RADAR: Annotation and prioritization of variants in the post-transcriptional regulome for RNA-binding proteins

Jing Zhang[1,2#], Jason Liu[1,2#], Donghoon Lee[1], Lucas Lochovsky[2], Jo-Jo Feng[2], Shaoke Lou[1,2], Michael Rutenberg-Schoenberg[1,3], Mark Gerstein[1,2,4*]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA
[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA
[3]Chemical Biology Institute, Yale University, West Haven, CT 06516, USA
[4]Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA

* To whom correspondence should be addressed. Tel: +1 203 432 6105; Fax: +1 360 838 7861; Email: pi@gersteinlab.org

Jing Zhang and Jason Liu are co-first authors.

# Equal contributors

Abstract

RNA Binding proteins (RBP) play key roles in post-transcriptional regulation. Their binding sites cover more nucleotides than coding exons, but most current methods ignore such RBP-mediated regulation. Here, we integrate the entire ENCODE eCLIP experiments to deeply annotate the RBP regulome. We propose a variant impact scoring framework called RADAR, which uses conservation, RNA structure, network centrality, and motif information to provide a baseline impact score. Then RADAR further incorporates tissue-specific inputs to highlight disease-specific variants. Results on somatic and germline variants demonstrate that RADAR can successfully pinpoint deleterious variants, such as splicing-disruptive ones that cannot be highlighted by current prioritization methods.

Keywords

RNA binding protein, post-transcriptional regulation, variant prioritization, variant functional impact

Background

Dysregulation of gene expression is a hallmark of many diseases, including cancer [1]. In recent years, the accumulation of transcription-level functional characterization data, such as transcriptional factor binding, chromatin accessibility, histone modification, and methylation, has brought great success to annotating and pinpointing deleterious variants. However, beyond transcriptional processing, genes also experience various delicately controlled steps, including the conversion of premature RNA to mature RNA, and then the transportation, translation, and degradation of RNA in the cell. Dysregulation in any one of these steps can alter the final fate of gene products and result in abnormal phenotypes[2-4]. Furthermore, the post-transcriptional regulome covers an even larger amount of the genome than coding exons and demonstrates significantly higher cross-population and cross-species conservation. Unfortunately, variant impact in the post-transcriptional regulome has been barely investigated, partially due to the lack of large-scale functional mapping.

RNA binding proteins (RBPs) have been reported to play essential roles in both co- and post-transcriptional regulation[5-7]. RBPs bind to thousands of genes in the cell through multiple processes, including splicing, cleavage and polyadenylation, editing, localization, stability, and translation[8-12]. Recently, scientists have made efforts to complete these post- or co-transcriptional regulomes by synthesizing public RBP binding profiles[13-16], which have greatly expanded our understanding of RBP regulation. Since 2016, the Encyclopedia of DNA Elements (ENCODE) consortium started to release data from various types of assays on matched cell types to map the functional elements in post-transcriptional regulome. For instance, ENCODE has released large-scale enhanced crosslinking and immunoprecipitation (eCLIP) experiments for hundreds of RBPs[17]. This methodology provides high-quality RBP binding profiles with strict quality control and uniform peak calling to accurately catalog the RBP binding sites at a single nucleotide resolution.

Simultaneously, ENCODE performed expression quantification by RNA-Seq after knocking down various RBPs. Finally, ENCODE has quantitatively assessed the context and structural binding specificity of many RBPs by Bind-n-Seq experiments[18].

In this study, we aimed to construct a comprehensive RBP regulome and a scoring framework to annotate and prioritize variants within it. We collected the full catalog of 318 eCLIP (for 112 RBPs), 76 Bind-n-Seq, and 472 RNA-Seq experiments after RBP knockdown from ENCODE to construct a comprehensive post-transcriptional regulome. By combining polymorphism data from large sequencing cohorts, like the 1,000 Genomes Project, we demonstrated that the RBP binding sites showed increased cross-population conservations in both coding and noncoding regions. This strongly indicates the purifying selection on the RBP regulome. Furthermore, we developed a scoring scheme, named RADAR (***R***N***A*** Bin***D***ing Protein regulome ***A***nnotation and p***R***ioritization), to investigate variant impact in such regions. RADAR first combines RBP binding, cross-species and cross-population conservation, network, and motif features with polymorphism data to quantify variant impact described by a baseline score. Then, it allows tissue- or disease-specific inputs, such as patient expression, somatic mutation profiles, and gene rank list, to further highlight relevant variants (Fig. 1). By applying RADAR to both somatic and germline variants from disease genomes, we demonstrate that it can pinpoint disease-associated variants missed by other methods. In summary, RADAR provides an effective approach to analyze genetic variants in the RBP regulome, and can be leveraged to expand our understanding of post-transcriptional regulation. To this end, we have implemented the RADAR annotation and prioritization scheme into software for community use (radar.gersteinlab.org).

Results

Defining the RBP regulome using eCLIP data

We used the binding profiles of 112 distinct RBPs from ENCODE to fully explore the human RBP regulome (Supplementary Table S1), which has been previously under-investigated. Many of these RBPs are known to play key roles in post-transcriptional regulation, including splicing, RNA localization, transportation, decay, and translation (Supplementary Fig. S1).

