 63: [50] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

. Mix membership model is a hierarchical Bayesian topic model framework and can help to uncover the underlining semantic structure of a document collection.

Page 63: [51] Moved to page 63 (Move #7) **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

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 (

Page 63: [52] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

has been described in Blei 2003.

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.

To explore how well the LDA mixed-membership analysis on TF regulatory network, we extend our dataset

Page 63: [53] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

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

Page 63: [54] Moved to page 63 (Move #8) jingzhang.wti.bupt@gmail.com 3/23/18 4:23:00 PM

. 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

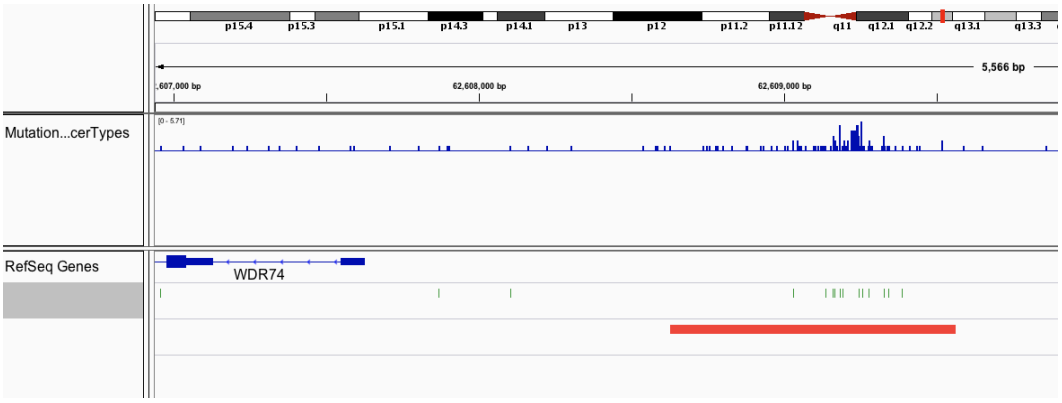
Page 65: [55] Deleted jingzhang.wti.bupt@gmail.com 3/23/18 4:23:00 PM

Page 66: [56] Deleted jingzhang.wti.bupt@gmail.com 3/23/18 4:23:00 PM

Page 66: [57] Deleted jingzhang.wti.bupt@gmail.com 3/23/18 4:23:00 PM

Excerpt From Revised Manuscript	

Page 67: [58] Deleted jingzhang.wti.bupt@gmail.com 3/23/18 4:23:00 PM



Page 67: [59] Deleted jingzhang.wti.bupt@gmail.com 3/23/18 4:23:00 PM

BCL6 mutations were found in promoter region.

XK, TG
@@@7mar - yuck!
Are any SVs associated with BCL6?

Page 68: [60] Moved to page 56 (Move #5) jingzhang.wti.bupt@gmail.com 3/23/18 4:23:00 PM

1.

Page 68: [61] Moved to page 56 (Move #6)	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

2.

