Tags:

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

@@@ : assignment

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

---

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:

-- Editor 0.1 – Overall comments on the paper --

$$$Presentation @@@MG &&&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've tried to respond to extensively revise our manuscript in our new version. In summary, (JZ2MG: this is to the editor, not to referees, but still can be seen by referees, how much detail should we go to)
Excerpt From Revised Manuscript	

-- Editor 0.2 – Overall comments on the paper --

$$$Presentation @@@MG @@@JZ &&&compl)

Referee Comment	The referees also find that the current manuscript provides limited context with prior studies using similar approaches for use of prior ENCODE and Epigenome Roadmap datasets in cancer genomics. They detail the need for clearer presentation in context of prior studies as well comparisons to demonstrate advance.
Author Response	We thank the referees for this comment and have clarified the unique aspects of our paper.
Excerpt From Revised Manuscript	

-- Editor 0.3 – Overall comments on the paper --

$$$Presentation @@@MG @@@JZ&&&compl

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	*(Core of the argument)* It is unique in the following aspects.
Excerpt From Revised Manuscript	We disagree with the reviewer with regarding the new dataset and novelty of this manuscript. **1. Regarding the separate dataset to the main ENCODE manuscript** First, unlike previous roll-outs, ENCODE 3 does not associate specific data sets with specific papers. In addition, there are no dependencies between any of the papers in this package. All the ENCODE data is open to the public and is not associated with, for instance, the encyclopedia paper or a particular companion paper. In addition, while the encyclopedia paper considers annotations across cell-types (currently the center-piece of ENCODE), it does not take advantage of the cell lines rich in data. The ENCODEC paper takes a complementary approach by constructing cell-type specific annotations from cell lines rich in assay data. This ENCODEC Paper is unique *in its inclusion of replication timing data, STARR-seq and Hi-C data, rich annotations, and extensive network information*. None of these aspects are discussed in the main ENCODE manuscript. **2. Regarding the novelty of the manuscript** In the initial submission, the BMR calculations in the original manuscripts only occupy two sub figures of six total figures, but received the most criticism. We thank the referees pointing out a serie of references on this topic, especially the Martincorena et al 2017 paper. Note that this paper come out in Nov 2017 and we did our submission on Aug 22 2017. There are lots of other important discoveries in the paper as listed below. We also take the comments in heart and did a major revision to expand the novelty part. *2a) extensive regulatory networks for various cancer types*

In our revised ENCODEC manuscript, we provided a universal TF and RBP regulatory network based on xxx ChIP-seq and eCLIP experiment. Combining with RNA-seq data from TCGA, we proved that our network is more accurate than previous ones based on pure computational predictions.
Using these networks, we prioritized known regulators such as TP53 and ESR1, and used both ENCODE and public TF knockdown datasets as validations. We also pinpointed out a potential novel oncogene SUB1. It serves as a RNA binding protein to bind to the far most of 3' UTRs to up-regulate its target gene expressions. We also found that targets of SUB1 have a slower decay rate, indicating its important roles in regulating stability of mRNAs. In addition to looking at universal (not cell type specific) ChIP-Seq networks, we also look at network changes on a large-scale, tissue-specific manner. We feel that the rewiring of networks is best exemplified in cancer cells.

2b) More accurate annotations after integrating new types of assays
Our ENCODEC manuscript takes a complementary approach by constructing cell-type specific annotations from cell lines rich in assay data. These annotations are important in power calculations related to recurrent mutations. This highly accurate annotation takes advantage of next generation assays such as STARR-seq and elements linked by ChIA-PET and Hi-C. This is not possible obviously in the general and co-annotation but it's extremely useful on the cancer context.

2c) Replication timing data
Although a major feature of ENCODE is replication timing, none of the other papers use it. Previous work [[cite]] on mutation burden calculation usually selected replication timing data from HeLa cell line due to the limited amount of data available. The wealth of the ENCODE replication timing data will greatly help to parametrize somatic mutation rates. We will highlight this in our revised manuscript.

2d) Structural variations
One unappreciated aspect of ENCODE is that next generation functional assays, in addition to characterizing functional elements in the genome, enable one to determine structural variants. This has been the case for the Hi-C experiments, but there are many other experiments done by experimentalists that have given rise to a large number of structural variants. These structural variations of course are most applicable to the cancer cell lines that many of the ENCODE assays have been run on. We have referenced these structural variations in the earlier version of the paper but admittedly have not really highlighted them or talked about them as much. Since ENCODE provides novel SV data and inclusion of SV analysis was suggested by some of the referees, we have greatly expanded our analysis of SVs in the context of cancer. We will include some new figures as well as add a variety of new data sets that have been designed specifically for this project.

2e) TF/RBP knockdown/knockout experiments
ENCODE has 77 CRISPRi based TF knockout and and 533 shRNA based TF/RBP knockdown experiments, which serves a great resource to investigate network perturbations after disruption of a regulator. The ENCODEC paper is the only paper that focuses on such data. In our current manuscript, we have already used some of such

	knock down data to validate effects of key regulations in multiple cancer types. We will highlight the usage of such experiments in our revised version.

Referee #1 (Remarks to the Author):

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

$$$NoveltyPos

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

-- Ref 1.2 – Scope of the paper --

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

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

Author Response

We thank the reviewer for pointing out the existence of other literature that relates to the significant problems we address.

WM
Certainly, several groups have used epigenetic data before to model gene regulatory networks and background mutation rates. The present work is novel in that it offers substantially wider and deeper integration of diverse, specialized functional assays than does any prior work, including use of the specialized functional assays STARR-seq, Repli-Seq, and eCLIP and an analytically more sophisticated construction of gene regulatory networks.

Our significant findings include the following:
1. Our compact enhancer annotations and extended gene definitions increase our power to detect significantly burdened genes. This would have permitted, for example, the discovery of CANCER_GENE_X with a Y% smaller cohort.
2. Network-driven estimates of the contributions of all relevant TFs to the expression levels of all cancer genes. For example, we find that TF1 and TF2 most upregulate EGFR expression in lung adenocarcinoma.
3. The regulatory network changes from the cell-of-origin to cancer cell bring cancer cells to a more stem-like state.

##
However, we disagree with the reviewers regarding the novelty of this paper.

The main focus of our work is to perform large scale data integration in order to tailor the ENCODE resource for cancer research. For example, in our manuscript we showcase the value of ENCODE data in improving BMR to aid in driver mutation detection. As mentioned by the reviewer, an improved cancer driver detection method is an important topic that been the subject of several high profile papers (see, for example, Martincorena, 2017, published 3 months after our submission). Such prior results do not compromise the novelty of our paper and signify the value of our data. Our purpose is not to propose a better cancer driver detection method, but to highlight the value of ENCODE data applied to cancer research.

There are many elements of our initial submission that we have been expanded in this revision. These elements are summarized in the section above. As an ENCODE resource paper, we hope the deliverables, including processed raw signal files, a more accurate and more compact genome annotation, and extensive tissue-specific and universal regulatory networks provide value to the cancer community.

Excerpt From Revised Manuscript	

-- Ref 1.3 – BMR --

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

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 have to say why these papers don't do as accurate estimation We thank the reviewer for identifying these references. We recognize that epigenetic features have been previously been used to estimate BMR and improve driver mutation detection. Our aim was not to produce novel BMR estimation models, but rather to showcase how ENCODE data can help improve the performance of such models. 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). In addition, were able to use cell-type matched feature data across our BMR analysis. This includes more commonly used features for BMR modification, like the 932 histone modification features we used, but also many other features, especially the 51 replication time data, that have proven useful but are less frequently incorporated into BMR models.

	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, only 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. ~~Admittedly, we agree that this part is less novel as compared to other sections. We have moved two related sub panels to the supplement and~~
Excerpt From Revised Manuscript	

-- Ref 1.4 – BMR --

$$$BMR $$$Text @@@WM @@@JZ&&&compl

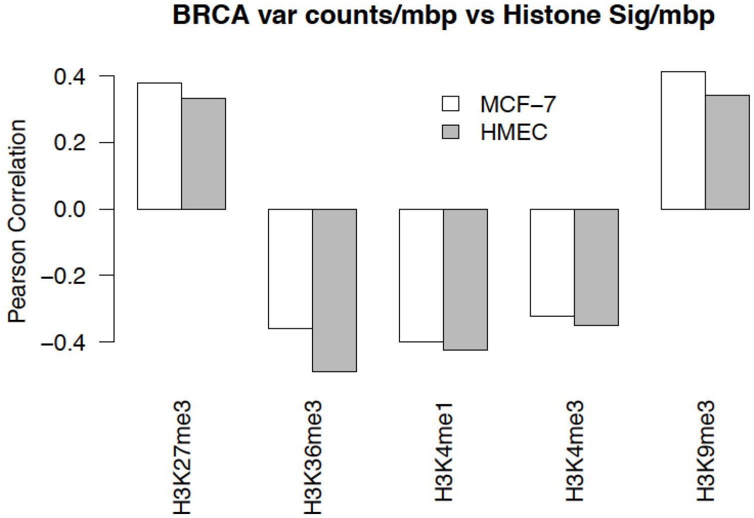
Referee Comment	Most large-scale cancer genome sequencing papers also have models at various levels sophistication, most of them including the issue of proper tissue-type matching. Importantly, Polak et al, 2015 (Cell-of-origin chromatin organization shapes the mutational landscape of cancer, Nature) 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. Thus, that "matched" cell lines are better than unmatched or addition of more epigenetic features results in some improvement is almost trivial at this point. Which marks contribute to this is also not new.
Author Response	We agree that it is not novel to say "matched" cell lines are better predictors of mutation rates, as we also cited the Polak 2015 paper in our initial submission. However, our point is not to provide a "novel" driver detection method, but rather to highlight the value of

