HIGHLIGHTS IN YELLOW ARE CURRENTLY IN PROGRESS

-- Ref1.0 – General comments –

Reviewer' comment:

	0 - Neither the software nor a test instance was available for review.

Author's response:

-- Ref1.1 – General comments –

Reviewer' comment:

	1 - The abstract is vague. In my view, the authors lose a critical opportunity by not reporting the significance of previously studied cases of genetic variants that affect RBP function or how their new method can help to sort the important genetic variants from the rest (DNA vs RNA).

Author's response:

JL2JZ&MG: need to discuss this comment and make changes to abstract. Do we need more analysis?

We thank the reviewer for pointing this out. We agree that it should be further emphasized how genetic variants affecting RBP function are an important part of studying disease. To this end, we have revised our abstract to reflect how our method, RADAR, explores mutations in the RBP regulome and how they can be separated from mutations affecting DNA.

-- Ref1.2 – General comments –

Reviewer' comment:

	2 - What is the rational to only show comparison among RADAR, FunSeq2 and CADD? See for example, https://www.ncbi.nlm.nih.gov/pubmed/29340599 (A benchmark study of scoring methods for non-coding mutations). Please motivate your choice.

Author's response:

[JZ2JZ: need to add this FATHMM-MKL later]

The reviewer makes a valid point. To address this concern, we include two points on the motivation for comparing to FunSeq2 and CADD below. In addition, we also include a further comparison to FATHMM-MKL.

- RADAR vs FunSeq2: the idea to use Shannon entropy in our scoring scheme for RADAR is adapted from FunSeq2. However, FunSeq2 primarily uses annotations such as TFBS, DHS, enhancer, etc. while RADAR focuses on the RBP regulome. We believe that RADAR is therefore complementary to FunSeq2 and that by comparing them, we can see that RADAR is able to pinpoint variants not prioritized by FunSeq2 (two examples in the manuscript).

- RADAR vs CADD: we make the comparison between these two scores in order to show that while RADAR does consider coding regions intersecting the RBP regulome, it also considers many of the noncoding regions no prioritized by CADD.

-- Ref1.3 – General comments –

Reviewer' comment:

	3 - The relevance of RBPs on RNA splicing is not considered at all.

Author's response:

We thank the reviewer for making the suggestion to include RNA splicing considerations in our manuscript. However, there are xxx splicing factors inside the RBP list already. We want to further highlight several points here:

To this end, we have:

1 We Included a table in our supplement (shown below) categorizing each RBP by their function, many of which are splicing related.

· Included a download link on our website of eCLIP data annotated by each RBPs specific function, which can easily be filtered for splicing related RBPs.

· Updated Figure 2, with a heatmap showing the clustering of RBPs by the similarity of their target gene sets. Splicing and non-splicing RBPs are seen to form clusters, perhaps due to functional similarity. This heatmap is shown below in Figure R.XXX.

· Included a motif similarity diagram between RBPs of splicing functionality (Figure R.XXX).

