-- Ref1.0 – Software –

Reviewer' 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. We feel it is easy to use in this revised version. The main changes include:

1. We provided both online and downloadable version of the software with detailed documentation

2. We put everything on radar.gersteinlab.org and at the same time on https://github.com/gersteinlab/RADAR

3. We provided short test instance for user to check.

More details please check the excerpt in the revised version as below.

Excerpt from the revised Supplement:

We have included a downloadable ZIP file at radar.gersteinlab.org which contains the RADAR source code (radar.py) and a directory containing all data files needed by RADAR (http://radar.gersteinlab.org/#!page-downloads/). This website also provides software documentation, usage information, an example. We also provided a web version of the software that can be used to run RADAR directly through the site: simply upload a variant file, select any tissue-specific scores, and provide a cancer type and the website will print the a list of scored variants.

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

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

After a few minutes (RADAR takes around 2-3 minutes to score ~1 million variants), there will be a file in the provided output directory called [input BED file name].radar_out.bed with all of the requested scores. -kg, -mr and -rp are optional parameters that are used to indicate whether these tissue-specific scores (key genes, mutation recurrence, and RBP regulation power) are requested. Cancer type information is required if any tissue-specific scores are requested.

The RADAR source code can be found at https://github.com/gersteinlab/RADAR.
Here is an step-by-step walkthrough of using RADAR to score Alexandrov breast cancer variants.

First, we must download the required software: 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). Follow the installation instructions for each one. You can confirm you have successfully installed each piece of software by attempting to run `bedtools` on the command line, which should print out documentation (screenshot below is curtailed).

```
[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]
```

User can confirm that they have Python and pybedtools installed by running the Python shell using the `python` command and attempting to import the pybedtools module with `import pybedtools`. If there are no errors, the prerequisite software was installed successfully. Note the Python version number on the first line after running the `python` command.

```
[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
>>>
```

Next, download the RADAR package in ZIP format from the RADAR website Downloads page (http://radar.gersteinlab.org/#!page-downloads)

Download RADAR and all necessary resource files in a ZIP format here.

Prerequisite Software

The following software are required to run RADAR.

1) BEDTools

2) Python (tested on Python 2.7.11)

3) pybedtools

Unzip the file at the command line. The resulting radar/ directory contains a .py file (the executable script) and a resources/ directory that contains all data files needed by the RADAR script to produce scores.

```
[yf95@farnam1 example]$ unzip radar.zip
Archive:  radar.zip
   creating: radar/
  inflating: radar/radar.py
   creating: radar/resources/
  inflating: radar/resources/all_RBP_peaks_unmerged_labeled_sorted.bed
  inflating: radar/resources/significant_peaks
   creating: radar/resources/main_scores/
  inflating: radar/resources/main_scores/chr13_scored
  inflating: radar/resources/main_scores/chr20_scored
  inflating: radar/resources/main_scores/chr22_scored
  inflating: radar/resources/main_scores/chr21_scored
  inflating: radar/resources/main_scores/chr14_scored
  inflating: radar/resources/main_scores/chr9_scored
  inflating: radar/resources/main_scores/chr11_scored
  inflating: radar/resources/main_scores/chr4_scored
  inflating: radar/resources/main_scores/chr8_scored
  inflating: radar/resources/main_scores/chr16_scored
  inflating: radar/resources/main_scores/chrY_scored
  inflating: radar/resources/main_scores/chr18_scored
  inflating: radar/resources/main_scores/chr5_scored
  inflating: radar/resources/main_scores/chr7_scored
  inflating: radar/resources/main_scores/chr19_scored
  inflating: radar/resources/main_scores/chr2_scored
  inflating: radar/resources/main_scores/chr15_scored
  inflating: radar/resources/main_scores/chr12_scored
  inflating: radar/resources/main_scores/chr6_scored
  inflating: radar/resources/main_scores/chr17_scored
  inflating: radar/resources/main_scores/chrX_scored
  inflating: radar/resources/main_scores/chr10_scored
  inflating: radar/resources/main_scores/chr3_scored
  inflating: radar/resources/main_scores/chr1_scored
  inflating: radar/resources/regulator_pval.txt
  inflating: radar/resources/rbp_peak_significance
[yf95@farnam1 example]$ ls
radar  radar.zip
```

