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

Table of Contents

List of supplementary figures

1 Defining the RBP regulome using eCLIP data

1.1 *Functional Annotation of RBPs*

eCLIP is an enhanced version of the crosslinking and immunoprecipitation (CLIP) assay, and is used to identify the binding sites of RNA binding proteins (RBPs). We collected all available eCLIP experiments from the ENCODE data portal (encodeprojects.org). There were 178 experiments from K562 and 140 experiments from HepG2, totaling 318 eCLIP experiments from all available ENCODE cell lines (released and processed by July 2017).

These experiments targeted 112 unique RBP profiles. eCLIP data was processed by ENCODE 3 uniform data processing pipeline. The eCLIP peak calling method and processing pipeline were developed by Gene Yeo's lab at the University of California, San Diego (https://github.com/YeoLab/clipper, CLIP-seq cluster-identification algorithm[1]). For each peak, the enrichment significance was calculated against a paired input, and we filtered those peaks with a flag of 1000, which are considered to be the statistically significant peaks.

We summarized the list of available RBPs in Table S1 (in separate data package) and provided detailed annotation as we can. We also summarized different categories of RBPs in Figure S 1.

Figure S 1 Annotation summary of RBPs

- 61.9% RNA Binding
- 0.7% RNA Splicing Regulat
- 0.7% tRNA Splicing
- 1.5% RNA Editing
- 1.5% Structural Constituent
- 2.2% tRNA Binding
- 3% Regulation of mRNA Stabilit
- 3% mRNA Polyadenylation
- 3.7% rRNA Processing
- 1.6% Splicing via Spliceosome
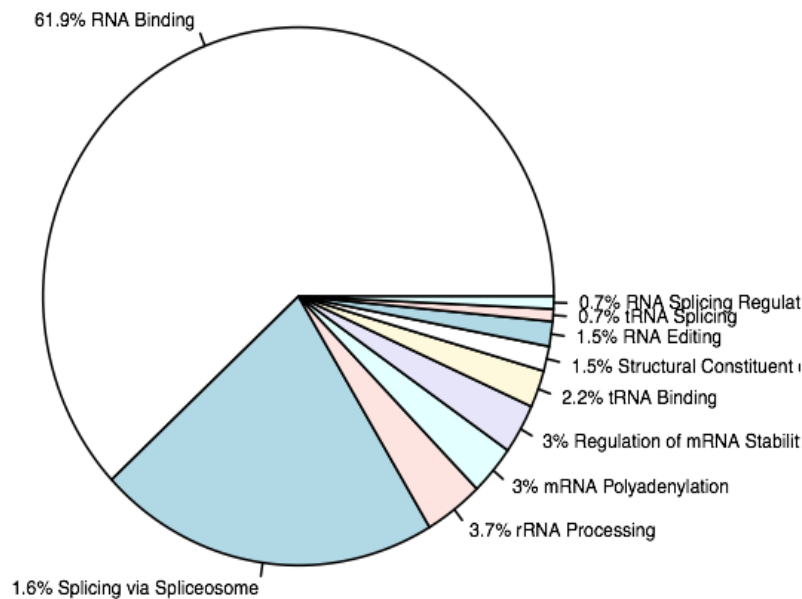
1.2   *Functional Annotation of RBP binding sites*

From the raw peaks from ENCODE, we further removed the ones overlapped with either blacklist regions from ENCODE (https://www.encodeproject.org/annotations/ENCSR636HFF/, select hg19) or gap regions like Telomere and Centromere from ucsc (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/gap.txt.gz). In total, over 99% of the binding locations are preserved after blacklist removal.

We further tried to annotate these peak regions by dividing them into different annotation categories from Gencode V19. Specifically, we extracted 7 different annotation categories, including coding exons, 3'UTR, 5'UTR, 3'UTR extended (1000bp downstream), 5'UTR extended (1000bp upstream), nearby intron (up to 100bp to the exon/intron junctions), and deep introns. For any region that might overlap two annotation categories, we only preserve one in the order mentioned above. The raw number of nucleotides in each annotation category is given in Table S2.

Table S2. RBP binding peaks within annotated regions

Annotation Type	Nucleotides
Coding Exon	156069
3' UTR	65447
5' UTR	28339
3' UTR extended	39985
5' UTR extended	45036
Nearby Intron	102892
Deep Intron	312424

1.3   *Inference of cross-population conservation of RBP binding sites*

We tried to infer the cross-population conservation level of the RBP binding sites from polymorphism data in large sequencing cohorts like the 1,000 Genomes Project. Specifically, for each RBP we divided all the binding peaks into coding and noncoding regions separately and then calculated the number of common (n_c) and rare variants (n_r) in these two categories. Then a one-sided binomial test of $n_c, n_r,$ vs. the genome background f was calculated to evaluate the enrichment of rare variants.