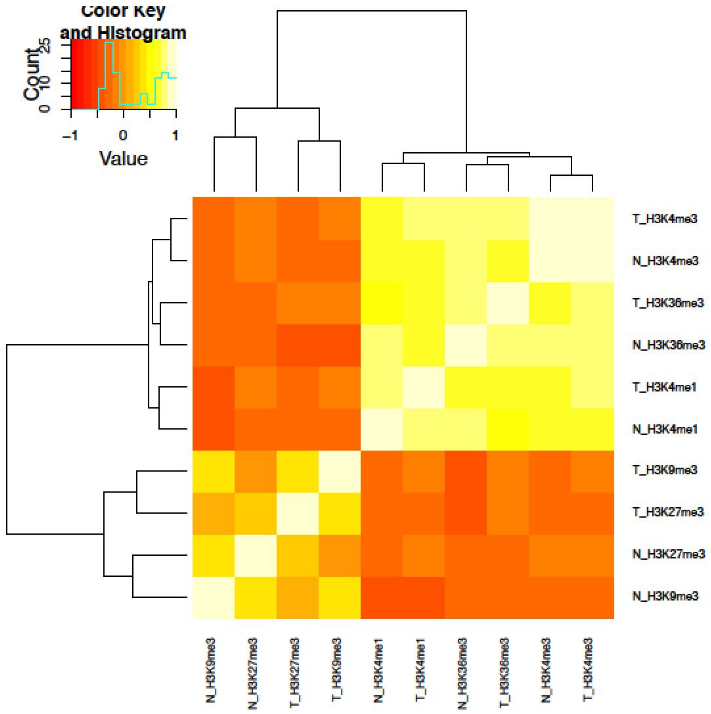
	various types of ENCODE data, some of which are unique features, such as replication timing.
Excerpt From Revised Manuscript	

-- Ref 1.5 – BMR Match between Tissues & Cell lines --

$$$BMR $$$Calc @@@JZ @@@JL&&&compl

Referee Comment	Stepping back, it is not obvious to me that using the ENCODE cell lines, despite the availability of more epigenetic data, is the best approach to calculating the background rate in the first place—they briefly mention that using cell lines (rather than tissues) can be problematic, but do not explore this further. If this were a regular research paper, the authors would have to shown how the proposed approach is different and how it is better than methods already available.
Author Response	Thanks for pointing out the Polak 2015 paper. (Note we did cite this in our manuscript.) 1. First we want to emphasize some specific type of data from ENCODE such as Hi-C and replication timing. By pointing out that using data from a matched cell line is better, is not used as a novel conclusion (as we also cited the Polak 2015 paper), but rather to emphasize the value of our data. Take replication timing as an example, a lot previous work ([jz cite]) actually use replication timing data from Hela cell line due to the limited choice. In our revised manuscript, we described that there are xxx high quality replication timing data, which is quite valuable for cancer genomics. 2. Regarding the cell line data, we still think they are quite useful to predict the mutation rates. Even in the the Polak 2015 paper, it is not always the case that cell-of-origin can be predicted perfectly using the epigenetic features (Fig. 4 b). 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). As seen from the following figure, MCF-7 provides similar (and sometimes even better correlation with mutation counts). We also found that histones from tissue and matched cell lines are actually quite correlated in a larger scale (see heatmap below). * pls note that polak et al don't consider cell line catalog * also not at all clear that cancer lines aren't better proxy for tumor mut. Than normal tissue . see below .. make a suppl. Figure... * replication timing is often the best feature & is only in cell lines





3.In general, tissue data is always more difficult and there are less such data. On the contrary, the cell line functional characterization data has lots of advantage in terms of assay richness. For some specific cancer types, such prostate cancer, cell lines like LNCap might further help.

Excerpt From Revised Manuscript	

-- Ref 1.6 – Difference between ENCODEC and FunSeq --

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

Referee Comment	The rest of the sections (and their corresponding supplement sections) are variable in significance and quality. That ENCODE data helps in 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	Variant/regulator prioritization is one of the most important applications of the ENCODEC resource. We want to clarify that our current approach is completely different from the Funseq approach (as shown in Fig 6 in the initial submission). Funseq takes all the broad annotations from ENCODE2 from various cell types to prioritize SNVs/indels. However, with the increased number and novel types of assays from ENCODE3, our current prioritization scheme follows tissue specific manner. It adopts a top-down scheme: 1) first combine cohort level expression level to prioritize key regulators; 2) then combine patient expression profiles and epigenomic features to prioritize key regulatory elements; 3) the pinpoint the SNVs after incorporating final scale features like motif breaking, conservation, and etc. The tissue-specific features, network perturbations, and integration of external expression/mutation features are all new in our current proposal.
Excerpt From Revised Manuscript	

-- Ref 1.7 – Novelty and presentation of the paper --

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

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. Personally, I wonder whether a review paper that gives an update to the ENCODE database and state the illustrative examples succinctly might be more appropriate than several studies, in which more work/descriptions are needed to show novelty, packaged together?
Author Response	We thank the referee for praising the new STARR-seq assays and we incorporated more STARR-Seq data to the revised manuscript. We also added new data types, like several whole genome sequencing of the cell lines in our revised manuscript. We incorporated more TF/RBP knockdown data to validate our prioritized known and novel key regulators, such as TP53, ESR1, ZNF687, and SUB1. We wish to point out that this is not designed as a paper with novel findings about cancer genomics but rather an illustration of powerful resource
Excerpt From Revised Manuscript	

Referee #2 (Remarks to the Author):

-- Ref 2.1 – Novelty of the paper --

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

<table>
<tr><td>Referee Comment</td><td>The manuscript does not report new datasets in addition to the ENCODE release and offers limited conceptual novelty.</td></tr>
<tr><td>Author Response</td><td>
<p>We thank the referee for pointing out the dataset problem, which we believe is more of a presentation issue. In addition to the massive traditional assays such as ChIP-seq, DNAse-Seq, and RNA-seq, we incorporated a list of new data types as summarized below.</p>
<table>
<tr><td>Assay</td><td>More info</td></tr>
<tr><td>STARR-seq</td><td>K562, MCF-7, LNCaP, HepG2</td></tr>
<tr><td>Hi-C</td><td>K562, MCF-7, LNCaP, HepG2, etc ...</td></tr>
<tr><td>Replication timing</td><td>Xxx cell lines</td></tr>
<tr><td>CRISPERi based knockout</td><td>77</td></tr>
<tr><td>shRNA based knowck down</td><td>533</td></tr>
<tr><td>SV/SNV call set</td><td>Xxx cell lines</td></tr>
<tr><td>Bionano</td><td>Xxx cell lines</td></tr>
<tr><td>WGS</td><td>Xxx cell lines</td></tr>
</table>
<p>In addition, we wish to point out that unlike previous roll-outs, ENCODE 3 does not associate specific data sets with specific papers. In addition, there are no dependencies between any of the papers in this package. All the ENCODE data is open to the public and is not associated with, for instance, the encyclopedia paper or a particular companion paper.</p>
<p>We thank the reviewers about the comments on presentation of this manuscript. We want to emphasize that the main goal of ENCODEC is about ENCODE resource for cancer community, instead of novel scientific cancer discoveries. By integrating the novel types of assays with massive traditional assays, we provide the following list of resources.</p>
<ul>
<li>Ready to use signal files that can help BMR estimation, including the ones that are quite limited in existing methods such as replication timing and Hi-C</li>
</ul>
</td></tr>
</table>

	● Accurate and compact annotation for assay rich cell lines for somatic mutation hotspot detection. ● Accurate enhancer gene linkage supported by multiple type of advanced assays like STARR-seq and Hi-C ● Universal and tissue-specific experimental based TF/RBP networks ● Imputed TF networks for more than 20 cancer types ● Paired tumor to normal networks to investigate network perturbations ● High quality SNV and SV calls from WGS and other types of assays
Excerpt From Revised Manuscript	