Here is a head of the example input file we will be using (publicly accessible data from Alexandrov et al breast cancer variants). This file (called Breast.bed) is downloadable at http://radar.gersteinlab.org/#!page-example.

```
[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
```

Now we are ready to run the software and score our variants. Move into the recently unzipped radar/ directory, where the radar.py file exists (using `cd radar/`). Locate the path to the Breast.bed file (in the example command below, we will assume it exists in the directory above radar.py). Also identify a directory into which you would like the output file to be written (in this example, we will write the output to the same directory that contains Breast.bed). Here, we will produce all of the tissue-specific scores for BRCA. At the command line, we can run the following command to score these variants.

```
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 head 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
chr1    14841   14842   G       T       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
chr1    17764   17765   G       A       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
chr1    30527   30528   C       T       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
chr1    69522   69523   G       T       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
```

-- Ref1.1 – Abstract –

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.

Revised Abstract from Manuscript:

-- Ref1.2 – Comparison of methods –

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:

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 variant scoring methods mainly focusing on the noncoding regions. In the initial submission, we selected FunSeq2 and CADD because:

- RADAR shares a lineage to FunSeq2 in some ways, such as adapting the Shannon entropy scoring scheme, we believe that the comparison the two is natural, in order 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 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."

RADAR.supplementary.table-04282018.xlsx

-- Ref1.3 – RBP Splicing –

Reviewer' 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 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 R2. *Summary of RBP functions, including splicing, polyadenylation, etc.*

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 the most common function aside from RNA binding (see radar.gersteinlab.org).

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.XXX. *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 via Spliceosome	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
RNA Splicing Regulation	RBFOX2
mRNA Polyadenylation	CPSF6, CSTF2, GRSF1, MTPAP
Regulation of mRNA Stability	FMR1, KHSRP, PUM2, SERBP1
rRNA Processing	DKC1, RPS11, RPS5, SBDS, XRN2
Structural Constituent of Ribosome	RPS11, RPS5
RNA Editing	DKC1, PUS1

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

```
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 scoring a variant for each component of the score (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, shown in comment 1.4c below.

Excerpt from the Manuscript:

Please see new methods section.

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

Reviewer' comment: regarding the method section

```
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' 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.

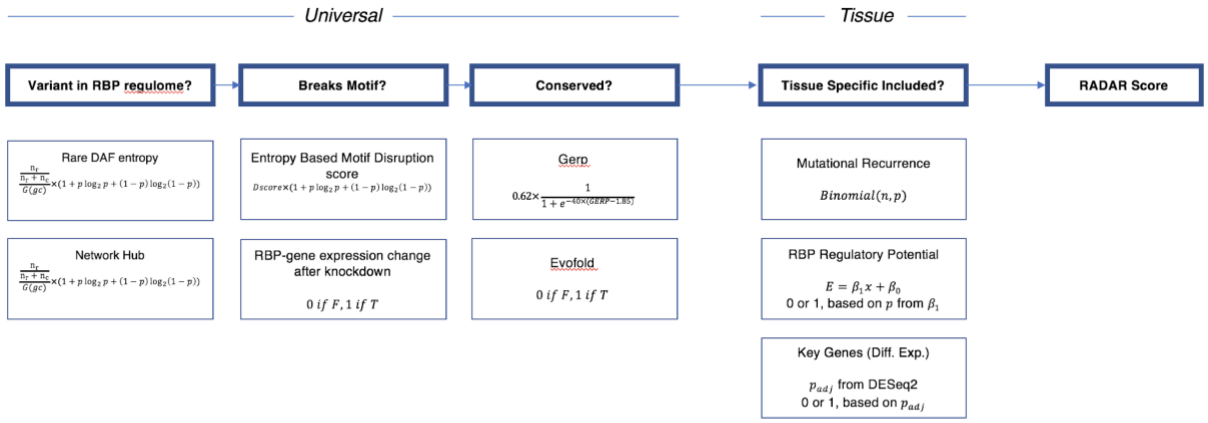
Author's response:

To further clarify the scoring of a given variant in addition to an updated methods section, we also provide a flowchart for scoring, shown below in Figure R.XXX, 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:

Addition of multiple boxes (no flow, rather independent objects added together)

Add a variable for the GERP equation, find out from FunSeq2



-- 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.