However, in our analysis we found that GC content might be a potential bias in such calculation. As in Figure S 2, the background rare variant percentage f demonstrates noticeable changes with GC percentage. One possible explanation is that GC content usually affects read coverage in high-throughput sequencing experiments, which is a sensitive parameter in the downstream variant calling process. Therefore, to remove such bias, we calculated the GC

adjusted background rare variant percentage by dividing the coding/noncoding regions into 500bp bins, and grouping these bins at GC resolution of 0.02. For each RBP, when calculating the background, we only select the bins with closest GC percentage. The comparison of foreground and background rare variant percentage for every RBP in coding and noncoding regions are given in Figure S 3.

Figure S 2. Background Rare variant percentage vs. GC
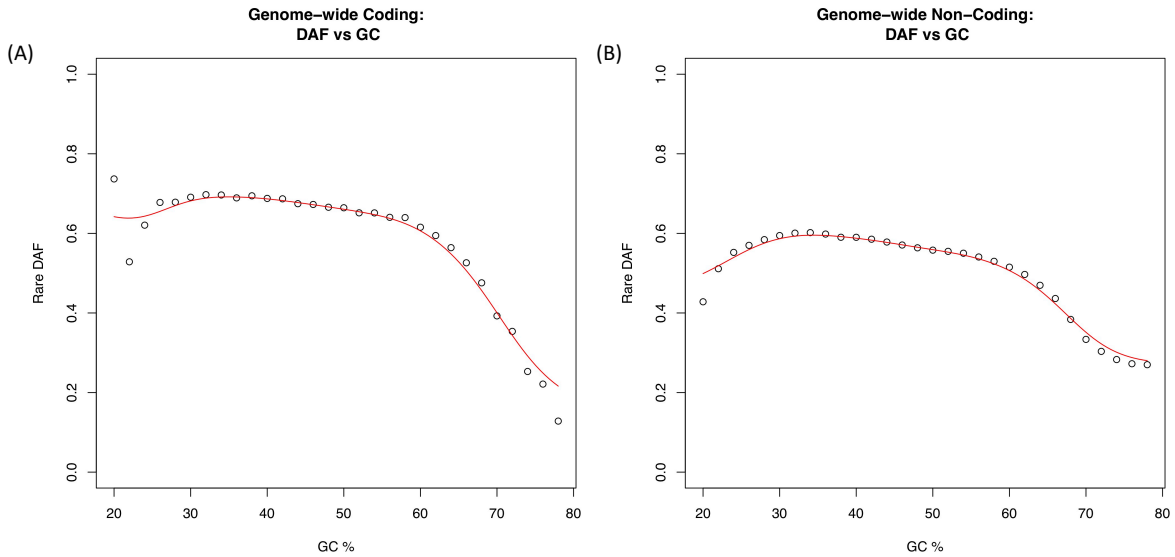
For some RBPs, if there are no coding/noncoding rare/common variants in their binding sites, the f value for binomial test will be missing. We provided the full raw calculation of GC corrected rare variant enrichment for each RBP in Table S3.

Figure S 3. Rare variant enrichment after GC correction in coding and noncoding regions respectively. The dashed blue/red line is the genome average without GC correction for coding and noncoding regions, and the solid blue/red line is the background after GC correction. Blue/Red star on top of each bar indicate significantly enriched in rare variants after GC correction in one-sided binominal test against the coding/noncoding average.