Our definition of the RBP regulome covers 52.6 Mbp of the human genome after duplicate and blacklist removal (Fig. 2A). It is 1.5 and 5.9 times the size of the whole exome and lincRNAs, respectively. In addition, only 53.1% of the RBP regulome has transcription-level annotations, such as transcription factor binding, open chromatin, and active enhancers. 55.1% of the RBP regulome is in the immediate neighborhood of the exome regions, such as coding exons, 3' or 5' untranslated regions (UTRs), and nearby introns (Fig. 2B; see methods section for more details). Furthermore, we observed significantly higher cross-species conservation score in the peak regions versus the non-peak regions in almost all annotation categories, providing additional evidence of regulatory roles of RBPs (Fig. 2C). In summary, the large size of the regulome, the limited overlap with existing annotations, and the elevated conservation level highlight the necessity of computational efforts to annotate and prioritize the RBP regulome.

Using universal features for baseline RADAR score

To annotate and prioritize variants in RBP binding sites, we built a baseline score framework for RADAR that includes three components: (1) sequence and structure conservation; (2) network centrality; and (3) nucleotide impact from motif analysis.

Sequence and structure conservation in the RBP regulome

Cross-species sequence comparisons have been widely used to discover regions with biological functions [19, 20]. For example, GERP score maps the human genome to other species to identify nucleotide-level

evolutional constraints[21, 22]. Therefore, we used the GERP score in our RADAR baseline framework to detect potentially deleterious mutations in the RBP regulome (see methods section for more details).

Since the enrichment of rare variants indicates a purifying selection in functional regions in the human genome[19, 23, 24], here we inferred the conservation of RBP binding sites by integrating population-level polymorphism data from large cohorts (i.e. the 1,000 Genomes Project)[25, 26]. GC percentage may confound such inference by introducing read coverage variations, which is a sensitive parameter in the downstream variant calling process[27, 28]. Therefore, we calculated the fraction of rare variants, defined as those with derived allele frequencies (DAFs) less than 0.5%, within the binding sites of each RBP. Then we compared them with those from regions with similar GC content as a background (see methods section for more details). In total, 88.4% of the RBPs (99 out of 112) showed elevated rare variant fraction in coding regions after GC correction (Fig. 3A). Similarly, in the noncoding part of the binding sites, 93.8% of RBPs (105 out of 112) exhibited an enrichment of rare variants. This observation convincingly demonstrates the accuracy of our RBP regulome definition (Supplementary Table S3).

Some well-known disease-causing RBPs demonstrate the largest enrichment of rare variants. For example, the oncogene XRN2, which binds to the 3' end of transcripts to degrade aberrantly transcribed isoforms, showed significant enrichment of rare variants in its binding sites[29]. Specifically, it demonstrates 12.7% and 10.3% more rare variants in coding and noncoding regions, respectively (adjusted P values at 1.89×10^{-9} and 2.85×10^{-118} for one-sided binomial tests)[30]. Hence, we used the enrichment of rare variants to infer the selection pressure in RBP binding sites and adjust the baseline variant scores in such regulator regions (see methods for more details).

RNA secondary structures have been reported to affect almost every step of protein expression and RNA stability[31]. We incorporated structural features predicted by Evofold, which uses a phylogenetic stochastic context-free grammar to identify functional RNAs in the human genome that are deeply conserved across species[32]. We found that the RBP binding sites demonstrated significantly higher conversation after intersecting with conserved structural regions defined by Evofold. Thus, we used the Evofold regions in our baseline scoring system.

Highlighting variants in binding hubs

It has been reported that genes within network hubs demonstrate higher cross-population conservation — a sign of strong purifying selection[23, 24, 33]. We hypothesized that RBP binding hubs could show similar characteristics because once mutated they might introduce larger regulation alterations. To test this, we separated the regulome based on the number of associated RBPs. Most regulome regions (62%) were associated with only one RBP (Supplementary Fig. 6). As the number of RBPs increased, we observed a clear trend of larger rare variant enrichment (Fig. 3D). For instance, noncoding regions with at least five or 10 RBPs exhibited 2.2% or 13.4% more rare variants, respectively (top 5% and 1%, Fig 3D). This observation supports our hypothesis that the RNA regulome hubs are under stronger selection pressure and, therefore, should be given higher priority when evaluating functional impact of mutations.

Emphasizing genes differentially expressed after RBP knockdown

RNA-seq expression profiling before and after shRNA mediated RBP depletion from ENCODE can help to infer the gene expression changes introduced by RBP knockdown. Variants with disruptive effects on RBP binding may affect or even completely remove the RBP binding and hence affect gene expressions in a similar way. Therefore, we extracted the differentially expressed genes from RNA-Seq before and after shRNA-mediated RBP depletion. Then, we up-weighted all variants that were located near the differentially

expressed genes and simultaneously disrupted the binding of the corresponding RBPs (schematic in Supplementary Fig. S9).