As seen in the following table, 5 out of 6 of the universal scores components are different from FunSeq2 and all 3 parts of the tissue specific score are quite different from FunSeq2.

Excerpt from the revised Supplement:

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' 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:

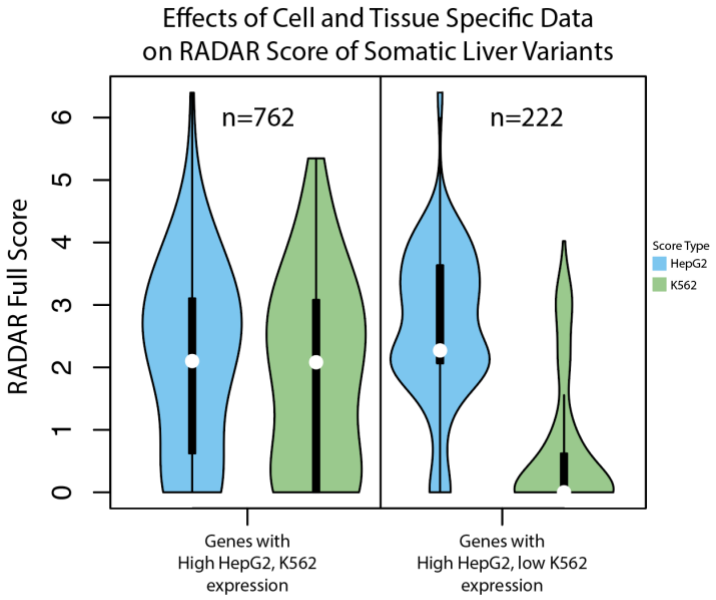
We thank the referee for this comment and and we feel that it is a good comment. We agree that it is important to run the tissue specific score comparison. As suggested, we have completed built the RADAR model using the two different cell specific data, creating a HepG2 and K562 score (baseline and tissue specific in each). We give two examples below to show how using cell type specific data could influence the RADAR score.

Our conclusion is that for 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 \cite{23945592} 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.

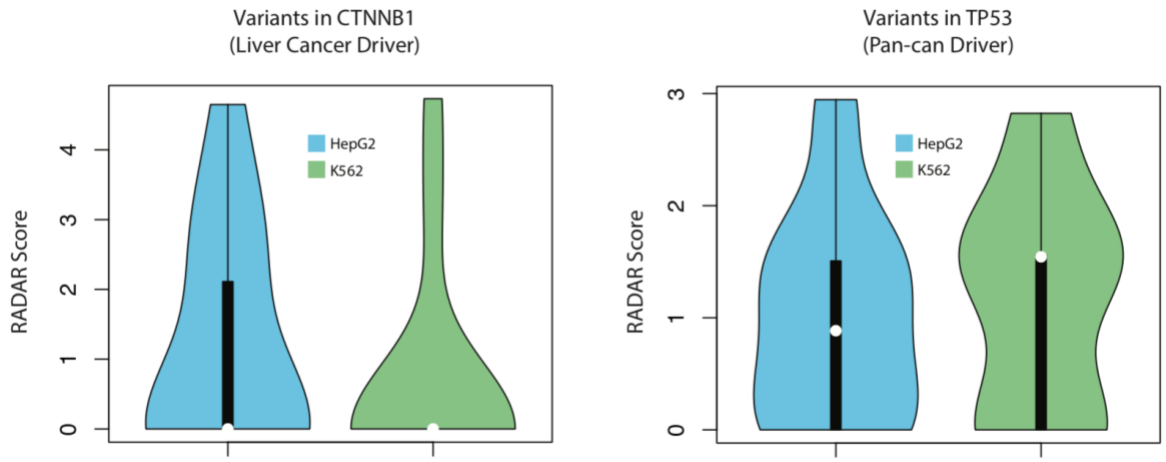


Example 2: scoring on somatic variants from tumor-specific and pancan driver genes

We compare the HepG2 and K562 scores for a set of Liver cancer variants available publicly from the Alexandrov et al paper \cite{23945592}.

