# Tags:

Use comma for seperation between tags

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

PLEASE NOTE $$$ @@@ &&& %%% are reserved as shown above.

PLEASE USE ### only for all other tags.

Usage example:

<ID>REF 0.0 - Overall comments on the paper

<TYPE>$$$BMR

<ASSIGN>@@@MG,@@@JZ,@@@DL,@@@JL,@@@WM,@@@PDM,@@@Peng,@@@TG,@@@XK,@@@STL,@@@MTG

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

# Format:

Referee Comment: Courier New, 10pt

Author Response: Helvetica Neue, 12pt

Excerpt 1.1-A (Revised MS): Times New Roman, 10pt

# Referee expertise:

Referee #1: cancer genetics, mutational processes

Referee #2: statistical genetics

Referee #3: human genetics

Referee #4: gene expression

Referee #5: cancer genomics

# 

# Cover Letter

Dear Orli,

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

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

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

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

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

Yours sincerely,

marK

# Editor:

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

<TYPE>$$$Presentation

<ASSIGN>@@@MG

<PLAN>

<STATUS>%%%65DONE

|  |  |
| --- | --- |
| Referee  Comment | The referees have raised a range of technical concerns on the analyses, including for the background mutation rate, the need to include statistical significance to support many of the claims, and the limitations of this data including cell lines used. |
| Author  Response | We have tried to revise our manuscript to completely and definitively address all of the referee's comments. We felt many of them are good suggestions, so we expanded upon them extensively while keeping the focus of our manuscript. In particular, we have expanded the manuscript to address suggestions related to  - Highlight the overall value of this resource to cancer genomics  - Extend analysis of genes’ effects on somatic and germline SNVs or SVs  - Normal-tumor-stem comparisons from network and expression profiles  - Discuss SUB1 as an example to highlight the cancer network biology  - SVs’ effects on networks and extended genes  - CRISPR-based validations on SV effects  *Regarding the misunderstanding on the BMR section*  One misunderstanding we wish to clarify is that the main goal of the BMR section is to demonstrate how the richness of ENCODE data can improve BMR estimation, and not so much to discover novel drivers genes. Hence, we feel that detailed cancer driver comparisons are outside the scope of our manuscript.  Another point we want to emphasize is the necessity of including many features due to the heterogeneous nature of tumor data, which **was also accurately pointed out by referee 4**. Usually, there are numerous non-cancerous cells, such as immune, fibroblasts, and blood cells, within and around the tumor cells, which may play important roles in cancer \cite{xxx}. We have shown that ENCODE dramatically increases the available genomic data by more than a factor of 10 compared to the current methods (2,069 vs 169). We want to further point out that the majority of such data are actually from real tissues (1,339 out of 2,069). We have shown that the inclusion of more data noticeably improves BMR estimation. |

## 

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

<TYPE>$$$Presentation

<ASSIGN>@@@MG,@@@JZ

<PLAN>

<STATUS>

|  |  |
| --- | --- |
| Referee  Comment | The referees also find that the current manuscript provides limited context with prior studies using similar approaches for use of prior ENCODE and Epigenome Roadmap datasets in cancer genomics. They detail the need for clearer presentation in context of prior studies as well comparisons to demonstrate advance. |
| Author  Response | We thank the referees for this comment, and we have tried to provide better context with prior work in our revised manuscript. We note that we have cited many of these works in our initial submission. Some papers came out well before we submitted our paper in Aug 2017. Martincorena et al 2017, was published in Nov 2017 (this was work from the lab of Peter Campbell, and we excluded him due to a conflict of interest in our initial submission).  We want to further point that the main focus of this work from Dr. Peter Campbell’s lab was not at all on BMR estimation, but rather selection patterns in coding regions in cancer (abstract below). BMR estimation and noncoding regions are not even mentioned in the abstract or the main manuscript associated with that work.  As suggested, we now cite this paper in our revised manuscript, and we make it clear how our paper is different from this one. However, we feel that it may not be entirely reasonable to carry out detailed comparisons with that work. In fact, after our submission, several new studies were released that linked the noncoding genomes to cancer, such as Zhang et al 2018. We strongly believe that our ENCODEC resource would benefit such analyses, so we have updated our reference list in this revised version.  *“Universal Patterns of Selection in Cancer and Somatic Tissues: Cancer develops as a result of somatic mutation and clonal selection, but quantitative measures of selection in cancer evolution are lacking. We adapted methods from molecular evolution and applied them to 7,664 tumors across 29 cancer types. Unlike species evolution, positive selection outweighs negative selection during cancer development. On average, <1 coding base substitution/tumor is lost through negative selection, with purifying selection almost absent outside homozygous loss of essential genes. This allows exome-wide enumeration of all driver coding mutations, including outside known cancer genes. On average, tumors carry 4 coding substitutions under positive selection, ranging from <1/tumor in thyroid and testicular cancers to >10/tumor in endometrial and colorectal cancers. Half of driver substitutions occur in yet-to-be-discovered cancer genes. With increasing mutation burden, numbers of driver mutations increase, but not linearly. We systematically catalog cancer genes and show that genes vary extensively in what proportion of mutations are drivers versus passengers.* |

## 

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

<TYPE>$$$Presentation

<ASSIGN>@@@MG,@@@JZ

<PLAN>&&&DisagreeFix

<STATUS>

|  |  |
| --- | --- |
| Referee  Comment | The referees also recommended that the current manuscript does not represent a distinct advance to the main ENCODE manuscript, as it does not report separate new datasets, methods, or clear novel findings. Some referees also recommended that this may be more suitable as Perspective in a specialized journal that further highlights the use on the current ENCODE datasets for cancer genomic studies. |
| Author  Response | We thank the referees for pointing out potential sources of confusion about whether this is a novel biology paper or a resource paper, as well as for raising their questions regarding the relationship between our paper and the whole ENCODE package. In our revised version, we have tried to make these points more explicit.  ***Regarding the objectives of our paper and how to relate it to the whole package:***   * this paper should be considered as a "*resource*" paper, not a novel biology paper * this work is the main integrative paper that provides deep annotation for several cell types, while the main encyclopedia paper is focused on broad and universal annotations (for all cell types) based on 4 assays. * this is the only paper in ENCODE that provides comprehensive networks from ENCODE3 and this is the only paper that incorporate novel data types from the ENCODE functional characterization center   ***Regarding data in this paper***   * our paper is the only one that incorporates multiple novel assays in ENCODE3, such as STARR-seq, Hi-C, TF knockouts * it is the only one with unique validations that have been carried out with various techniques, such as luciferase assays, CRISPR engineering, and knockout experiments * ENCODE 3 "data" are not explicitly tied to any paper. Unlike previous rollouts, ENCODE 3 does not associate particular data sets with specific papers (as codified in an agreement with NHGRI.)   ***Regarding the new methods in this paper***  As summarized below, we have many under-appreciated methods for integrating multiple assays for deep annotations. We have tried to make these more clear in our revised version:   * Multiple methods regarding enhancer predictions   + CASPER: Pattern recognition-based enhancer prediction that integrate more than 10 histone modification marks   + ESCAPE: Enhancer predictors based on STARR-seq methods   + CARE: Compact and AccuRate Enhancer prediction by integrating STARR-seq and genomic features * A method for enhancer-gene linkage predictions: JEME+Hi-C * A gene community-based method to analyze network rewiring * A integrative new method to prioritize regulators based on burdening, rewiring and expression regulations * A new pipeline for variant prioritization |

# 

# Referee #1 (Remarks to the Author):

## <ID>REF1.0 – Preamble

<TYPE>$$$Text

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%65DONE

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

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

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

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

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

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

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

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

**2. Regarding the BMR section**

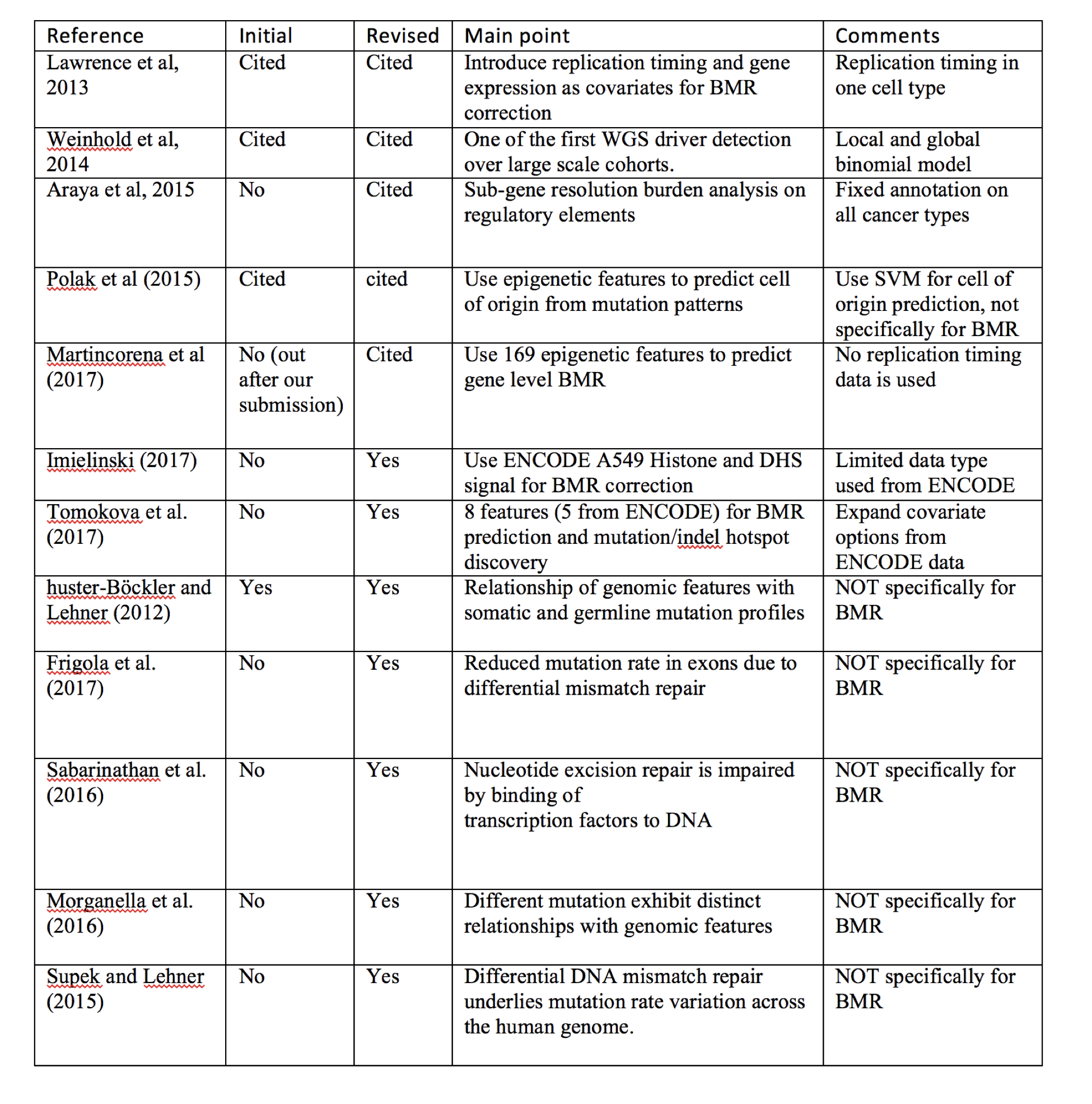
With respect to the BMR estimation part in particular, the reviewer noted that there have been several prior publications focusing on applications such as cancer driver detection. We thank the referee for pointing out this body of related work.

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

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

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

Table R1. status of the related references



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

<TYPE>$$$NoveltyPos

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

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

## 

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

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

<ASSIGN>@@@JZ,@@@WM,@@@PDM

<PLAN>&&&OOS

<STATUS>%%%95DONE

|  |  |
| --- | --- |
| Referee  Comment | Just to take the first application as an example, the problem of estimating background somatic mutation rate accurately in order to better identify cancer drivers has been studied extensively in the literature. One paper, “Mutational heterogeneity in cancer and the search for new cancer-associated genes” (Nature 2013), is cited in the current manuscript, but there are many others. For instance, Weinhold et al, 2014 (Genome-wide analysis of noncoding regulatory mutations in cancer, Nat Genetics), Araya et al, 2015 (Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations, Nat Genetics), and similar non-coding mutation identification papers all include steps to account for epigenetic features in their background rate calculation. |
| Author  Response | We thank the referee for pointing out these works. Modelling background mutation rate has been an important topic of inquiry, as even modest improvements can be of great benefit. As suggested, we have cited all the references mentioned above, and we have tried to provide better context of previous work in the revised manuscript.  In our revised manuscript, we have explicitly clarified how the new ENCODE data can be useful for BMR estimation. Our contribution is to provide data in a ready-to-use format that is considerably more expansive than those in previous works (2,069 features vs. 169 in Matincorina et al 2017). We have shown that this scale of data can benefit previous models to better characterize BMR. |
| Excerpt 1.2-A (from MS) | (Wait for main text) |

## <ID>REF1.3 – BMR: Matching

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

<ASSIGN>@@@JZ,@@@WM

<PLAN>&&&DisagreeFix

<STATUS>%%%50DONE

|  |  |
| --- | --- |
| Referee  Comment | Most large-scale cancer genome sequencing papers also have models at various levels sophistication, most of them including the issue of proper tissue-type matching. “matched” cell lines are better than unmatched or addition of more epigenetic features results in some improvement is almost trivial at this point. Which marks contribute to this is also not new. |
| Author  Response | We thank the referee for this comment, and we have tried to better clarify our main goal in our revised manuscript. We made it very clear that we are not claiming to have developed the use of negative binomial regression with epigenetic features on BMR estimation. Instead, our key points are that:   * The ENCODE3 rollout dramatically expands the number of high quality genomic data available for this type of regression by more than an order of magnitude (**2,069** compared to 169 in Matincorina et al 2017), many of which are from real tissue samples or primary cells.   We have also moved this to fig `1 to emphasize its connection w the data  There is significant technical challenge in processing this scale of data to create a ready-to-use resource that may be applied to BMR estimation.   * This expanded data provides a significantly larger pool to find the best match for a given cancer type * More data are useful due to tumor heterogeneity.   While it is valuable to match cancer to its cell of origin, tumors are highly heterogeneous (as clearly pointed out by referee 4 also), so a combination of different data sets provide the best overall fit to mutation rates. We have shown this in the updated version of Figure 2 (see Excerpt 1.3-A and 1.3-B).  [[WUM+PDM2all: WUM suggested that a clear demonstration of the value of increasing numbers of features, would be to compare the accuracy of using only the 169 features used by Inigo et al. to the accuracy we achieve with all 2,069 features. PDM agrees that this would be useful and clear.]] |
| Excerpt 1.3-A (from MS and fig.) | The 2,017 uniformly processed histone modification and 51 replication timing data may serve as a resource to significantly improve BMR estimation accuracy.  We also showed that BMR estimation can be improved dramatically by selecting appropriate combination of multiple features from ENCODE. |
| Excerpt 1.3-B (from suppl. - cross validation) | 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>REF1.4 – BMR: cell of origin features vs. many features

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

<ASSIGN>@@@JZ,@@@JL

<PLAN>&&&DisagreeFix,&&&More

<STATUS>%%%70DONE

|  |  |
| --- | --- |
| Referee  Comment | Importantly, Polak et al, 2015 (Cell-of-origin chromatin organization shapes the mutational landscape of cancer, Nature) in fact show that cell-of-origin chromatin features are much stronger determinants of cancer mutations profiles than chromatin feature of matched cancer cell lines, and that cell type origin can be predicted from the mutational profile.  Building on that model, but more data |
| Author  Response | We thank the referee for raising this point about features from cells-of-origin, and we have expanded upon the relevant discussion in our revised manuscript. In summary, we have made the following changes.   1. We have added more to the discussion section that accurate cell-of-origin definitions are challenging. Distinct subtypes of tumor cells may derive from different 'cells of origin' \cite{21248838}. (see Excerpt 1.4-A) 2. **In contrast to the results of Polak et al., we suggest that linear combinations of cancer cell lines and tissues may provide a basis for a more accurate determination of cancer mutation profiles than either cell-of-origin, or a single matched cancer sample - perhaps because of the intrinsic heterogeneity of tumor samples. [[Consistent with the stemness discussion etc., would need to flesh out argument or provide suggestive evidence.]]** |
| Excerpt 1.4-A  (from MS - disc. sect.) | Recently work has pointed out the effect from cell-of-origin on tumor from multiple aspects, such as mutational process and tumor classifications. However, to accurately define tumor cell-of-origin is sometimes challenging. For example, even different subtypes of tumor from the same organ may originate from different cell types. The richness of ENCODE data provides us a larger pool to find the best representative cell of origin. |

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

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

<ASSIGN>@@@JZ,@@@JL

<PLAN>&&&Rewrite

<STATUS>%%%50DONE

1) We do better because of the cell line data - we have more data. We've made a suppl. Fig. to show that the cell-line features are as strong as the tissue ones and that that aggreg. Of all does better.