-- Ref 2.2 – Comment on utility of the resource --

$$$NoveltyPos

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

-- Ref 2.3 – Comparison of negative binomial to other methods --

$$$BMR $$$Text $$$Calc @@@JZ&&&compl

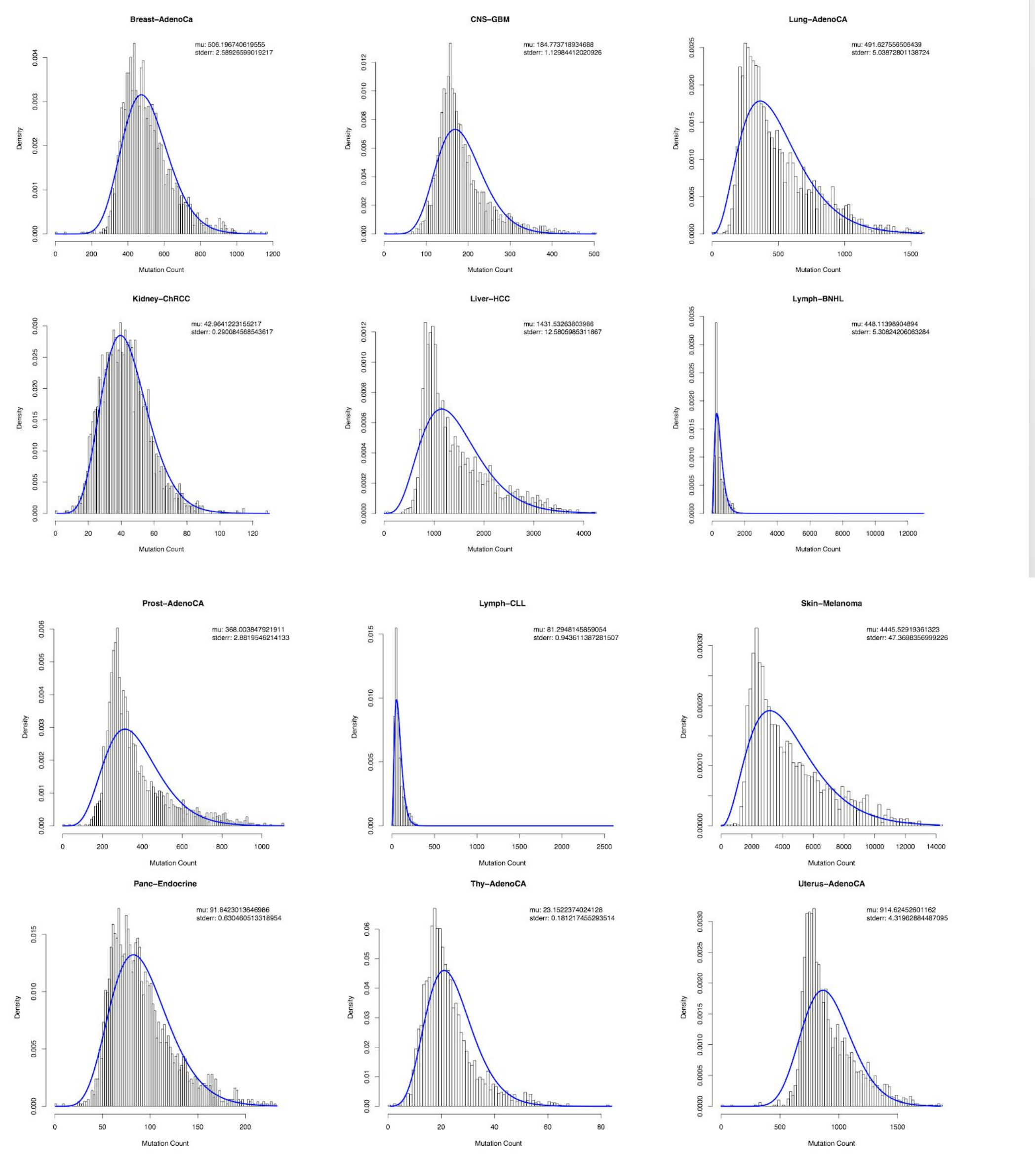
Referee Comment	1) The negative binomial regression (Gamma-Poisson mixture model) was introduced in Nik-Zainal et al. Nature 2016 and Marticorena et al., Cell 2017. Why was not this available

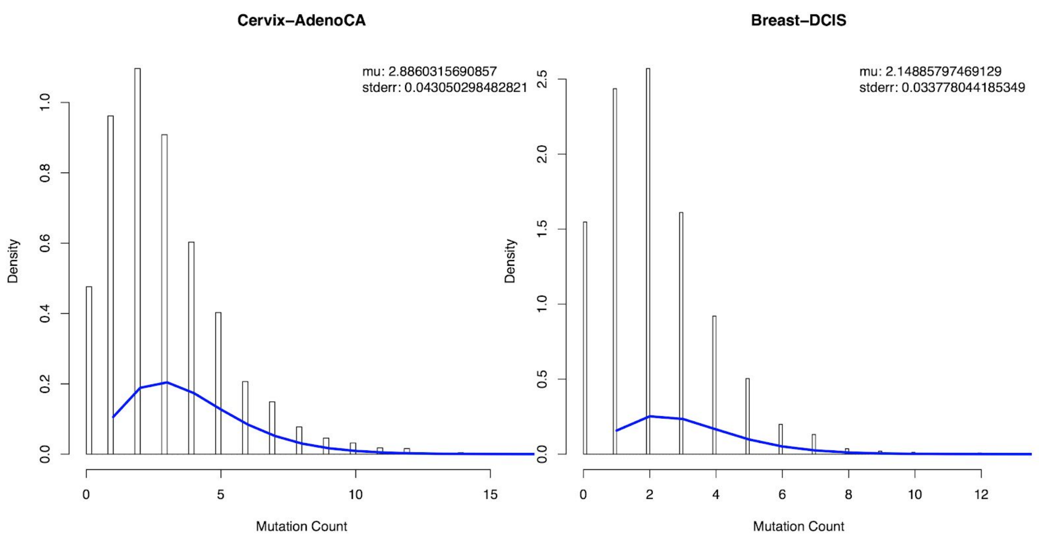
	method applied, and what is the benefit for the procedure used by the authors?
Author Response	We thank the reviewer to point out these references and they are also good models. Actually one of mentioned paper (Marticorena, 2017) was published on Nov 2017, almost three months after our submission. We want to emphasize that the goal of this paper is not to propose novel cancer driver detection method, but rather than highlight that ENCODE data can help BMR estimation, also in those model. In our revised manuscript, we tuned down this part by moving two sub-panels (A & B) in Figure 2 to the supplementary figures. We also added these references and clarified our point by proper acknowledgement.
Excerpt From Revised Manuscript	

-- Ref 2.4 – Questions about the Goodness of fit of the Gamma-Poisson Model --

$$$BMR $$$Calc (little) @@@JZ&&&compl

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


Breast−AdenoCa
mu: 506.196740619555
stderr: 2.58926599019217
Density
0.000 0.001 0.002 0.003 0.004
0 200 400 600 800 1000 1200
Mutation Count
CNS−GBM
mu: 184.773718934688
stderr: 1.12984412020926
Density
0.000 0.002 0.004 0.006 0.008 0.010 0.012
0 100 200 300 400 500
Mutation Count
Lung−AdenoCA
mu: 491.627556506439
stderr: 5.03872801138724
Density
0.0000 0.0005 0.0010 0.0015 0.0020 0.0025
0 500 1000 1500
Mutation Count
Kidney−ChRCC
mu: 42.9641223155217
stderr: 0.290084568543617
Density
0.000 0.005 0.010 0.015 0.020 0.025 0.030
0 20 40 60 80 100 120
Mutation Count
Liver−HCC
mu: 1431.53263803986
stderr: 12.5805985311867
Density
0.0000 0.0002 0.0004 0.0006 0.0008 0.0010 0.0012
0 1000 2000 3000 4000
Mutation Count
Lymph−BNHL
mu: 448.11398904894
stderr: 5.30824206063284
Density
0.0000 0.0005 0.0010 0.0015 0.0020 0.0025 0.0030 0.0035
0 2000 4000 6000 8000 10000 12000
Mutation Count
Prost−AdenoCA
mu: 368.003847921911
stderr: 2.8819546214133
Density
0.000 0.001 0.002 0.003 0.004 0.005 0.006
0 200 400 600 800 1000
Mutation Count
Lymph−CLL
mu: 81.2948145859054
stderr: 0.943611387281507
Density
0.000 0.005 0.010 0.015
0 500 1000 1500 2000 2500
Mutation Count
Skin−Melanoma
mu: 4445.52919361323
stderr: 47.3698356999226
Density
0.00000 0.00010 0.00020 0.00030
0 2000 4000 6000 8000 10000 12000 14000
Mutation Count
Panc−Endocrine
mu: 91.8423013646986
stderr: 0.630460513318954
Density
0.000 0.005 0.010 0.015
0 50 100 150 200
Mutation Count
Thy−AdenoCA
mu: 23.1522374024128
stderr: 0.181217455293514
Density
0.00 0.01 0.02 0.03 0.04 0.05 0.06
0 20 40 60 80
Mutation Count
Uterus−AdenoCA
mu: 914.62452601162
stderr: 4.31962884487095
Density
0.0000 0.0005 0.0010 0.0015 0.0020 0.0025 0.0030
0 500 1000 1500
Mutation Count