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.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.*



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

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.

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 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 and 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' comment:

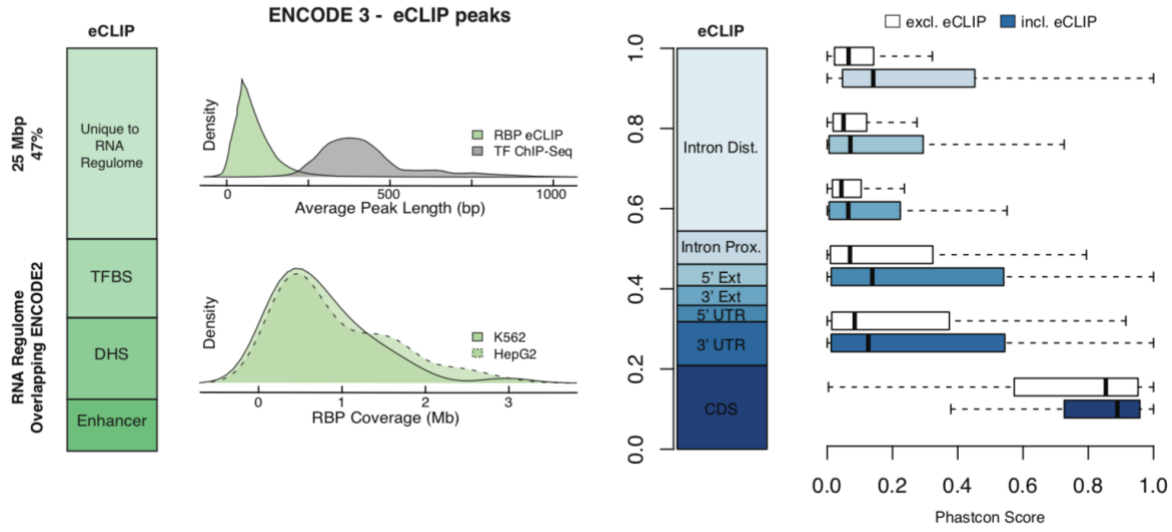
> Similarly, although the 'RBP regulome' appears smaller than that for TFs, it is unclear whether this is simply because the average peak size for eCLIP is significantly smaller than the average CHIP-seq peak due to differences in method and peak callers (likely, as most known RBP and TF motifs are of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

Excerpt from the Manuscript:

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.*



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

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. 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:

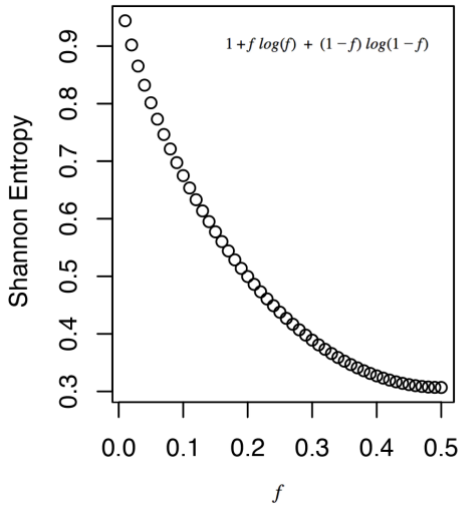
We agree with the reviewer's comment that some RBPs bind more broadly than others. When weighting different RBPs, we are careful to not bias our score results as to only prioritize those variants that fall in broadly binding peaks. In order to account for this, we used a scoring scheme based on Shannon entropy as well as rare DAF. 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 the Figure 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.XXX. Changes in Shannon entropy as f changes.