2) Also we do have lots tissue data, which summarize in a new suppl. Table

3) another unique feature of encode is having the replication timing data. Replication timing is the single strongest feature in most contexts and the one most mechanistically related to mutation rate.

|  |  |
| --- | --- |
| Referee  Comment | Stepping back, it is not obvious to me that using the ENCODE cell lines, despite the availability of more epigenetic data, is the best approach to calculating the background rate in the first place—they briefly mention that using cell lines (rather than tissues) can be problematic, but do not explore this further. If this were a regular research paper, the authors would have to shown how the proposed approach is different and how it is better than methods already available. |
| Author  Response | We thank the referee for raising this question about cell line data usage in our paper, and we feel this is a good opportunity to clarify that ENCODE is not just about cell lines. In our revised manuscript, we have extensively discussed the use of different types of data from multiple aspects in both the main manuscript and the supplements:  *Regarding the cell line data in the BMR part*   * Certain data types, like TF ChIP-seq, are only predominantly available in cell lines (Excerpt 1.5-C). **Although whole tissue data could theoretically provide a closer match, this data is not obtainable due to current technical limitations. Cell line data reflects the current best possible data for these data types.**We added a table to clarify that the features extracted from ENCODE data are not just from cell lines. The majority are from tissues or primary cells (Excerpt 1.5-A).   *Regarding the robustness of using cell line inference on real patient data*   * added a whole new external validation section to compare with our conclusions drawn from cell lines (Excerpt 1.5-E). [[**Cells + tissues only from cells side by side comparisons]]**   **Subset data** |
| Excerpt 1.5-A (from suppl.) | In total, we have used 2,017 histone ChIP-seq and 51 replication timing Repli-chip and Repli-seq features to predict BMR. We did a PCA of the signals these features and selected the best combination of 20 PCs for BMR prediction. It is worth pointing out that the majority of our data is from real tissue or primary cells. A summary of cell types of these features were given below.  **[[WUM’s comment: Could we show a back-of-the-envelope power analysis that shows the improved capability of identifying a rare driver variant based on marginal improvements in BMR.]]**  Table S1. Summary of ENCODE histone ChIP-seq data **[[WUM suggests and PDM agrees that this data may be more clearly presented as a pie chart]]**    Table S2. Summary of ENCODE3 Replication timing data   |  |  |  | | --- | --- | --- | | **Cell Type** | **Repli-seq** | **Repli-chip** | | cell line | 101 | 10 | | in vitro differentiated cells | 0 | 35 | | primary cell | 12 | 5 | | stem cell | 6 | 11 | | induced pluripotent stem cell line | 0 | 2 |   Table S3. Summary of 51 replication timing features from Repli-chip and Repli-chip   |  |  | | --- | --- | | **Cell State** | **Repli-chip/Repli-seq** | | Pluripotent | 8 | | DE | 3 | | Liver/Pancreas | 6 | | Neural crest/Early mesoderm | 7 | | Late mesoderm | 6 | | NPC | 2 | | Myeloid/Erythroid | 5 | | Lymphoid | 5 | | Cancer | 9 | |
| Excerpt 1.5-B (from suppl. - mutation rate vs. cell line & tissue) | We calculated the pearson correlation of the breast cancer mutations count per Mbp vs. various histone modification features in tissue and cell line. Cell line data provides comparable (and sometimes better) correlation with mutation counts. |
| Excerpt 1.5-C (from MS - disc. sect.) | Some features, like TF binding events, have been shown to affect somatic mutation rates but the majority of such data are mainly available in cell lines. Hence, we systematically investigated the RNA-seq and TF ChIP-seq data and found that many of the cancer transcriptome/TF binding landscape are quite similar to each other, as compared to the initial of primary cells. This has also been mentioned by previous reports, such as Lotem et al. 2005 and Hoadley et al. 2014. The fact that cancer cells lose diversity and showed a distinct pattern from the primary cells highlights the values of cell line data. |
| Excerpt 1.5-D (from fig. 5) | We performed RCA/PCA analysis on ENCODE RNA-Seq, shRNA/siRNA knockdown, and TF ChIP-seq data and found that cancer cells demonstrate a consistent pattern to be more similar to stem cells, as compared to their primary cells of origin. **Relevance? Added new figures and analysis to “Appendix”** |
| Excerpt 1.5-E (validation of cell line data) | We predicted the regulatory activities of transcription factor (TF) MYC using a ChIP-seq profile in MCF-7 cells. We found that the MYC regulatory activity is highly correlated with the MYC expression across TCGA breast tumors. For most TFs, their regulatory activities predicted using ENCODE ChIP-seq profile in cell lines are significantly correlated with their expression levels across breast tumors. Moreover, using the same MCF-7 ChIP-seq profile, the MYC regulatory activity predicted for lung tumors is also significantly correlated with MYC expression level in TCGA lung cancer. These results indicate that the ChIP-seq profiles from a particular cell line can capture regulatory targets in human tumors from diverse cancer types. To select ChIP-seq or eCLIP profiles that are representative of the regulatory targets in human cancers, we only reported the results of TFs or RBPs whose regulatory activities are significantly correlated with their gene expression level in each TCGA cohort.    **Supplementary Figure X. The clinical relevance of ENCODE cell line data in human primary tumors**.  **(a)** The correlation between *MYC* expression level and regulatory activity across tumors. The MYC regulatory activity in each tumor was predicted using the ChIP-seq profile in MCF-7 cell line. The Pearson correlation between MYC gene expression level and regulatory activity were computed across tumors in each cancer type. The statistical significance of Pearson correlation was tested by the two-sided student t-test. BRCA: breast invasive carcinoma. LUSC: lung squamous carcinoma.  **(b)** The distribution of correlation *p*-values in TCGA breast cancer. For each TF, we tested the statistical significance of Pearson correlation between TF expression levels and regulatory activities predicted across tumors through two-sides student t tests as panel a. For TCGA breast cancer cohort, most *p*-values are very significant with a few non-significant values.  The fraction of regulators with statistically significant correlations in different cancer types for ChIP-seq and eCLIP networks. In each TCGA cancer type, we computed the correlations between regulator expression levels and regulatory activities across tumors for all regulators (TFs, or RBPs). We selected regulators with statistically significant correlations through two-sided student t test (FDR < 0.05). |

## 

## 

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

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

<ASSIGN>@@@JZ

<PLAN>&&&DisagreeFix

<STATUS>%%%90DONE

Go out of driver gene discovery, but emphasize regulators

|  |  |
| --- | --- |
| Referee  Comment | That ENCODE data helps in prioritization of non-coding variants has been well demonstrated already (including by some of the authors on this paper), and so the value of the described analysis less clear. |
| Author  Response | **The prioritization of non-coding variants is a major frontier in genomics and cancer genomics, and these prior publications suggest the importance of this topic** We have tried to clarify that the uniqueness of our method lies in that fact that   * It not only prioritizes non-coding variants, but also elements and regulators, which is not included in the other papers. We have highlighted this in revised Fig. 3 (Excerpt 1.6-A). * [[we have performed careful small scale validation on all levels - regulator, element & SNV via KD, crispr & luciferase to validate this * and performed targeted validations on key regulators (Excerpt 1.6-B). [[WUM+PDM2all: Is this related to the **prioritization of regulators in MCF-7 (REF 1.5)?]]** * Uses an approach that uses the full scale of the current ENCODE data which is beyond that in the previous publications - eg it uses RBP networks, the starr-seq, KD, the many tissue-specific dataset, network dynamics - none of these features have been exploited previously. * For variant prioritization, we added discussions to emphasize the integration of various novel assays in a tissue-specific manner, which was not possible in previous works (Excerpt 1.6-C). The fact that we coupled this with successful validation demonstrates the considerably greater value of the integrated ENCODE data. [PDM+WUM2all: This analysis could use more specifics on what was done, and for what reason -- Excerpt 1.6C itself is about the same length as this summary point, and provides more specifics.]] |
| Excerpt 1.6-A (from fig. 3) | ((  New legend of figure 3.  Figure to put here  Ask Feng’s group to write up here!  [JZ2MG: wait]  )) |
| Excerpt 1.6-B (from suppl. - regulator validation) | **[[PDM+WUM2all: The following text could use more explanation as to why this analysis is relevant. It currently reads like an excerpt from a methods section, and the figure has no accompanying caption.]]**  To detect predicted common target gene of MYC and SUB1, shRNA plasmids containing 4 targets sites of each gene were used to transfected to HepG2 cell using Lipofectamine(™) 3000 following the manufacturer's instructions (Invitrogen) (target sites for each gene are listed in Sup table 1). Briefly, 0.12 M HepG2 cells were seeded in each well of one 24-well plates 24 hours before transfection. 500 ng plasmids containing either single shRNA or 4 shRNA plasmids as pool were mixed with 0.75 uL Lipofectamine(™) 3000 in Opti-MEM I medium (Invitrogen) and loaded to HepG2 cells in each well. Blank plasmids without shRNA target sequence was used as control. To improve transfection efficiency, 2 ug/mL puromycin was used to select successful transfected cells. 72 hours after transfection, total RNA was extracted using RNeasy Mini Kit (Qiagen) and followed by cDNA generation using SuperScript III (Invitrogen). Knockdown efficiency and target gene expression level were quantified and compared to BACTIN by qPCR using KAPA SYBR® FAST qPCR Master Mix (2X) Kit (Sigma). The qPCR primers were listed in Sup table 2. |
| Excerpt 1.6-C (from MS - disc. sect.) | In particular, our prioritization framework takes into account the STARR-seq data, the connections from Hi-C, the better background mutation rates, and the network rewiring data, which is only possible in the context of the highly integrated and their data available on certain cell lines. |

# Referee #2 (Remarks to the Author):

## <ID>REF2.0 – Preamble

<TYPE>$$$Text

<ASSIGN>@@@MG,@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

We greatly appreciate the referee's feedback, especially the positive comments regarding the overall value of our resource, the extended gene, and the network rewirings. As suggested, we have tried to address the reviewer’s comments, and we further extend and reorganize our analyses to illustrate the value of the resources in this paper.

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

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

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

[WM suggested replacement text: Referee 2 provided helpful feedback on our statistical approach and description thereof. This feedback was an opportunity for us to critically reflect on our statistical choices and improve the clarity of our writing. In some cases, we maintained our original approach but with new justification of appropriateness, and in other cases we have modified our approach based on analyses we performed in response to Referee 2’s comments. In the revised manuscript, we have taken care to implement or address each point raised by the reviewer. We feel that this has strengthened our paper, and we thank the reviewer for his or her attention and expertise.

In reading Referee 2’s comments, we realized that we also need to clarify that we are part of the ENCODE data release. This means that our manuscript would be the first publication of many of the ENCODE data sets analyzed herein, listed in table R2.X.]

## <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,$$$Calc

<ASSIGN>@@@JZ

<PLAN>&&&OOS

<STATUS>%%%85DONE

|  |  |
| --- | --- |
| Referee  Comment | 1) The negative binomial regression (Gamma-Poisson mixture model) was introduced in Nik-Zainal et al. Nature 2016 and Marticorena et al., Cell 2017. Why was not this available method applied, and what is the benefit for the procedure used by the authors? |
| Author  Response | We thank the referee for pointing out the previous efforts on cancer driver detection by negative binomial regression. We certainly agree with the reviewer that negative binomial regression is a standard technique to handle overdispersion in count data.[[We agree with the we have in fact applied a very similar method in NZ et al. in our paper (see point below REF2.4 for the small difference). We have clear in the revised version.]] A number of earlier works (such as Imielinski et al 2016) also used negative binomial regression. In our revised manuscript, we have cited those works and tried to provide a better context of related work.  We also try to make it more clear that we are not claiming to provide a novel negative binomial regression-based driver detection method, but rather to use this as a showcase for the value of ENCODE data.  We did, in fact, use very similar methods to Marticorena et al. these are well established stat methods and there's lots of R packages for this. |
| Excerpt 2.2-A (from MS) | ... |

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

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

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix,&&&OOS

<STATUS>%%%90DONE

|  |  |
| --- | --- |
| Referee  Comment | Also, does Gamma-Poisson model fits data for most cancers well or is it just an approximation? One can use non-conjugate priors but this is probably beyond the scope of this work. |
| Author  Response | We thank the referee for mentioning the goodness-of-fit of the Gamma-Poisson model. As suggested, we now provide more figures in our supplement to investigate this.  For most cancer types, fitting a Gamma-Poisson is pretty good (as seen in the figures below). However, we agree that it is interesting to investigate other non-conjugate priors. As the referee mentioned, this is out of scope, but we have noted this in the text.  Add more summarize stat saying we used NB in XXX cases |
| Excerpt 2.3-A (added in Supp.) |  |

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

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

<ASSIGN>@@@JZ,@@@JL

<PLAN>fix it up but we have time we'll 2 new calc. And maybe even split into 2

<STATUS>%%%70DONE

|  |  |
| --- | --- |
| Referee  Comment | 2) It seems that the Poisson model was not rejected for cancers with very low mutation counts (liquid tumors). Is this a power issue rather than the property of the mutation process? |
| Author  Response | We thank the reviewer for mentioning this, and we feel this is a good point. We carried out further analyses in our revised manuscript.at   * We added a new plot to show the average mutation rate vs. the overdispersion parameter (Excerpt 2.4-A). We show with these plots that often it is not nec. To use nb model. The rejection of NB may because of power issue * We added a new supplementary figure of the Q-Q plot using Poisson and NBR, and we found that they provide similar results. (Excerpt 2.4-B). * Other papers shwo that it possible to use poisson regression with good covariates and kmer sep. (eg https://www.biorxiv.org/content/early/2017/12/19/236802).   In summary, we feel is simpler to avoid introducing additional parameters when not necessary for a good fit.  observe that a higher mutation rate is often associated with overdispersion, which is suggestive of a power issue, but we cannot rule out a biological component as well. |
| Excerpt 2.4-A (from suppl.) | 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. |
| Excerpt 2.4-B (from suppl. - Q-Q plot of Poisson vs. NBR) | We have used both Poisson and Negative Binomial distribution on the mutation burden calculation with exactly the same covariate set. Q-Q plot of p values on breast cancer CDS region were given below. Q-Q plots from these distributions look similar. Similar to the conclusion by wadi et al, it is necessary to first check whether covariate effects have been corrected and local kmer context has been calibrated and then test the level of overdispersion. |

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

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

<ASSIGN>@@@JZ

<PLAN>MG to do a text & JZ to fix the text, WM & PDM

<STATUS>%%%75DONE,%%%CalcDONE

Add the cross validation in this response section

Step2: add the PCA validation

We did see the increasing value of features

|  |  |
| --- | --- |
| 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 | You need many features, we are not building a lower degree model  out the limited contribution from the higher-order principal components. In the revised version, we have tried to better illustrate our main point: the wealth of the ENCODE data for BMR estimation. In summary, we have   * revised figure 2 by directly using a combination of features via forward selection (Excerpt 2.5-A), and we have moved the PCA part into the supplement. * added a supplementary figure of cross validations (Excerpt 2.5-B)   Explain more why we are doing PCA  Can validate in the end and will be OK at each stage |
| Excerpt 2.5-A (from MS and extended fig. 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. |
| Excerpt 2.5-B (added in Supp.) | To avoid overfitting problem, we performed 5 fold cross validation using the selected model for each cancer type and listed the performance as below. |

## 

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

<TYPE>$$$Power,$$$Calc

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

- JL - suppl. Table arguing that TF sites should be shrunk like RBP ones & ask RARE daf, and a comparison of enhancers (MTG) and rare daf

- JW pot. Do a suppl. Calc. incorporating uncertainty (put in after

<STATUS>%%%80DONE

[JZ2JZ: more equations to come]