Using motif analysis to determine nucleotide impact

Mutations that change the RBP binding affinity may alter RBP regulation via motif disruption. We quantified the difference of position weight matrix (PWM) scores of the mutant allele against the reference allele. RADAR consists of two sources of motifs. First, we used the motifs identified from RNA Bind-n-Seq experiments from ENCODE because it has been reported that many RBP binding events *in vivo* can be captured by binding preferences *in vitro*. Second, we used the *de novo* motifs discovered directly from binding peaks using the default settings in DREME (see details in methods). For each variant, we quantified the nucleotide effect using the highest motif score from these two sources.

Incorporating user-specific features to reweight variant impact

Variant Prioritization can be improved if informative priors can be appropriately incorporated into the scoring system. Therefore, our RADAR framework allows various types of user-input to help identify disease-relevant variants. Specifically, we adopted a top-down scheme to incorporate regulator and element level information to up-weight factors that are possibly associated with disease of interest.

Highlighting key regulators through expression profiles

Key regulators are often associated with disease progression, so variants that affect such regulation should be prioritized[34]. RADAR finds such key regulators by combining the RBP regulatory network information with TCGA expression profiles. Specifically, we first constructed the RBP network from eCLIP binding peaks and defined the gene differential expression status from disease and normal cell types (see Methods). Then for each RBP, we quantified its regulation potential by associating its network connectivity with aggregated disease-to-normal differential expressions from many samples. We applied

this approach on 19 cancer types from TCGA and the regulation potentials are given in Fig. 4. The values of the regulation potential (β_1, 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).

In our RADAR framework, we further highlight variants that are associated with RBPs with high regulation potential in their corresponding cancer types by adding extra score to their disease-specific scores (see more details in methods). We can easily extend such analysis for other diseases by incorporating differential expression profiles from others cohorts such as GTEx[35, 36].

Up-weighting key elements from either prior knowledge or mutational profiles

RADAR reconsiders the functional impact difference among RBP peaks by their associated genes. For example, users can input a prioritized gene list, such as well-documented risk genes for the disease of interest. Alternatively, genes that undergo significant expression or epigenetic changes are mostly cell-type-specific and can be used to highlight more relevant variants. Our RADAR framework will up-weight all the RBP peaks that are close to these genes.

In addition, RADAR can incorporate somatic variant recurrence, which has been widely used to discover key disease regions, to reweight different RBP peaks. Peaks with more somatic mutations than expected are often considered to be disease-driving[37-39]. Here, we first defined a local background somatic mutation rate from a large cohort of cancer patients to evaluate the mutation burden in each RBP peak (see details in methods). Variants that are associated with burdened elements are given higher priority in our scoring scheme.

Prioritizing variants with a RADAR weighted scoring scheme

By integrating the pre-built and user-specific data context described above, our entropy-based scoring scheme evaluated the functional impacts of variants that are specific to post-transcriptional regulation (Fig. 1 and Table 1). First, RADAR added up the (universal) score of variants for all pre-built features, which include sequence and structure conservation, network binding hub, RBP-gene association, and motif information. Then, depending on user inputs, RADAR further up-weighted variants' score based on tissue specific information from mutations in the key RBP binding sites, nearby genes with differential expression, or the RBP regulatory potential.

Table 1. Features used by RADAR

Category	Feature	Source	Scoring Scheme
Universal	Cross-population conservation	eCLIP	Adjusted-entropy
	Cross-species conservation	Gerp	Sigmoid Function
	Structural conservation	Evofold	Fixed Value
	RBP Binding hub	eCLIP	Adjusted-entropy
	RBP-gene association	shRNA RNA-seq	Fixed Value
	Motif disruption	Bind-n-Seq	Adjusted-entropy
		DREME	
User-specific	RBP regulatory potential	Expression	Fixed Value

	Differentially expressed Genes	Prior knowledge	Fixed Value
	Mutation Recurrence	Mutation profiles	Fixed Value

Applying RADAR to pathological germline variants

We calculated baseline RADAR scores on all pathological variants from HGMD (version 2015) and compared then with variant scores from 1,000 Genomes variants as a background. As expected, the HGMD variants scored significantly higher than somatic mutations (Supplementary Fig. S10). For example, the mean RADAR score for HGMD variants is 0.589, while it is only 0.025 for 1,000 genomes variants (P value $<2.2\times10^{-16}$ for one sided Wilcoxon test). We further compared the baseline RADAR scores of HGMD and 1,000 genomes variants within the RBP regulome to remove potential bias since HGMD variants may be more likely to be within or nearby to exons. We still observed significantly higher baseline RADAR score in the HGMD ones (1.871 vs. 1.337, P value $<2.2\times10^{-16}$ for one sided Wilcoxon test, Supplementary Fig. S10).