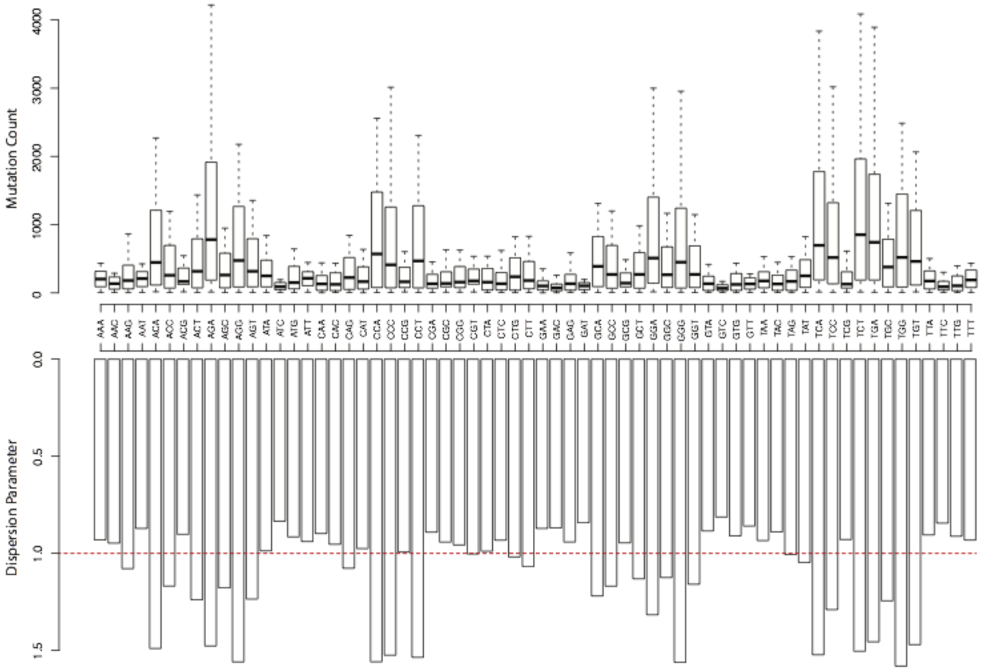
	
Excerpt From Revised Manuscript	

-- Ref 2.5 – Was the Poisson Model used for low mutation cancers --

$$$BMR $$$Text $$$Calc (little) @@@JZ @@@JL&&&compl

Referee Comment	2) It seems that the Poisson model was not rejected for cancers with very low mutation counts (liquid tumors). Is this a power issue rather than the property of the mutation process?
Author Response	We thank the reviewer for pointing this out. To answer this question, we plotted the overall mutation count under different 3mer context vs. the estimated overdispersion parameter (using the AER package) in R in the following figure. On one side, it is obvious that for 3mers with higher number of variants, there is a tendency of larger overdispersion. However, we also think it is due to variance-to-mean relationship. A larger variation usually accepts the Negative binomial distribution. We admit that it is possible that with data sets with lower count of variants Poisson model might be more likely to get rejected. However, it is also related to the heterogeneity of the data.

<table>
<tr><td></td><td>Many other methods (such as Marticorena, 2017) directly use Negative Binomial regression. It is simpler to not introduce additional parameters. But we think it is better to check how heterogeneous the count data is even after correcting various covariate effects.



</td></tr>
<tr><td>Excerpt From Revised Manuscript</td><td></td></tr>
</table>

-- Ref 2.6 – Cross validation analysis to do model selection --

$$$BMR $$$Calc @@@JZ&&&compl

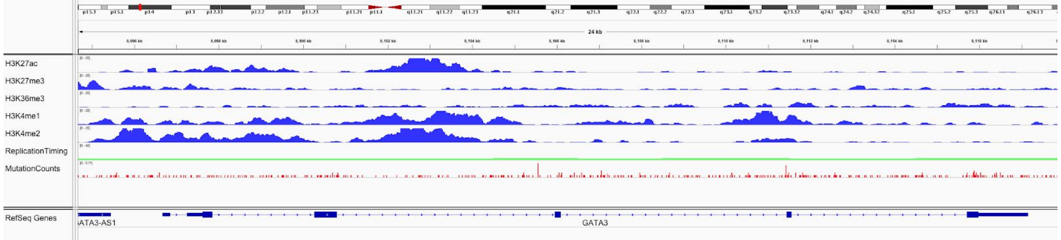
Referee Comment	3) The approach with principal components used for the BMR estimation does not seem to work well. Starting with the second PC most components have roughly the same prediction power. One possibility is that higher principle components do not capture the additional signal and reflect noise 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 referees for understanding/agreeing with our point - a lot of data helps PCs are not part of our model - it is just for the demonstration purpose. And we did not use it in our final BMR estimation. In the revised version, we actually used forward selection to show that adding more data will greatly help with the BMR prediction. Although we did see PCs used in other literature (Marticorena, 2017). It is mainly for a less computationally intensive feature selection procedure to use on all cancer types. In our analysis, we believe that various cancer types might come from completely different origins. To maximize the BMR estimation accuracy, it is better to use tumor-specific features separately. Hence, we used forward selection in our analysis and in the revised manuscript, we made it very clear.
Excerpt From Revised Manuscript	

-- Ref 2.7 – Comments on the power analysis and compact annotations --

$$$Power $$$Calc (from JZ presentation) @@@JZ&&&TBC

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

Author Response	We thank the referee for accurately pointing out this problem. The current power analysis, which is also mentioned in previous literatures [[cite. Jz2add]] assumes that all the functional sites are within the test regions, is a fairly strong assumption and usually far away from the truth. In some cases, we do feel that the true functional sites might be allocated across various coding/noncoding elements. One example is the GATA3 case, there might be some mutational hotspots outside of the coding regions only. Some kind of joint test might increase detection power.  24 kb H3K27ac H3K27me3 H3K36me3 H3K4me1 H3K4me2 ReplicationTiming MutationCounts RefSeq Genes ATA3-AS1 GATA3 Actually this is the reason why we are proposing testing the extended gene regions. To illustrate this concept, we added a whole section of new power analysis in our supplementary file to discuss cases when and how power can be increased by joint testing.
Excerpt From Revised Manuscript	

-- Ref 2.8 – Q-Q plots --

$$$BMR $$$Calc $$$Thinking @@@JZ&&&TBC

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 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	

-- Ref 2.9 – Novelty of the paper --

$$$NoveltyPos

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

-- Ref 2.10 – BMR effect on local context --

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

Referee Comment	However, it is unclear whether the analysis takes into account complexities of the mutation model in regulatory regions. The influence of tri- or even penta-nucleotide context can be significant.
Author Response	In the main figure, we did not show how local context effect may affect BMR in order to highlight the effect of accumulating features. However, in the supplementary file where we described our method, we separate the 3mers to run negative binomial regression. We showed that in Supplementary figure xxx that local context effect is huge - usually up to several order of effect on BMR. We made this point more clear in our revised manuscript.

Excerpt From Revised Manuscript	

-- Ref 2.11 – Confounding factors --

$$$BMR&&&compl

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.
Excerpt From Revised Manuscript	

-- Ref 2.12 – Power analysis of extended genes --

$$$Power $$$Calc (JZ presentation) @@@JZ&&&TBC

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

Excerpt From Revised Manuscript	

-- Ref 2.13 – Minor Comment: Burden --

$$$Presentation $$$Minor $$$Text @JZ

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

-- Ref 2.14 – Minor Comment: Terminology --

$$$Presentation $$$Minor $$$Text&&&compl

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 to xxx in our revised manuscript.
Excerpt From Revised Manuscript	

Referee #3 (Remarks to the Author):

-- Ref 3.1 – Presentation --

$$$Presentation

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

-- Ref 3.2 – BMR --

$$$BMR

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	

-- Ref 3.3 – Presentation --

$$$Presentation

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	

-- Ref 3.4 – Untitled --

$$$Presentation

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	

-- Ref 3.5 – Untitled --

$$$BMR

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	

-- Ref 3.6 – Untitled --

$$$Annotation

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	
Excerpt From Revised Manuscript	

-- Ref 3.7 – Untitled --

$$$Network

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	

-- Ref 3.8 – Hieratchy --

$$$Hierarchy

Referee Comment	For Figure 4, what does the star symbol (*) mean in the legend? Did the authors use a different grey color to show the connection between TFs? I'm not able to read the grey gradient for the edges.
Author Response	
Excerpt From Revised Manuscript	

-- Ref 3.9 – Untitled --

$$$Network

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	

Referee #4 (Remarks to the Author):

-- Ref 4.1 – Strengths of the Paper --

$$$NoveltyPos&&&compl

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.
Excerpt From Revised Manuscript	

-- Ref 4.2 – Changing the presentation of the supplement --

$$$Presentation $$$Text @@@JZ @@@DC&&&TBC

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've tried to fix the presentation
Excerpt From Revised Manuscript	

-- Ref 4.3 – Trimming and editing parts of the manuscript (wait after updated version of manuscript)--

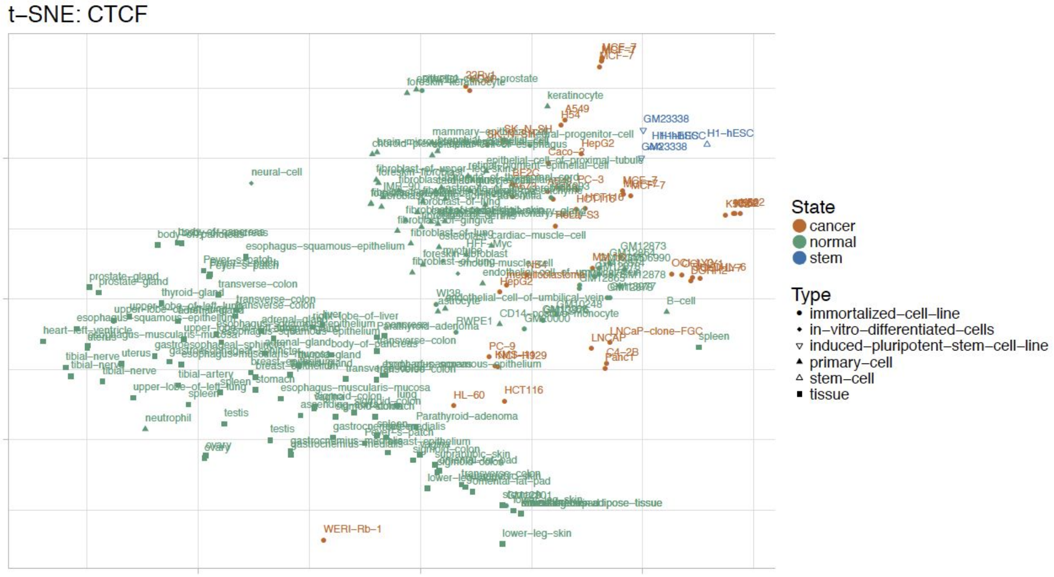
$$$Presentation $$$Text $$$Later

Referee Comment	1) The manuscript is quite complex and efforts are needed to improve clarity. Some of the text can seem to be somehow redundant or not needed (for instance, general comments about the ENCODE project; or the Step-Wise prioritization scheme (page7; other parts at page 7, for instance).
Author Response	As requested, we've trimmed & edited
Excerpt From Revised Manuscript	