Figure R.XXX. *Similarity of target genes of splicing and non-splicing RBPs.*

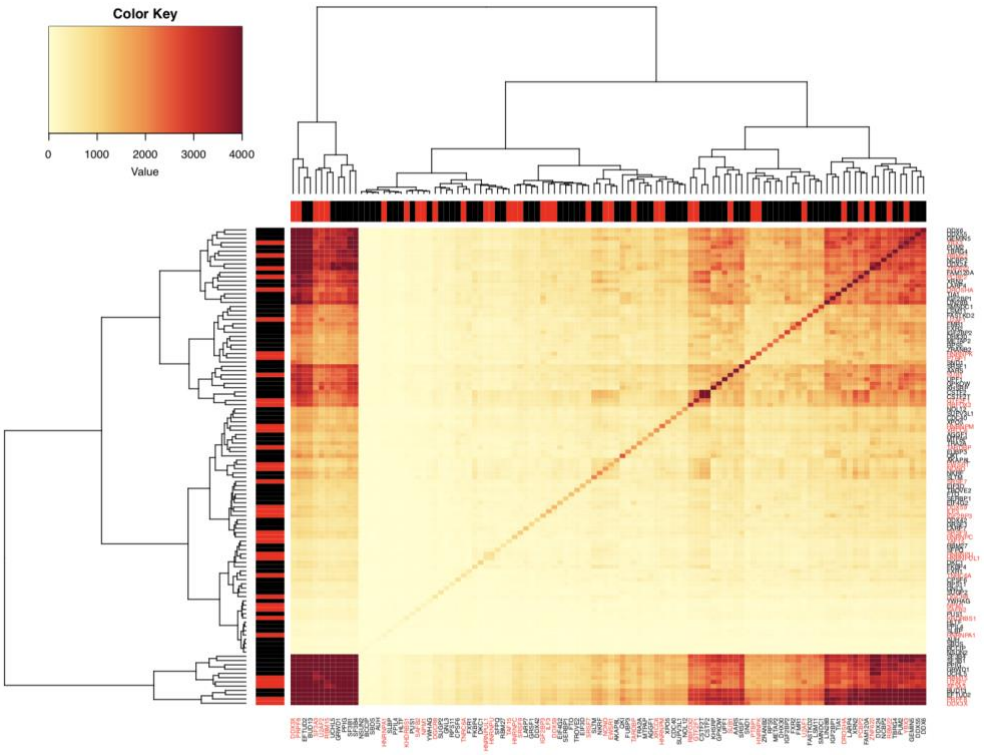
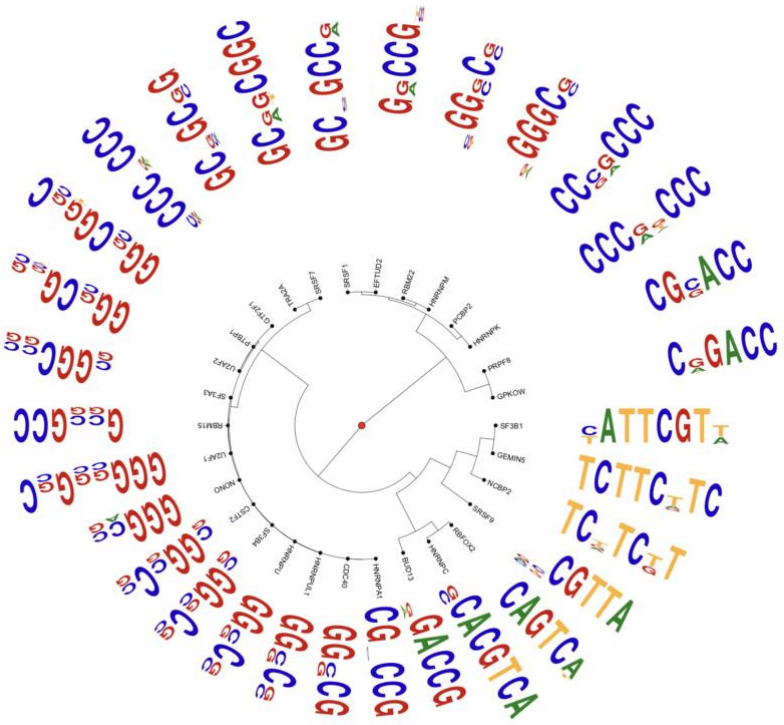


Figure R.XXX. *Motif similarity between different splicing RBPs.*



-- Ref1.4 – General comments –

Reviewer' comment:

	4- The basic and tissue-specific scoring is not well explained. The method section is mixed with results (eg. In Regulatory Power from Linear Regression). Please separate results from methods. I would like to see a clear presentation on how a RADAR score is computed for a given variant from basic and user-specific contributions in mathematical terms.