JZ: RBP resolutions, ChIP-exo

|  |  |
| --- | --- |
| 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 | MG Dictation:  I so we agree with the referee that the manger that we agree with  referee that the compact sanitation formally makes sense however it  makes sense only with the assumption that were reaching for functional  sites we believe there are compact annotation is in fact enriching for  functional sites however we cannot be certain and we tried to be asked  about the syntax we do provide quite a number of additional  supplementary exhibits to argue that the compact annotation does in  fact rich for function sites these are outlined below  We thank the referee for this feedback. we certainly agree with the referee. As suggested, we have expanded our power calculations under various assumptions. In summary, we have now included:  If the assumptions is   * an entirely new section on power analysis and the effect of test region functional site ratios (Excerpt 2.6-A) * more discussion (in the main text) about the pros and cons of merging test regions (Excerpt 2.6-B) * We have added a number of case studies in the supplement showing the value of the approach (Excerpt 2.6-C) * a new supp. section of showing the quality of our annotation and how we believe that it is more highly enriched in true functional sitess(Excerpt 2.6-D) * A new table that shows the justification for compactifying TF bidning site annotations |
| Excerpt 2.6-A (from suppl.) | Suppose that we define the following parameters.    [[[ ... 2 pages ...]]]Then under the null hypothesis, the probability to observe at least one mutation per patient is    Under the alternative hypothesis,    We did a simulation by starting from a very noisy test region with pretty low true risk loci percentage. We have showed that by trimming the nosie loci, statistical power can be increased. But after we have removed the noise and start to trim the true functional loci, the statistical power drops quickly.  [WM2all: Our response here is not good enough. Jing’s upward sloping then downward sloping plot only AFFIRMS refs objections, does not say how we are ROBUST to their objection or intelligently handle it.  \* The way to show robustness is to show the effects of uncertainty. How does compact gene fare power-wise on expectation if we lob off nucleotides with a 50% depletion of functional impact per nucleotide compared to core, with a 10% depletion of functional impact per nucleotide compare to core, with a 90% depletion …  \* Let’s use ref2’s weighting scheme! Define annotation-credence scores as probabilities of functional affect for nucleteotide in and near an element. Count a 50%-uncertain-functional-effects nucleotide for 0.5 points in the numerator if mutation but only 0.5 points in the denominator. We should do this analysis. Why not hit all of ref2’s points?] |
| Excerpt 2.6-B (from MS) | In summary, our claim is that first we provide compact annotations to pick up functional nucleotides and remove noisy ones through the guidance of many functional characterization assays. Then we hope to join the distributed functional sites together to increase statistical power. |
| Excerpt 2.6-C (from suppl.) | We provided two examples to explain the motivation of our compact and extended gene annotations and why we feel our assumptions for the power analysis is reasonable.  1) Enhancers: Traditionally, enhancers were called as a 1kb peak regions, which admittedly introduced a lot of obviously nonfunctional sites. We believe we can get functional region more accurately by trimming the enhancers down using the exact shapes of many histone marks and further integration with STARR-seq and Hi-C data.  2) TFBS hotspots around the promoter region of WDR74. Instead of testing the conventional up to 2.5K promoter region, we can trim the test set to a core set of the promoter region where many TFs bind, which perfectly correlates with the mutation hotspots (red block) for this well-known driver site (blue line for pan-cancer and green line for liver cancer). |
| Excerpt 2.6-D (from suppl.) | With the ensemble method, for example, we can get more accurate annotation and pin-point to sequences where transcription factors would actually bind to. To estimate the false positive rate would not be very practical at this stage as there is no gold-standard experiment that could assert an predicted enhancer is definitely negative. Here we took the FANTOM enhancer data set and assess the overlap percentage of our enhancer annotation in each ensemble step. We showed that each ensemble step indeed increases the percentage of overlap between our annotation and the FANTOM enhancer set. The overlap percentage for our annotation is much higher than that of the Roadmap annotation, and is also higher than the main encyclopedia enhancer annotation annotation (ccRE).     * *Regarding the quality of enhancer-gene linkages:*   To show how our JEME+Hi-C approach captures enhancer-gene linkages compared to existing linkages, we used published chromHMM derived enhancer-gene linkages (cite chromhmm) as the comparison dataset and GTEx whole blood eQTLs as the benchmark. We found the linkages, which the enhancer has an eQTL that changes the expression of the target gene significantly. After finding all the eQTL supported linkages for chromHMM and JEME+Hi-C, we calculated the fraction of enhancer-gene linkages that has eQTL support for various types of linkages in chromHMM and in JEME+Hi-C. As can be seen in figure below, JEME+Hi-C has higher fraction overlapped with eQTL-gene linkages. |

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

<TYPE>$$$NoveltyPos

<ASSIGN>@@@WM

<PLAN>&&&AgreeFix,&&&MORE - WM fix workding

<STATUS>%%%95DONE

Put …

Condense

|  |  |
| --- | --- |
| 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 of the extended gene.  Based on this, we have tried to highlight this more in the revision.  We also tried to make it more clear that the is useful in many contexts in applying encode to cancer.   * GWAS germline variant enrichment analysis across different annotations in the main figure (Excerpt 2.7-A) * A new figure panel to stratify patient expression levels based on the mutation status from various annotations. (Excerpt 2.7-B) * A new figure in the supplement to show variant effect in extended gene regions on regulator activities (Excerpt 2.7-C) * A CRISPR based validation of oncogene activation based on extended genes (Excerpt 2.7-D)   .] |
| Excerpt 2.7-A (from MS) | We extracted all the breast cancer GWAS variants from GWAS Catalogue and only kept those with European ancestry. Then we extracted all the LD SNPs within 500kb of the GWAS SNP (r2>0.8) to calculate variant enrichment in different annotations sites. The R package VSE was used (https://cran.r-project.org/web/packages/VSE/vignettes/my-vignette.html). We found that extended gene regions showed significantly larger variant enrichment than the CDS regions and TSS regions. |
| Excerpt 2.7-B (from MS) | For a given gene, we tried to separate patients into groups with or without mutations under certain annotations, such as CDS, UTR, TF/RBP binding sites, enhancers, and our extended gene. We then tried to test difference of gene expressions (FPKM) from these two groups based on two-sided Wilcoxon. We found that our extended gene annotation provides better expression separation between these two groups. As an ex to illustrate the value of the extended gene for expression analysis, we show a well-known splicing factor SRSF2, which has been recently reported to drive liver cancer development \cite{28082404}, gives the strongest p-value for stratifying expression out of all genes in liver cancer. |
| Excerpt 2.7-C (from suppl.) | We analyzed the association between TF mutations in extended gene region and TF regulatory activity in three cancer types (breast, liver, and leukemia). Between each pairs of mutation type (e.g., ENH1, TF, eCLIP, UTR) and cancer type, we tested the association between mutation status and TF regulatory activity by two-sided rank-sum test and converted the *p*-values into FDRs by Benjamini-Hochberg procedure. Only the combination between liver cancer and ENH1 mutation has statistically significant results (FDR < 0.25, panel a). A mutation in the enhancer region of DPF2 or RELA indicates a lower TF regulatory activity (panel b). These results indicate that mutations in enhancers may cause TF loss-of-function in certain cancer types.    **Supplementary Figure X. Mutations in level one enhancers affects the activity of nearby TFs.** (a) The association between TF regulatory activity and mutation in enhancer regions. For each cancer type, the association between TF regulatory activity computed using ChIP-seq data and mutation status of nearby enhancer region was tested by two-sided rank-sum test. Only liver cancer has significant associations (FDR < 0.25) for TF DPF2 and RELA, and the results for liver cancer are shown with volcano plot. X-axis represents the z-score of rank-sum test and Y-axis represents the negative log p-values. (b) The regulatory activities of significant TFs in panel a in tumors with mutated or wild-type TF genes. The comparison between two groups was done by two-sided rank-sum test. |
| Excerpt 2.7-D (from MS) | ((Ask Feng’s group for text and wait for figure to come in)) |

## <ID>REF2.8 – Q-Q plots

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

<ASSIGN>@@@JZ

<PLAN>&&&Defer

<STATUS>%%%95DONE

Move back

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

## 

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

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

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%95DONE

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

## <ID>REF2.10a – Confounding factors

<TYPE>$$$Other

<ASSIGN>@@@JZ,@@@WM

<PLAN>&&&AgreeFix - WM to fix the wording

<STATUS>%%%85DONE

|  |  |
| --- | --- |
| Referee  Comment | Next, TF binding and nucleosome occupancy is known to interfere with the activity of DNA repair system. |
| Author  Response | We thank the referee for raising this important point. Actually many of the current background mutation rate estimation methods assume a constant rate in a fairly large region, such as within a gene (including long introns) or up to Mbp fixed bins. At such large scale, it is difficult to small scale features such as TF binding, nucleosome occupancy, histone modification (which changes sharply in less kbps).  There is no proven method to account for TF binding on DNA repair, making it outside the scope of our paper, but we have now included a discussion about how this confound limits the interpretation of some of our results  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. (see excerpt below) |
| Excerpt 2.10-A (from MS) | However, most of the current BMR models are focused on larger scale mutation rate variations by integrating many features at 50 kb to 1 Mb resolution while ignoring small scale perturbations introduced by TF binding and nucleosome occupancy. Improvement of such finer scale features in the future could further improve BMR estimation. |

## <ID>REF2.10b –

<TYPE>$$$Other

<ASSIGN>@@@ [[WM]]

<PLAN>&&&AgreeFix - WM to fix the wording

<STATUS>%%%85DONE

**REF2.10b…... (split)**

It would be great to see a formal analysis about how extended genes increase power of cancer driver discovery.

We have done this in conj. w/ the compact gene analysis . See excerpt XX & respond YYYY above.

We appreciate the referee’s comments that our extended gene concept and its power advantage is an important concept in our paper, and we now give it a formal treatment in the revised manuscript. We conclude that a doubling the length of a test region by merging two equivalent segments maximally doubles the power [MADE UP] in the setting where both elements are subject to the same strength and direction of selection, although the effect trails off in certain conditions[[FIX to right nubmers]]

## 

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

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

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

|  |  |
| --- | --- |
| Referee  Comment | 1) I would not use the term “burden test”. This usage is slightly confusing because this term is commonly used in human genetics where it refers to a case-control test. |
| Author  Response | We thank the referee to point out his confusion about the term “burden test”. This is where some of the confusions of this paper come from. Originally we intended to use this term because we want to emphasize that our resource is not just for somatic variant analysis such as cancer driver detection. We have other applications such as case-control GWAS variant interpretation. We have re-organized our analysis to better convey our idea. Please check details to the response in REF 2.7 above. |

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

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

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

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

# 

# Referee #3 (Remarks to the Author):

## <ID>REF3.0 – Preamble

<TYPE>$$$Text

<ASSIGN>@@@MG,@@@JZ,@@@PDM

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

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

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

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

## <ID>REF3.1 – Presentation of the paper

<TYPE>$$$Presentation

<ASSIGN>@@@PDM

<PLAN>&&&AgreeFix

<STATUS>%%%25DONE

|  |  |
| --- | --- |
| Referee  Comment | It is difficult to understand the significant novel findings in this paper (compared to the main ENCODE paper).  [ WM2all: This first sentence deserves to be a standalone point. We need to respond with a succinct summary of the novel discoveries and innovation of this paper: we have the extended gene concept, network rewiring, etc, and how they are importantly enabled by the new ENCODE data; in addition some concrete discoveries: SUB1 as a possible cancer driver gene, convergent patterns of regulation as diverse cell types become cancerous; ... ]  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 thank the referee to raise this comment about our supplementary file. Our original thinking was some of the contents are distributed in multiple sections. For example, each step in the final prioritization scheme is corresponding to a separate part in the supplements. As suggested, we have added the specific sections in our revised manuscript to make it easier to check the technical details.  [[PDM2all: Here’s some possible text for this response:  ---    We appreciate these points of feedback, and agree that the clarity of presentation of our results could be improved. There are accordingly a number of modifications we have made to improve the communication of our results to the reader.  In order to emphasize specific results and validations that demonstrate the value of the ENCODEC resource, we now include section headings that highlight these results. We have also refocused our introduction and discussion to emphasize our novel findings and the novelty of the ENCODEC resource (see excerpt below).  We also now provide more detailed referencing to supplementary information. Specific sections or subsections are now included with all references to supplementary information (e.g. see Supplement, section 7.2). All supplementary tables and figures are also specifically referenced by number where they appear in the main text.  ---  ]] |
| Excerpt 3.1-A (from MS) | [[### Add several sentences/paragraphs from main text that clearly communicate our main results and the value of the ENCODEC resource]] |

## <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 | In our revised version, we have re-organized and updated Figure 2 to better illustrate our key idea - the scale of data from ENCODE helps to interpret genome variations in cancer. We have tried to make it clearer by better legends.  For the original question,as shown below in the new figsupports the claim because replication timing from MCF-7 outperforms that from HeLa to predict BMR in breast cancer. We have added a sentence in the supplementary document and moved this panel to supplement.  [[PDM: Some possible response text:  The results in Figure 2A provide the support for this statement. We compared the correlation of observed vs. predicted background mutation rate (BMR) for several cancer types. We found that using Repli-seq data from a cancer-type matched cell line to model BMR provided stronger correlation to observed BMR than using similar Repli-seq data derived from HeLa-S3. HeLa-S3 derives from cervical cancer and is not a direct match to any of the three cancer types we examined – breast adenocarcinoma, chronic lymphocytic leukemia, and liver hepatocellular carcinoma. This analysis was intended as a straightforward demonstration of the value of using matched cell lines from ENCODE to model cancer-type specific processes.  …  Etc.  Could add some details about about the new figure 2 here… How it illustrates a similar point as original figure 2 but with greater clarity etc.  ]] |
| Excerpt 3.2-A (from fig. 2) | Wait for new figure 2 |

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

<TYPE>$$$Presentation

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

|  |  |
| --- | --- |
| Referee  Comment | In Figure 1, “top tier” should point to cell types that is mentioned in the content. However, we also see SNV, SV, Mutation, etc. |
| Author  Response | We thank the referee for this comment. In fact, by integrating many assays such as whole genome sequencing and Irys, we called the SNV and SVs for several top tier cell lines, and release them together with our resource (see excerpt 2). In the revised figure 1, we have made it clearer that our resource include these SVs and SNVs.  [[PDM2all: There seem to be two things going on -- 1. The reviewer thinks we mislabelled our figure because data types != cell lines. 2. We actually intentionally associated a number of data-types with the top-tier cell lines. Good if our response can cover both, e.g.:  The reviewer is correct that ‘top tier’ refers only to the 6 cell lines that we have identified as having significant associated assay data (K562, HepG2, A549, MCF-7, HeLa-S3, and H1-hESC). The association of the labels ‘SNV, SV, mutation, etc.’ with these cell lines was intended to indicate that we are releasing these data-types for the top-tier cell lines. We have revised our figure layout and labelling to emphasize that ‘top-tier’ only refers to those 6 cell-lines. The data we have processed and provisioned in association with these cell-lines are now labeled separately.  ]] |
| Excerpt 3.3-A (from fig. 1) | Wait for updated Fig 1 |
| Excerpt 3.3-B (from suppl.) | JZ2DL: could you pls make a table from Feng’s data and deposit it to our resource? DL: todisc |

## 

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

<TYPE>$$$Presentation

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

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

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

<TYPE>$$$BMR

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

|  |  |
| --- | --- |
| Referee  Comment | For Figure 2B, what does ‘regression coefficients of remaining features’ mean? Does that means beta\_0 or the remaining regression noise? From Figure 2B, the coefficient to regression is rounded to -0.001 and 0.001. How should we understand these values? If the coefficients are for the main features, we would be expecting higher coefficients, wouldn't we? In this case, does it means the lower the better? |
| Author  Response | To better illustrate the value of ENCODE data and our extended gene annotation, we reorganized our analysis to provide a new figure and moved this to the suppl. We have also fixed the text to describe our method and specifically answer the referee's questions (details in the excerpt below).  WM suggested replacement text: The plotted regression coefficients are the non-intercept (i>0) beta\_i’s. The coefficients are much more substantial than they appear in the original manuscript. This is because, in the original manuscript, we arbitrarily chose to measure the response variable in units of mutations per nucleotide, for which observed values for the response variable are much less than 0.001. We now recast the regressand in new units (mutations per megabase) which make the beta values easier to interpret. For example, a 1 standard deviation increase in H3K9Me3 intensity predicts an X% increase in the mutation rate in HMEC. |
| Excerpt 3.5-A (from suppl.) | For the genomic features, we calculated the average signal per 1Mbs and transformed it into Z scores. It is worth mentioning that we also had an offset parameter, which means we are trying to estimate the point mutation rate (~10E-6 in some cases), so 0.001 is not a small value. Regarding the interpretation of the regression coefficient, a larger absolute value means better BMR estimation. |