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

$$$CellLine &&&More $$$Calc @@@PE @@@DL @@@JZ @@@Peng&&&TBC

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). 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	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 (DL maybe) We thank referee for bringing this point. As we stated in the manuscript, we agree with the referee that immortalized cell lines may not be the best representation of normal and cancerous counterparts of primary cells and tissues. One of the strengths of ENCODE release 3 is massive expansion of functional genomic data into various primary cells and tissue types. In this revision, we have extensively explored the chromatin landscape and expression patterns across all of available ENCODE primary cells and tissues, and compared with existing immortalized cell lines with deep annotations. We have chosen CTCF ChIP-seq, which has the most abundant number of cell types in ENCODE, as an example to highlight that ENCODE cell lines are not far different from primary cells.

	 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.
Excerpt From Revised Manuscript	

-- Ref 4.5 – Relationship of H1 to other stem cells --

$$$Stemness $$$Calc &&&More @@@PE @@@DL&&&TBC

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, <u>H1 may not necessarily be the best cells to compare with tumor phenotype</u>. Authors should discuss/defend of further elaborate on this approach. I believe that a key analysis should be done

	against other stem cells (like tissutal stem cells, etc.).
Author Response	> PE's imputed network stuff > histones DHS &&&&&& explicit imputed network Expand the resource - Tissue-specific networks, not in any other paper We admit that H1-hESC may not be the most ideal stem cells to compare with tumor phenotype. We have chosen H1-hESC because it offers the broadest ChIP-seq coverage and has the most amount of other assays in ENCODE. However, 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.
Excerpt From Revised Manuscript	

-- Ref 4.6 – Fixes for Figure 1 --

$$$Presentation $$$Later @@@DL&&&TBC

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

	Less data, more on overview schematic) 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	

-- Ref 4.7 – How do SVs affecting BMRs & Network --

$$$BMR $$$NETWORK $$$Calc &&&More @@@DL (rewire) @@@XK + @@@TG (expression & elements vs SV) @@@STL (mechanism) &&&TBC

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

-- Ref 4.8 – Aspects of heterogeneity related to cell liens --

$$$CellLine $$$Text @@@WM @@@JZ @@@MRS @@@PDM&&&compl

Referee Comment	6) Most cancers are not necessarily represented by a single cell type used to obtain genomics data in this study, but contains numerous types of cells with different mutations, as well as normal cells, infiltrating cells, all in a three dimensional structure, often producing metastatic colonizing other organs. However, this study focuses only on comparisons between cells. These limitations should be better discussed, also to put in perspective future studies on single cells.
Author Response	###JZ: strength of cell line, no heterogeneity, emphasize this, co-expression network ### Can mention something related to single cells ### Some clinically significant changes will occur in ###WUM text### The referee is correct that tissue heterogeneity represents a source of complexity not directly modeled in our resource, a limitation which we now discuss with greater emphasis. Nonetheless, some of our analyses *indirectly* model tissue heterogeneity, and some other of our analyses that do not model tissue heterogeneity should be particularly robust to this potential source of variance. One way in which we indirectly model tissue heterogeneity is by incorporating the patient's tumor's transcriptome when constructing patient-specific regulatory networks. Paracrine signalling by stromal tissue can trigger a signalling cascade that results in altered TF expression and therefore potentially global gene regulation in a patient sample. We empirically take such consequences into account by adding or removing regulatory network edges from patient-specific regulatory networks based on patient-specific TF expression levels, which implicitly takes into account the role of normal cell signalling on those TF levels in cancer cells. An example of an analysis that should be particularly robust to the presence and activities of stromal and infiltrating cells is our BMR calculations. 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 the coming years, we might be able to better model this complexity making use of new single-cell epigenetic data, which is just beginning to emerge. https://www.nature.com/articles/s41467-018-03149-4

Another possibility for future improvements that we mention in our updated discussion section is the potential to model regulatory networks and the BMR separately for each major subclone present in a patient cancer sample, whose differential mutations can be approximately inferred using existing computational tools.
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003665

###PDM text###
As the reviewer correctly states, genomic and epigenomic heterogeneity in tumor cells, as well as heterogeneity in the tumor microenvironment (e.g., immune cell infiltrates, hormonal factors, normal cell populations, etc.) are significant factors in tumor growth and development. Nonetheless, we feel there remains value in single-cell comparisons between tumor and normal cells.

Among the strengths of cell-line comparisons is the ability to perform well-controlled analyses of cancel cell function in a way that is not possible with whole tumor specimens. For example, the detailed gene co-expression network analyses we highlight in our manuscript (see section XXX), were made possible by a homogenous cancer-cell population with robust and uniform expression signal. Such an analysis in whole-tumor specimens would be challenging due to the need for deconvolution of expression signals originating from various cell types present in tumors.
Apart from the advantage of single-cell analyses of enabling examination of complex cancer cell biology, there is, moreover, reason to believe that single-cell analyses may capture important tumor biology present *in vivo*. Cancers that result from a single progenitor cell, or homogenous progenitor population, provide a justification for the use of single-cell analyses and comparisons. There is evidence that a number of cancers may develop according to the cancer stem-cell model, which posits that it is only a small population of stem-like cells that are responsible for tumor development and observed intratumoral heterogeneity (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.

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

Excerpt From Revised Manuscript	

-- Ref 4.9 – lncRNAs and BMR--

$$$BMR $$$Calc @@@JZ&&&TBC

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	

-- Ref 4.10 – (Minor) updates to figure numbering in suppl. --

$$$Presentation $$$Minor &&&TBC

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

-- Ref 4.11 – (Minor) Figure legends--

$$$Presentation $$$Minor&&&TBC

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	

Referee #5 (Remarks to the Author):

-- Ref 5.1 – Positive comment of the paper --

$$$NoveltyPos

Referee Comment	While the resources provided in this manuscript are potentially interesting for the cancer genomics community and comprise an extensive body of work
Author Response	We thank the referee for the positive comment.
Excerpt From Revised Manuscript	

-- Ref 5.2 – Untitled --

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

Referee Comment	it is not clear what are the main findings in the paper and their statistical and biological significance. The manuscript seems to be somewhat confused between a perspective piece or a guide to ENCODE data for the cancer community (which should be published in a more specialized journal), and a genomics study with clear findings.
Author Response	We thank the referee for pointing this out. We have made explicit that (1) this paper is to be considered as a "resource" paper, not a novel biology paper. Our goal is to integrate a number of assays (e.g. replication timing, STARR-seq and Hi-C), to provide deep, integrative annotations and various networks across many cell types.

Excerpt From Revised Manuscript	

-- Ref 5.3 – Novelty of the paper --

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

Referee Comment	As it is, the manuscript falls short of the novelty characteristic of publications in Nature. The main concepts presented in this manuscript have been explored extensively before; albeit not with the same amount of ENCODE data specifically (e.g. Martincorena et al (2017); Lawrence et al (2013); Polak et al (2015); Imielinski (2017); Roadmap Epigenomics). The cancer genome community has been using ENCODE and Roadmap data in various ways, including in papers such as Tomokova et al. (2017), Schuster-Böckler and Lehner (2012), Frigola et al. (2017), Sabarinathan et al. (2016), Morganella et al. (2016), Supek and Lehner (2015). There is no clear comparison to prior work and no demonstration of improved results compared to those in the literature.
Author Response	We thank the referee to point out many related references. We tried to cite some of the in our manuscript. But note that some important reference, such as Martincorena 2017, came out after our submission in Aug 2017. We agree with the reviewer that the concept of using genomics features can help to estimate BMR. However, our goal in this manuscript is to demonstrate that ENCODE data is quite useful for a variety of models, rather than to develop a novel cancer driver detection method. The BMR part takes only two sub-panels of Fig. 2, and we do have many other aspects in the manuscript to go beyond this point. For example, 1. We provided accurate noncoding annotation by integrating multiple novel assays such as Hi-C and STARR-seq, which may increase power in somatic mutation burden test. 2. We integrated more than 1000 ChIP-seq/eCLIP experiments to provide detailed TF/RBP networks. By combining cohort RNA-seq data, we identified both known (TP53 and ESR1) and novel (SUB1) cancer-associated regulators 3. Through whole genome sequencing data, we provided high-quality SV calls in top cancer cell lines, and investigate their effects on enhancers and networks. 4. For the first time, we have incorporated thousands of ChIP-seq experiments to directly observe the tumor-to-normal network perturbations and quantify it such changing events