We further compared RADAR scores of HGMD variants with other methods. In total, 720 HGMD variants were explained by our methods but could not be highlighted by other methods (see details in methods, Supplementary Table S9). Many of these variants are located nearby the splice junctions. An example is shown in Fig. 5. This variant is located 4 base pairs away from splice junction in BRCA1. eCLIP experiments showed strong binding evidence in 5 RBPs (Fig. 5). Specifically, the T to C mutation strongly disrupts the binding motif of PRPF8, increasing the possibility of splicing alteration effects. Our finding is not highlighted in previous methods for variant prioritization, such as FunSeq, CADD, and FATHMM-MKL.

Applying RADAR baseline score to somatic variants in cancer

We next aimed to leverage our baseline RADAR scheme to evaluate the deleteriousness of somatic variants from public datasets. Due to the lack of a gold standard, we evaluated our results from two perspectives. First, we reasoned that since hundreds of cancer-associated genes are known to play essential roles through various pathways[40, 41], variants associated with these genes are likely to have the higher functional impact[23]. To test this hypothesis, we first selected variants within the 1kb region of the COSMIC[42] genes and compared them with other variants. We tested four cancer types, breast, liver, lung, and prostate cancer, and found in all cases that variants associated with COSMIC genes showed significant enrichment, with a larger RNA level functional impact (Fig. 5 and Supplementary Fig. S11). For example, we found a 4.58- and 8.75-fold increase in high-impact variants at a threshold level of 3 and 4, respectively, in breast cancer patients ($P < 2.2\times10^{-16}$, one-sided Wilcoxon).

In our second approach, we hypothesized that variant recurrence could be a sign of functionality and may indicate an association with cancer[19, 23, 24]. Thus, we compared the variants' scores from RBP binding peaks with or without recurrence. Specifically, we separated the RBP peaks with variants from more than one sample from those that were mutated in only one sample, and then compared the baseline RADAR scores. We found that in most cancer types, peaks with recurrent variants were associated with a larger fraction of high-impact mutations. For example, in breast cancer recurrent elements demonstrated a 1.67-, and 2.57-fold more high-impact variants with RADAR greater than 3.0 and 4.0, respectively, resulting in a P value of 2.2×10^{-16} in one-sided Wilcoxon test. We observed similar trend in most of the other cancer types (Supplementary Fig. S11).

A case study on breast cancer patients using disease specific features

We applied our method to a set of breast cancer somatic variants from 963 patients released by Alexandrov *et al.* [43]. We used COSMIC gene list, expression and mutational profiles as additional features. In total, we found that around 3% of the 687,517 variants could alter post-transcriptional regulation to some degree. We incorporated the above disease-specific features and demonstrated how they could help to reweight the variant scoring process on a coding variant in Fig. 6. This variant is located within an RBP binding ultra-hot region and showed high sequence conservations (7% more rare variants for its binding RBP). It also demonstrated strong motif disruption effect (PPIG in Fig 6). All such features resulted in a baseline RADAR score of 3.67, which is ranked 296 out of all variants. However, we found that it is located in the exon region of the well-known tumor suppressor TP53 (orange track in Fig.6), and its binding peaks demonstrated more than expected somatic mutations (purple in Fig. 6). Besides, 3 out of the 6 RBPs binding there showed high regulation potential in breast cancer (green in Fig. 6). Hence, these additional features boost its overall RADAR score to 6.67, which is ranked 38 out of all variants). In comparison, this variant only shows moderate scores for FunSeq2 (3) and CADD (7.46), and while it is scored in the top but showed much lower rank than RADAR.

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.

Discussion

In this study, we integrated the full catalog of eCLIP, Bind-n-Seq, and shRNA RNA-Seq experiments from ENCODE to build an RNA regulome for post-transcriptional regulation. Our defined RBP regulome is remarkably larger than one may think and covered up to 52.6 Mbp of the genome (Fig. 2A) and the majority of it was not covered by ith previous transcription-level annotations, such DHS, TFBS, and enhancers. We found that the RBP regulome demonstrated a noticeably higher conservation from two aspects: higher cross-species conservation in almost all annotation categories (Fig. 2C) and higher cross-population conservation by showing significant enrichment in rare variants (Fig. 3). These two sources of evidence support the notion that the RBP regulome is under strong purifying selection and carries out essential biological functions. In addition, these results signify the necessity of computational tools to annotate and prioritize variants in the RBP regulome.

By integrating a variety of regulator-, element-, and nucleotide-level features, we propose an entropy-based scoring frame, RADAR, to investigate the impact of somatic and germline variants. The variant prioritization framework of RADAR contains two parts. First, by incorporating eCLIP, Bind-n-Seq, shRNA RNA-seq experiments with conservation and structural features, we built a pre-defined data context to quantify the baseline variant impact score. This approach is suitable for multiple-disease analysis or cases where no other prior information can be used. We applied this RADAR baseline score to HGMD pathological variants and highlighted many candidates that cannot be highlighted by other methods. Besides, our RADAR framework provided detailed explanations of the underlying disease-causing mechanism (Fig.

5). In addition to the baseline score, RADAR also allows user-specific inputs such as prior knowledge, patient expression and mutation profiles for a re-weighting process to highlight relevant variants in a disease-specific manner. As an example, we applied the RADAR disease-specific scores to variants from several cancer types and showed that RADAR could identify relevant variants in key cancer-associated genes (Fig. 6).