Page 72: [62] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

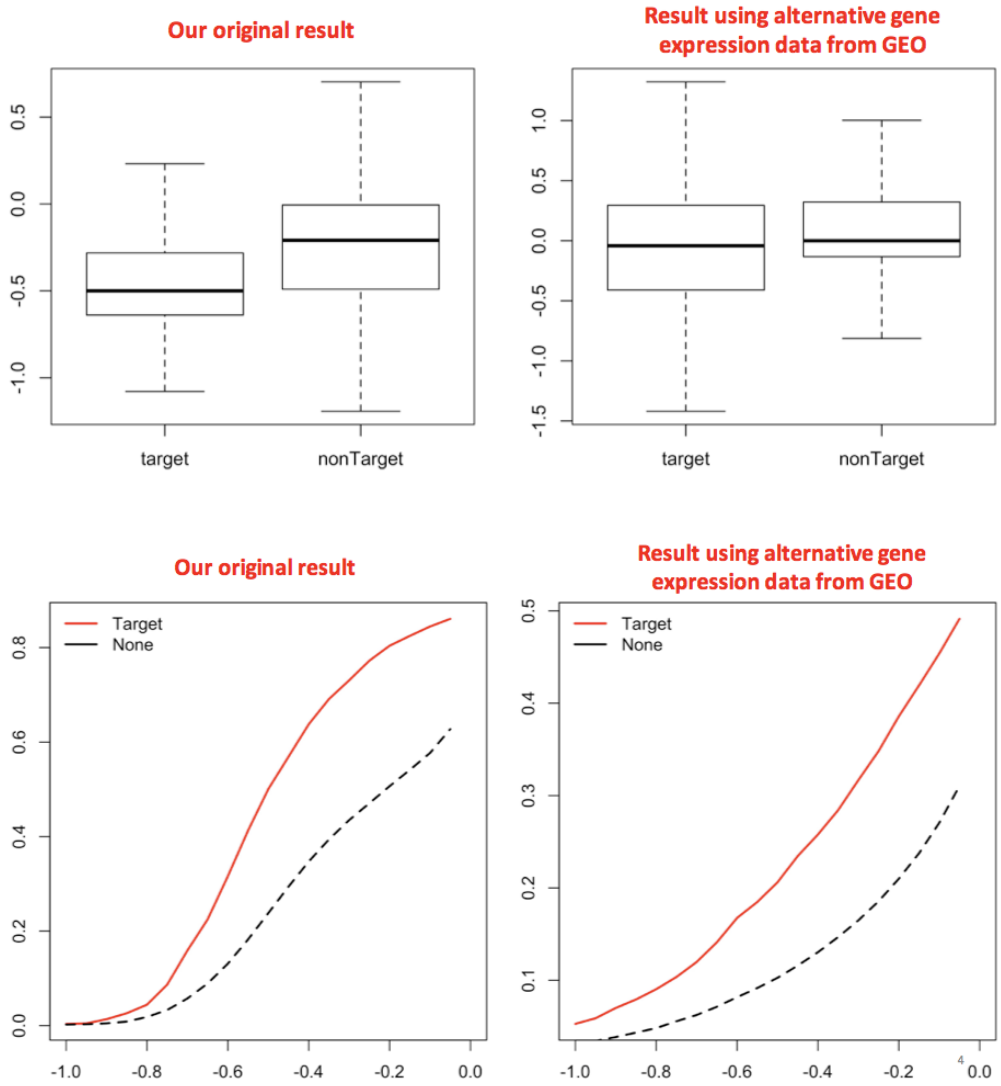
line.

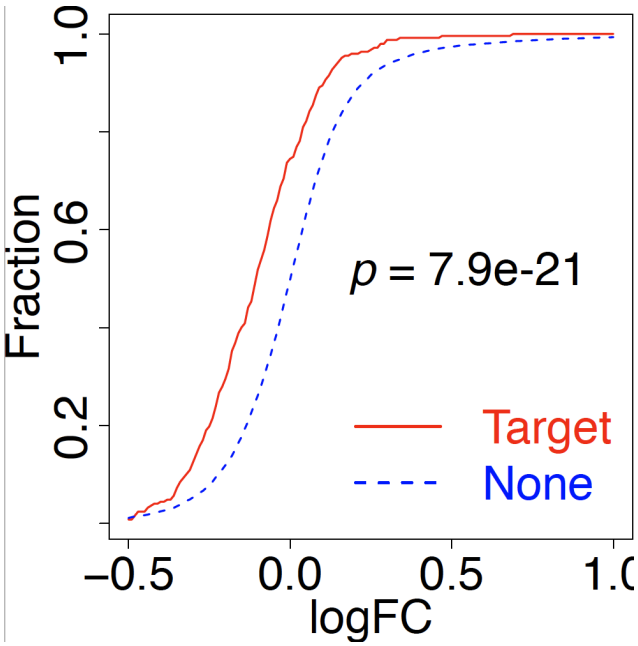
We note that, even though these alternative analyses were conducted on a different cell line, the results we obtain (shown below in the right panels, and now made available in the supplementary materials) validate the behavior of the network, and they are consistent with our previous results (in which gene expression was measured in the MCF7 cell line).