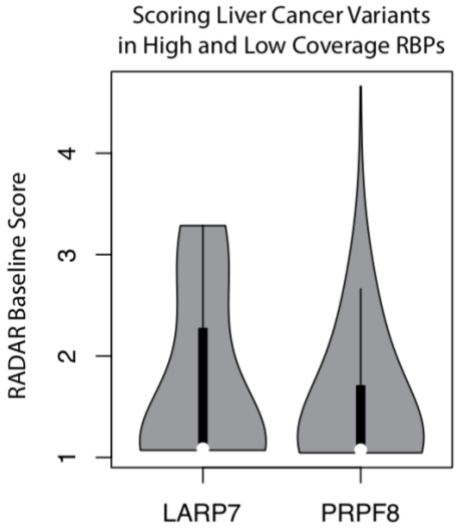
The second component of the score if the rare DAF. Given an RBPs binding peaks, which contains r rare mutations and c common mutations, the rare DAF is given as:

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

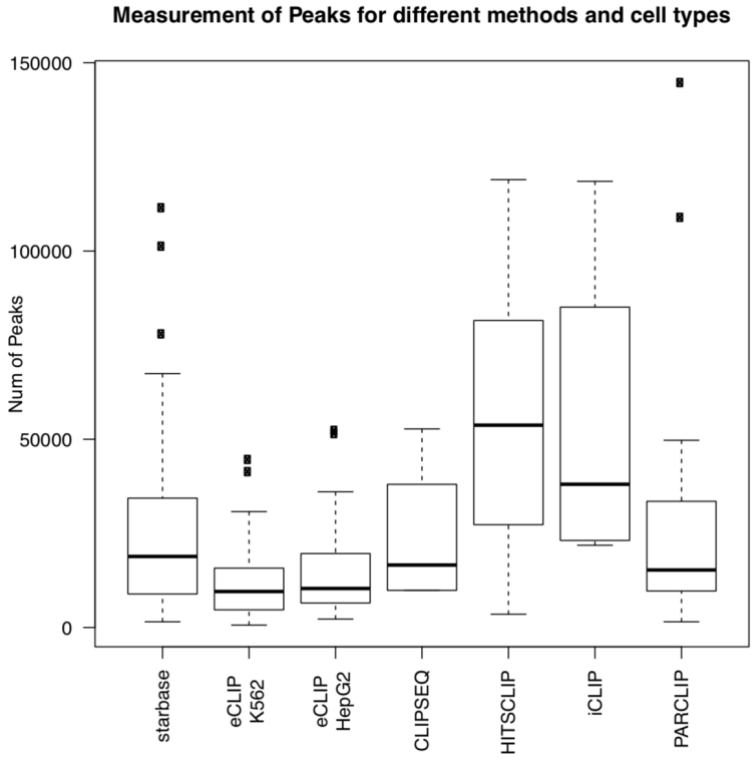
In this equation, since both r and c depend on how broadly an RBP binds, we have a measure that is independent of the coverage of the RBP.

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.



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.



Excerpt from the Supplement:

See description in the “author’s response”.

-- Ref2.1b – General comments –

Reviewer’s comment:

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

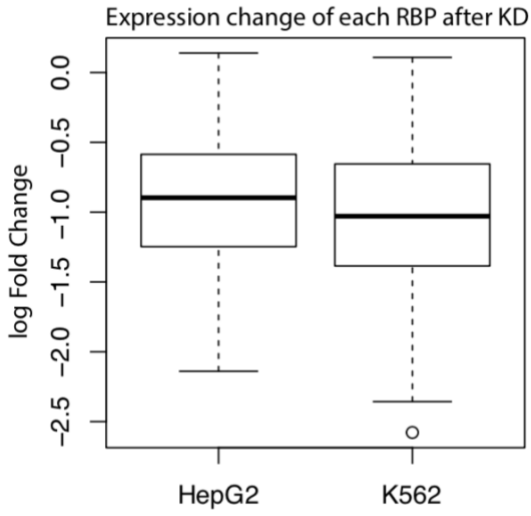
Author’s response:

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 could 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.XXX 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:

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


Expression change of each RBP after KD
log Fold Change
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
HepG2
K562