# Response letter to

"*RADAR: An integrative framework for variant annotation and prioritization in post-transcriptional regulome of RNA binding proteins*"

## -- Ref1.0 – Software –

***Reviewer's comment:***

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

***Author's response:***

We thank the referees for pointing this out. In this round, we significantly improved the interface of our software with extensive testing, which we feel is easy to use. The main changes include:

- We provided both RADAR online and command line versions with documentation and data (https://github.com/gersteinlab/RADAR and http://radar.gersteinlab.org)
- We provided a short test instance for users to check.

For more details please check the excerpt in the revised version as below.

***Excerpt from the revised Supplement:***

We include a downloadable ZIP file at http://radar.gersteinlab.org/#!page-downloads which contains the RADAR source code (radar.py) and a directory containing all data files needed by RADAR (resources/). Individual files are listed below for download. This website also provides software documentation, usage information, performance benchmarks, and test examples. We also provided a web version of the software that can be used to run RADAR directly through the site. A variant file (BED format) can be uploaded, with the option to select any tissue-specific features. The output contains each scored feature as well as the full RADAR score.

RADAR can also be run from the command line after unzipping radar.zip and downloading the necessary dependencies (Python, BEDTools and pybedtools). Users can run the software using the following command:

```
python radar.py -b [BED file containing variants to be scored] -o
[output directory] -c [cancer type] [-kg -mr -rp]
```

The -kg, -mr and -rp are optional parameters that are used to indicate whether tissue-specific scores (key genes, mutation recurrence, and RBP regulation power) should be computed. These options require specifying a TCGA cancer abbreviation. After running the software, the output scores can be found in

```
[input BED file name].radar_out.bed
```

The RADAR source code can be found at https://github.com/gersteinlab/RADAR.
Included below is a step-by-step walkthrough of using RADAR to score the *Alexandrov et al* breast cancer variants.

```
[yf95@farnam1 ~]$ python
Python 2.7.13 (default, Jun  1 2017, 16:52:45)
[GCC 5.4.0] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> import pybedtools
>>>
```

```
[yf95@farnam1 ~]$ bedtools
bedtools is a powerful toolset for genome arithmetic.

Version:   v2.27.1
About:     developed in the quinlanlab.org and by many contributors worldwide.
Docs:      http://bedtools.readthedocs.io/
Code:      https://github.com/arq5x/bedtools2
Mail:      https://groups.google.com/forum/#!forum/bedtools-discuss

Usage:     bedtools <subcommand> [options]
```

Dependencies for RADAR include BEDTools (which can be found at http://bedtools.readthedocs.io/en/latest/content/installation.html), Python (which can be found at https://www.python.org/downloads/, our tests were conducted with version 2.7), and pybedtools (http://daler.github.io/pybedtools/main.html).

The RADAR package can be downloaded at http://radar.gersteinlab.org/#!page-downloads. The unzipped file includes a RADAR directory contains python executable script and a resources directory containing all data files necessary for the RADAR script to produce scores.

```
[yf95@farnam1 example]$ head Breast.bed
chr1    13506   13507   G       A       TCGA-EW-A1OZ-01A-11D-A142-09
chr1    14841   14842   G       T       PD5935a
chr1    16995   16996   T       C       PD7201a
chr1    17764   17765   G       A       PD5935a
chr1    17764   17765   G       A       PD7216a
chr1    28587   28588   G       T       PD4962a
chr1    30527   30528   C       T       PD5935a
chr1    61396   61397   G       A       PD4967a
chr1    69522   69523   G       T       TCGA-BH-A0BP-01A-11D-A10Y-09
chr1    83442   83443   C       T       PD4072a
```

The software can be run from command line as follows:

```
python radar.py -b ../Breast.bed -o .. -c BRCA -kg -mr -rp
```

```
[yf95@farnam1 radar]$ python radar.py -b ../Breast.bed -o .. -c BRCA -kg -mr -rp
[yf95@farnam1 radar]$ ls ..
Breast.bed  Breast.radar_out.bed  radar  radar.zip
```