Author's response:

We thank the reviewer for this suggestion. We have restructured our manuscript to remove any results in the methods section. The methods section now contains specific details on scoring a variant for each component of the score (6 basic, 3 user-specific). Equations used in each part of the score have been carefully added to the appropriate sections.

To further clarify the scoring of a given variant, we also provide a flowchart for scoring, shown below in Figure R.XXX.

Figure R.XXX. *Schematic flowchart of scoring a variant.*

-- Ref1.5 – General comments –

Reviewer' comment:

	5 - Please assess the individual relevance of the features listed in Table 1 for RADAR. Especially, the data types that are not modelled by the preceding software FunSeq2 (see Figure 1).

Author's response:

In our model we use a total of 6 features in the baseline score and 3 features in the user-specific score. Of the baseline features, five of them are different. All three tissue specific features are unique to RADAR and not included in FunSeq2. The overarching difference of RADAR in comparison to other variant prioritization methods is that RADAR considers only regions intersected by eCLIP peaks. In FunSeq, the main goal is to prioritize noncoding regions associated with TFBS, DHS, and enhancer regions. However, 47% of eCLIP peaks are do not intersect those regions. So while funseq may prioritize variants associated with transcriptional regulation, RADAR focuses on variants associated with the post-transcriptional regulation.

More specifically, our features also differ technically from those in FunSeq. A summary of these points is below:

Although we share a similar entropy scoring scheme to FunSeq2 for some of our features, RADAR improves on the methodology through its correction of GC bias as well as its focus on the RBP

regulome. Also, we include RNA structure as a feature, which considers the 3D structure of binding. For our network hubs, we go a step further than previous models by showing its positive correlation with selection pressure, rather than simply making the assumption that network hubs correlate with functional significance. We also include knockdown data from shRNA-seq to show the important regulatory effects that RBP can have on genes.

One of the most important aspects that RADAR differs from many other methods, such as FunSeq2, is its incorporation of tissue specific components. In particular, the tissue components consider the differential expression of tumor to normal samples from TCGA, tissue specific mutational burden on eCLIP elements, and regression based methods to determine the regulatory effects of RBPs on its target genes in specific tissues. These data and features are not covered in FunSeq and we believe are what make our method, RADAR, unique.

-- Ref1.6 – General comments –

Reviewer' comment:

	6 - Please use the cell-line specific aspect of ENCODE to assess the performance of your method. I believe that cell-specific information for K562 and HepG2 cell lines are available, such as shRNA-seq, eCLIP. Variant information might be also available for both cell lines as I have seen whole genome sequencing data in NCBI's SRA. Please train / build the model on one cell type ("Baseline) and evaluate on the other ("specific component"). This could be as convincing as an experimental validation.

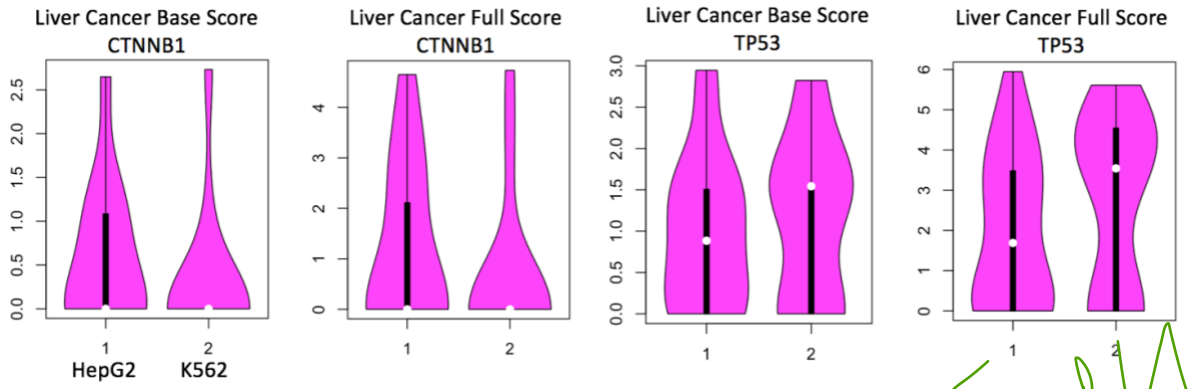
Author's response:

We have completed built the RADAR model using the two different cell specific data, creating a HepG2 and K562 score (baseline and tissue specific in each). We give two examples below to show how using cell type specific data could influence the RADAR score.