It is important to note that as compared to ChIP-Seq experiments which generate peaks with up to kbp resolution, eCLIP experiments provide higher resolution functional site annotation (even single nucleotide resolution). Such accurate and compact annotation can greatly improve our variant function interpretations. We also want to mention that most of the current eCLIP peak calling approaches call peaks on the annotated transcribed regions. With the development of computational approached for eCLIP peak calling, we hope that the size of our annotated RBP regulome can be further expanded.

Conclusions

In summary, we have shown that RADAR is a useful tool for annotating and prioritizing post-transcriptional regulome for RBPs, which has not been covered by most of the current variant impact interpretation tools. Our method provides additional layers of information to the current gene regulomes. Importantly, the RADAR scoring scheme can be used in conjunction with existing transcriptional-level variant impact evaluation tools, such as FunSeq [23, 24], to quantify variant impacts. Given the fast-expanding collection of RBP binding profiles from additional cell types, we envision that our RADAR framework can better tackle the functional consequence of mutations from both somatic and germline genomes.

Methods

eCLIP Data Processing and Quality Control

We collected 318 eCLIP experiments of 112 unique RBPs from the ENCODE data portal (encodeprojects.org, released and processed by July 2017). eCLIP data was processed through the ENCODE 3 uniform data processing pipeline and peaks with score 1,000 were used in our analysis. We then removed peaks overlapped with blacklisted regions. We further separated the peaks into coding regions and the noncoding regions in our analysis to infer the selection pressure. 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).

Universal RADAR Score

Cross-population conservation inference

The cross-population conservation score consists of two components. The Shannon entropy considers the length effect of the RBPs while the selection pressure inference aims to determine the conservation of regions. For the Shannon entropy, for each RBP, we define f to be

$$f = \frac{n_{in}}{n_{total}} \tag{1}$$

where n_{in} represents the number of 1KG variants falling in that RBP, and n_{total} to be the total number of 1KG variants (fixed number). In this way, f takes into account the coverage of an RBPs binding site, since a larger coverage is more likely to have a larger value of n_{in}. The Shannon entropy is therefore equal to

$$S_f = 1 + f \times \log_2 f + (1 - f) \times \log_2 (1 - f) \tag{2}$$

We then calculate the selection pressure from the enrichment of rare germline variants from the 1,000 Genomes Project. Our analysis at each step is separated into coding and noncoding parts. For a given RBP, we suppose its binding peaks contain n_r rare variants ($DAF \leq 0.005$) and n_c common variants. The percentage of rare variants in that RBP's binding peaks is defined as

$$\rho = \frac{n_r}{n_r + n_c} \quad (1)$$

The value of ρ is often confounded by factors such as GC content or sequencing depth. In order to correct for potential GC content bias, we bin the genome into 500 base pair bins and group them according to their GC percentage. Then we compute the background rare variant percentage using the same rare and common variants from 1,000 Genomes Project for each group (See Supplement). For a given RBP with GC percentage g, we select the background group with closest GC, to obtain a background rare variant percentage ρ_b^g. Therefore, after adjusting for GC bias, the enrichment of rare variants is defined to be

$$\rho_{adj} = \rho / \rho_b^g \quad (3)$$

RBPs with a ρ_{adj} larger than 1 suggests a higher than expected selection pressure. We then adjust the species conservation entropy score as follows

$$S_{population\_conservation} = \rho_{adj} \times S_f \quad (4)$$

Given a variant falling in the RBP regulome that intersects a set of RBP eCLIP peaks, set P, the cross-species conservation score of that variant is equal to the maximum $S_{population\_conservation}$ for all RBPs in set P.

Cross-species conservation using GERP

We use GERP score as a way to measure the cross-species conservation. For each position, a GERP of greater than 2 is often used to define bases that are conserved. The transformation of GERP to a RADAR

component score, is adapted from Fu *et al*. Therefore, a sigmoidal transformation is used to fit the GERP scores between 0 and 1, and the parameters used force the curve to be sharp at GERP equal to 2 (see Supplement).

Structural Conservation

We use the output of Evo-fold as an indicator of cross species RNA structure. A variant falling in a region given by Evo-fold as conservative receives a score of 1 while a variant that does not fall in such region receives a score of 0.

RBP Binding Hubs and Networks

We define the number of RBPs binding at a position to as *H*. We first separated the RBP peaks into coding and noncoding regions and then grouped regions on the genome by *H*. For each group, we calculated the GC-corrected enrichment of rare variants ρ_{adj} for each group in coding and noncoding regions by equation (2). We also compute *f* for the set of regions associated with a fixed *H* (From equation 2). For each value of *f*, we compute the Shannon entropy from equation 2 to be S. We determine the hub numbers associated with hot and ultra-hot regions to be the number such that only 5% and 1% of 1KG variants fall in such regions, respectively. These cutoffs are defined to represent rare and ultra-rare events. Our values of ρ_{adj} are altered in such a way to reflect this phenomenon. The ρ_{adj} associated with hub scores less than that of the hot regions are converted to 0. The ρ_{adj} associated with hot regions demonstrate a mostly increasing trend and are smoothed using a kernel smoother. The ρ_{adj} of the ultra-hot region is kept constant, equal to the max ρ_{adj} of the hot region.