## <ID>REF3.6 – Definition of the extended gene

<TYPE>$$$Annotation

<ASSIGN>@@@JZ,@@@PDM

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

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

## <ID>REF3.7 – Validations

<TYPE>$$$Annotation

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

|  |  |
| --- | --- |
| Referee  Comment | For the Figure 2D and its description on the third paragraph of page 4 (as well as Figure 3A), did the authors validate all the genes systematically? Is there any validation rate showing the precision rate of the method? |
| Author  Response | We thank the referee for raising the question of validations.  For Figure 2D, it is about the somatically burdened genes. We fully agree with the referee that it is useful to compare our BMR to established benchmarks. We are aware of community efforts and are very involved with the PCAWG effort to do whole genome cancer analysis. One of our authors is the co-leader of the non-coding annotation group. PCAWG, which is a hybrid of TCGA and ICGC, has not developed any explicit BMR benchmark. Validation we can do is to provide literature support for our discovered genes and we have added them into a supplementary table (Excerpt 3.7-A).  Please note that we do have explicit validation for the prioritized SNVs and SVs in the paper.  \* For instance, Figure 2C shows a validation of extended gene that initiate oncogene transcription (Excerpt 3.7-B).  \* For Fig. 3A, We have used TF/RBP knockdown experiments to validate several key regulators, such as MYC and SUB1. We have also used external data to validate our conclusion. These analysis were added into our revised supplements (Excerpt 3.7-C).  [WM2all: To demonstrate the validity of conclusions drawn from an extended gene analysis, we only need to externally validate as relevant in cancer those genes that were found burdened only using the extended gene approach but not found burdened on the basis of the CDS or TSS mutation distributions alone; namely: USFP2-LIHC, ALB-LIHC, PLA2G6-BRCA, LTB-CLL, ST6GAL1-CLL, CXCR4-CLL, XBP1-BRCA, BACH2-CLL, BTG2-CLL, MKS1-CLL, and BCL6-CLL (or so I thought - why isn’t SUB1 among these?). This is surely a manageable number and we could have tested each of them. Maybe we still can using the secret weapon. In fact, I suspect these are mostly false positives. We may have to say that we now have a multi-pronged demonstration of the value of the extended genes: such as by correlating gene expression and mutation count, cancer GWAS enrichment, and ability to re-discover known cancer driver genes with fewer patients than would be required for CDS-only (or TSS-only)]  [[PDM2all: We can phrase this response and contextualize it a bit better.  The reviewer is only asking us about a validation of the genes in figures 2D and 3A. Any validations of other unrelated genes, regulators, SNVs/SVs are great, but not what the reviewer asked.  Fortunately, there appears to be some intersection of our validations, and the genes in Figures 2D/3A.  We could say e.g.:  Yes, although not all regions associated with all genes in Figures 2D and 3A could be validated, we prioritized genes for validation according to their [[target expression, regulatory status, etc.]]. We then performed TF/RBP knockdown for these genes. The results are shown in Figure ####.  In the particular case of the gene #### that initiates oncogene transcription, we performed targeted validation with CRISPR knockdown. [[is this what Feng is doing?]].  ]]  Regarding the validation rate, we have prioritized SNVs at the end of our manuscript, 6 out of 8 SNVs were shown to affect gene expressions (Excerpt 3.7-C). |
| Excerpt 3.7-A (from suppl - extended fig. 2D) | We have listed the literature supporting our discovered genes with higher than expected mutations.  **BRCA**   |  |  |  |  | | --- | --- | --- | --- | | Gene | Cancer Type | Literature Support (PMID) | Known Cancer Gene (CGC) | | CBFB | Breast | 22722202, 16959974, 20668451 | YES  TSG | | HIST1H2BF | Breast | 26113056 |  | | HIST1H2AD |  |  |  | | HINT3 |  |  |  | | HIST1H3D | Breast | 26113056 |  | | PIK3CA | Breast | 26028978, 29636477, 25176561, 27358378 | YES  Oncogene | | TP53 | Breast | 11879567, 12619115, 8013000 | YES  TSG/Oncogene |   **LIHC**   |  |  |  |  | | --- | --- | --- | --- | | Gene | Cancer Type | Literature Support (PMID) | Known Cancer Gene (CGC) | | TERT | Liver | 26336998, 25267585, 28947783 | YES | | KRTAP5-11 |  |  |  | | NFE2L2 | Liver | 22459801 | YES | | SETDB1 | Liver | 26471002, 26481868, 27334461 |  | | ARID2 | Liver | 21822264, 26169693, 22095441 | YES TSG | | DUSP22 |  |  |  | | IFI44L | Liver | 27254796 |  | | PHLDB2 | Liver | 22681909 |  | | AL590714.1 |  |  |  | | APOB | Liver | 23723369 |  | | APOA2 |  |  |  | | PLCXD2 |  |  |  | | ZNF595 |  |  |  | | ALB | Liver | 24663086 |  | | CTNNB1 | Liver | 26715116 | YES Oncogene | | TP53 | Liver | 17401425 | YES  TSG/Oncogene |   **CLL**   |  |  |  |  | | --- | --- | --- | --- | | Gene | Cancer Type | Literature Support (PMID) | Known Cancer Gene (CGC) | | NXF1 | CLL | 27060156 |  | | ATM | CLL | 26113859, 22952040 | YES  TSG | | SYVN1 |  |  |  | | WDR74 |  |  |  | | LTB | CLL | 12801841 |  | | SF3B1 | CLL | 25371178 | YES | | BTG2 |  |  |  | | RPL11 | CLL | 12200376 |  | | BCL7A | CLL | 23043359 | YES Oncogene | | CXCR4 | CLL | 24855209, 20501831 | YES  Oncogene | | BACH2 |  |  |  | | BCL2 | CLL | 27069256 | YES  Oncogene | | TP53 | CLL | 27742075 | YES TSG/Oncogene | | BCL6 | CLL | 19367498 | YES  Oncogene | |
| Excerpt 3.7-B (from MS and fig. 2C) | (Add Feng’s text to b) |
| Excerpt 3.7-C (from suppl. - fig. 3) | To detect predicted common target gene of MYC and SUB1, shRNA plasmids containing 4 targets sites of each gene were used to transfected to HepG2 cell using Lipofectamine™ 3000 following the manufacturer's instructions (Invitrogen) (target sites for each gene are listed in Sup table 1). Briefly, 0.12 M HepG2 cells were seeded in each well of one 24-well plates 24 hours before transfection. 500 ng plasmids containing either single shRNA or 4 shRNA plasmids as pool were mixed with 0.75 uL Lipofectamine™ 3000 in Opti-MEM I medium (Invitrogen) and loaded to HepG2 cells in each well. Blank plasmids without shRNA target sequence was used as control. To improve transfection efficiency, 2 ug/mL puromycin was used to select successful transfected cells. 72 hours after transfection, total RNA was extracted using RNeasy Mini Kit (Qiagen) and followed by cDNA generation using SuperScript III (Invitrogen). Knockdown efficiency and target gene expression level were quantified and compared to BACTIN by qPCR using KAPA SYBR® FAST qPCR Master Mix (2X) Kit (Sigma). The qPCR primers were listed in suppl. table 2. |
| Excerpt 3.7-D (from fig. 6B) | Figure 6B. Validation of candidate regions |

## <ID>REF3.8 – novel oncogenes

<TYPE>$$$Annotation

<ASSIGN>@@@JZ,@@@PDM

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

|  |  |
| --- | --- |
| Referee  Comment | Are there any novel oncogenes detected by the method? |
| Author  Response | We have tried to make it clear that the main goal of this paper is to illustrate the value of ENCODE data and the usefulness of our deep annotations. We did find interesting genes that are associated with cancer, such as SUB1 (which is also mentioned by REF5 a potential novel oncogene). To our knowledge, this is the first work to claim SUB1 to be associated with cancer as an RBP. There are other work mentioning this gene, but not from the RBP aspect. We have added many follow up analysis on SUB1 in our revised version.  [WM2: Furthermore we show that up to 10% [MADE UP!] of known cancer genes could have been identified as significantly mutated with a 10% smaller patient cohort than would be necessary using traditional annotations. The effect size was amplified in more rarely mutated cancer genes (10% findable with 30% smaller cohorts [MADE UP!] As the cancer genomics community attempts to identify the thin tail of rare cancer drivers, which are the ones remaining to be discovered, innovative annotation approaches such as the extended gene concept are expected to increase the community’s power to detect novel cancer genes] |
| Excerpt 3.8-A (from suppl.) | **Supplementary Figure X: eCLIP peaks of SUB1.** (a) The composition of SUB1 peaks over different gene regions is shown for each replicate. (b) For each gene region, the relative enrichment (fraction of SUB1 peaks / fraction of all peaks) of SUB1 peaks is shown. (c) The distribution of SUB1 peaks over 3’UTR regions is shown. The mean across all RNA binding proteins profiled by eCLIP experiments are shown as background with standard deviation as error bars. |
| Excerpt 3.8-B (from suppl) | We found that SUB1 targets are enriched in cancer associated genes, such as genes in Cancer Gene Census (P=1.8e-16 by Fisher’s exact test), and such genes showed larger down regulation upon SUB1 knockdowns. Among many of such genes, we have shown some IGV examples together with SUB1 binding sites on the 3’ UTRs. |
| Excerpt 3.8-C (from suppl.) | Using ENCODE eCLIP data and TCGA tumor profiles, we applied RABIT framework to identify RNA binding proteins (RBP), whose target genes are differentially regulated in cancer. (a) The fractions of patients with target genes up or down regulated are shown for each combination of RBP and cancer type. (b) The patient fractions with target genes differentially regulated are shown for all cancer types and RBPs whose fraction values are larger than 50% in at least one cancer. (c) All lung adenocarcinoma patients are divided to two groups according to SUB1 activity predicted by RABIT. The overall survival was shown by KM plot. The association between SUB1 activity and survival was tested through Cox-PH regression. (d) In the left panel, the cumulative distributions of gene expression after SUB1 knock down in HepG2 cell are shown for predicted SUB1 targets and none targets. In the right panel, the cumulative distributions of mRNA decay rates in HepG2 cell are shown. The comparison between two categories is done through Wilcoxon rank-sum test. |
| Excerpt 3.8-D (in Suppl.) | Among genes whose 3’UTR regions have *SUB1* eCLIP sites, we observed significant enrichment of functional categories including *MYC* targets and spliceosome. *MYC* activation induces an increase in total precursor messenger RNA synthesis, which increases the burden on the core spliceosome to process pre-mRNA1. Also, *MYC* activation can stimulate oxidative phosphorylation, which fulfills the bio-energetic demands of cancer cells 2. These results together indicate that *SUB1* may stabilize the *MYC* target genes and pathways to promote the malignant growth of cancer cells. |

## <ID>REF3.9 – Logic gates

<TYPE>$$$Network

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

|  |  |
| --- | --- |
| Referee  Comment | Are circuit gates necessary for Fig 3B? There are OR, AND and NOT gates used.​ For Figure 3C(i), what is the meaning of the values between the green and yellow dots (MYC and \*)?The figure legends are not explaining the figure very well and many details are omitted. |
| Author  Response | * Not necc. But useful * We have redrawn the figure to make it clearer. * The circuit gates represent how MYC and NRF1 work together. * The value of green and yellow means the number of genes under different situations. Specifically, <-113-> means in our network there are 113 genes regulate MYC and at the same time, are the target of MYC. <-1487- means there are 1487 genes regulating MYC, and -2135-> means there are 2135 genes being regulated MYC, but not regulate MYC. * Figure legend have been updated |
| Excerpt 3.9-A (from fig. 3 and fig. legend) | ((Wait for Figure 3)) |

## 

## <ID>REF3.10 – Network hierarchy

<TYPE>$$$Hierarchy

<ASSIGN>@@@DL,@@@PDM

<PLAN>&&&AgreeFix

<STATUS>%%%99DONE

|  |  |
| --- | --- |
| Referee  Comment | For Figure 4, what does the star symbol (\*) mean in the legend? Did the authors use a different grey color to show the connection between TFs? I’m not able to read the grey gradient for the edges. |
| Author  Response | We thank referee for pointing out this issue.  First, we have updated figure legend to make it clear what the star symbol (\*) means in the revised manuscript. The star symbol (\*) indicates a statistically significant difference between  [[... PDM2all: It’s not clear to me what we are testing here. Wilcoxon rank sum test between what two conditions? My guess is a rank sum test of the expression correlations of genes with their target genes that was performed pairwise among levels of the hierarchy. We are probably best to be specific here.]]  we have performed Wilcoxon rank sum test to show the significance of regulators placed in different network hierarchy.  Second, we also improved the presentation of the network hierarchy figure. For the cell type specific network, we highlighted gained and lost edges  [[PDM2all: gained vs. lost is defined according to what metric? Target expression correlation? There also appears to be some discretization here -- arrows are either green or red or grey (but not intermediate, e.g. grey-green. What are these thresholds? Also, what is the difference between an arrow being red, vs. a circle being red?]]  with green and red arrows, added labels colors to represent gainers and losers. |
| Excerpt 3.10-A (from fig. 4) | **Figure 4. Regulatory network rewiring and hierarchies.** …  … (C) Cell-type specific network using K562 and GM12878 ...  … If a p-value is less than 0.05, it is flagged with one star (\*). If a p-value is less than 0.01, it is flagged with two stars (\*\*). If a p-value is less than 0.001, it is flagged with three stars (\*\*\*). |

## <ID>REF3.11 – Network rewiring

<TYPE>$$$Network

<ASSIGN>@@@DL

<PLAN>&&&AgreeFix

<STATUS>%%%99DONE

|  |  |
| --- | --- |
| Referee  Comment | For Figure 5B, what does the vertexes and edges represent? I guess they represent genes and their network connection, respectively? How did you select the genes and why are some of them "thick" while others "thin"? |
| Author  Response | We thank referee for pointing this issue out.  Vertices represent genes (regulators) and edges represent regulatory linkage between TFs and genes.  We have used edge colors to show regulatory rewiring between cell types.  Green line - gained edge  Grey line - retained edge  Red line - lost edge  We have used node thickness (size) to show overall rewiring status.  Think node - high rewiring  Thin node - low rewiring  We have redrawn the figure and added a legend to make it clearer (see excerpt 3.11-A).  (Reference shadow figure) |
| Excerpt 3.11-A (from fig. 4) | **Figure 4. Regulatory network rewiring and hierarchies.** |