RADAR has generated the output file, Breast.radar_out.bed, which contains the list of scored variants. A header of the output file is shown below (note that the header takes up one line in the file, but is broken onto two lines in this screenshot):

```
[yf95@farnam1 radar]$ head ../Breast.radar_out.bed
chr     start   stop    ref     alt     cross_species_conservation      RBP_binding_hub GERP    Evofold motif_disruption        RBP_gene_association
total_universal     key_genes       mutation_recurrence     RBP_regulation_power    total_tissue_specific   total_score
chr1    13506   13507   G       A       0       0       0       0       0       0       0       0       0       0       0       0
chr1    14841   14842   G       T       0       0       0       0       0       0       0       0       0       0       0       0
chr1    16995   16996   T       C       0       0       0       0       0       0       0       0       0       0       0       0
chr1    17764   17765   G       A       0       0       0       0       0       0       0       0       0       0       0       0
chr1    28587   28588   G       T       0       0       0       0       0       0       0       0       0       0       0       0
chr1    30527   30528   C       T       0       0       0       0       0       0       0       0       0       0       0       0
chr1    61396   61397   G       A       0       0       0       0       0       0       0       0       0       0       0       0
chr1    69522   69523   G       T       0       0       0       0       0       0       0       0       0       0       0       0
chr1    83442   83443   C       T       0       0       0       0       0       0       0       0       0       0       0       0
```

## -- Ref1.1 – Abstract –

***Reviewer's 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:***

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. The main concern is the 100 word limit for a software paper and we feel it is also important to emphasize another uniqueness of RADAR - post-transcriptional regulation.
To this end, as suggested 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. Please see the details from the excerpt below.

***Revised Abstract from Manuscript:***

RNA-binding proteins (RBPs) play key roles in post-transcriptional regulation and disease. Their binding sites cover more of the genome than coding exons; nevertheless, most noncoding variant-prioritization methods only focus on transcriptional regulation. Here, we integrate the portfolio of ENCODE-RBP experiments to develop RADAR, a variant-scoring framework. RADAR uses conservation, RNA structure, network centrality, and motifs to provide an overall impact score. Then it further incorporates tissue-specific inputs to highlight disease-specific variants. Our results demonstrate RADAR can successfully pinpoint variants, both somatic and germline, associated with RBP-function dysregulation, that cannot be found by most current prioritization methods (e.g. variants affecting splicing).

## -- Ref1.2 – Comparison of methods –

***Reviewer's 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:***

We thank the referee for this comment, and we agree that it is important to explain our the motivation of selecting other methods for comparisons. Our original thinking was to compare the RADAR results with popular methods focusing on the noncoding regions. Specifically,

- RADAR shares a lineage to FunSeq2, such as adapting the Shannon entropy scoring scheme. We believe that the comparison is natural, to see how prioritizing variants from a transcriptional versus post-transcriptional perspective would differ.
- We also compare RADAR to CADD, due to the popularity that CADD has gained, in the field of variant prioritization.

As suggested by the referee, our new revision includes additional comparisons to other methods mentioned in the Benchmark paper. Specifically, we have added a comparison to FATHMM-MKL. We did not include GWAVA since the installation is not applicable and runs with an error. Another reason is that while GWAVA does have an online interface to score variants, it only scores common germline variants, unlike the other methods.

***Excerpt from the Manuscript:***

RADAR aims to prioritize variants relevant to the post-transcriptional regulome, while FunSeq2, FATHMM-MKL, and CADD focus on those that affect the transcriptional regulome. Therefore, we do find many variants that demonstrate a high overall RADAR score, but only show moderate FunSeq2, CADD, and FATHMM-MKL scores. For example, 13 coding and 41 noncoding variants that are ranked within the top 1% of overall RADAR scores are not in the top 10% of CADD, FunSeq2, or FATHMM-MKL scores (Supplementary Table S10 and Table S11). Many of such variants are located in RBP binding hubs, and undergo strong purifying selection, demonstrated strong motif disruptiveness, and are regulated by key RBPs that are associated with breast cancer from multiple sources of evidence. We believe the discovery of such events demonstrates the value of RADAR as an important and necessary complement to the existing transcriptional-level function annotation and prioritization tools.
Supplemental tables: http://radar.gersteinlab.org/resources/RADAR.supplementary.table.xlsx

## -- Ref1.3 – RBP Splicing –

***Reviewer's comment:***

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

***Author's response:***

We agree with the reviewer that RNA splicing is an important factor to consider in the RBP regulome and we indeed considered splicing in our initial submission. We have tried to make this point more clear in our revised manuscript. We now further highlight the splicing factors in supplementary tables. We also have now included a downloadable link on our website of eCLIP data annotated by each RBPs specific function, which can easily be filtered for splicing-related RBPs, and found at http://radar.gersteinlab.org/splicing.zip and http://radar.gersteinlab.org/non_splicing.zip.