Example 1: JL working on new analysis here

Example 2: We compare the HepG2 and K562 scores for a set of Liver cancer variants available publicy from the Alexandrov et al paper. Here we see that variants that fall in CTNNB1, a well known cancer driver gene unique to liver cancer are scored much higher when using the HepG2 version of the score compared to the K562 version. As a control, we look at the scores of variants falling in TP53, a well known cancer driver, but not specific to liver cancer. The results are shown in Figure R.XXX below.

Figure R.XXX. *Difference in RADAR cell type specific score (HepG2 and K562) when scoring liver cancer variants in CTNNB1, a known driver gene unique to liver cancer, and in TP53, considered to be a driver in multiple cancer types.*



-- Ref2.0 – General comments –

Reviewer' comment:

	One major concern appears to be whether the observed results are reflective of true biology or simply artifacts of various algorithms. For example, figure 2 and lines 21-32 discuss the overlap between eCLIP peaks and annotations. However, the description of the CLIPper algorithm in Lovci et al (2013) used in the ENCODE pipeline suggests that clusters are identified only within transcripts, which would then trivially localize all eCLIP peaks to transcript annotations. Similarly, although the 'RBP regulome' appears smaller than that for TFs, it is unclear whether this is simply because the average peak size for eCLIP is significantly smaller than the average CHIP-seq peak due to differences in method and peak callers (likely, as most known RBP and TF motifs are of similar sizes).

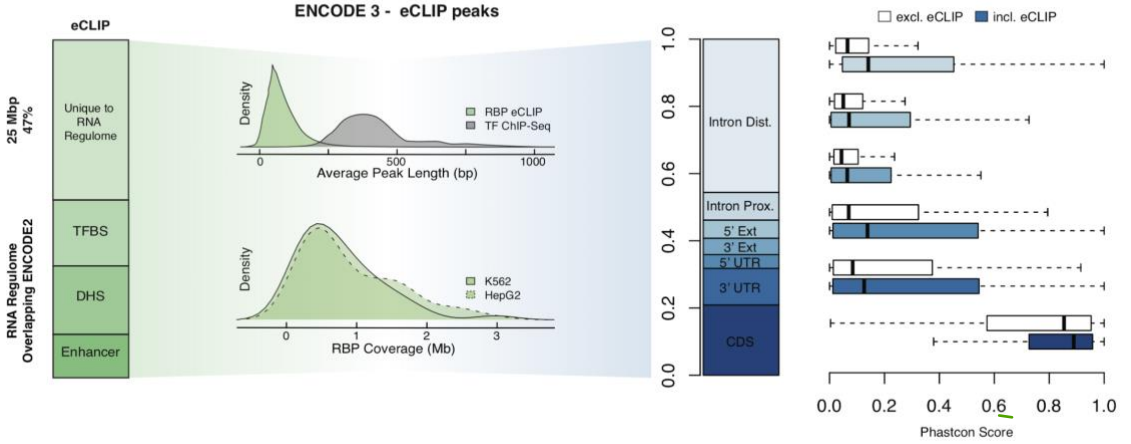
Author's response:

We thank the reviewer for the comments and address the reviewer in a point by point fashion below:

"For example, figure 2 and lines 21-32 discuss the overlap between eCLIP peaks and annotations. However, the description of the CLIPper algorithm in Lovci et al (2013) used in the ENCODE pipeline suggests that clusters are identified only within transcripts, which would then trivially localize all eCLIP peaks to transcript annotations."