Excerpt From Revised Manuscript	

-- Ref 5.4 – BMR --

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

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	Similar to comment to referee 2
Excerpt From Revised Manuscript	

-- Ref 5.5 – TCGA benchmark on the gene level --

$$$BMR $$$Calc $$$Later @@@WM @@@JZ (hard comparison)&&&TBC

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	>SK: what are the available in the TCGA BMR benchmark Compare with PCAWG results, not in the main manuscript Ask Li Ding for TCGA benchmark Hey we have a new method for BMR estimation. We want to compare it against Do you know of any good ones? Non driver TCGA gene (remove cancer genes) Calc bmr and compare with benchmark? * 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...
Excerpt From Revised Manuscript	

-- Ref 5.6 – Addressing improvements of the BMR --

$$$BMR $$$Calc @@@JZ&&&TBC

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
Excerpt From Revised Manuscript	

-- Ref 5.7 – Power analysis --

$$$Power $$$Calc @@@JZ&&&TBC

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	JZ's presentation
Excerpt From Revised Manuscript	

-- Ref 5.8 – Comparing power analysis to other work --

$$$Power $$$Text @@@JZ&&&TBC

Referee Comment	5. "Increased" power of the combined strategy is suggested in the manuscript, yet comparison to prior work is missing.
Author Response	We thank for the referee to point this out. In our revised manuscript, we have added a whole new section in the supplementary file to discuss this problem. In summary, previous power calculations was based on the assumption that all functional sites are within the test region, hence it is better to have short and accurate annotations. However, we found that this assumption is pretty strong and is not realistic for some cases. Instead, we added a whole section where some functional sites are allocated across multiple regions and then a combined strategy is better.
Excerpt From Revised Manuscript	

-- Ref 5.9 – Calculation of power --

$$$Annotation $$$Calc @@@JZ&&&TBC

Referee Comment	6. The authors claim that reduction of functional elements increases power to discover recurrently mutated elements. This point needs quantitative support in the main manuscript (some analysis is given in the supplemental). For example, in the enhancer list derived from the ensemble method, what fraction of enhancers are estimated to be false positives?
Author Response	(JZ's presentation)
Excerpt From Revised Manuscript	

-- Ref 5.10 – Assessing quality of enhancer gene linkage annotation --

$$$Annotation $$$Text @@@KevinYip @@@SKL&&&TBC

Referee Comment	7. The authors claim superior quality of gene-enhancer links and gene communities derived from their machine learning approach. The method should at least be outlined in the main text, and accompanied by data supporting its accuracy and better performance compared to existing approaches.
Author Response	We thank the referee for the comments. We have done as suggested: We have added a few sentences to the main text better desc. The methods and we have created suppl. Section XXX that shows the performance of JEME + Hi-C
Excerpt From Revised Manuscript	

-- Ref 5.11 – What data sets are used --

$$$BMR $$$Punt @@@JZ&&&TBC

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	JZ: punt
Excerpt From Revised Manuscript	

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

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

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	

-- Ref 5.13 – Additional QQ plots --

$$$BMR $$$Text&&&compl

Referee Comment	10. The significance analysis of cancer cohorts (Figure 2) should highlight known cancer genes versus those newly found in this study. A QQ-plot should be included to confirm that the algorithm accurately models the background expectation.
Author Response	Yes, we have provided the QQ plot in the supplementary file in our initial submission.
Excerpt From Revised Manuscript	

-- Ref 5.14 – Sequence coverage --

$$$BMR $$$Text&&&compl

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	

-- Ref 5.15 – Power analysis for Compact Annotation --

$$$Power $$$Calc @@@JZ&&&TBC

Referee Comment	How do the new “compact annotations” lead to improved results over traditional annotations?

Author Response	We demonstrate through power analysis in our supplementary file. When all the functional sites are within the test region, a shorter or "compact" annotation can significantly reduce noise level and increase statistical power.
Excerpt From Revised Manuscript	

-- Ref 5.16 – BCL6 Questions --

$$$Annotation $$$Calc @@@XK @@@TG&&&TBC

Referee Comment	11. The authors mention that BCL6 would have been missed in an exclusively coding analysis. In which part of the extended annotations were recurrent BCL6 mutations found? If near the promoter, is the BCL6 5′ region a known AID off-target? Are BCL6 mutations in CLL associated with translocations?
Author Response	BCL6 mutations were found in enhancer region. XK, TG Are any SVs associated with BCL6?
Excerpt From Revised Manuscript	

-- Ref 5.17 – ChIP-seq vs other computational based networks --

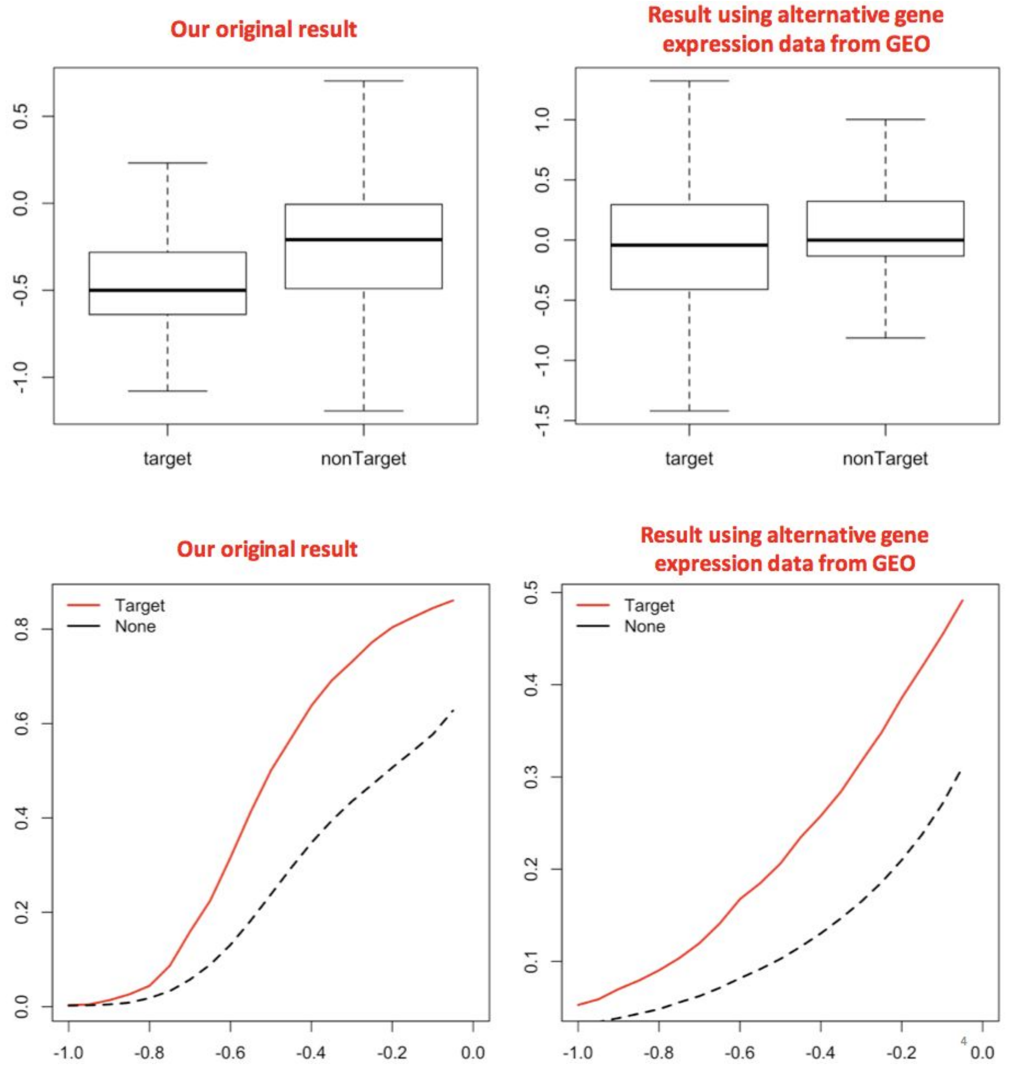
$$$Network $$$Calc @@@Peng @@@JZ&&&TBC

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	
Excerpt From Revised Manuscript	

-- Ref 5.18 – KD in MYC --

$$$Network $$$Text @@@DC&&&compl

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	dc & jz to do the comparison new arrays Search for other MYC KD data
Excerpt From Revised Manuscript	