Table R1. *Summary of splicing vs non-splicing RBPs*

| Splicing Related RBPs | | Non-Splicing Related RBPs | |
|---|---|---|---|
| HepG2 | K562 | HepG2 | K562 |
| 25 | 24 | 44 | 63 |

***Excerpt from the revised Manuscript:***

We also provide versions of the eCLIP peaks that are annotated by RBP's function, such as splicing – which is one of the most common categories of our RBPs (see http://radar.gersteinlab.org/#!page-downloads, Splicing related RBP peaks).

***Excerpt from the revised Supplement:***

We included a table in our supplement (extracted from the supplement and shown below in supplementary file S3 categorizing each RBP by their function, many of which are splicing related.

Table R.2 *Specific RBPs and their functions.*

| Category | RBPs |
|---|---|
| RNA Binding | DDX3X, DDX59, DGCR8, DROSHA, EWSR1, HNRNPA1, HNRNPC, HNRNPK, HNRNPM, HNRNPU, HNRNPUL1, IGF2BP3, ILF3, KHDRBS1, NONO, NPM1, PCBP2, PRPF8, PTBP1, RBFOX2, RBM15, RBM22, SAFB2, SF3A3, SRSF7, SRSF9, TAF15, TARDBP, TNRC6A, U2AF1, U2AF2, AARS, AUH, CPSF6, CSTF2, CSTF2T, DDX24, DDX42, DDX55, DDX6, DHX30, DKC1, EIF4G2, FAM120A, FASTKD2, FMR1, FUBP3, FXR1, FXR2, GEMIN5, GRSF1, IGF2BP1, IGF2BP2, KHSRP, LARP4, LARP7, LIN28B, LSM11, MTPAP, NOL12, NSUN2, PPIL4, PUM2, PUS1, QKI, RBM27, RPS11, RPS5, SERBP1, SF3B4, SFPQ, SLBP, SLTM, SMNDC1, SRSF1, SUGP2, SUPV3L1, TIA1, TRA2A, TROVE2, UPF1, XPO5, ZRANB2 |
| tRNA Binding | AARS, NSUN2, XPO5 |
| tRNA Splicing | CSTF2 |
| Pre mRNA Splicing | GTF2F1, HNRNPA1, HNRNPC, HNRNPK, HNRNPM, HNRNPU, HNRNPUL1, NONO, PCBP2, PRPF8, PTBP1, RBM15, RBM22, SF3A3, SRSF7, SRSF9, U2AF1, U2AF2, BUD13, CDC40, CSTF2, EFTUD2, GEMIN5, GPKOW, NCBP2, SF3B1, SF3B4, SRSF1, TRA2A |
| Splicing Regulation | RBFOX2 |
| Polyadenylation | CPSF6, CSTF2, GRSF1, MTPAP |
| mRNA Stability | FMR1, KHSRP, PUM2, SERBP1 |
| rRNA Processing | DKC1, RPS11, RPS5, SBDS, XRN2 |

| | |
|---|---|
| Ribosome Structure | RPS11, RPS5 |
| RNA Editing | DKC1, PUS1 |

## -- Ref1.4a – Basic and tissue RADAR score explanation –

***Reviewer's comment:***

> *4a- The basic and tissue-specific scoring is not well explained.*

***Author's response:***

We thank the reviewer for this suggestion. We have restructured our methods section, which now contains specific details on how to score a variant from each feature (6 basic, 3 user-specific). Equations used in each part of the score have been added to the appropriate sections and numbered. We have also included a simplified flowchart of the scoring scheme with relevant mathematical equations in our supplement, as shown in comment 1.4c below.

***Excerpt from the Manuscript:***

Please see updated methods section.