Page 72: [63] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

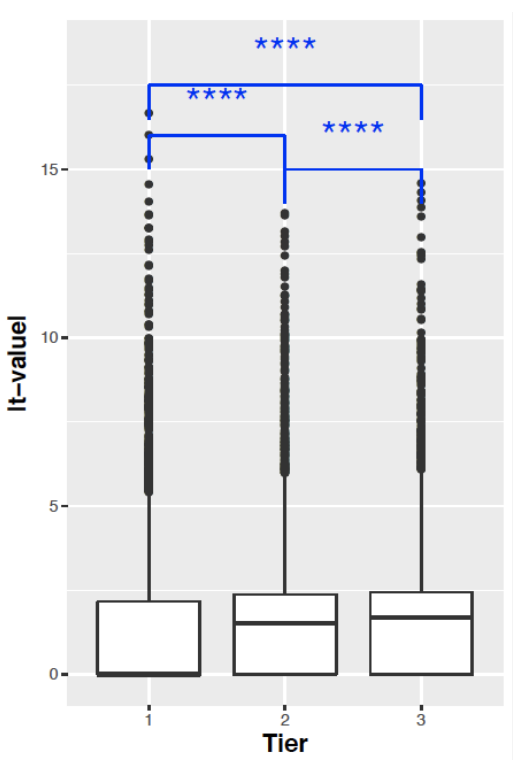
We also found another array based MYC knockdown data the results correlate well

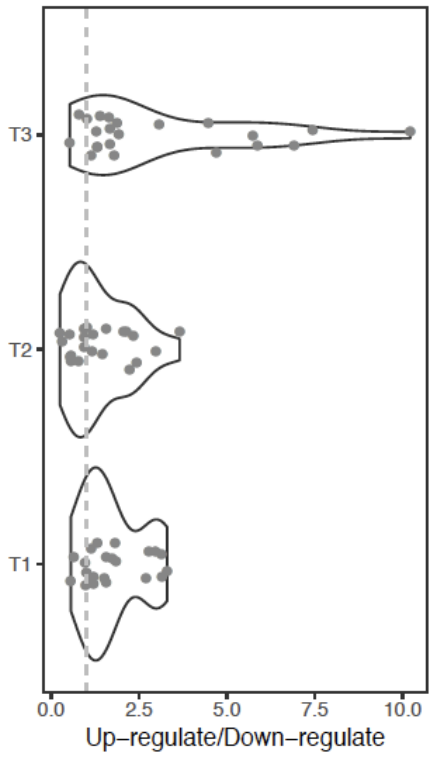
Page 72: [64] Deleted	jingzhang.wti.bupt@gmail.com	3/23/18 4:23:00 PM