-- Ref 5.19 – SUB1 analysis --

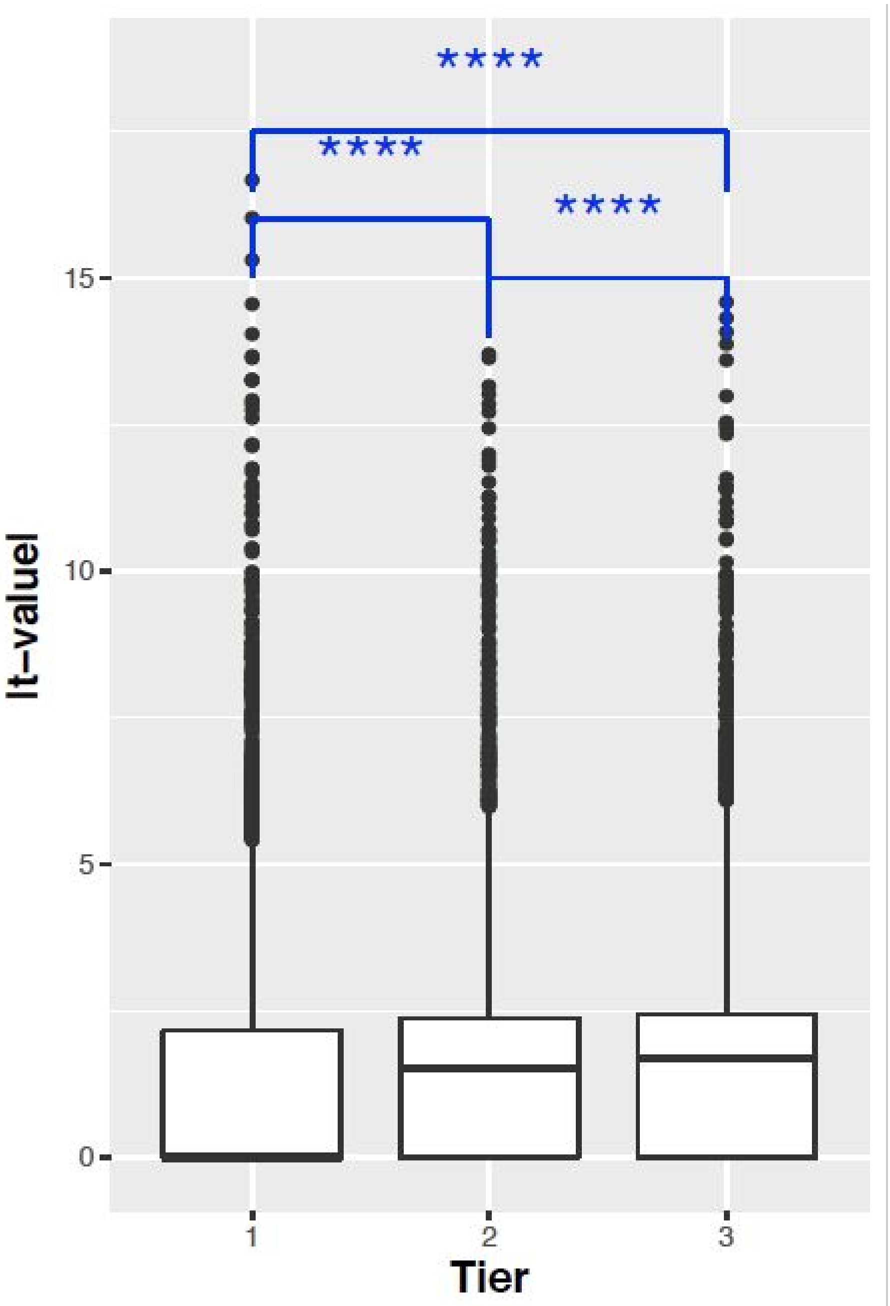
$$$NoveltyPos $$$Calc &&&More @@@MRS @@@JL @@@YY&&&TBC

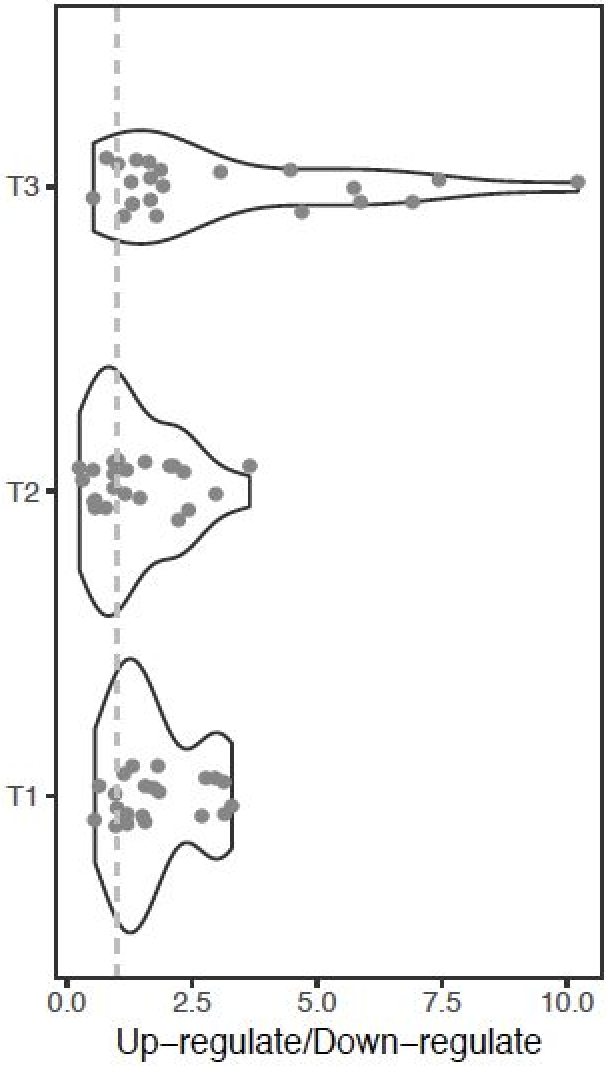
Referee Comment	14. SUB1 is a potentially interesting new cancer gene. The authors should further explore the biology of this gene.
Author Response	We thank the referees for the positive comments. We did follow up with SUB1 in this round of revision. &&&& we've done more with the network
Excerpt From Revised Manuscript	

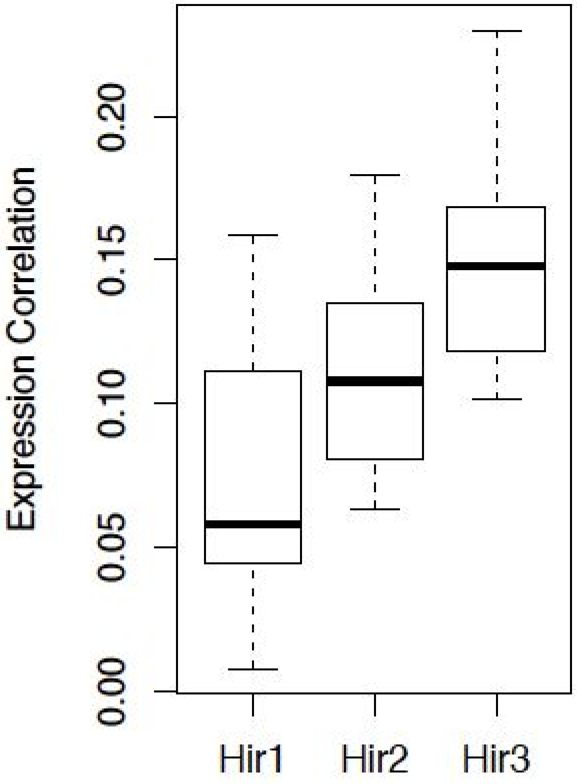
-- Ref 5.20 – Significance of network hierarchy --