## -- Ref1.4b – Separation of results and methods sections –

***Reviewer's comment:***

> *4b- The method section is mixed with results (eg. In Regulatory Power from Linear Regression). Please separate results from methods.*

***Author's response:***

We thank the reviewer for this comment. In the revised version, we have removed all results from the methods section, so that the methods section now clearly illustrates only the models and equations used to score variants.

***Excerpt from the Manuscript (Results section):***

The values of the regulation potential (, see Methods) for all cancer types and RBPs are provided in supplementary Table S7. We found that among the RBPs with larger regulation potential, many have been reported as cancer-associated genes (Supplementary Table S8). For RBPs with high regulation potentials from aggregated expression analysis, we also performed a patient-wise regulation potential inference, where the differential expression of a gene is determined as the normalized difference between

an individual patient's tumor and normal expressions. Then, we tried to associate this individual regulation potential with disease prognosis. We downloaded the patient survival data from TCGA and performed survival analysis using the survival package in R (version 2.4.1-3). Interestingly, the regulatory power of two key RBPs PPIL4 and SUB1 were found to be significantly associated with patient survival (Fig. 4C).

## -- Ref1.4c – Clear presentation of the scoring system –

***Reviewer's comment:***

> *4c- 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.*

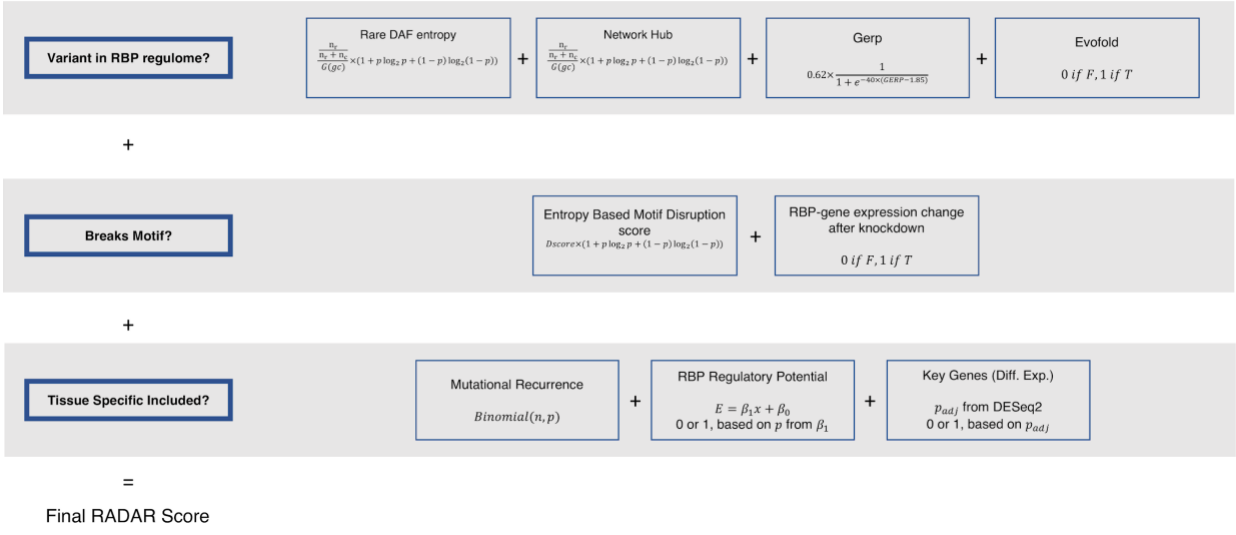
***Author's response:***

To further clarify the scoring of a given variant in addition to an updated methods section, we also provided a flowchart for scoring, shown below in Figure R.1, extracted from the supplement. We hope this flowchart, when used in conjunction with the detailed methods section with mathematical equations, will clarify how variants are scored, in both the basic and user-specific contributions.

***Excerpt from the revised Supplement:***

A simple flowchart of calculations for scoring a variant using RADAR:

Figure R.1 *Simple flowchart of RADAR scoring.*



## -- Ref1.5 – Relevance of features of RADAR –

***Reviewer's 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:***

We thank the reviewer for the comment and suggestion. We have addressed the importance of features in RADAR in a detailed text below, as well as provide a table below, extracted from the revised supplement.