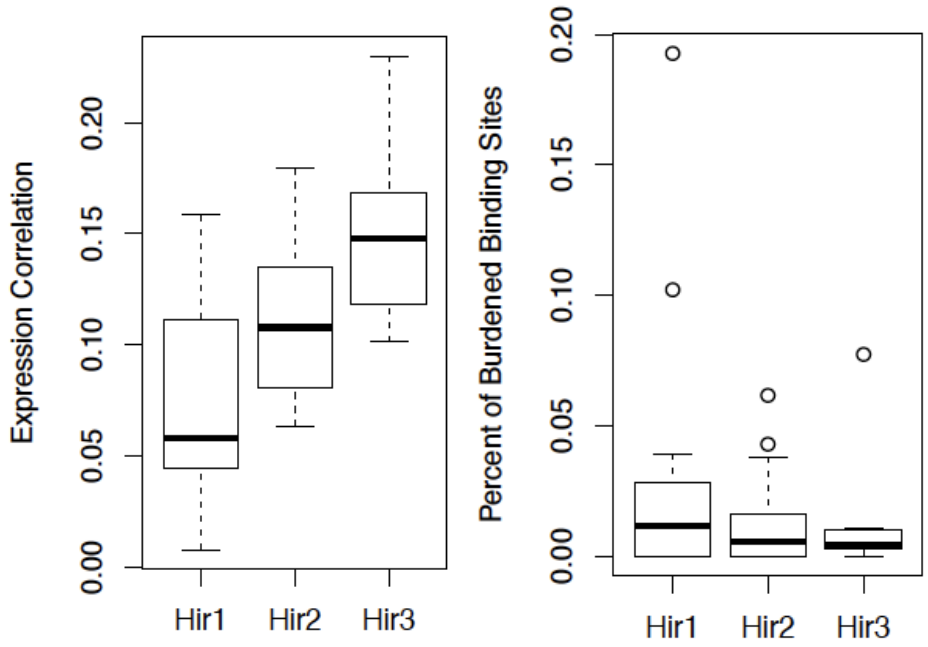




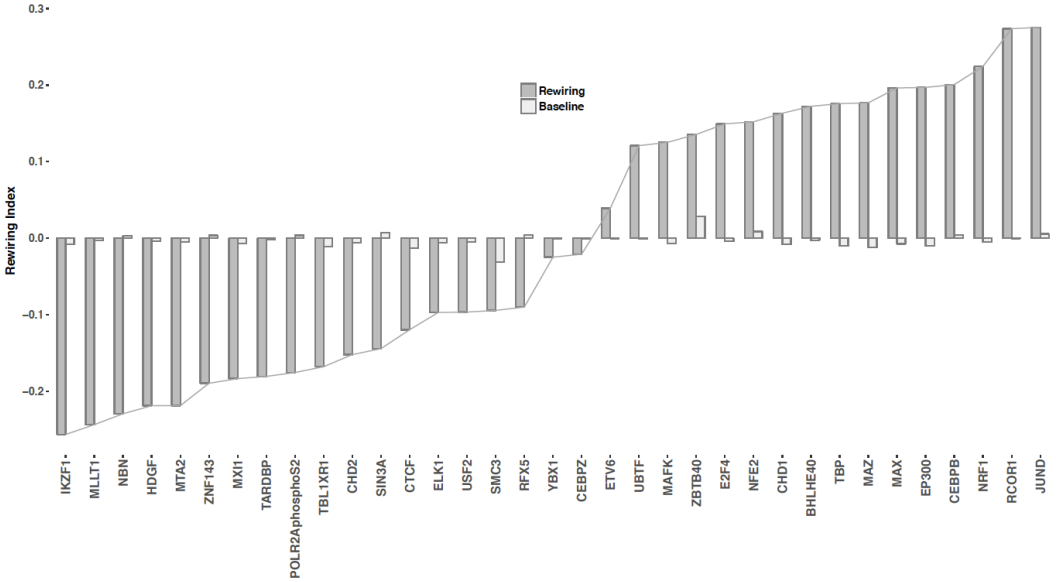
Page 76: [65] Deleted jingzhang.wti.bupt@gmail.com 3/23/18 4:23:00 PM

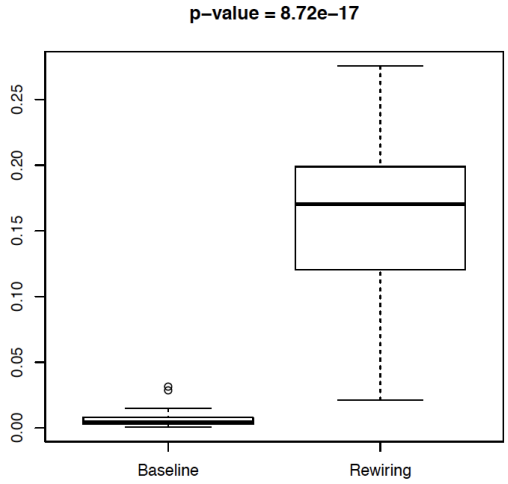






Page 78: [66] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**





Page 78: [67] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Page 79: [68] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

####7mar we truly thank referee. Took referee's comment to heart, made hugh improvement

Page 79: [69] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

to do - same as 16
False positive rate analysis
Think about test of significance (have some more analysis) DL/JZ disc.

Page 80: [70] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 81: [71] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 82: [72] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 82: [73] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	

Page 83: [74] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

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!

Page 83: [75] Deleted **jingzhang.wti.bupt@gmail.com** **3/23/18 4:23:00 PM**

Excerpt From Revised Manuscript	