Finally, we multiply the values of the new ρ_{adj} by its respective Shannon entropy, S. A variant falling in the regulome with H would have a score equal to the $\rho_{adj}(H) * S$.

Motif Analysis and Disruption

We used the changes of PWMs introduced by a variant to quantify the motif disruptiveness effect through motiftools (https://github.com/hoondy/MotifTools). Specifically, we defined the D-score as in equation (3) to represent the difference between sequence specificities in reference to an alternative sequence.

$$Dscore = motifscore_{ref} - motifscore_{alt} = -10 \times \log_{10}\left(\frac{P_{ref}}{P_{alt}}\right) \quad (6)$$

where P_{ref} and P_{alt} are the PWM scores from the reference and alternative allele. To quantify a motif breaking event, we require that the P value for the reference allele is at least 5×10^{-4}. There are two motif sources in our analysis. First, we identified RBP motifs using DREME software (Version 4.12.0) directly from RBP peaks. Then, we also incorporated motifs from RNA Bind-N-Seq (RBNS) [18] to characterize sequence and structural specificities of RBPs. For each variant that affected multiple RBP binding profiles, we used the max score. A threshold of Dscore > 3 is used to describe a disruption event that is significant, and a variant having a Dscore less than this threshold receives a score of 0. For variants receiving a Dscore larger than the threshold, we additionally compute the Shannon entropy given for a variant with Dscore, *D*, as

$$S_{motif} = 1 + f(v, D) \times \log_2 f(v, D) + (1 - f(v, D)) \times \log_2(1 - f(v, D)) \quad (7)$$

where *f(*v, D) represents the number of 1KG variants, *v*, that have a Dscore greater than *D* divided by the total number of 1KG variants.

RBP-gene association using shRNA RNA-seq

To determine if an RBP, *R*, is associated with a gene, *g*, we intersect the peaks of *R*, with the transcript annotation of *g*. If an intersect exists, we form a linkage between the intersected peak of *R* and *g*. If some variant falls in that specific peak of *R*, the variant significantly disrupts the motif of RBP *R*, and gene *g* demonstrates at least a 2.5-fold change in its expression after KD of RBP *R*, we give the variant an additional score of 1.

Tissue Specific Score

RBP Regulatory Potential

RADAR allows inputs in addition to the pre-built context to calculate the disease-specific variant score. In this paper, we used the TCGA expression profiles as an example on the cancer variant prioritization. Specifically, we downloaded expression profiles of 19 cancer patients of 24 types from TCGA. In order to get a robust differential expression analysis, we excluded several cancer types that have less than 10 normal expression profiles and used DESeq2 [44] to find tumor-to-normal differentially expressed genes (corrected P from DESeq2 <0.05). Let y_i^k represent the differential expression status of gene i of the k^{th} cancer type.

We inferred the regulatory power of each RBP, R, through a regression approach of the above differential expression and RBP network connectivity as

$$\vec{\boldsymbol{y}}^{k,R} = {\beta_0}^{k,R} + {\beta_1}^{k,R}\vec{\boldsymbol{x}}^R + \varepsilon^{k,R} \quad (8)$$

where $\vec{\boldsymbol{x}}^R$ is the binary connectivity vector for all genes and *R* (1 if the gene is a target, else 0). We used the absolute value of ${\beta_1}^{k,R}$ to indicate the regulation potential of each RBP, R, in cancer type k. If a variant

falls in a region with at least one RBP binding, and at least one of the p-values associated with ${\beta_1}^{k,R}$ is significant, then we consider variants falling in that particular RBP to have an additional score of 1.

Recurrence in Somatic Mutations

We prioritized variants in RBP binding sites are with more-than-expected somatic mutations. In order to evaluate the somatic mutation burden, we first separated the genome into 1Mbp bins and calculated a local background mutation rate in each window. Then for each eCLIP peak, we counted the number of somatic mutations, and compared it to the nearest local 1Mbp context using a one-sided binomial test. If a specific RBP binding site was enriched for somatic mutations, the variant falling in that site was given a higher priority.

Differentially Expressed Key Genes

For each peak of each RBP, we find the associated gene of that peak by intersecting with the Gencode gene definitions. Using the DESeq2 results, we consider genes with q-values that are less than 0.05 to be differentially expressed. If an RBP peak is associated with a gene that is significantly differentially expressed in a tissue type, we increase the score of the variant falling in such peak by 1.

Accessibility Availability of data and materials

We have made this RNA variant annotation and prioritization tool available as an open-source software at radar.gersteinlab.org. The website contains details on usage, examples, resources, and dependencies. We also provided a genome-wide pre-built RADAR baseline score for every base pair on the genome (hg19 version of genome). Users can directly query the annotation and functional impact score from radar.gersteinlab.org. Also, we released all pre-built data context and genome-wide baseline scores at radar.gersteinlab.org.