# 

# Referee #4 (Remarks to the Author):

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

<TYPE>$$$NoveltyPos

<ASSIGN>@@@MG,@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

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

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

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

<ASSIGN>@@@DC,@@@JZ,@@@PDM

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

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

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

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

<ASSIGN>@@@DC,@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

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

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

<TYPE>$$$CellLine,$$$Validation

<ASSIGN>@@@JZ,@@@DL,@@@Peng,@@@DC

<PLAN>

<STATUS>%%%90DONE

|  |  |
| --- | --- |
| Referee  Comment | One of the limitations of the analysis are the cells that are central in the ENCODE, that are immortalized, including cancer cells and “normal” immortalized counterparts. Most of these cell lines have been kept in culture for decades and further selected for cell growth very extensively. Many of the cell lines may have/have accumulated further mutation and rearrangements, if compared to what cancer cells are at the moment that they leave the human body. The authors accurately acknowledge, in the discussion, stating that it is difficult to match cancer cells with the right normal counterpart; it may also be even more difficult to define what are they really ...  It would be appropriate to (computationally) verify at least a small part of the data in other systems, taking from published studies including normal cells control and primary cancers. |
| Author  Response | [[PDM2all: A possibly somewhat clever thing to do would be to highlight where we have already made these comparisons. E.g. We have already made such a comparison for BMR, haven’t we? A big part of our paper is tumor vs. normal comparison isn’t it? We could highlight some pre-existing analyses, alongside any new investigations.]]  We agree that it is important to verify the discoveries from cell lines in primary cancers. We have added such comparisons in our revised version. Specifically, we added a supplementary section to show that TF regulatory activities predicted from ENCODE TF regulatory networks compared with their expression levels are highly correlated in breast and lung cancer (see exceprt below). |
| Excerpt 4.4-A (from suppl.) | We predicted the regulatory activities of the transcription factor (TF) MYC using a ChIP-seq profile in MCF-7 cells. We found that the MYC regulatory activity is highly correlated with the MYC expression across TCGA breast tumors (Supplementary Figure Xa). For most TFs, their regulatory activities predicted using ENCODE ChIP-seq profile in cell lines are significantly correlated with their expression levels across breast tumors (Supplementary Figure Xb). Moreover, using the same MCF-7 ChIP-seq profile, the MYC regulatory activity predicted for lung tumors is also significantly correlated with MYC expression level in TCGA lung cancer (Supplementary Figure Xa). These results indicate that the ChIP-seq profiles from a particular cell line can capture regulatory targets in human tumors from diverse cancer types. To select ChIP-seq or eCLIP profiles that are representative of the regulatory targets in human cancers, we only reported the results of TFs or RBPs whose regulatory activities are significantly correlated with their gene expression level in each TCGA cohort (Supplementary Figure Xc).    **Supplementary Figure X. The clinical relevance of ENCODE cell line data in human primary tumors**.  **(a)** The correlation between *MYC* expression level and regulatory activity across tumors. The MYC regulatory activity in each tumor was predicted using the ChIP-seq profile in the MCF-7 cell line. The Pearson correlation between MYC gene expression levels and regulatory activity were computed across tumors in each cancer type. The statistical significance of the Pearson correlation was tested by the two-sided student t-test. BRCA: breast carcinoma. LUSC: lung squamous cell carcinoma.  **(b)** The distribution of correlation *p*-values in TCGA breast cancer. For each TF, we tested the statistical significance of Pearson correlation between TF expression levels and regulatory activities predicted across tumors through two-sided student t tests as for panel a). For the TCGA breast cancer cohort, most *p*-values are very significant with few non-significant values.  The fraction of regulators with statistically significant correlations in different cancer types for ChIP-seq and eCLIP networks. In each TCGA cancer type, we computed the correlations between regulator expression levels and regulatory activities across tumors for all regulators (TFs, or RBPs). We selected regulators with statistically significant correlations through a two-sided student t test (FDR < 0.05). |

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

<TYPE>$$$CellLine

<ASSIGN>@@@JZ,@@@DL

<PLAN>&&&MORE

<STATUS>%%%95DONE

|  |  |
| --- | --- |
| Referee  Comment | I have seen data in other studies, showing that many of cancer cell transcriptome are quite similar to each other, if compared to initial or primary cells, showing that in particular cancer cells lose diversity |
| Author  Response | The reviewer is correct that many cancer transcriptomes de-differentiate and lose diversity during tumorigenesis. In relation to this and other points, we have expanded our analysis on stemness in the revised manuscript and made a new figure, which is shown in the response to the Excerpt 4.6-A. |

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

<TYPE>$$$Stemness$$$Calc

<ASSIGN>@@@DL,@@@PE,@@@DC, PDM to put fixes in

<PLAN>&&&AgreeFix,&&&MORE

<STATUS>%%%75DONE

|  |  |
| --- | --- |
| Referee  Comment | 3) One of the conclusions, deriving from the analysis of H1-hESC is the some cancer are “moving away from stemness”. However, while it is true that the cancer cells pattern diverge from the H1 cells, H1 is a human embryonic stem cells: although interesting, H1 may not necessarily be the best cells to compare with tumor phenotype. Authors should discuss/defend of further elaborate on this approach. I believe that a key analysis should be done against other stem cells (like tissutal stem cells, etc. ). |
| Author  Response | We thank the referee for this comment, which we found insightful. In fact, one of the virtues of ENCODE is the large number of different tissues and cell types available. Thus, we have responded to the referee's comment and actually expanded on this point by showing all the cancer types in relation to a number of stem cells available within ENCODE. We have now included an additional figure.  We initially focused on H1 because it is one of the top-tier ENCODE cell lines with broadest cell type coverage. In developing this figure, we were able to use the ENCODE knockdown data as a validation to observe overall pattern from the effect of oncogenes. See excerpt for more details. |
| Excerpt 4.6-A (from fig. 5 and suppl.) | We have highlighted the de-differentiation of cancerous cell types into stem-like cell types using proximal regulatory network (CTCF ChIP-seq) and distal regulatory network (ccRE ELS hotspots), and we show that our findings are in agreement with previous findings using gene expression (RNA-seq).  We performed PCA analysis (reference component analysis (RCA) for gene expression; {\cite: Li, Huipeng, et al. "Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors." Nature Genetics 49.5 (2017): 708.}) using uniformly processed poly A long RNA-seq, CTCF ChIP-seq, and candidate cis-regulatory element from ENCODE encyclopedia. We consistently found that cancer cells tend to cluster together, closer to the stem-like cell cluster, in contrast to their normal counterparts.  Figure 5. PCA (RCA) of regulatory networks and gene expression. |
| Excerpt 4.6-B (from suppl.) | We find that stem-like cells in ENCODE, including top-tier H1-hESC cell line, form a cluster and their regulatory patterns and expression profiles are distinct from differentiated normal cell types and tissues. This highlights that pluripotent embryonic stem cells like H1-hESC maybe not far distinct from other stem-like cells and cell-of-origin.  For the proximal network, we built a simple regulatory network based on CTCF binding peaks. Our preliminary network consists of 14,536 TSS (2.5kb up/downstream) with CTCF peaks across 207 cell types. We filtered for recurrent CTCF binding in at least 20 different cell types to subset the network, and finally, we used 9,506 CTCF hotspots near TSS across 207 cell types to perform PCA analysis.  For distal network, we built 990,079 merged ccRE ELS sites across 609 ccRE annotation. We used two filters to select recurrent distal element. First, we selected ccRE ELS sites that are 100kb away from TSS, and second, we selected ccRE ELS sites seen in more than 20 different cell types. We finally used 13,497 ccRE ELS hotspots across 134 cell types and performed PCA analysis.  For the gene expression, we simply used replicate-merged FPKM of 20,345 protein coding genes across 329 cell types to run RCA (reference component analysis). |

## <ID>REF4.7 – Fixes for Figure 1

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

<ASSIGN>@@@DL

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

|  |  |
| --- | --- |
| Referee  Comment | 4) I have difficulties to fully understand Fig.1, in particular the patient cohort (PC) at the bottom of the “depth approach” (just above the green box of cell –specific analysis). The two rows are at the bottom of the columns report mutation and expression, but they belong to the columns of the cell lines (K562, HepG2, etc). I just simply do not understand that part of the figure, in particular the relation between cell lines and the patient cohort (the figure legend does not help, and also supplementary material did not help). |
| Author  Response | In the revised manuscript, we have modified the figure 1 to make it more clear. We understand that numbers in the mutation and expression rows can be confusing, so we have moved cohort-based data matrix out of cell-type data matrix to the supplement and tried to expand the description of it make it clearer (see excerpt).  [[PDM2all: We might also want to say what the numbers represent (I am not totally sure…).  E.g. The patients cohort represents individual samples from TCGA for which we have called mutations and gene expression levels. They were placed in the columns of the cell lines, because the associated cancer types of the cell lines (e.g., breast adenocarcinoma and MCF 7, liver hepatocellular carcinoma and HepG2) is the same as for these patients.  However, we see how this could have created confusion… etc.]] |
| Excerpt 4.7-A (fig. 1) | (to be continued for fig 1) |

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

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

<ASSIGN>@@@DL,@@@XK, @@@TG,@@@STL,@@@JL

<PLAN>&&&AgreeFix,&&&MORE

<STATUS>%%%30DONE

|  |  |
| --- | --- |
| Referee  Comment | 5) The analysis assumes that genomes of all the cells discussed are essentially the same. However, for many of the cancer genomes, there have been rearrangements, often dramatic like Chromothripsis. How is this affecting the BMR and the linking of non-coding elements to the target genes? How many of the cells analyzed were dramatically rearranged? |
| Author  Response | [[PDM2all: We should probably be very careful about how we answer this. They are implying that our analyses may be invalid because we did not account for the effect of rearrangements. I can only really see two ways of answering this:   1. We checked: rearrangements occur infrequently and/or are not of major consequence to our analyses. 2. We checked: rearrangements must be accounted for and we have done so.   We appear to be leaning towards choice 2) right now (although not clear that we have done more than show that rearrangements matter -- would also need to demonstrate that we clearly account for their effect in all calculations).  Right now we show case examples where SVs affect our calculations. This is a bit like proving to the reviewer we should have taken SVs into account in all our previous work.  Further, it looks like we’re going to have a fairly strong and multi-faceted approach to answering the first question (about BMR and regulatory networks in relation to SVs). However, it looks like we have not yet answered the second question. Not sure how difficult this will be to do, but it would be useful to have some answer like, e.g. of the six top-tier cell lines, ### have more than ### SVs]]  The referee asked us to comment on the relationship of structural variants, BMR, and networ. We think these are very useful suggestions. In the revision, we have responded to and extended the referee’s suggested in multiple respects as listed below. In particular we show how SVs relate to the mutation rate & we have a validation relating them to ETG connections.  We have also created a small sjupp table showing the nNumner of somatic SVs per cell line |
| Excerpt 4.8-A (from suppl. - SNV and SV in top-tier cell lines.) | We have called SV and SNVs from multiple ENCODE cell lines by integrating various assays as shown in the following table.  JZ2DL: add Feng’s table |
| Excerpt 4.8-B (from suppl. - SNV density around SVs.) | We compared the SNV/InDel density near the SV boundaries in strictly matched ENCODE cell lines and found that there are noticeably elevated SNV/InDel rates around SVs. |
| Excerpt 4.8-C (from suppl. - SV vs. histone marks) | We extracted SV events in K562 and compared them with several histone modification marks. We found clear patterns as below.  [JZ2STL: please add more text and the exact procedure below]  [[PDM2all: We were kind of going over this on the call, but what is the significance of this?]] |
| Excerpt 4.8-D (from suppl. - SV vs. gene expressions) | We have shown in the follow figure several examples of SVs near promoter regions that may affect gene expression.  [JZ2TG: please add more text to describe your procedure here. Also please add x axis labels] |
| Excerpt 4.8-E (from suppl. - SV and rewiring) | Figure 4. Rewiring panel D  (JZ2DL: pls describe what you have done here)  We examined the fraction of rewired edges affected by SNVs and SVs. Larger fraction of gained edges were affected by SNVs while larger fraction of lost edges were affected by SVs. |
| Excerpt 4.8-F (from suppl - SV and oncogene activation) | Ask Feng to write a text |

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

<TYPE>$$$CellLine,$$$Text

<ASSIGN>@@@WM,@@@JZ,@@@MRS

<PLAN>&&&AgreeFix

<STATUS>%%%65DONE

|  |  |
| --- | --- |
| Referee  Comment | 6) Most cancers are not necessarily represented by a single cell type used to obtain genomics data in this study, but contains numerous types of cells with different mutations, as well as normal cells, infiltrating cells, all in a three dimensional structure, often producing metastatic colonizing other organs. However, this study focuses only on comparisons between cells. These limitations should be better discussed, also to put in perspective future studies on single cells. |
| Author  Response | We thank the referee for bringing this up. The reviewer is correct thatgenomic and epigenomic heterogeneity in tumor cells, as well as heterogeneity in the tumor microenvironment (e.g., immune cell infiltrates, hormonal factors, normal cell populations, etc.) are significant factors in tumor growth and development. In our revised manuscript, as suggested we have tried to   * Added more discussion in main text about the limitation of anlayses conducted in single cells, and how future techniques can help (Excerpt 4.9-A). * Specifically in relation to BMR, we now clearly point out that most cancers can not be represented by a single cell type and that is exactly why we used multiple genomic features across cell types [[PDM2all: don’t we somehow mean multiple cells?]] to characterize BMR. ENCODE data expanded features by more than a factor of 10 as compared to other related work published recently). * Regarding rewiring, we have better introduced the concept of a composite normal and discussed the limitations of [[PDM: of this technique?]] current technique |
| Excerpt 4.9-A (from MS - single-cell sequencing in discussion) | One limitation of the current ENCODE data is that most of the current release of data is performed over a small number of cells. However, genomic and epigenomic heterogeneity in tumor cells, as well as heterogeneity in the tumor microenvironment (e.g., immune cell infiltrates, hormonal factors, normal cell populations, etc.) are significant factors in tumor growth and development. We believe that the development of single-cell sequencing technologies may capture important tumor biology and provide new insights in cancer. |
| Excerpt 4.9-B (from MS - heterogeneity & BMR) | While it is valuable to match cancer to its cell of origin, tumors are highly heterogeneous and there are usually multiple normal cell types are around and inside tumor cells, so a combination of different data sets provide the best overall fit to mutation rate. |
| Excerpt 4.9-C (from suppl. - heterogeneity & BMR) | [[PDM2all: any changes we make to Ref. 1.5A should be reflected here. However, here we also need to establish a stronger connection to tumor heterogeneity. It’s sort of implied that all this data allows us to make more accurate BMR predictions, but it’s not clear how this text excerpt connects to the idea of tumor heterogeneity being inadequately modeled in a single cell line.]]  The ENCODE3 rollout dramatically expands the genomic data available for this type of regression by more than a factor of 10 (2,069 vs. 169), many of which are from tissue or primary cells. In total there are 2,017 histone ChIP-seq and 51 replication timing Repli-chip and Repli-seq features to predict BMR. We did a PCA of the signals from these features and selected the best combination of 20 PCs for BMR prediction. It is worth pointing out that the majority of our data is from tissue or primary cells. A summary of cell types for these features is given below.  Table S1. Summary of ENCODE histone ChIP-seq data    Table S2. Summary of ENCODE3 Replication timing data   |  |  |  | | --- | --- | --- | | **Cell Type** | **Repli-seq** | **Repli-chip** | | cell line | 101 | 10 | | in vitro differentiated cells | 0 | 35 | | primary cell | 12 | 5 | | stem cell | 6 | 11 | | induced pluripotent stem cell line | 0 | 2 |   Table S3. Summary of 51 replication timing features from Repli-chip and Repli-chip   |  |  | | --- | --- | | **Cell State** | **Repli-chip/Repli-seq** | | Pluripotent | 8 | | DE | 3 | | Liver/Pancreas | 6 | | Neural crest/Early mesoderm | 7 | | Late mesoderm | 6 | | NPC | 2 | | Myeloid/Erythroid | 5 | | Lymphoid | 5 | | Cancer | 9 | |

## 

## <ID>REF4.10 – lncRNAs and BMR

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

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%90DONE

|  |  |
| --- | --- |
| Referee  Comment | 7) When analyzing the BMR in cancer, did the author estimate the mutation rate in the lncRNAs?  Is there any other interesting lesson from the analysis of the non-coding regions and their mutations rate?  NO |
| Author  Response | We thank the referee to point out this. Our BMR model captures the mutation rate over the whole genome. Thus, we are able to calculate the mutation burden of lincRNAs. We have added results on lncRNAs in our revised supplements (see excerpt below). |
| Excerpt 4.10-A (from suppl. - burden test on lincRNAs) | We also calculated the mutation burden on lincRNAs. We have found well-known cancer associated lncRNAs to be burdened, such NEAT1 in liver cancer, MALAT1 in breast cancer. Results and Q-Q plots were given in Supplementary Table X. |

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

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

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

|  |  |
| --- | --- |
| Referee  Comment | In the supplementary material, there is room to improve figures (some numbers are too small). |
| Author  Response | We thank the referee for pointing this out and we have made revisions to the supplementary figures in our revised manuscript to improve interpretability. |

## 

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

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

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

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

# 

# Referee #5 (Remarks to the Author):

## <ID>REF5.0 – Preamble (remove)

<TYPE>$$$Text

<ASSIGN>WM @@@MG,@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

This table was a product of a long fruitful convo btw DL + WM

It is incomplete, may need help

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Our approach** | **Theoretical advantage of our approach** | **Empirical support for effectiveness** |
| Extended genes |  | Aggregates signal for increased power |  |
| Compact annotations | Pare regulatory elements to functional core using ENCODE peaks | Purifies signal for increase power |  |
| Network analysis |  |  |  |
| BMR |  | Debiases signal for increased accuracy | R^2 of Sanger = 0.7; R^2 us = 0.73 (MADE UP!) |

## **\*\* develop integrative accurate** **ext gene Annotation that are very useful finding key somatic variants, germline, etc ....**

### \*\* develop Network both TF & RBP & show regulator collaboration , network rewiring :

### \*\* show how cancer epigenetics can be related to stem state :

### \*\*\* prioritization of regulators, elements & variants

### \*\*\* do careful small scale of Validations:

In the revised we try to hgihtlight stats & comparison to prev & we go that in the pt.-by-pt

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

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

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

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

We have also tried to illustrate the usefulness of the above resource to prioritize both key regulators and genomic variations (single nucleotide and structural variations) using various techniques, such as luciferase assays, CRISPR, and knockdowns. We hope that all the above aspects serve as good examples to illustrate the value of our resource to cancer genomics.