Basically, RADAR and FunSeq2 focused on different regulatory levels: post-transcriptional versus transcriptional regulation, although they share some similarities in the entropy-based scoring system. Their basic building blocks are different. RADAR is based on eCLIP, shRNA RNA-Seq, and RNA Bind-n-Seq, while FunSeq2 is based on ChIP-Seq, DHS, and enhancers.

***Excerpt from the revised Supplement:***

Below we show the comparison of RADAR to FunSeq2 and also describe the relevance of each feature to variant prioritization on the RBP regulome.

Table R.3 *Features of RADAR*

| Universal Features | Same as FunSeq2? | Relevance to RADAR |
|---|---|---|
| Cross-Population Conservation in eCLIP | N | Conservation of post-transcriptional regulome |
| Cross-Species Conservation | Y | Important for considering cross-species conservation. |
| Structural Conservation (Evofold) | N | RNA secondary structure |
| RBP Binding Hubs | N | Binding hubs are more conserved |
| RBP-gene associations | N | Gene expression changes caused by motif disruption |
| Motif Disruption | N | Disrupts binding of RBPs |
| **Tissue Specific Features** | **Same as FunSeq2?** | **Relevance to RADAR** |
| RBP Regulatory Potential | N | RBPs regulate gene networks |
| Differential Expression of Key Genes | N | DE is a hallmark of regulation |
| Mutational Recurrence | N | Recurrence in specific tissues demonstrate unique hotspots |

## -- Ref1.6 – Cell specific validation –

***Reviewer's 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.*

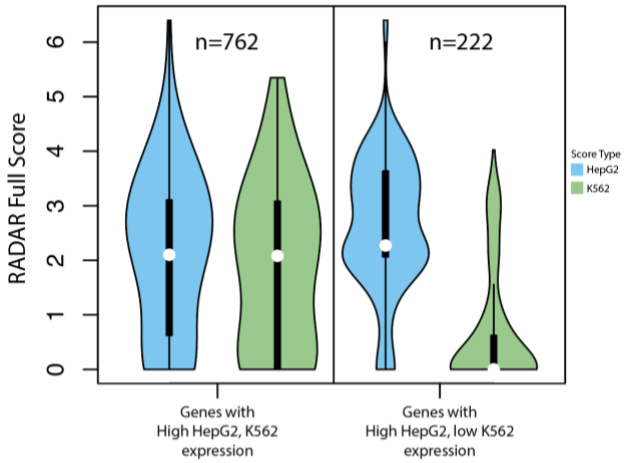
***Author's response:***

We thank the referee for this comment and we feel that it is a good comment. We agree that it is important to run the tissue-specific score comparison. As suggested, 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.
We conclude that for the universal score the commonly expressed genes showed comparable HepG2 and K562 scores, while HepG2 specific genes demonstrated much higher scores. We also found that tissue-specific features in our second scoring system greatly helped to distinguish cell type information. We added this part in the results and discussion sections.

***Excerpt from the Supplement:***

*Example 1: Comparison of full RADAR scores on variants in common and differentially expressed genes in HepG2 and K562.*
Here we show that somatic Liver cancer variants from the 2013 *Alexandrov et al* paper falling in genes with both high expression in K562 and HepG2 (top 10% expression from total RNA-seq) demonstrate comparable scores when using matched cell type scoring schemes. Those variants falling in genes with high expression in HepG2 (top 10%) but low in K562 (FPKM<1) demonstrate scores that are much lower when using the K562 scoring scheme.

Figure R.2 *Effects of Cell and Tissue Specific Data on RADAR Score of Somatic Liver Variants*