List of abbreviations

RBP: RNA binding Protein; ENCODE: the Encyclopedia of DNA Elements; PWM: position weight matrix; eCLIP: enhanced crosslinking and immunoprecipitation; RADAR: ***R***N***A*** Bin***D***ing Protein regulome ***A***nnotation and p***R***ioritization; UTR: untranslated regions; DAF: derived allele frequency; RBNS: RNA Bind-N-Seq.

Competing interests

The authors declare that they have no competing interests

Authors' contributions

JZ, JL and MG conceived the study and wrote the manuscript. JZ, JL, DL, LL, JF wrote the framework and performed the method evaluation. DL developed the website. LL and MR carried out studies associating RNA structure. DL and SL participated in motif analysis. All authors have read and approved the final manuscript.

Acknowledgements

This work was supported by the National Institutes of Health (grant number 1U24HG009446-01), AL Williams Professorship, and in part by the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center. We also thank Brenton Graveley, Gene Yeo, Peter Freese, and Eric Van Nostrand for useful discussion on eCLIP and RBNS data.

References

1. Croce CM: **Causes and consequences of microRNA dysregulation in cancer.** *Nat Rev Genet* 2009, **10:**704-714.
2. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G: **Epigenomics: Roadmap for regulation.** *Nature* 2015, **518:**314-316.
3. Yang G, Lu X, Yuan L: **LncRNA: a link between RNA and cancer.** *Biochim Biophys Acta* 2014, **1839:**1097-1109.
4. Schmitt AM, Chang HY: **Gene regulation: Long RNAs wire up cancer growth.** *Nature* 2013, **500:**536-537.
5. Gerstberger S, Hafner M, Tuschl T: **A census of human RNA-binding proteins.** *Nat Rev Genet* 2014, **15:**829-845.
6. van Kouwenhove M, Kedde M, Agami R: **MicroRNA regulation by RNA-binding proteins and its implications for cancer.** *Nat Rev Cancer* 2011, **11:**644-656.
7. Swinburne IA, Meyer CA, Liu XS, Silver PA, Brodsky AS: **Genomic localization of RNA binding proteins reveals links between pre-mRNA processing and transcription.** *Genome Res* 2006, **16:**912-921.
8. Dreyfuss G, Kim VN, Kataoka N: **Messenger-RNA-binding proteins and the messages they carry.** *Nat Rev Mol Cell Biol* 2002, **3:**195-205.
9. Fu XD, Ares M, Jr.: **Context-dependent control of alternative splicing by RNA-binding proteins.** *Nat Rev Genet* 2014, **15:**689-701.
10. Zheng D, Tian B: **RNA-binding proteins in regulation of alternative cleavage and polyadenylation.** *Adv Exp Med Biol* 2014, **825:**97-127.

11. Fossat N, Tourle K, Radziewic T, Barratt K, Liebhold D, Studdert JB, Power M, Jones V, Loebel DA, Tam PP: **C to U RNA editing mediated by APOBEC1 requires RNA-binding protein RBM47.** *EMBO Rep* 2014, **15:**903-910.
12. Glisovic T, Bachorik JL, Yong J, Dreyfuss G: **RNA-binding proteins and post-transcriptional gene regulation.** *FEBS Lett* 2008, **582:**1977-1986.
13. Li JH, Liu S, Zhou H, Qu LH, Yang JH: **starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data.** *Nucleic Acids Res* 2014, **42:**D92-97.
14. Blin K, Dieterich C, Wurmus R, Rajewsky N, Landthaler M, Akalin A: **DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation.** *Nucleic Acids Res* 2015, **43:**D160-167.
15. Anders G, Mackowiak SD, Jens M, Maaskola J, Kuntzagk A, Rajewsky N, Landthaler M, Dieterich C: **doRiNA: a database of RNA interactions in post-transcriptional regulation.** *Nucleic Acids Res* 2012, **40:**D180-186.
16. Hu B, Yang YT, Huang Y, Zhu Y, Lu ZJ: **POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins.** *Nucleic Acids Res* 2017, **45:**D104-D114.
17. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, et al: **Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP).** *Nat Methods* 2016, **13:**508-514.
18. Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB: **RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins.** *Mol Cell* 2014, **54:**887-900.
19. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants.** *Nat Genet* 2014, **46:**310-315.
20. Ritchie GR, Dunham I, Zeggini E, Flicek P: **Functional annotation of noncoding sequence variants.** *Nat Methods* 2014, **11:**294-296.
21. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A: **Distribution and intensity of constraint in mammalian genomic sequence.** *Genome Res* 2005, **15:**901-913.
22. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: **Identifying a high fraction of the human genome to be under selective constraint using GERP++.** *PLoS Comput Biol* 2010, **6:**e1001025.
23. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M: **FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer.** *Genome Biol* 2014, **15:**480.
24. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, et al: **Integrative annotation of variants from 1092 humans: application to cancer genomics.** *Science* 2013, **342:**1235587.
25. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015, **526:**68-74.
26. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491:**56-65.
27. Garner C: **Confounded by sequencing depth in association studies of rare alleles.** *Genet Epidemiol* 2011, **35:**261-268.