**2. Regarding BMR estimation**

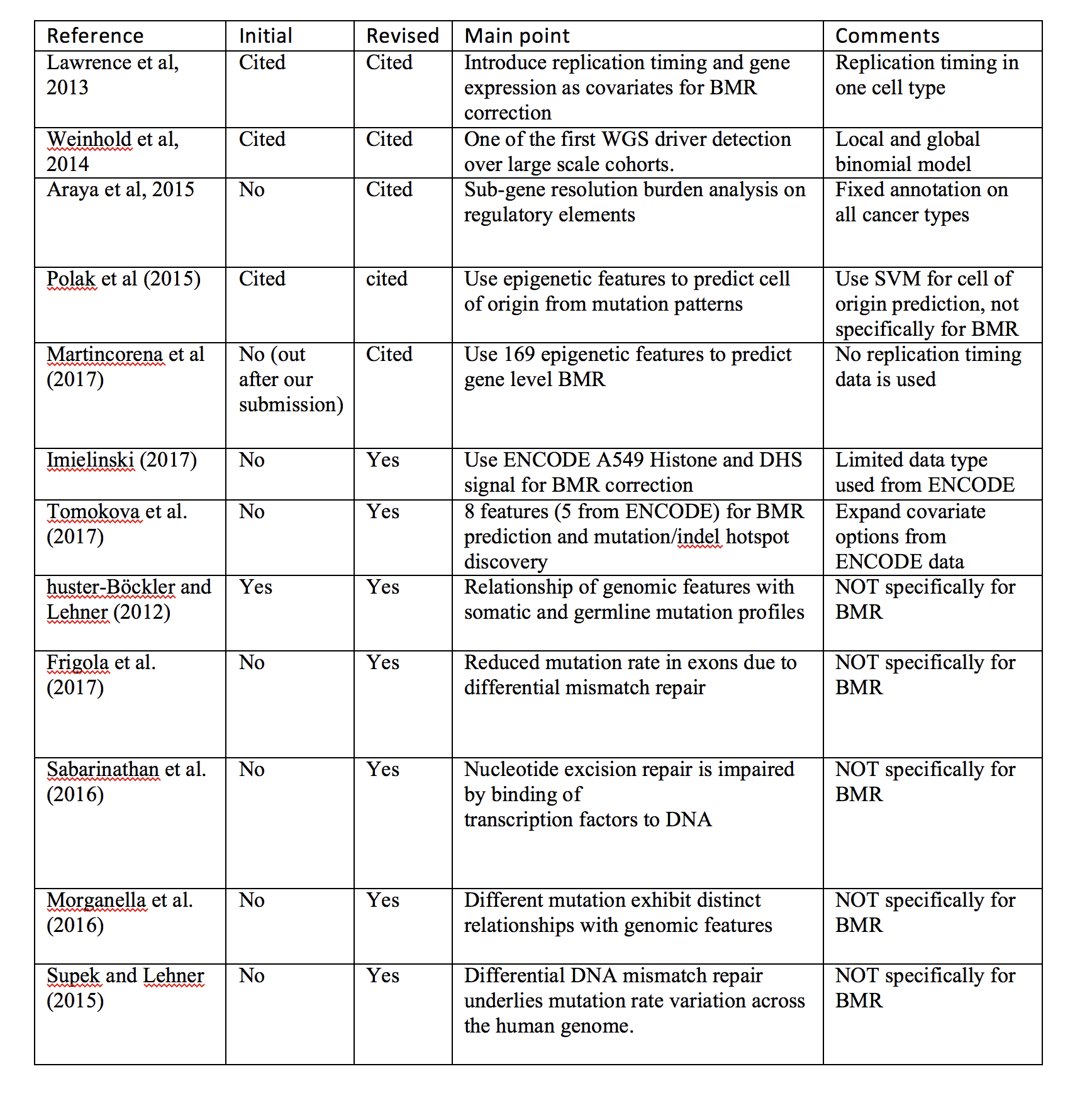
With respect to the BMR estimation part in particular, the reviewer noted that there had been many existing publications focusing on applications such as cancer driver detection.

[[PDM2all: Good to say e.g., ‘Because this is an important area of investigation in cancer genomics...]]

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

Second, we would also like to emphasize that the main goal of our paper is not to present novel methods of driver discovery, but rather to illustrate that the richness of the ENCODE data can be leveraged to noticeably improve the accuracy of BMR estimation. Hence, we feel it is slightly outside the scope for our ENCODE resource paper to make detailed comparisons with driver gene discovery. [[PDM2all: could even potentially go further and highlight the unrealistic expectation of driver gene discovery in a single paper -- it’s a multistep, multistudy process]] In the revised version, we have clearly highlighted the value of ENCODE data in our updated Fig. 1.

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



## <ID>REF5.0 – Overall comments

Ref’s preamble: [[The authors present a cancer-genomics focused resource, named ENCODEC, based on data generated by the ENCODE consortium. This resource contains three contributions: tissue-specific models for the background mutation density; a refined list of functional genomic regions for mutational recurrence testing; and gene-enhancer links. The authors describe the methods to generate these, and explore possible novel cancer driver candidates through application of their resource with TCGA data.

Respond that there is a network part!!!

BMR is only a small part of the paper

While the resources provided in this manuscript are potentially interesting for the cancer genomics community and comprise an extensive body of work, 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. 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.]]

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

<TYPE>$$$Text

<ASSIGN>@@@MG,@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

|  |  |
| --- | --- |
| Referee  Comment | While the resources provided in this manuscript are potentially interesting for the cancer genomics community and comprise an extensive body of work, it is not clear what are the main findings in the paper and their statistical and biological significance. |
| Author  Response | We thank the referee partially for the positive comment.  We detail w/ some of this criticism in preamble.  [[PDM2all: Add text indicating, we have clarified our principal findings as well as the statistical evidence that underpin them wherever possible in our revised manuscript.]]  [[PDM2all: One exercise that could be beneficial for us during this revision process is to make a short list of each of the claims of superiority or novelty we have made, and then list benchmarks or statistical test supporting those claims where appropriate.]] |

## 

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

<TYPE>$$$Text

<ASSIGN>@@@JZ ; @@@PDM

<PLAN>&&&AgreeFix PDM to fix wording

<STATUS>%%%85DONE

|  |  |
| --- | --- |
| Referee  Comment | 1. The manuscript does not clearly state innovation and novelty over previously published data and methods. Several published studies have used epigenomic data types, including replication time and histone modifications from ENCODE and other sources, to model background mutational background density and define genomic elements of interest. The use of the Negative Binomial/gamma-Poisson distributions to model mutational background in cancer has also been published (Imielinski et al 2016; Martincorena et al, 2017). |
| Author  Response | We thank the reviewer for identifying relevant references. In the revised manuscript, we have tried to provide a better context of related work.  We agree that BMR methods have been published & the NBR has been used. The point is to show that you can do better with this approach using the the full set of encode data.  We have also tried to make it clear that BMR accuracy can be improved by using ENCODE3 data. Negative binomial regression is a standard statistical technique that serves this goal. We have made the following changes to attempt to make this clear and fully address the reviewer’s comments.  [JZ2MG: this is a key question they are looking for, so I prefer to summarize it in the following bullet points. Other questions, I can put them into Excerpt 5.2-A (about xxx) for a more concise doc. Pls comment ]   * A new supplementary table to summarize our 2,069 features (vs. 169 in Martincorena et al., 2017) (Excerpt 5.2-A) * We added several references to provide context for previous work (Excerpt 5.2-B). * We have showed how more features with careful feature selection can improve BMR estimation (Excerpt 5.2-C). * We have stated clearly in the main text: more data are helpful, and we have added discussions about the motivation for this (Excerpt 5.2-D). |
| Excerpt 5.2-A (more features in ENCODE3, in Suppl) | Table S1. Summary of ENCODE3 histone ChIP-seq data   |  |  | | --- | --- | | **Cell Type** | **Histone ChIP-seq** | | tissue | 818 | | primary-cell | 521 | | cell-line | 339 | | in-vitro-differentiated-cells | 179 | | stem-cell | 114 | | induced-pluripotent-stem-cell-line | 46 |   Table S2. Summary of ENCODE3 Replication timing data   |  |  |  | | --- | --- | --- | | **Cell Type** | **Repli-seq** | **Repli-chip** | | cell line | 101 | 10 | | in vitro differentiated cells | 0 | 35 | | primary cell | 12 | 5 | | stem cell | 6 | 11 | | induced pluripotent stem cell line | 0 | 2 | |
| Excerpt 5.2-B (better context of previous work) | Many methods have incorporated effects from multiple genomic features by techniques such as negative binomial regression and poisson regression. |
| Excerpt 5.2-C (updated main text  Fig.) | The 2,017 uniformly processed histone modification signal tracks and 51 replication timing data may serve as a resource to significantly improve BMR estimation accuracy.  We also found that BMR estimation can be improved dramatically by selecting an appropriate combination of multiple features from ENCODE. |
| Excerpt 5.2-D (more text in discussion) | Recent work has focused on the effect of cell-of-origin on tumor attributes such as mutational process and tumor classifications. However, to accurately define tumor cell-of-origin is sometimes challenging. For example, even different subtypes of tumor from the same organ may originate from different cell types. The richness of ENCODE data provides a larger pool from which to draw the most representative cell of origin. |

## <ID>REF5.3 – BMR: TCGA benchmark

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

<ASSIGN>@@@JZ,@@@WM

<PLAN>&&&MORE

<STATUS>%%%60DONE,%%%CalcDONE WM to try to put together

|  |  |
| --- | --- |
| Referee  Comment | Throughout, the main manuscript lacks data and statistics supporting the claims made. For example, the performance of tissue-specific background mutation models applied to TCGA data needs to be evaluated against known results and benchmarks from TCGA. It seems that some of these are presented in the extensive supplement and should be moved to the main manuscript. |
| Author  Response | We thank the referee for this comment. As suggested, we have added explanations for some things claim of significance by moving results from the supplement to the main text. But please note Too long 100+ pages, can not move back all  Parametric way is better to interpret  Specifically for the BMR part, we fully agree with the referee that it is useful and important to compare our BMR to established benchmarks.We have tried to benchmark our BMR to other datasets as suggested. In particular, we are aware of community efforts and are very involved with the PCAWG effort to do whole genome cancer analysis.in fact One of our authors is the co-leader of the non-coding driver group.  PCAWG, which is a hybrid of TCGA and ICGC, has not developed specific BMR benchmark. Instead, what they have done is to develop several randomization schemes accepted by multiple groups. Hence, we tried to compare our estimated BMR with such randomizations. Please note that the TCGA Pancan paper is not appropriate here since it is the whole exome and we focus on noncoding.  1. Using a permuted breast cancer dataset, we performed BMR estimation and calculated somatic mutation burden on the CDS regions of ~20k protein coding regions. We found no gene burdening in this randomized dataset (QQ plot given below).  Figure R 2. Q-Q plot of observed vs. uniform p values from permuted breast cancer data set. Diagonal shown in red.    2. We downsampled the simulated dataset. We used half of the data for training and compared the rest with our predictions in the promoter regions. The reason why we picked this particular comparison is because most of other published TCGA benchmarks interrogated protein coding regions, where the relative rates of synonymous and nonsynonymous mutations can be used to calibrate BMRs. This particular calibration is not possible in noncoding regions.  Specifically, we split the PCAWG Liver-HCC somatic SNV set equally into training and testing sets. We applied the Sanger permutation approach used in PCAWG on the training set and used this to predict mutation rates for each of 14,000 promoters, and calculated the residuals between these predictions and the withheld testing data. Similarly, we calculated predicted mutation rates for those same promoters using the ENCODEC model for liver tissue, and calculated the residuals of these predictions from the testing set promoter mutation rates. Overall, the residuals from the ENCODEC predictions are comparable to the PCAWG-derived predictions.  Figure R X. Down sampling of PCAWG data on promoter regions    In summary our meth does sim. Torandom but is more interpr & more efficeient computationaly  Finallynally, Please also note that this work is comparing to accepted PCAWG benchmarks, which are not fully published yet, so we only include them in this response. If these papers come out before the ENCODE package, we can certainly move sections of this response to the text of the paper. |

## <ID>REF5.4 – Power analysis: (1) compact annoations

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

<ASSIGN>@@@JZ,@@@DC

<PLAN>&&&MORE

<STATUS>%%%85DONE

JZ2JZ: add more

|  |  |
| --- | --- |
| Referee  Comment | 4. How do the new “compact annotations” lead to improved results over traditional annotations? |
| Author  Response | We had some discussion in the alreay the supplement in the inti on the value of compact annotation.  To anwser this better in the revision, We have updated Fig. 2. And expanded some other sections. In short, compact annotation helps with burden calculations..  :   * an entirely new section on power analysis and the effect of test region functional site ratios (Except 5.4-A) * more discussion (in the main text) about the pros and cons of merging test regions (Except 5.4-B) * real examples in the supplement (Except 5.4-C) * a new section of quality metrics of the compact annotations to capture functional sites and rm noise(Except 5.7-A) |
| Excerpt 5.4-A (power analysis on compact annotations) | Suppose that we define the following parameters.    Then under the null hypothesis, the probability to observe at least one mutation per patient is    Under the alternative hypothesis,    We did a simulation by starting from a very noisy test region with pretty low true risk loci percentage. We have showed that by trimming the noise loci, statistical power can be increased. But after we have removed the noise and start to trim the true functional loci, the statistical power drops quickly. |
| Excerpt 5.4-B (added in disc. sect) | In summary, our claim is that first we provide compact annotations to pick up functional nucleotides and remove noisy ones through the guidance of many functional characterization assays. Then we hope to join the distributed functional sites together to increase statistical power. |
| Excerpt 5.4-C (more examples in Suppl on compact annotation) | We provided two examples to explain the motivation of our compact and extended gene annotations and why we feel our assumptions for the power analysis is reasonable.  1) Enhancers: Traditionally, enhancers were called as a 1kb peak regions, which admittedly introduced a lot of obviously nonfunctional sites. We believe we can get functional region more accurately by trimming the enhancers down using the exact shapes of many histone marks and further integration with STARR-seq and Hi-C data.  2) TFBS hotspots around the promoter region of WDR74. Instead of testing the conventional up to 2.5K promoter region, we can trim the test set to a core set of the promoter region where many TFs bind, which perfectly correlates with the mutation hotspots (red block) for this well-known driver site (blue line for pan-cancer and green line for liver cancer). |