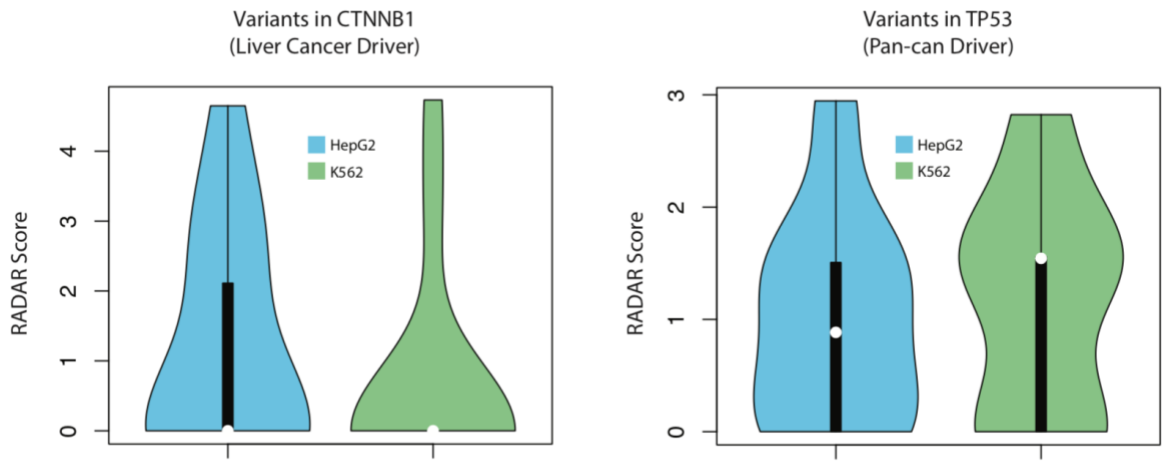



*Example 2: scoring on somatic variants from tumor-specific and pancan driver genes*

We compare the HepG2 and K562 scores for a set of Liver cancer variants available publicly from the 2013 *Alexandrov et al* paper.

Here we observed 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.3 below.

Figure R.3. *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.*




**-- Ref2.0a – eCLIP versus transcript annotations –**

***Reviewer's 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.*

***Author's response:***

We thank the reviewer for the comment and we agree that the peak calling is an important factor in the scoring system. Different from ChIP-Seq data peak calling, the normalization issue in eCLIP data needs more thinking since the definition of the transcribed regions is not as obvious. Extending the null model to the whole genome might introduce numerous false positives.
We hope that in the future as the development of computational algorithms the peaks will be called more accurately, which directly helps the scoring system. At the moment, we prefer to use the more conservative peak calling on the annotated transcribed region. But we added this point into the discussion section.

***Excerpt from the Manuscript:***

It is important to note that different from ChIP-Seq data peak calling, the normalization issue in eCLIP data is more complex. The definition of the transcribed region is not as obvious and extending the null model to the whole genome might introduce false positives. As a conservative approach, we use the results that are more conservative, where peak calling is done on only the annotated transcribed region.

## -- Ref2.0b – Relative size of the RBP regulome –

***Reviewer's comment:***

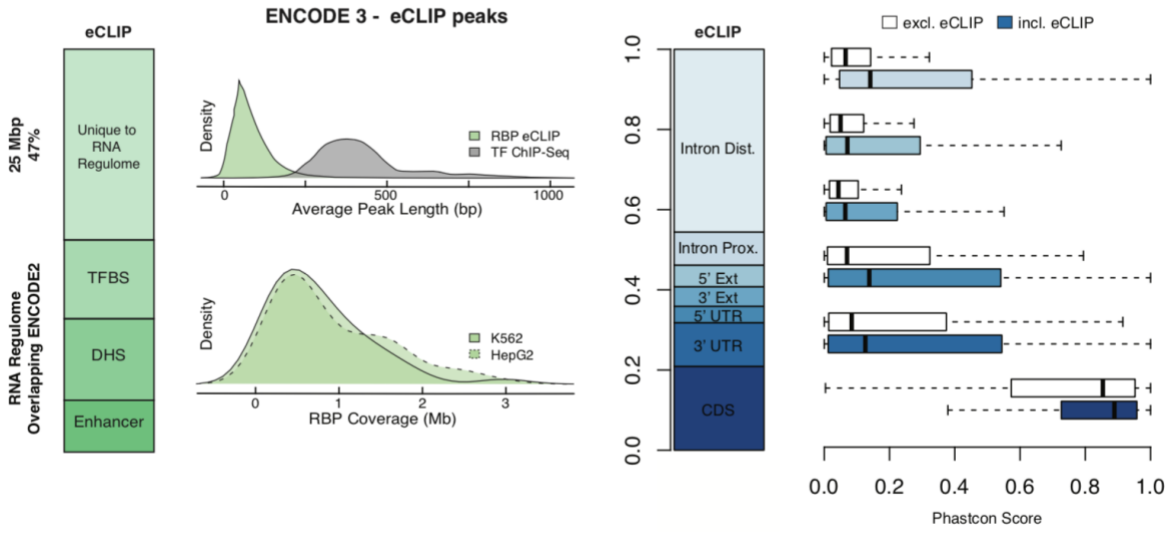
*Similarly, although the 'RBP regulome' appears smaller than that for TFs, it is unclear whether this is simply because the average peak size for eCLIP is significantly smaller than the average CHIP-seq peak due to differences in method and peak callers (likely, as most known RBP and TF motifs are of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