5

Rare DAF%, Non-Coding
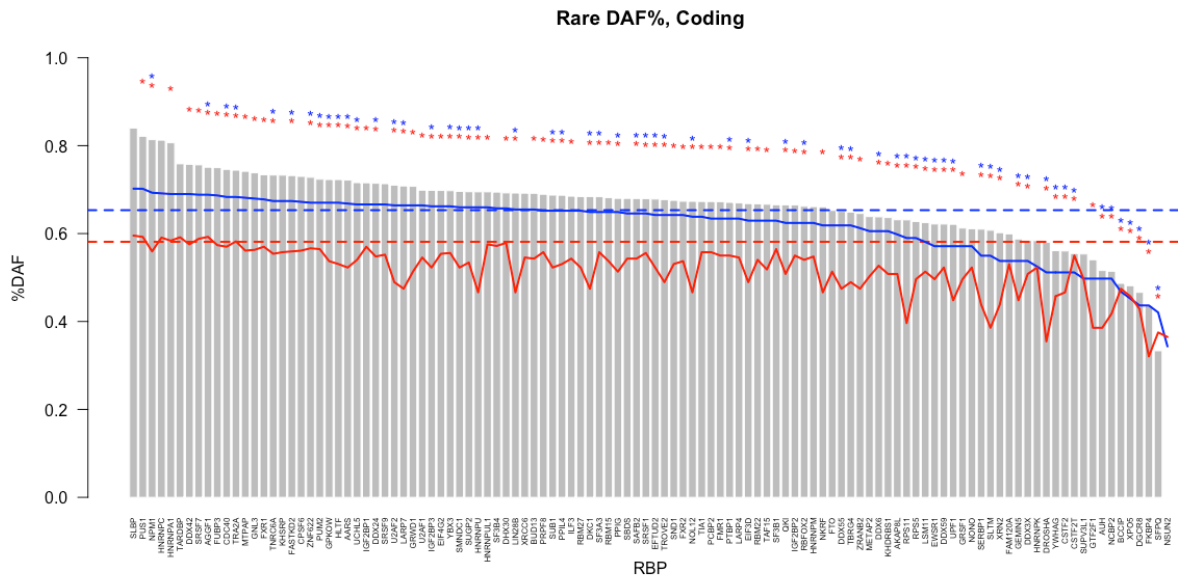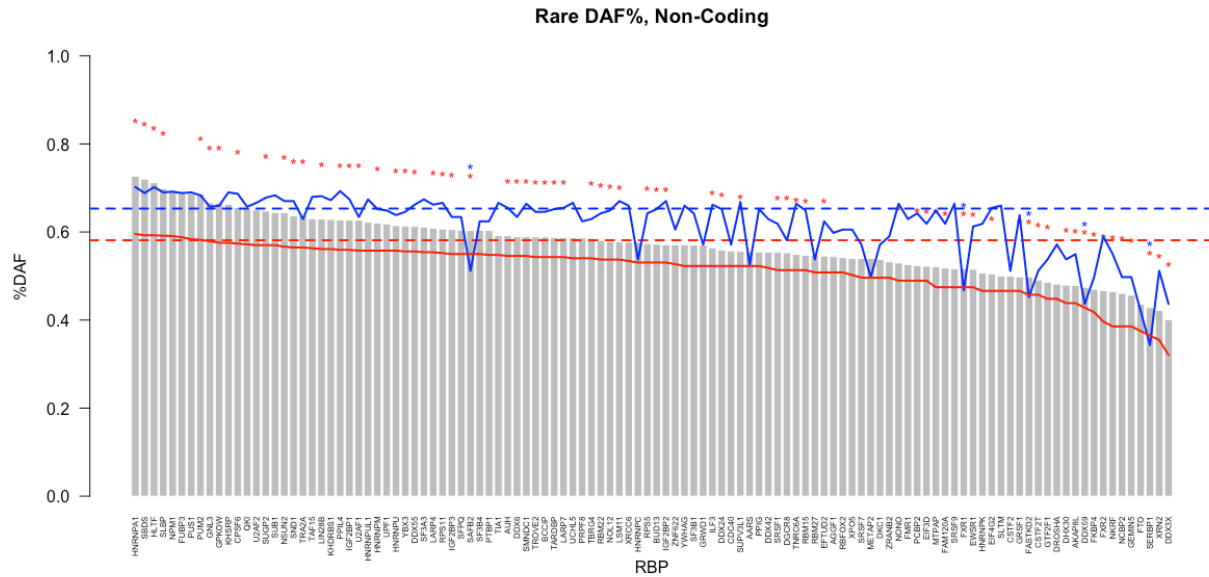
Inference of cross-species conservation of RBP binding sites

PhastCons conservation scores were downloaded from UCSC genome browser. For each annotation category (coding exons, 3'UTR, 5'UTR, nearby introns), we separate the annotation into regions covered by RBP peak and those not covered. After deduplication and merging of the bed files, we then calculated the average PhastCons score in each region using bigWigAverageOverBed (downloaded from UCSC genome browser). Then the boxplots of peak vs. nonpeak regions were given in Figure S 2 in the main manuscript.

1.4   Inference of structure conservations

We downloaded the Evofold bed files for hg19 from UCSC Genome Browser and used it as a feature for our analysis. Specifically, we found that after requiring that any RBP peaks should also be with conserved structure in Evofold, these binding sites significantly increases its population-level conservations (as shown in Figure S 4).
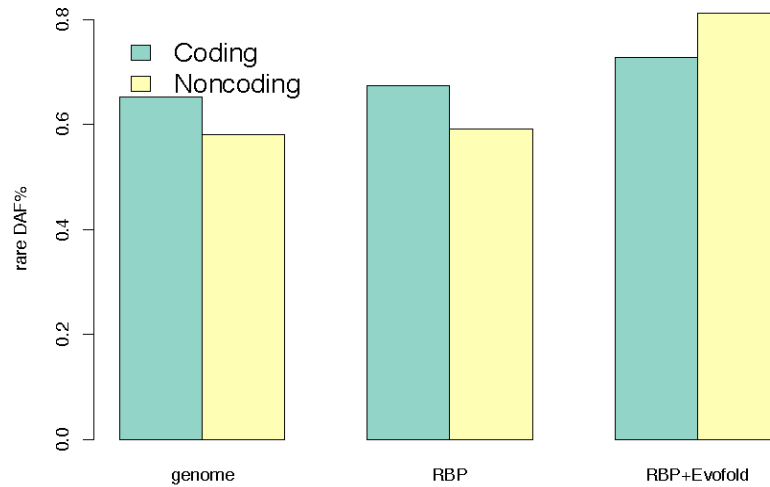
2 RBP binding network analysis

We also investigated the RBP binding events interactions from two aspects: co-binding analysis and RBP binding hub analysis. Details are given in the following sections.