## 

## <ID>REF5.5 – Power analysis: (2) adding more reference

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

<ASSIGN>@@@JZ -

<PLAN>&&&MORE - JZ to paste in

<STATUS>%%%95DONE

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

## <ID>REF5.6 – BMR: driver detection comparison

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

<ASSIGN>@@@JZ, @@@DC

<PLAN>&&&MORE,&&&OOS

<STATUS>%%%25DONE

|  |  |
| --- | --- |
| Referee  Comment | Again, sensitivity/specificity analyses of driver discovery with large sets, or long vs. reduced element size need to be added. An improvement of background mutation rate is suggested in the manuscript. But concrete comparisons of discovered drivers with previous work, highlighting how the presented approach is more sensitive or improves specificity, are missing. |
| Author  Response | Mark’s dictation:  First of all we've tried to be clear I in this paper that driver we we  regard discovering new cancer drivers is out of scope for this paper  the main aim is to build a useful annotation from the encode data and  also develop a Mauser background mutation rate from the credit we show  how these I can be use as a resource for helping driver discovery and  we are trying to label I am in figure XXL X how the discovered drivers  from this I do take into account the compact annotation on but we  think I detailed analysis of this is out of scope for the paper  furthermore we tried to point out in the paper that the compact and  extended annotation has many other uses for cancer beyond a somatic  driver discovery in particular it's useful for to us and for finding  him differentially expressed genes we've modified the text in excerpt  XXL why whites reflect this  We have also tried to make it clear that the main purpose of our BMR analysis is not to make novel driver discoveries but to test the hypothesis that the richness of the ENCODE data can noticeably improve BMR estimation accuracy. We feel it is out of the scope of this paper to make a detailed comparison of cancer driver discovery rates.  For the driver discovery part, we have now labeled known driver genes in our calculations with supporting literature and further compared our results with established methods.  **We have also tried to make clear** that the main pt . of the annotation is not solely somatic driver but rather a variety of things related to cancer - expr, gwas, &c. The main goal of Fig.2 is to demonstrate the usefulness of the extended gene annotations. Hence, we have also tried to re-organize all of our related analysis from the supplement to serve this goal, which includes   * Better annotation disease associated germline variants (Excerpt 5.6-A). * Better stratify gene expression level by mutational status (Excerpt 5.6-B). * CRISPR based validation of oncogene activation by SV events (Excerpt 5.6-C). |
| Excerpt 5.6-A (extended gene in GWAS SNPs) | [[PDM2all: the analyses here and in 5.6B are useful and interesting. However, they asked us about driver genes, so my impression is that we have to answer that question foremost. Otherwise, it looks a bit like we’re dodging the question.]]  We extracted all breast cancer and leukemia GWAS variants from the EMBL-EBI GWAS Catalog. We removed studies with irrelevant phenotypes such as BMI after chemotherapy and only kept studies with European ancestry. Then we extracted all LD SNPs within 500kb of the GWAS SNP with r2>0.8 in 1000 Genomes Phase 3 data to calculate variant enrichment in different annotations categories. The R package VSE was used (https://cran.r-project.org/web/packages/VSE/vignettes/my-vignette.html). We found that   * Adding more associated annotations significantly improved the GWAS SNP enrichment (Distal+Proximal+CDS > Proximal+CDS> CDS). * Tissue specific annotations work better then annotations from distant cell types (for breast cancer MCF-7 > K562, and for leukemia K562 > MCF7) |
| Excerpt 5.6-B (extended gene in expression analysis) | For a given gene, separated patients into groups with or without mutations in certain annotations, such as CDS, UTR, TF/RBP binding sites, enhancers, and our extended gene. We then tested differences in gene expression (FPKM) between groups based on a two-sided Wilcoxon rank sum test. We found that our extended gene annotation provides better expression separation between these groups. Specifically, we found a well-known splicing factor SRSF2, which has been recently reported contribute to liver cancer development \cite{28082404}, gives the strongest p-value for stratifying expression out of all genes in liver cancer.    JZ2JL: please update using DL’s new figure |
|  |  |

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

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

<ASSIGN>@@@JZ,@@@MTG

<PLAN>&&&AgreeFix

<STATUS>%%%95DONE

|  |  |
| --- | --- |
| Referee  Comment | 6. The authors claim that reduction of functional elements increases power to discover recurrently mutated elements. This point needs quantitative support in the main manuscript (some analysis is given in the supplemental).  **For example, in the enhancer list derived from the ensemble method, what fraction of enhancers are estimated to be false positives?** |
| Author  Response | Reducing the number of elements cuts the multi. Test burden  We have made this clearer in the main text by some text from suppl.  We have also made clearer taht we have reduced the number of FP annotation by stringent quality controls. As the ref. Suggests we have tried to quant. The number of FP.  we have revised our manuscript to discuss the quality of annotations, including:   * Excerpt 5.7-A * Enhancer-gene linkages (Excerpt 5.8-A) * TF regulatory networks (Excerpt 5.14-A,B,C )   Enhancer prediction part, we have summerized part of our the method after our initial submisison into a manuscript avaialble from bioarcv . [[2 or 3 sentences w/ numbers on perform. Add a figure ]]  .  Specifically, we have systematically assessed CASPER/Matched Filter annotations. The method of CASPER/Matched Filter is described in detail in another paper under review in Nature Methods. In there the method is extensively evaluated in drosophila cell lines, mouse tissues and human cell lines. For example, we looked at the area under the ROC curve (AUROC) and PR curve for predictions in six different tissues in mouse. The average AUROC is around 0.8 and the average AUPR is around 0.37. We also compare our final compact annotation set to the other published annotations using FANTOM enhancer dataset, the details of which is presented below. |
| Excerpt 5.7-A (enhancer QC) | With the ensemble method, we can get more accurate annotation and pin-point to sequences where transcription factors would bind to. To estimate the false positive rate is challenging as there is no gold-standard experiment that could assert that a predicted enhancer is negative.  Here we took the FANTOM enhancer dataset and assessed the overlap percentage of our enhancer annotation in each ensemble step. We showed that each ensemble step indeed increases the percentage of overlap between our annotation and the FANTOM enhancer set. The overlap percentage for our annotation is higher than that of the Roadmap annotation and is also higher than the main encyclopedia enhancer annotation (ccRE). |

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

<TYPE>$$$Annotation,$$$Text

<ASSIGN>@@@KevinYip,@@@SKL,@@@GG,@@@DC [[DC to fix text]]

<PLAN>&&&MORE

<STATUS>%%%95DONE -

|  |  |
| --- | --- |
| Referee  Comment | 7. The authors claim superior quality of gene-enhancer links and gene communities derived from their machine learning approach. The method should at least be outlined in the main text, and accompanied by data supporting its accuracy and better performance compared to existing approaches. |
| Author  Response | We have outlined the linkage and community methods in the main more as suggested.  We have  Added data supporting its accuracy and better performance in the supp.. in particualr   * Adding a section to the supplement to show our JEME+Hi-C enhancer targets are better than the chromHMM ones (Excerpt 5.8-A) * Adding a comparison of our gene community method with others such as NMF showing that our method improves preservation of the original data structure of ChIP-seq experiments (Excerpt 5.8-B) |
| Excerpt 5.8-A (QC of enhancer-gene linkage) | Previously, we developed a computational approach JEME to predict enhancer-gene linkages. We have done extensive benchmark against other methods, such as IM-PET, Prestige, and Targetfinder. Details can be found in \cite JEME.  [[...]]sing the significant Hi-C interactions that are found using the method FitHiC (ref Fithic). This 2-step filtering provides confidence that our enhancer-target gene linkages are likely to have physical interactions between them.  To show how our JEME+Hi-C approach captures more accurate enhancer-gene linkages compared to existing linkages, we used published chromHMM derived enhancer-gene linkages (cite chromhmm) as the comparison dataset and GTEx whole blood eQTLs as the benchmark. We found the linkages, which the enhancer has an eQTL that changes the expression of the target gene significantly. After finding all the eQTL supported linkages for chromHMM and JEME+Hi-C, we calculated the fraction of enhancer-gene linkages that has eQTL support for various types of linkages in chromHMM and in JEME+Hi-C. As can be seen in figure below, JEME+Hi-C has higher fraction overlapped with eQTL-gene linkages.  Figure R X. Overlapping the gene-target linkages with GTEx eQTLs. |
| Excerpt 5.8-B (gene community method comparison) | [[PDM2all: if we could make this text a bit more concise it could communicate the point a bit more clearly.]]  Mixed membership model is a hierarchical Bayesian topic model framework and can help to uncover the underlying semantic structure of a document collection. The core of topic models is Latent Dirichlet Allocation (LDA), which cast the mixed-membership (topics) problem into a hidden variable model of documents. The LDA model has been widely used to analyze a wide variety of data types, including but not limited to text and document data, genotype data, survey and voting data. The advantage of LDA over other algorithms (like SVD, PLSI) used in semantic analysis has been described in Blei 2003. In particular, this paper mentioned that LDA allow document to belong to multiple topics simultaneously, and the topic mixture weight was treated as k-hidden random variable to reduce overfitting problem rather than a set of individual parameters that explicitly link to the training set.  With regards to the referee’s question, there is no ready-made answers since the data type (TF target network) and problem-definition of our study are both specific. Fundamentally the LDA method is an unsupervised method, therefore there are no labels on the dataset and accuracy metrics are not applicable. If we treat the LDA mixed-membership analysis as a dimensionality reduction problem, it is possible to compare how well of a model can reproduce the information of original data, as described in paper (Guo, Y., & Gifford, D. K. (2017). Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. BMC Genomics, 18(1), 45.). The correlations of the original target gene vectors between two TFs are compared with those of dimension reduced vectors. The better method should be much close to original vectors correlations.  To explore how well the LDA mixed-membership analysis on TF regulatory network, we extend our dataset from 122 GM12878 and K562 samples to all the 862 TF ChIP-seq assays included in ENCODE data portal. In order to get a reliable correlation, we also increase the number of topic to 50 as the number of TF sample increases. The non-negative matrix factorization (NMF) and Kmeans clustering are used for comparison because the nature of regulatory network requires a non-negative decomposition. The same target dimension K =50 was used to NMF and target number of clusters K=50 for Kmeans. The Euclidean distance between each data the centroids are used to calculated 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. |

## 

## <ID>REF5.9 – What data sets are used

<TYPE>$$$BMR

<ASSIGN>@@@JZ,@@@WM

<PLAN>&&&Defer - WM to make the table

<STATUS>%%%75DONE

|  |  |
| --- | --- |
| Referee  Comment | 8. From the main manuscript, it is not clear which cancer data sets were analyzed with the new background mutation rate estimates and functional regions. Datasets and sample size should be mentioned explicitly. |
| Author  Response | We thank the referee for bringing out this point. We provide it here in the table and summarized it in a line in the main text.  [[List the DBGAP accessions]] |
| Excerpt 5.9-A | Wait for the main text  JZ2JZ |

## <ID>REF5.10 – Mutational signatures

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

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%85DONE

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

## <ID>REF5.11 – Additional QQ plots

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

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

|  |  |
| --- | --- |
| Referee  Comment | 10. The significance analysis of cancer cohorts (Figure 2) should highlight known cancer genes versus those newly found in this study. A QQ-plot should be included to confirm that the algorithm accurately models the background expectation. |
| Author  Response | We thank the reviewers for pointing this out. We have updated Fig. 2 to label the known cancer genes (Except 5.11-A).  We have provided the Q-Q plot in the supplementary file in our initial submission and we have extracted some of Q-Q plot in the excerpt below. The Q-Q plot below indicates no obvious P value inflation, which indicates our BMR estimation is should be OK. |
| Excerpt 5.11-A (updated Fig.2) | Figure 2 |
| Excerpt 5.11-B (in suppl.) | Q-Q plot for breast cancer on various annotations. |

## <ID>REF5.12 – Sequence coverage

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

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%95DONE

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

## <ID>REF5.13 – BCL6 Questions

<TYPE>$$$Annotation,$$$Calc

<ASSIGN>@@@XK,@@@TG,@@@PDM

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

[JZ2JZ: more investigations]

JZ2MG: wait, not yet updated

|  |  |
| --- | --- |
| Referee  Comment | 11. The authors mention that BCL6 would have been missed in an exclusively coding analysis. In which part of the extended annotations were recurrent BCL6 mutations found? If near the promoter, is the BCL6 5’ region a known AID off-target? Are BCL6 mutations in CLL associated with translocations? |
| Author  Response | JZ2JZ: check  We thank the referee for this comment. As suggested, we found that the there is a mutation hotspot near the first intron of BCL6.  It's hyper mut  Not sure if target or not |
| Excerpt 5.16-A |  |

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

<TYPE>$$$Network,$$$Calc

<ASSIGN>@@@Peng,@@@JZ,@@@DL,@@@PDM [[PDM]], DL

<PLAN> &&&AgreeFix

<STATUS>%%%95DONE

|  |  |
| --- | --- |
| Referee  Comment | 12. The manuscript notes that the new networks presented contain “more accurate and experimentally based” gene links. This claim should be supported with comparisons with existing networks and statistical evaluation. How many of the derived networks are false positives? How many networks are derived in total? |
| Author  Response | [[\*\*Rework point counterpoint]]We thank the referee for bringing this point up, and we believe that this is one of the core strength of ENCODE. We also feel that it is important to make comparisons with existing networks with more statistical evaluation. We have made the following revisions in the updated manuscript.  **1. Regarding the proximal regulatory element network:**  (Excerpt 5.14-A). -  *Comparison with Biogrid and String:* our networks can capture a higher fraction of gold-standard, literature supported interactions than networks such as Biogrid and String  *\* Comparison with DHS-based imputed networks:* our networks provide better correlation with gene expression following TF knockdown than the DHS-based imputed network provided in Neph et. al. 2012 (Excerpt 5.14-B).  *\* False positive rate:* ENCODE enforces strict data quality standards for all ENCODE produced ChIP-seq experiments, which allows for rigorous false positive control (Excerpt 5.14-C).  **2. Regarding the distal regulatory element network:**  With the ChIP-seq, DNase-seq, STARR-seq, ChIA-PET, and Hi-C experiment, ENCODE has a distal TF-enhancer-gene network of high quality. We feel this is a unique aspects of our resource.  *2.1 High quality of integrative enhancer definitions:* (Excerpt 5.7-A).  *2.2 High quality of enhancer-gene linkages*: (Excerpt 5.8-A). |
| Excerpt 5.14-A (comparison with Biogrid and String network) | To evaluate the quality of ENCODE transcriptional regulatory networks, we utilized the TRRUST database, which manually curated transcriptional regulations from Pubmed articles (Han et al., 2018). We defined the TRRUST interactions as the gold standard and tested the fraction of gold standard interactions that other networks can recapitulate. The ENCODE network captures a higher fraction of these benchmark interactions than protein physical networks, including Biogrid and String experimental interactions (Supplementary Figure X). Moreover, the fraction of gold standard network interactions that ENCODE network recapitulated is consistently higher than random. These results supported the higher relevance of ENCODE networks on transcriptional regulation compared to other networks. We also constructed another post-transcriptional network between RBPs and target genes through linking the RBP binding sites on gene 3’UTR regions. To the best of our knowledge, the current study is the first one to study RBP-gene interactions systematically; thus we are not aware of any previous resources that can provide gold standard regulations for comparison.    **Supplementary Figure X. ENCODE networks captured a higher fraction of curated regulations than other networks.** The TRRUST database manually curated 8,412 transcriptional regulatory interactions from Pubmed articles (Han et al., 2018). We computed the fractions of TTRUST interactions that other networks can recapitulate. Since each ENCODE ChIP-seq interaction has a regulatory potential (RP) score, we showed the fractions with different RP thresholds. The random fraction for ENCODE network was estimated through 100 perturbed TTRUST networks using the stub-rewiring method that preserved the gene network degrees (Milo et al., 2002). |
| Excerpt 5.14-B (comparison with imputed network) | Our new regulatory network edges are derived from ENCODE TF ChIP-seq experiments, and they provide more accurate gene linkages than imputed networks from other genomic features. To demonstrate the superiority of our new network, we have evaluated our experimentally derived ChIP-seq networks with DHS-based imputed networks from previous publications. We have used two types of ChIP-seq networks. The first one is based on proximity to TSS and the second one based on target identification from profiles (TIP) method. For the imputed network, we used Neph et. al. 2012 (Neph, Shane, et al. "Circuitry and dynamics of human transcription factor regulatory networks." Cell 150.6 (2012): 1274-1286.) TF-to-TF network imputed from DNase I hypersensitive footprints. In addition to the Neph et. al. DHS network, we also built our own version of a similar DHS network, utilizing the ENCODE DNase-seq dataset. To test the gene linkages, we utilized ENCODE RNAi based TF knockdown and CRISPR-based TF knockout datasets to test how target gene linkages defined by various network definitions are affected by knockdown. Overall, target genes of ENCODE ChIP-seq networks had larger differential expression after knockdown (Supplementary figure X). Moreover, DHS-imputed network derived from ENCODE DNase-seq performed better than the previously published method (not shown here, available in Supplementary document).  Supplementary figure X. Comparison of the ENCODEC network with previously published regulatory network using ENCODE CRISPRi knockdown data. Target genes of ENCODEC ChIP-seq based networks have larger expression differential after knocking down. Examples of RFX5, SP2, and USF2 shown. More details with full figures comparing all variants of ENCODEC networks can be found in supplementary document.        1 |
| Excerpt 5.14-C (false positives) | [[PDM2all: this sounds like a commentary specifically on the quality control standards of ChIP-seq experiments in ENCODE3. Not sure how this directly relates to the reviewer’s question.]]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 were used to compile ENCODEC resources that have more than 2 replicates, which allows further quality control of the derived network. ENCODE used an 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 less than 2.    After extensive quality controls for the concordance between replicates, peaks are called using macs2 {"Zhang et al. Model-based Analysis of ChIP-seq ([MACS](http://liulab.dfci.harvard.edu/MACS/index.html)). *Genome Biol* (2008) vol. 9 (9) pp. R137"} with p-value cutoff of 0.01. |