***Author's response:***

We thank the reviewer for this suggestion. We agree with the referee that due to the different resolution of assays, the comparison in our original Figure 2 takes a simple approach, and while important, may suffer from biases. Therefore, we have moved the Figure 2 from our initial submission to the supplement (Main Manuscript Fig 2) and modified the new Figure 2 to better represent the novelty of eCLIP. Specifically, we have changed the focus to show that the RBP regulome covers a decent amount of the genome that is not overlapped with any existing annotations. While the eCLIP peaks do 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 in Figure R-2A.

***Excerpt from the Manuscript:***

Main Manuscript Figure 2. (A) Intersection of eCLIP peaks versus transcriptional level annotations, with 25Mbp unique to the RBP regulome; (B) Average length of binding peak for RBP eCLIP data versus TF ChIP-Seq and the similar distribution of RBP coverage between K562 and HepG2 cell lines; (C) Fraction of RBPs falling into each annotation category as well as boxplots of PhastCons scores of annotations intersecting peaks (blue) versus annotations with no intersections (white).



## -- Ref2.1a – Weighting of RBPs with different patterns of binding –

***Reviewer's 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. 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:***

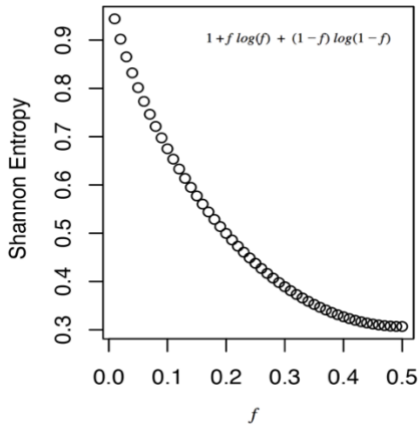
We agree with the reviewer's comment that some RBPs bind more broadly than others. When weighting different RBPs, we are careful to not bias our score results as to only prioritize those variants that fall in broadly binding peaks. In order to account for this, we used a scoring scheme based on Shannon entropy. The explanation below has been now inserted into our supplement.

For entropy: given, *f*, which is roughly the fractional coverage of an RBPs peaks on the genome, as the number of 1KG variants falling within all peaks of an RBP divided by the total number of 1KG variants, the entropy is given as:

$$1 + f\ log(f)\ +\ (1-f)\ log(1-f)$$

In this equation, an increase in $f$ will cause a decrease in the entropy (for $f < 0.5$) which is shown in Figure R.4 below. When $f > 0.5$, the opposite occurs, but because our RBP eCLIP annotations are much smaller compared to the size of the genome, $f$ remains less than 0.5. Therefore, broadly binding peaks are actually slightly weighted smaller than narrow binding peaks. This ensures that our results are more reflective of predictions on all RBPs rather than just those that bind broadly.

Figure R.4. Changes in Shannon entropy as $f$ changes.



The second component of the score if the rare DAF. Given an RBPs binding peaks, which contains $r$ rare mutations and $c$ common mutations, the rare DAF is given as:

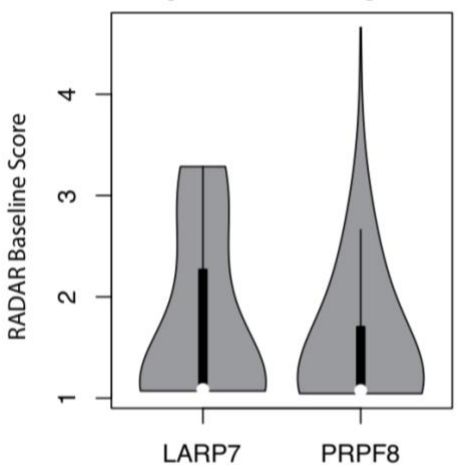
$$\rho = r / (r + c)$$

In this equation, since both $r$ and $c$ depend on how broadly an RBP binds, we have a measure that is independent of the coverage of the RBP.