2.1 *RBP co-binding analysis*

We defined the co-binding percent of each RBP pair by the ratio of overlapping nucleotides over the union of nucleotides in their binding peaks. Then we constructed a co-binding percentage matrix for all RBPs to measure their co-binding status. Then, we performed a hierarchical clustering of this matrix by the "pvrect" package in R with an alpha value of 0.02 to identify the co-binding pairs. The resulting clusters of RBPs with significance were found to follow patterns of functional co-binding found in literature and results are given in Figure S 5.
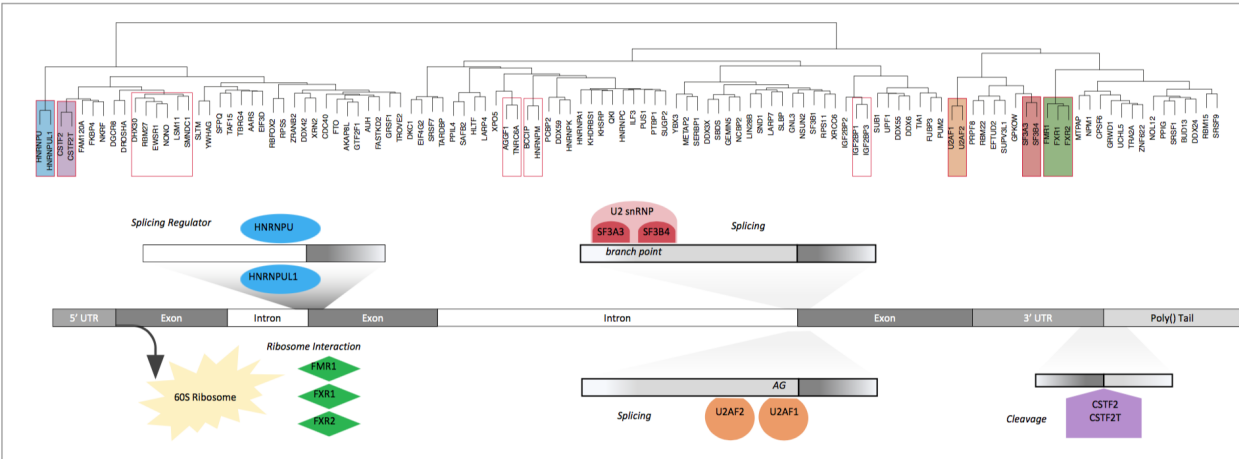
Using a significance level threshold of 0.05, we found several pairs of well-known regulatory partners with different binding preferences. For example, the famous heterogeneous nuclear ribonucleoprotein (hnRNP) family protein HNRNPU and its paralog HNRNPUL1 were found to bind together in the nearby intron region, probably regulating the pre-mRNA splicing process[2]. F3A3 and SF3B4, which encode two units of splicing factor 3a protein, were also found to co-bind in the nearby intron region in our data[3, 4]. The SR family protein U2AF1 and U2AF2 are found to co-bind near the intron/exon junctions to jointly control splicing events[4, 5]. Two cleavage stimulation factor (CSTF) complex proteins, CSTF2 and CSTF2T, were found to bind near the 3' UTR, and were reported to be associated with 3' end cleavage and polyadenylation of pre-mRNAs. Consistent with previous report, three functional similar genes FMR1, FXR1, and FXR2 were found to co-express, and shuttle between the nucleus and cytoplasm and associate with polyribosomes, predominantly with the 60S ribosomal subunit[6, 7]. The discovery of the co-binding of such functional relevant proteins at various regions indicates the high quality of our regulome.

2.2 *RBP network hub analysis*

We also inferred the RBP binding hubs and hypothesized that they are under higher negative selection since once mutated, there is a higher chance to alter RBP regulations. Specifically, we calculated the number of RBPs that bind to each nucleotide and the distribution is given in Figure S 6. As expected, due to the specificity of RBP binding events, the majority (over 60%) of the RBP regulome was surrounded by only one RBP.
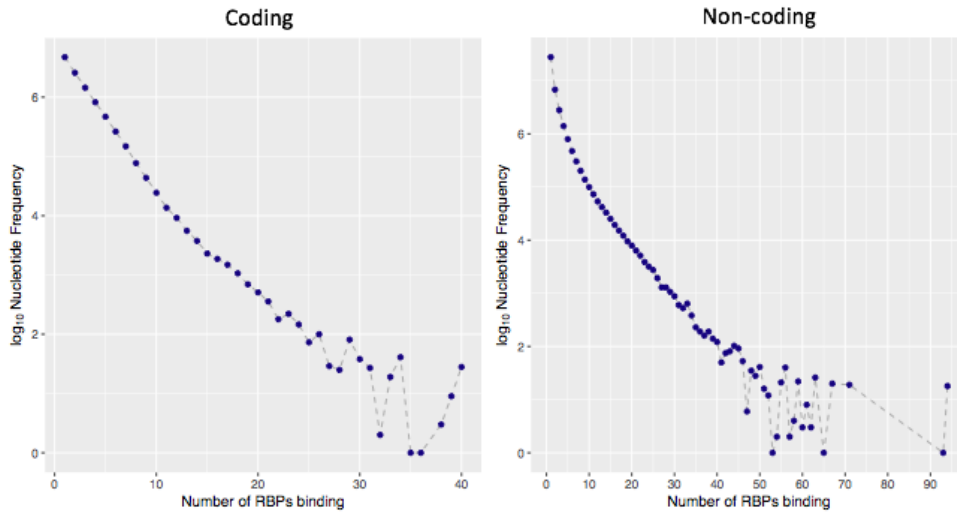
We then calculated the enrichment of rare variants for regions with at least 1, 2, 3, …, 112 RBPs. We corrected the GC bias in a similar way to section 1.3. As expected, as the number of RBPs increased, we observed an obvious trend of enrichment of rare variants (Figure S 7 and Figure S 8). For instance, in the noncoding region, around 5% of the regulome is surrounded with at least 5 RBPs, and they exhibited 3% more rare variants compared to the whole genome. For regions that are surrounded by at least 10 RBPs, which are around 1% of the whole regulome, we observed up to 12% more rare variants (Figure S 8). This observation significantly supports our hypothesis that the RNA regulome hubs are under stronger purifying selection, and should be given higher priority when evaluating the functional impacts of mutations.