28. Xu C, Nezami Ranjbar MR, Wu Z, DiCarlo J, Wang Y: **Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller.** *BMC Genomics* 2017, **18:**5.
29. Lu Y, Liu P, James M, Vikis HG, Liu H, Wen W, Franklin A, You M: **Genetic variants cis-regulating Xrn2 expression contribute to the risk of spontaneous lung tumor.** *Oncogene* 2010, **29:**1041-1049.
30. Davidson L, Kerr A, West S: **Co-transcriptional degradation of aberrant pre-mRNA by Xrn2.** *EMBO J* 2012, **31:**2566-2578.
31. Mortimer SA, Kidwell MA, Doudna JA: **Insights into RNA structure and function from genome-wide studies.** *Nat Rev Genet* 2014, **15:**469-479.
32. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D: **Identification and classification of conserved RNA secondary structures in the human genome.** *PLoS Comput Biol* 2006, **2:**e33.
33. Khurana E, Fu Y, Chen J, Gerstein M: **Interpretation of genomic variants using a unified biological network approach.** *PLoS Comput Biol* 2013, **9:**e1002886.
34. Jiang P, Freedman ML, Liu JS, Liu XS: **Inference of transcriptional regulation in cancers.** *Proc Natl Acad Sci U S A* 2015, **112:**7731-7736.
35. Consortium GT: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45:**580-585.
36. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, et al: **Genetic effects on gene expression across human tissues.** *Nature* 2017, **550:**204-213.
37. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al: **Mutational heterogeneity in cancer and the search for new cancer-associated genes.** *Nature* 2013, **499:**214-218.
38. Lochovsky L, Zhang J, Fu Y, Khurana E, Gerstein M: **LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations.** *Nucleic Acids Res* 2015, **43:**8123-8134.
39. Lochovsky L, Zhang J, Gerstein M: **MOAT: Efficient Detection of Highly Mutated Regions with the Mutations Overburdening Annotations Tool.** *Bioinformatics* 2017.
40. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nat Med* 2004, **10:**789-799.
41. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M: **Role of non-coding sequence variants in cancer.** *Nat Rev Genet* 2016, **17:**93-108.
42. Forbes SA, Beare D, Bindal N, Bamford S, Ward S, Cole CG, Jia M, Kok C, Boutselakis H, De T, et al: **COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer.** *Curr Protoc Hum Genet* 2016, **91:**10 11 11-10 11 37.
43. Alexandrov LB, Stratton MR: **Mutational signatures: the patterns of somatic mutations hidden in cancer genomes.** *Curr Opin Genet Dev* 2014, **24:**52-60.
44. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15:**550.

Figure Legends and Tables

Figure 1. RADAR workflow

There are two RADAR score components: (1) it first defines baseline component using pre-built data context including sequence and structural conservation, network and motif information; (2) it then incorporates user-specific input to further highlight tissue-specific variants.

Figure 2. RBP regulome and cross-species conservation.

(A): length of RBP binding sites vs. other coding/noncoding annotation categories; (B) Fraction of RBPs falling into each annotation category and boxplot of PhastCons scores in peak (blue) vs. nonpeak regions (white).

Figure 3. Cross-population conservation of RBP peaks and binding hubs.

(A-B): Rare variant percentage in coding/noncoding regions. The blue dot represents RBP peaks, and yellow dot represents genome average after GC correction. Shaded lines are the 95% confidence interval of the rare variant percentage of the RBP peaks; (C): an example of RBP binding hubs. Red/orange dots denote positions with top 1 and 5 RBPs binding; (D): corrected rare variant percentage at positions with different binding RBPs.

Figure 4. Regulation potential inference of RBPs

(A): schematic of RBP regulation potential calculation; (B): Heatmap of RBP regulation potential in 19 cancer types; (C): RBPs associated with patient survival.

Figure 5. Baseline RADAR score on germline and somatic variants

(A) an example of BRCA1 intron HGMD variant highlighted by RADAR baseline score. This variant breaks the motif of the splicing factor PRPF8 (red); (B) enrichment of high RADAR baseline score variants associated with COSMIC genes in breast cancer; (C) enrichment of high RADAR baseline score variants within RBP peaks with recurrent variants in breast cancer.

Figure 6. Example of breast cancer somatic variant with high overall RADAR score

We selected an exonic variant with high RADAR score on chromosome 17 as an example. It was inside an RBP binding hub with a high Gerp score and broke the motif of PPIG. It also has several tissue-specific features, like within the well-known cancer-associated gene TP53 (orange track) and its associated binding peaks were significantly burdened (purple). Besides, by adding expression profiles from TCGA, we found that 3 out of the 6 RBPs binding there demonstrated high regulation potentials in driving tumor-specific expression pattern. All these external pieces of evidence further boost this variant's tissue-specific score.