## <ID>REF5.15 – MYC KD Validation

<TYPE>$$$Network,$$$Text

<ASSIGN>@@@DC

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

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

## <ID>REF5.16 – SUB1 analysis

<TYPE>$$$NoveltyPos,$$$Calc

<ASSIGN>@@@MRS,@@@JL,@@@YY,@PDM [[PDM]]

<PLAN>&&&MORE

<STATUS>%%%95DONE

|  |  |
| --- | --- |
| Referee  Comment | 14. SUB1 is a potentially interesting new cancer gene. The authors should further explore the biology of this gene. |
| Author  Response | We thank the referee for this comment about SUB1, and also the related previous comment about MYC. This spurred us to really think about the biology of these key factors.  We foulllowing facts about the biology of the gene  We found out that SUB-1 actually has a plausible biological function in relation to cancer. We were able to figure out how it collaborates with other regulators, such as MYC, to demonstrate how our multi-networks, including the TF and RBP networks, potentially relate to underlying biology.  We updated Fig 3 by adding our new small-scale validation experiment to drill into the SUB-1 MYC connection and validate it partially on several important oncogenes. |
| Excerpt 5.16-A (SUB1’s enrichment in 3’UTR, in suppl.) | **[[PDM2all: See earlier comments on these figures/captions responding to ref 3.8A. However, these figures seem more suited to this reviewer’s comments (5.16) than for that response (3.8), so not all comments will apply.]]Supplementary Figure X: eCLIP peaks of SUB1.** (a) The composition of SUB1 peaks over different gene regions is shown for each replicate. (b) For each gene region, the relative enrichment (fraction of SUB1 peaks / fraction of all peaks) of SUB1 peaks is shown. (c) The distribution of SUB1 peaks over 3’UTR regions is shown. The mean across all RNA binding proteins profiled by eCLIP experiments are shown as background with standard deviation as error bars. |
| Excerpt 5.16-B (SUB1 target on famous oncogenes, in suppl.) | We found that SUB1 targets are enriched in cancer associated genes, such as genes in Cancer Gene Census (P=1.8e-16 by Fisher’s exact test), and such genes showed larger down regulation upon SUB1 knockdowns. Among many of such genes, we have shown some IGV examples together with SUB1 binding sites on the 3’ UTRs. |
| Excerpt 5.16-C (SUB1’s regulatory potential in different cancer types, in suppl.) | Using ENCODE eCLIP data and TCGA tumor profiles, we applied RABIT framework to identify RNA binding proteins (RBP), whose target genes are differentially regulated in cancer. (a) The fractions of patients with target genes up or down regulated are shown for each combination of RBP and cancer type. (b) The patient fractions with target genes differentially regulated are shown for all cancer types and RBPs whose fraction values are larger than 50% in at least one cancer. (c) All lung adenocarcinoma patients are divided to two groups according to SUB1 activity predicted by RABIT. The overall survival was shown by KM plot. The association between SUB1 activity and survival was tested through Cox-PH regression. (d) In the left panel, the cumulative distributions of gene expression after SUB1 knock down in HepG2 cell are shown for predicted SUB1 targets and none targets. In the right panel, the cumulative distributions of mRNA decay rates in HepG2 cell are shown. The comparison between two categories is done through Wilcoxon rank-sum test. |
| Excerpt 5.16-D (SUB1-MYC co-regulation in suppl.) | Among genes whose 3’UTR regions have *SUB1* eCLIP sites, we observed significant enrichment of functional categories including *MYC* targets and spliceosome. *MYC* activation induces an increase in total precursor messenger RNA synthesis, which increases the burden on the core spliceosome to process pre-mRNA 1. Also, *MYC* activation can stimulate oxidative phosphorylation, which fulfills the bio-energetic demands of cancer cells 2. These results together indicate that *SUB1* may stabilize the *MYC* target genes and pathways to promote the malignant growth of cancer cells. |
| Excerpt 5.16-E (SUB1-MYC co-regulation validation, in suppl.) | To detect predicted common target gene of MYC and SUB1, shRNA plasmids containing 4 targets sites of each gene were used to transfected to HepG2 cell using Lipofectamine(™) 3000 following the manufacturer's instructions (Invitrogen) (target sites for each gene are listed in Sup table 1). Briefly, 0.12 M HepG2 cells were seeded in each well of one 24-well plates 24 hours before transfection. 500 ng plasmids containing either single shRNA or 4 shRNA plasmids as pool were mixed with 0.75 uL Lipofectamine(™) 3000 in Opti-MEM I medium (Invitrogen) and loaded to HepG2 cells in each well. Blank plasmids without shRNA target sequence was used as control. To improve transfection efficiency, 2 ug/mL puromycin was used to select successful transfected cells. 72 hours after transfection, total RNA was extracted using RNeasy Mini Kit (Qiagen) and followed by cDNA generation using SuperScript III (Invitrogen). Knockdown efficiency and target gene expression level were quantified and compared to BACTIN by qPCR using KAPA SYBR® FAST qPCR Master Mix (2X) Kit (Sigma). The qPCR primers were listed in Sup table X. |
| Excerpt 5.16-F (New Fig. 3) | New Figure 3, JZ2DL please add |

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

<TYPE>$$$Network,$$$Calc

<ASSIGN>@@@DL,@@@JL

<PLAN>&&&AgreeFix

<STATUS>%%%99DONE

|  |  |
| --- | --- |
| Referee  Comment | 15. The manuscript claims that transcription factors placed at the top level of the network hierarchy are enriched in cancer-associated genes and drive expression changes. Both claims need to be supported with statistical tests. |
| Author  Response | DL2JZ: can you fill in XXX below with the actual p-value from HierNet analysis? I tried to look up from old data, but I couldn’t find exact pvals. Also could you add some descriptions to supplementary figures?  Mark’s dictation:  Wheat we actually had tried to put him's assistive significance into  the original network Harkey figure using Asterix however unfortunately  we don't we don't think this was clear Disney's people we were it to  be better written in the river in the revised main script we have I  tried to make this clear in the figure caption figure I see the  excerpts from other captions for figure 3 and figure four below also  we have the original I data at showing Cisco swimming it's now as a  supplementary figure shown below as well basically good  [[PDM2all: Replace with ‘We have updated our analyses to clearly convey the results of statistical testing’. Dangerous to point out the reviewers error in missing the results before -- prompts complaints about legibility, structure of the supplement, relation of supplement to main text, etc.]]  The right panel of Figure 4 shows results from Wilcoxon signed-rank test. If a p-value is less than 0.05 it is flagged with one star (\*). If a p-value is less than 0.01 it is flagged with two stars (\*\*). If a p-value is less than 0.001 it is flagged with three stars (\*\*\*). We find that the top-level of the generalized network was enriched with cancer-related TFs with p-value XXX and had larger correlation to drive target gene expression change (p-value XXX). |
| Excerpt 5.17-A (in suppl.) | Supplementary Figure X. |

## 

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

<TYPE>$$$Network,$$$Calc

<ASSIGN>@@@DL

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

|  |  |
| --- | --- |
| Referee  Comment | 16. In the tumor-normal network comparison, is the fraction of edge changes related to the total number of edges for a given TF? This analysis should further clearly state its null hypothesis (what changes are expected?). What happens when edges are randomly permuted?  [JZ2MG: we did not directly answer this question] |
| Author  Response | We thank the referee for pointing out this issue.    In short, we would like to clarify that the fraction edge changes is based on the fraction of regulatory edge changes between two cellular contexts.  We have added analyses in the revised supplement to estimate false positive rates of rewiring. |
| Excerpt 5.18-A (in suppl.) | … The rewiring index is then normalized across all regulatory proteins, and the sign reflects the direction of rewiring. Details of rScore derivation can be found in Supplementary 5.3. Given this, we assume a null hypothesis to be no change in regulatory edge across cell types. We expect no or minimal change in edges when two cellular contexts are similar. To demonstrate, we selected all available GM12878 ChIP-seq experiments that have at least two replicates, and we then calculated the same rewiring index between isogenic replicates of the same cellular context. We expect very small rewiring score given they are the same cellular context, and the edge changes between two networks will be simply a noise from ChIP-seq experiments.  As expected, when two cellular context are similar, as shown in “baseline”, minimal number of edges do change targets. However, in “rewiring”, TF do change targets extensively when compared across cancerous (K562) to normal (GM12878) cell lines. To put this into perspective, we calculated the fraction of regulatory edges that are due to noise. We estimate that, on average, 1.36% of observed regulatory edges could be false positives.  Supplementary Figure X1.    Supplementary Figure X2. |

## <ID>REF5.19 – Stemness rewiring analysis

<TYPE>$$$Stemness,$$$Calc

<ASSIGN>@@@DL,@@@JZ [DL, Jl}

<PLAN>&&&AgreeFix

<STATUS>%%%25DONE

|  |  |
| --- | --- |
| Referee  Comment | 17. The network change comparisons with the H1 stem cell models need statistical testing for significance. What fraction of the rewired edges are expected to be false positives? |
| Author  Response | We thank the referee for pointing this out. We agree with the referee’s suggestion and took this opportunity to significantly expand the statistical  aspects of rewiring and stemness analysis, which includes  *3. Regarding the new stemness analysis using PCA/RCA*: We ran Wilcoxon test to compare the tumor-stem and normal-stem distance (Excerpt 5.19-B,C) and found that tumor cells are more similar to stem cells, which is consistent with other findings \cite(TCGA i stemness).  *1. Regarding the false positives of the rewired edges:* approximately 1.36% of rewired regulatory edges are false positives (Excerpt 5.18-A).  *2. Regarding the statistical testing in the normal-tumor-stem analysis:* a section in the supplementary file on our original rewiring analysis (Excerpt 5.19-A) |
| Excerpt 5.19-A (in suppl.) | The H1 stem cell model uses fractional overlap of rewired edges between cancerous cell types vs. H1. Therefore we attempted to evaluate statistical significance of our model by measuring how much of H1 network changes are due to noise and use of other normal cell types to evaluate how much of rewired edges overlaps with H1.    Using replicates of H1-hESC ChIP-seq experiments, we made two independent H1 networks in addition to original replicate merged H1 network, and we made recalculated stemness of TF, whether they rewire toward or away from H1. We find that the results of all of stemness direction is reproduced using either replicate. |
| Excerpt 5.19-B (stemness in suppl.) | We performed PCA (RCA) analysis on RNA-seq, RNAi and CRISPR-based knockdown, and TF ChIP-seq data to find evidence for the hypothesis that clusters of cancerous cell types de-differentiate to a state that resemble more like stem-like cell types. We consistently found using different types of data that cancer cells’ regulatory status as well as gene expression profiles are closer in Euclidean distance to the stem state as compared to their primary cells of origin (Figure 5). We quantified and compared the L2 distance to stem-like clusters between cancerous cell types and normal cell types. We find that using both proximal network and gene expression profiles have statistically significant difference between normal-to-stem and cancer-to-stem distance (using Wilcoxon rank sum test, Suppl. Fig. A-B). We found observable difference in distal regulatory network but found no statistical significance.    Suppl. Fig. B    Suppl. Fig. C |
| Excerpt 5.19-C (stemness in Fig.5) | Figure 5. Proximal regulatory network, distal enhancer network, and gene expression profiles have been used to explore patterns across different cell types. As expected, stem-like cell types formed a cluster, suggesting stem-like cell types have a distinct regulatory profile from normal and cancerous cell types, and stem-like cells including H1 and iPSCs have similar regulatory patterns. We find that cancerous cell types have closer distance to a state closer to stem-like clusters, suggesting cancer cells de-differentiate to a stem-like state both in their regulatory programs and gene expression profiles. |

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

<TYPE>$$$Validation,$$$Text

<ASSIGN>@@@JZ,@@@DL

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

|  |  |
| --- | --- |
| 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 | The eight regions were selected from our integrative promoter and enhancer regulatory elements in MCF-7 cell lines. We prioritized these regulatory regions based on our integrative, stepwise variant prioritization as described in section 6.1 S. We have tried to make it more clear about the details of locations, surrounding genes, replicates and p values (Excerpt 5.20-A and Excerpt 5.20-B). |
| Excerpt 5.20-A (selection, replicate, and pvalues, in suppl.) | We selected the top ten regions from our proposed prioritization step and then tested their regulatory activities using a luciferase assay as described in section 6.2 S. Two of ten regions we tested failed due to issues with plasmid isolation. There were two biological replicates and three technical replicates for each biological replicate in designing luciferase assays validations. Error bars represent the 95% confidence interval across replicates. |
| Excerpt 5.20-B (in suppl.) | We have provided details for the surrounding genes and genomic features of all tested regions as below. [[PDM: Maybe so, but the plots are very difficult to read and interpret (size + labels). ]] |

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

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

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

|  |  |
| --- | --- |
| Referee  Comment | 19. The authors should consider moving the general overview diagrams that constitute much of the main figures to the supplement, and in turn present data-rich figures from there with the main manuscript. |
| Author  Response | We thank the referee for this comment. We have tried to revise the figures as requested. We have reduced the schematics in fig 1 and moved We have fixed figures 1 and xxx.  We have tried to rework our figures  [[PDM2all: Did we move anything from the maintext to the supplement, or from the supplement to the maintext? If so, we should say this.]] |
| Excerpt 5.21-A |  |

## 

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

<TYPE>$$$Validation,$$$Text

<ASSIGN>@@@JZ

<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 | The method that we used here is distinct from FunSeq.[PDM2all: PDM suggests drawing out more of the difference between our prioritization and FunSeq. Right now we don’t distinguish the ENCODEC prioritization from FunSeq at all (except implicitly).]] 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 withvalidation that we believe points to the great value of the integrated data.  [[use the response from ref 2]] |

## 

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

<TYPE>$$$Minor,$$$BMR

<ASSIGN>@@@JZ

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

|  |  |
| --- | --- |
| Referee  Comment | 21. When the authors describe recurrent events, are these significant? If so, please provide p-values (and q-values, when applicable). |
| Author  Response | We thank the referee to point this out. We have the values and q-values all deposited into our online resource and supplementary files. We have made this clearer in our revised manuscript. |
| Excerpt 5.23-A (in main text) | We have plotted the heatmap of p values for the recurrent analysis in three different cancer types. |

## 

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

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

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

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

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

<TYPE>$$$Minor,$$$CellLine

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

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

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

<TYPE>$$$Minor,$$$Stemness

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%50DONE

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

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

<TYPE>$$$Minor,$$$Validation

<ASSIGN>@@@DL

<PLAN>&&&AgreeFix

<STATUS>%%%90DONE

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

## 

## <ID>REF5.28 – Minor: SYCP2 and beyond

<TYPE>$$$Minor,$$$NoveltyPos

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

[JZ2JL: can you please do this quickly?]

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

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

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

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%TBC

[JZ2MG: is it OK for the text?]

|  |  |
| --- | --- |
| Referee  Comment | 27. The abstract mentions the usefulness of ENCODE data for interpretation of non-coding recurrent variants, yet this point is not explored much in the manuscript. |
| Author  Response | We thank the referee for this comment. Actually, we tried to show in Fig 6 how each data type has been integrated to evaluate the function of non-coding 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 5.29-A | Wait for abs |

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

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

<ASSIGN>@@@DL

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

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

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

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

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

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

## 

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

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

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%100DONE

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

## <ID>REF5.33 – Minor: Presentation

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

<ASSIGN>@@@DL

<PLAN>&&&AgreeFix

<STATUS>%%%95DONE

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

## <ID>REF5.34 – Minor: Supplementary document

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

<ASSIGN>

<PLAN>&&&AgreeFix

<STATUS>%%%75DONE

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