Figure S 7. Adjusted rare variant percentage vs. number of RBPs binding in coding regions. Regions with top 5% and 1% of RBPs binding are defined as the hot and ultra-hot regions.
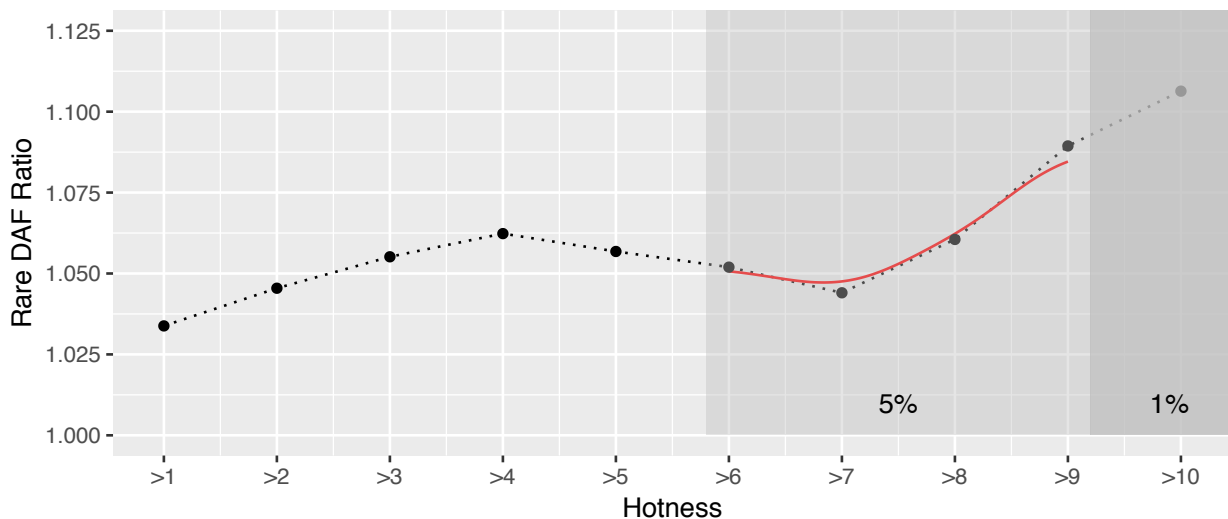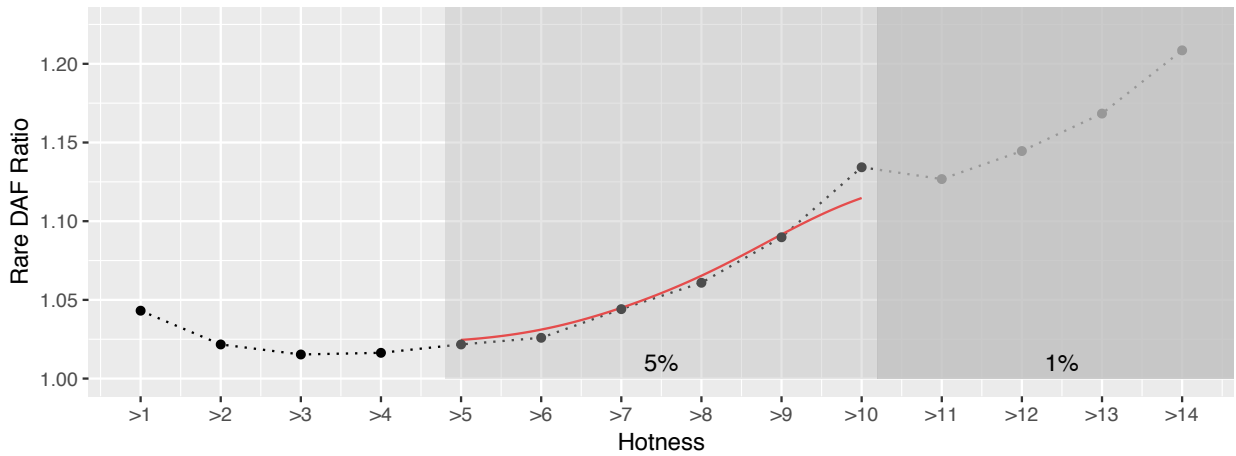
3 Motif analysis