We agree with the reviewer that the eCLIP peaks are in fact localized to the transcript regions. While the eCLIP peaks does show some overlap with previous transcript annotations such as TFBS, DHS, and enhancer regions, 47% of the eCLIP peak annotations do not intersect any of the previous ENCODE2 annotations and are unique to the RBP regulome. To illustrate this point better, we have modified our Figure 2 in the main figure pack, and extracted panel A, shown below as Figure R-2A.

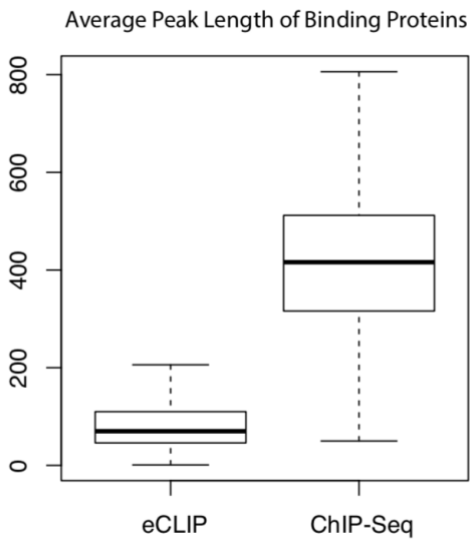
Figure R-2A. *Updated panel of Figure 2 showing eCLIP data as having a higher resolution than ChIP-Seq annotations, allowing for more accurate biological definitions of binding events.*



"Similarly, although the 'RBP regulome' appears smaller than that for TFs, it is unclear whether this is simply because the average peak size for eCLIP is significantly smaller than the average CHIP-seq peak due to differences in method and peak callers (likely, as most known RBP and TF motifs are of similar sizes)."

We also agree with the reviewer that the nature of eCLIP data is different than that of the TFs, a problem that stems from the differnce in eCLIP vs ChIP-Seq technologies. The resolution of ChIP-Seq is within +/- 300bp while eCLIP aims to give a single nucleotide resolution. Our updated Figure 2 shows that the average peak length of RBPs from eCLIP is much smaller than that of TFs from ChIP-Seq, and therefore provide a more accurate region of binding of RBPs. We also remove the comparison between the total coverage of TFs vs RBPs since we agree with the reviewer that it is an unfair comparison. Figure R.XXX below shows a boxplot comparing eCLIP and ChIP-Seq annotations.

Figure R.XXX. *Boxplots of relative average size of peaks (bp) between eCLIP and ChIP-Seq.*



"One major concern appears to be whether the observed results are reflective of true biology or simply artifacts of various algorithms."

We show now that RBP binding sites given by eCLIP are a higher resolution than TFBS, and therefore are more reflective of the true biology of binding proteins. In our analysis of scoring variants, we only rely on eCLIP annotations, and therefore minimize potential artifacts compared to scoring methods relying on annotations from ChIP-Seq. We believe one of RADAR's strengths is its usage of high quality eCLIP data to explore an accurate representation of the RBP regulome.

-- Ref2.1 – General comments –

Reviewer' comment:

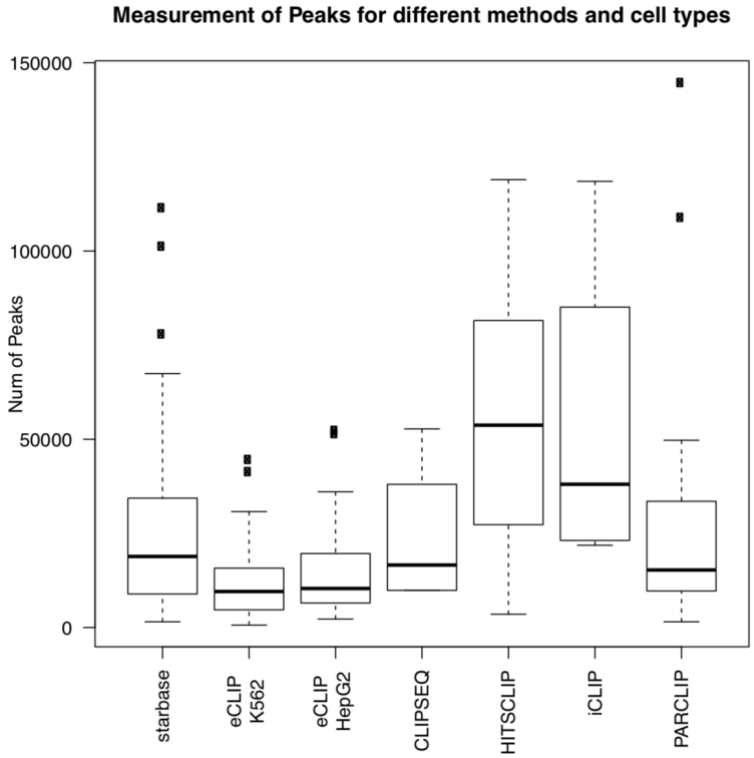
	One major question regards the weighting of eCLIP binding sites. The eCLIP data appears to contain not only narrow binding proteins, but also broad binding or coating proteins (such as POLR2G https://www.encodeproject.org/experiments/ENCSR820WHR/). Perhaps because of this, the number of significant peaks appears to range dramatically between datasets, from less than a hundred to tens of thousands. Similarly, knockdown of some proteins which are essential cause dramatically more gene expression changes than others. It is unclear from the manuscript how these are differently weighted in the end, and thus whether RADAR is simply reflecting predictions of a small number of broadly binding RBPs.

Author's response:

We thank the reviewer for the comments and address the comments in a point by point fashion below:

"The eCLIP data appears to contain not only narrow binding proteins, but also broad binding or coating proteins (such as POLR2G https://www.encodeproject.org/experiments/ENCSR820WHR/). Perhaps because of this, the number of significant peaks appears to range dramatically between datasets, from less than a hundred to tens of thousands."

We agree with the reviewer that the large variation in the number of binding peaks could potentially bias the results obtained in our analysis. However, we believe that, to the best of our knowledge, the eCLIP binding peaks are currently of the highest quality data in determining where RBPs will bind. In Figure R.XXX below we show boxplots of the number of binding peaks relative to the different CLIP methods. In particular, we can see that eCLIP provides the lowest variance in the number of binding peaks. To further minimize artifacts from the eCLIP data, we only select the binding peaks with a score of 1000, which are considered to be the most significant peaks. We remove any peak that has a low score (200), in hopes to remove any peaks that may be false positives.



"Similarly, knockdown of some proteins which are essential cause dramatically more gene expression changes than others."

We agree with the reviewer that some proteins that are knocked down result in greater changes than others. Variants falling in peaks of an RBP that are linked to genes showing high expression change after that RBP is knocked down are currently weighed higher than those variants falling in regions not linked to genes with significant changes. Therefore, it is appropriate to weigh a variants score higher when they fall in regions with more significant changes in expression after KD.

Figure R.XXX. *Summary of the KD data.*

"It is unclear from the manuscript how these are differently weighted in the end, and thus whether RADAR is simply reflecting predictions of a small number of broadly binding RBPs."

When giving a weight to the scores that rely on the eCLIP peaks, we incorporate the Shannon entropy as a function of the value f in our manuscript. The value of f is defined to be:

# of 1KG variants in an RBP's peaks / # of genome-wide 1KG variants

This value of f therefore takes into account the length effect of different RBPs.

Overall, while it is true that different RBPs could have different number of peaks, the eCLIP peaks are currently the highest quality data for representation of RBP binding sites, and therefore we take variations in the number of peaks between RBPs to be reflective of biology rather than artifact.