The product of the two components gives the final cross-population conservation score component. In both parts, we are careful to make sure that we are not confounding the score by the coverage of binding of RBPs.

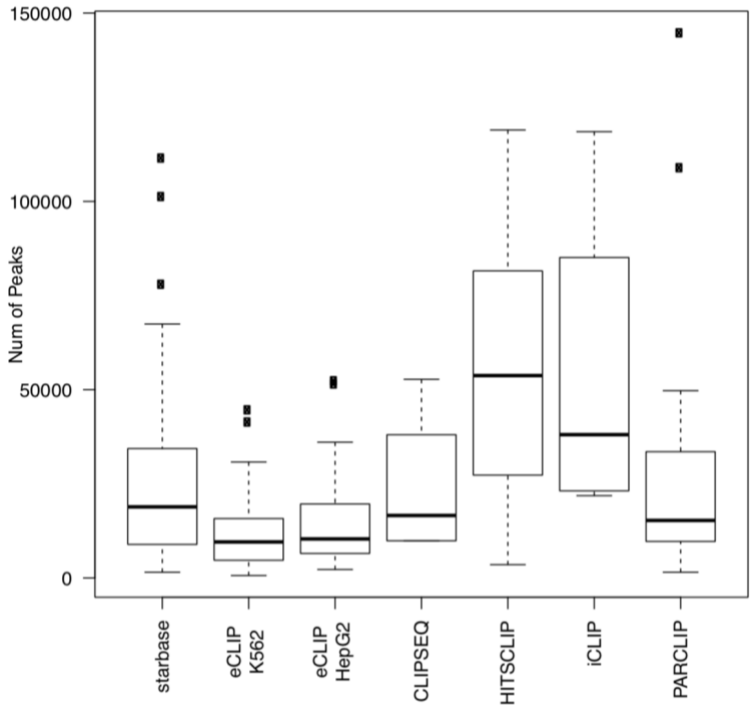
As a test, we actually checked the distribution of Liver cancer somatic variants falling in LARP7 and PRPF8. LARP7 peaks are in the bottom 10% of number of nucleotides covered by an RBPs peaks while PRPF8 has the largest number of nucleotides covered by an RBPs peaks. We do not observe significantly larger scores for variants falling in broadly binding peaks.

Figure R.5 Scoring Liver Cancer Variants in High and Low Coverage RBPs




Besides, we want to emphasize the quality of ENCODE eCLIP data. We showed a boxplot of the average number of peaks of different CLIP based methods for determining RBP binding peaks. It is important to note that although there is some variation in the coverage of different RBPs, we believe the eCLIP data is the most conservative and shows the lowest variance between RBPs compared to all other methods.

Figure R.6 Number of Peaks for different methods and cell types




**-- Ref2.1b – General comments –**

***Reviewer's comment:***

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

***Author's response:***

The score associated with the knockdown data does not share a concept with the Shannon entropy. Differences of expression after knockdown may be associated with differences in biology - some RBPs regulate isoforms while others regulate genes, and it may not be fair to compare these values. Other functions of regulation may include DNA decay or transportation. Therefore we include a conservative approach and do not weigh differences in expression after KD, since they may be associated with biological functions that are beyond the scope of this paper.

In addition, the expression changes are not just due to how important the RBP effect is, but coudl be significantly confounded by the fold change of expressions of the RBP itself during the knockdowns (which can vary quite a bit, see Figure R.7 below). In addition, expression changes could be caused by direct or indirect linkages in the RBP gene network, but at this stage, we only consider the direct links, as to have a more conservative approach. As the quality of KD data improves over time, a more accurate representation of changes in gene expression of networks related to RBPs will be possible.

***Excerpt from the Supplement:***

We check the quality of the KD data by seeing what the fold change expression of each RBP, after knock down of itself. While most of the fold changes are negative, the variance is high, suggesting efficiency of knockdown between different experiments may vary.

Figure R.7. Quality check of the KD data.