In our RADAR framework, we incorporated two sources of motifs: (1) motifs from RNA Bind-n-Seq experiments[8]; (2) *de novo* discoveries from RBP peaks by DREME [9]. For each variant, we used the changes of PWM scores to quantify the binding affinity alterations. If one variant breaks more than one PWM, RADAR will choose the maximum score for it.

3.1 *Motifs from RNA Bind-n-Seq experiments*
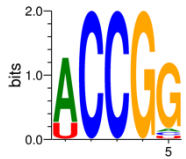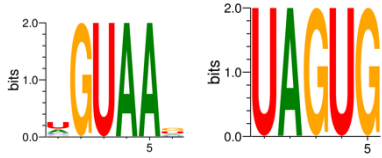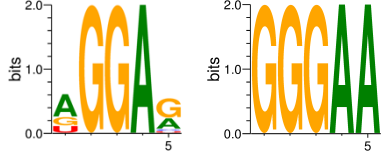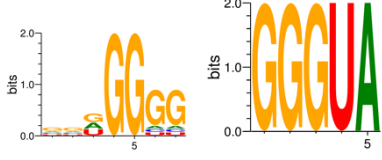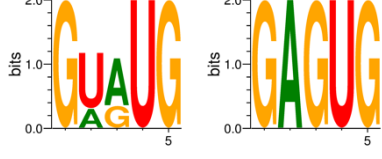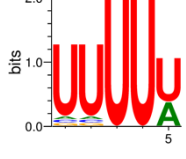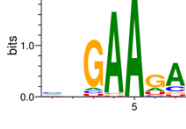
It has been reported that many of the RBPs' binding events *in vivo* can be captured by binding preferences *in vitro*. Hence, we utilized an in vitro RNA binding assay, RNA Bind-N-Seq[8] to characterize sequence and structural specificities of RBPs. We used RBNS motifs from 78 human RBPs to prioritize germline and somatic variants that could potentially disrupt an RNA-binding domain.

Briefly, we called on RBNS motifs based on an enrichment Z-score cutoff of 3. Some RBPs had up to four motifs, which ranged from 5-mer to 9-mers. In total, there are 17 RBPs overlapped with eCLIP RBPs, which are listed in Table S4 below. We treated all RBNS motifs independently from eCLIP-based *de novo* motifs.

Table S4 List of RBPs that have both eCLIP and Bind-n-Seq experiments

	RBP Name	RBNS motif
1	EIF4G2	

2	EWSR1	
3	FUBP3	
4	HNRNPC	
5	HNRNPK	
6	IGF2BP1	
7	IGF2BP2	
8	KHSRP	
9	PCBP2	
10	RBFOX2	

11	RBM22	
12	SFPQ	
13	SRSF9	
14	TAF15	
15	TARDBP	
16	TIA1	
17	TRA2A	

3.2 *Motifs from de novo discovery*

We collected the binding peaks for each RBP after blacklist removal. For any peak that is less than 150 bp in length, we extended it to 150 bp from both sides. For those longer than 150bp, we kept the original peak length. We then extracted sequence information from hg19 and performed *de novo* motif discovery DREME[9] with default settings (Version 4.12.0, http://meme-suite.org/tools/dreme).

3.3 *Motif disruption calculation using MotifTools*

We used D-score defined in MotifTools (https://github.com/hoondy/MotifTools) to evaluate the binding affinity alterations introduced by a variant. We only considered positive D-scores, which denote a variant that decreases the likelihood that a protein will bind the motif (motif-break). For variants that affect multiple RBP binding profiles, we used the max score over all D-scores.

4 RBP-gene association by RBP KD experiments

RNA-seq expression profiling before and after shRNA mediated RBP depletion from ENCODE can help infer the gene expression changes introduced by RBP knockdown. Variants with disruptive effect on RBP binding may affect or even completely remove RBP binding and hence affect gene expressions in a similar way. A schematic of our procedure is given in Figure S 9.

Specifically, we first collected 472 shRNA RNA-seq experiments (Table S5) and extracted the differentially expressed genes (Table S6) from such experiments. For example, in Figure S 9, we define the G1-RBP2 association from the RBP knockdown experiment. Then within the extended G1 region, we extracted all motif breaking variant effect for all possible RBPs (within peaks). If any variant breaks RBPs that has an association with G1, we give it an extra credit in our baseline score.

Figure S 9. Schematic of highlighting variants that breaks gene-RBP association from RBP knockdown experiments.