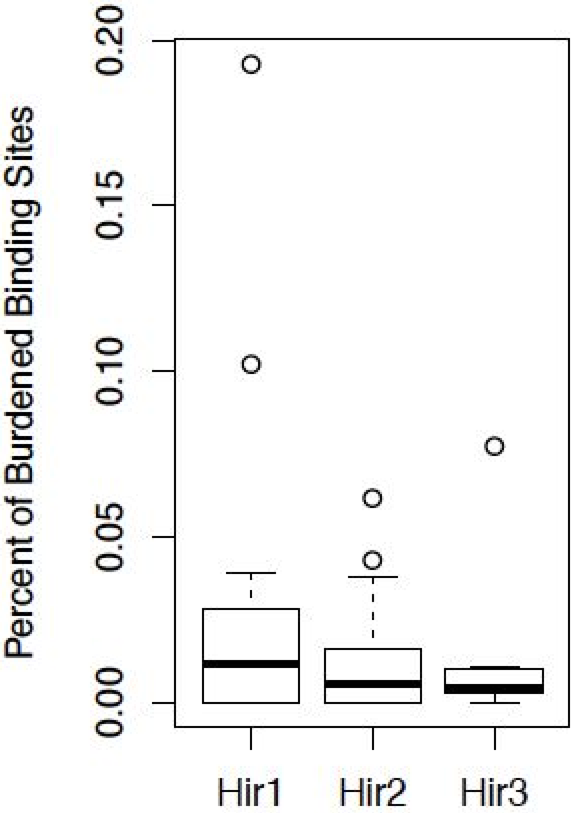
$$$Network $$$Text @@@DL&&&compl

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


****
****
****
15
10
5
0
lt−valuel
1
2
3
Tier

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

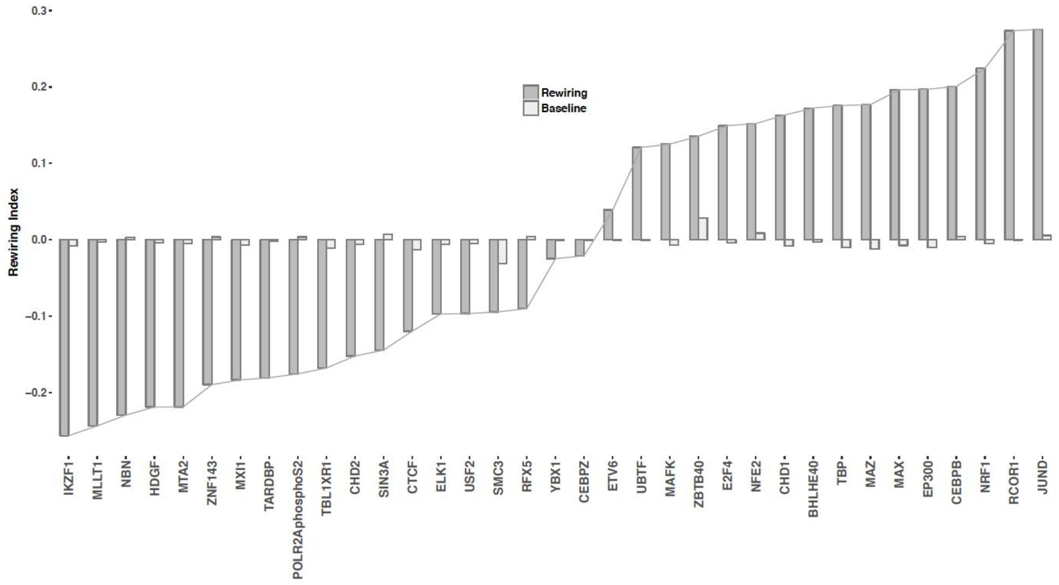

Expression Correlation
0.00
0.05
0.10
0.15
0.20
Hir1
Hir2
Hir3

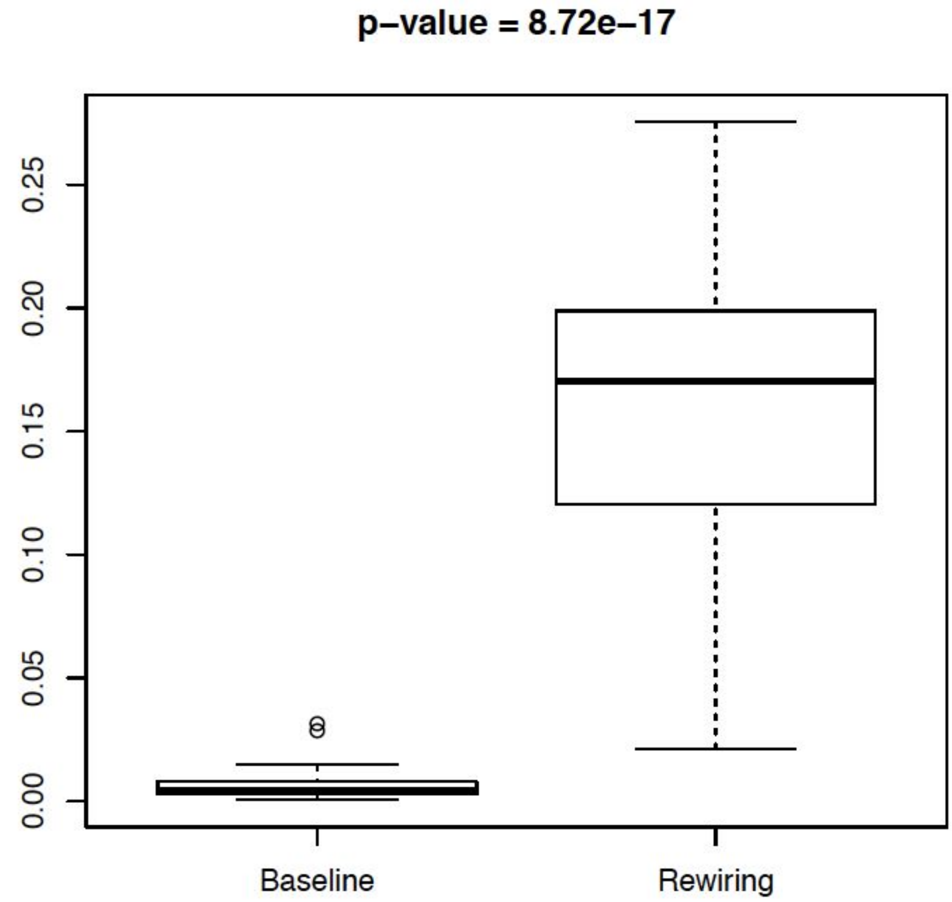
	
Excerpt From Revised Manuscript	

-- Ref 5.21 – Rewiring network --

$$$Network $$$Calc @@@DL&&&TBC

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

	 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	

-- Ref 5.22 – Rewiring analysis in the stem cells --

$$$Stemness $$$Calc @@@DL&&&TBC

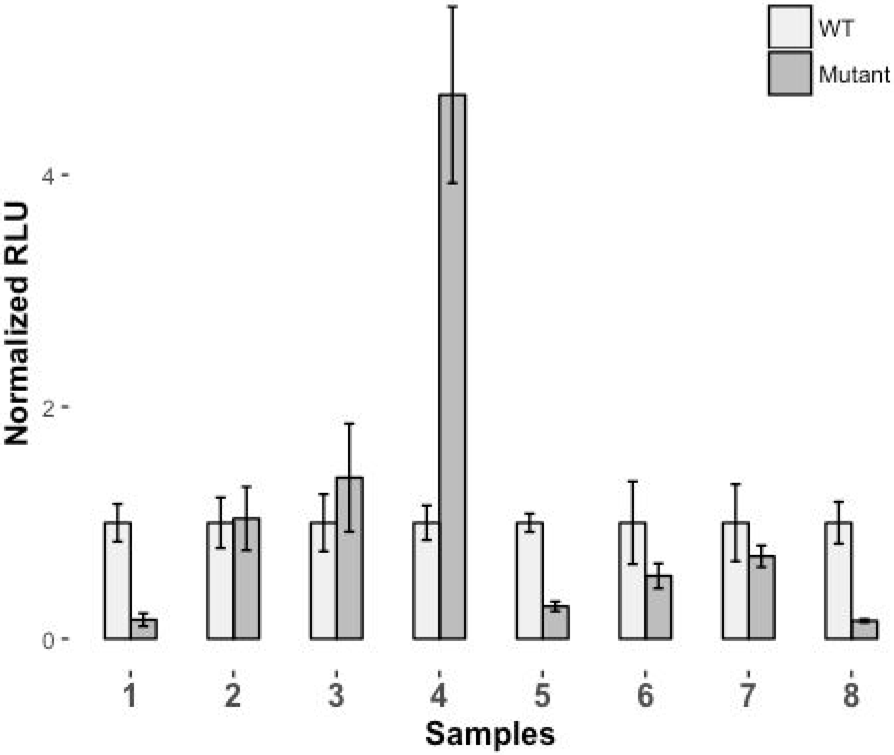
Referee Comment	17. The network change comparisons with the H1 stem cell models need statistical testing for significance. What fraction of the rewired edges are expected to be false positives?
Author Response	Statistical significance testing for H1 stem cell DL : to do - same as 16 False positive rate analysis Think about test of significance (have some more analysis) DL/JZ disc.
Excerpt From Revised Manuscript	

-- Ref 5.23 – Selection of regions for validation testing --

$$$Validation $$$Text @@@JZ @@@DL&&&compl

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	JZ, DL: we can answer 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 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

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


</td></tr>
<tr><td>Excerpt From Revised Manuscript</td><td></td></tr>
</table>

-- Ref 5.24 – (Minor) Presentation and revision to manuscript--

$$$Presentation $$$Text &&&TBC

Referee Comment	19. The authors should consider moving the general overview diagrams that constitute much of the main figures

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

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

$$$Validation $$$Text&&&TBC

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	How are we diff funseq BMR Rewiring Tissue specific
Excerpt From Revised Manuscript	

-- Ref 5.26 – (Minor) BMR --

$$$BMR&&&TBC

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

-- Ref 5.27 – (Minor) Untitled --

$$$Presentation&&&TBC

Referee Comment	22. Prior work using ENCODE chromatin data to define regulatory regions and gene enhancers links should be cited (referred to in the manuscript as "Traditional methods").
Author Response	
Excerpt From Revised Manuscript	

-- Ref 5.28 – (Minor) Untitled --

$$$CellLine&&&TBC

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

-- Ref 5.29 – Use of H1 for stemness calculation --

$$$Stemness&&&TBC

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

-- Ref 5.30 – Untitled --

$$$Validation&&&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	
Excerpt From Revised Manuscript	

-- Ref 5.31 – Untitled --

$$$NoveltyPos&&&TBC

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	

-- Ref 5.32 – Untitled --

$$$Presentation&&&TBC

Referee Comment	27. The abstract mentions the usefulness of ENCODE data for interpretation of non-coding recurrent variants, yet this point is not explored much in the manuscript.
Author Response	
Excerpt From Revised Manuscript	

-- Ref 5.33 – Untitled --

$$$Presentation&&&TBC

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

-- Ref 5.34 – Untitled --

$$$Presentation&&&TBC

Referee Comment	29. Figure 2d, q-values should be given for each identified driver gene.

Author Response	
Excerpt From Revised Manuscript	

-- Ref 5.35 – Presentation --

$$$Presentation&&&TBC

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

-- Ref 5.36 – Presentation --

$$$Presentation&&&TBC

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	

Excerpt From Revised Manuscript	

-- Ref 5.37 – Supplementary document --

$$$Presentation&&&TBC

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