13
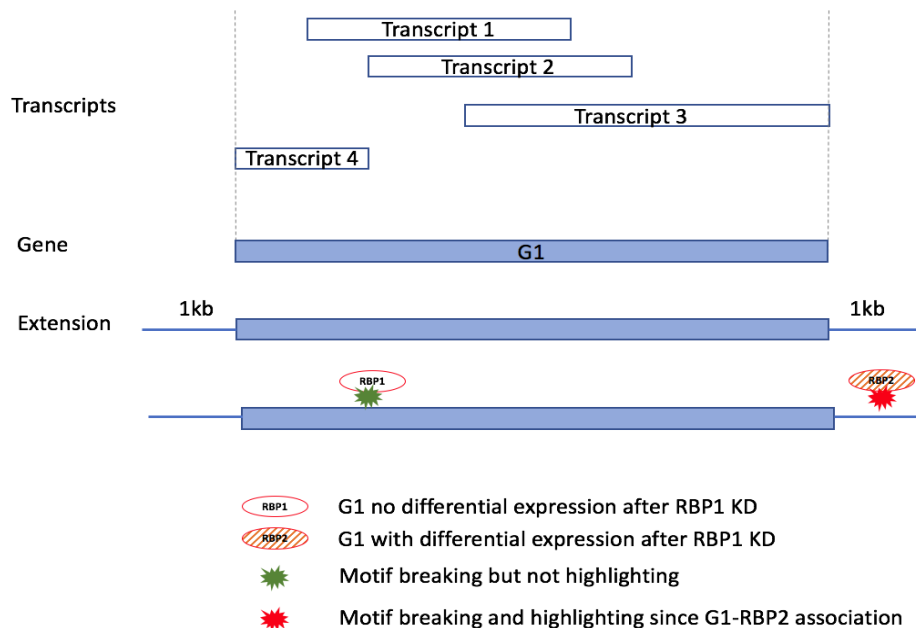
5. Highlighting key regulators through expression profiles

In order to detect the key RBP regulators that drive the disease-specific gene expression patterns, we constructed RBP regulatory networks and incorporated gene expression profiles to find RBPs that are associated with expression changes in patient cohorts.

Specifically, we first downloaded the full set of TCGA expression profiles for 24 cancer types. In order to get a robust differential expression analysis, we excluded 6 cancer types that have less than 10 normal expression profiles. For each cancer type, both tumor and normal expression were given to DESeq2[10] to identify tumor-specific gene differential expression status.

Then we tried to set up RBP regulatory network directly from the RBP peaks. We used the full set of protein coding genes in Gencode v19, and then extracted their 3'UTR regions. For any protein coding gene, a RBP is supposed to regulate this gene if this RBP has a binding peak intersecting the 3'UTR region.

We inferred the regulation power of each RBP by through a regression approach of the above differential expression status and RBP network connectivity. We used the absolute value of regression coefficient as the aggregated RBP regulation power. The full table of regulation powers in all 19 cancer types were given in Table S7. Interestingly, we found that for the RBPs with larger regulation power are those tends to be known to associated with cancer, as listed in Table S8.

For RBPs with high regulation powers, we also performed a patient-wise regulation power inference, where the differential expression is determined as the individual expression fold change. Then, we tried to use such individual regulatory power to predict 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).

6. Applying RADAR to pathological germline variants

HGMD variants (version 2015) were used in our analysis. For Figure 5, the signal tracks for the eCLIP experiments were directly downloaded from ENCODE. Funseq2 and CADD scores were directly calculated from their website. The list of highly prioritized variants discovered only by RADAR were provided in Table S9. The comparison of RADAR baseline scores of HGMD vs 1kg variants were given in Figure S 10. Since the majority of 1kG variants are located far away from the exon regions, we further extracted variants that are only inside the RBP regulome for both HGMD and 1kG variants and compared their RADAR baseline scores.
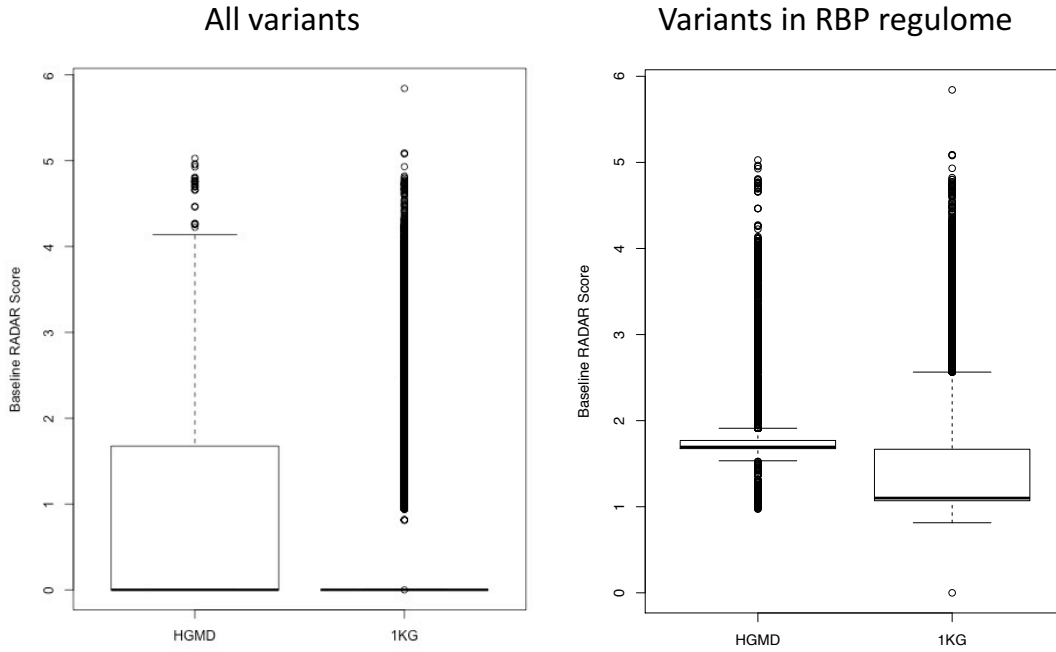
7. Applying RADAR to somatic variants in cancer

The breast, liver, lung, and prostate cancer variants were downloaded from the paper by *Alexandrov et al[11]*. We first calculated the baseline RADAR scores on these four cancer types. We found that in most cancer types, COSMIC genes and recurrent RBP peaks are associated with more high impact variants. Results are shown in Figure S 11.

We also used expression profiles were downloaded from TCGA and the mutational variants as disease-specific features to prioritize breast cancer variants. Several more interesting examples from breast cancer were given in the following figures. We have listed the somatic variants that are only highlighted by RADAR in Table S10 and Table S11.

Figure S 12. Highlighted breast cancer somatic variants in 3'UTR region
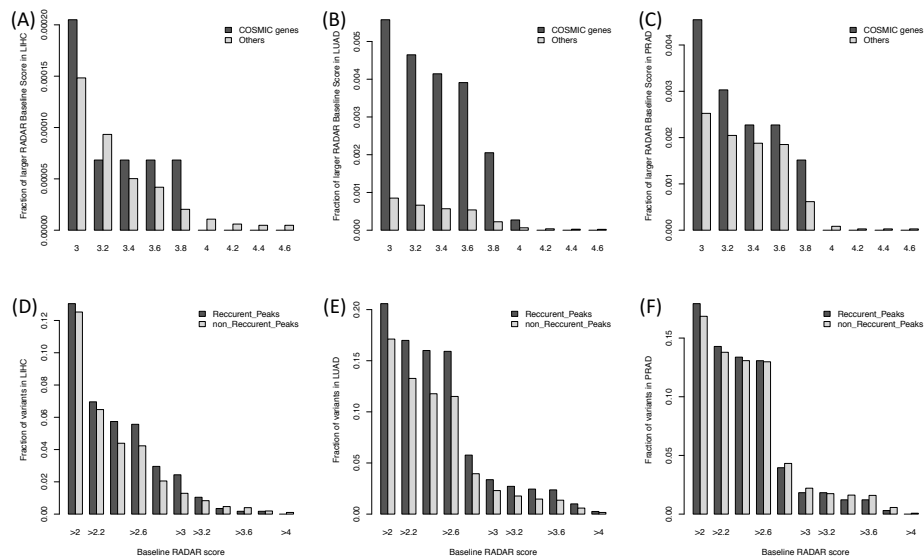
1. Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, et al: **Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges.** *Nat Struct Mol Biol* 2013, **20:**1434-1442.
2. Ye J, Beetz N, O'Keeffe S, Tapia JC, Macpherson L, Chen WV, Bassel-Duby R, Olson EN, Maniatis T: **hnRNP U protein is required for normal pre-mRNA splicing and postnatal heart development and function.** *Proc Natl Acad Sci U S A* 2015, **112:**E3020-3029.
3. van Roon AM, Oubridge C, Obayashi E, Sposito B, Newman AJ, Seraphin B, Nagai K: **Crystal structure of U2 snRNP SF3b components: Hsh49p in complex with Cus1p-binding domain.** *RNA* 2017, **23:**968-981.
4. Lin PC, Xu RM: **Structure and assembly of the SF3a splicing factor complex of U2 snRNP.** *EMBO J* 2012, **31:**1579-1590.
5. Obeng EA, Ebert BL: **Charting the "Splice" Routes to MDS.** *Cancer Cell* 2015, **27:**607-609.
6. Rousseau F, Labelle Y, Bussieres J, Lindsay C: **The fragile x mental retardation syndrome 20 years after the FMR1 gene discovery: an expanding universe of knowledge.** *Clin Biochem Rev* 2011, **32:**135-162.
7. Crawford DC, Acuna JM, Sherman SL: **FMR1 and the fragile X syndrome: human genome epidemiology review.** *Genet Med* 2001, **3:**359-371.
8. 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.
9. Bailey TL: **DREME: motif discovery in transcription factor ChIP-seq data.** *Bioinformatics* 2011, **27:**1653-1659.
10. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15:**550.
11. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al: **Signatures of mutational processes in human cancer.** *Nature* 2013, **500:**415-421.