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[1], Jo-Jo Feng[3], Shaoke Lou[1,2], Michael Schoenberg[1,2,3], Mark Gerstein[1,2,3]

[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]Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA

Abstract

RNA Binding proteins (RBP) play key roles in post-transcriptional regulation. Their binding sites cover a larger amount of the genome than coding exons, but most of the current variant prioritization methods ignore RBP regulations and only focus on transcriptional-level regulation. Here, we integrated the full catalog of eCLIP, RNA Bind-n-Seq, and shRNA RNA-Seq experiments from ENCODE to deeply annotate the RBP regulome. We proposed a variant impact scoring framework, RADAR (***R***N***A*** Bin***D***ing protein regulome ***A***nnotation and p***R***ioritization), which uses human and cross-species conservation, RNA structure, network centrality and RBP motifs to provide a baseline impact score. Then RADAR incorporated user-specific inputs, such as differential expression/mutational profiles or prior knowledge of regulators or genes to further highlight disease- and tissue-specific variants. Results on somatic and germline variants demonstrate that RADAR can successfully pinpoint intronic, splicing-disruptive variants in key genes such as TP53, which cannot be fully detected by current methods.

Introduction

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, this 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, the impact of variants 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 a full catalog of 318 eCLIP (for 112 RBPs), 76 Bind-n-Seq, and 472 shRNA RNA-Seq experiments 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 88 and 94 percent of RBPs showed increased cross-population conservations in both coding and noncoding regions, respectively, compared to the genomic average. This strongly indicates the purifying selection of 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 the variant impact in such regions. RADAR first combines RBP binding, conservation, network, and motif disruption features with polymorphism data to quantify variant impact described by a universal baseline score. Then, it allows tissue- or disease-specific inputs, such as differential expression, somatic mutation, and prior knowledge of genes, 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. Thus, RADAR provides an effective approach to analyzing 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 1), which has been previously underinvestigated. Many of these RBPs are known to play key roles in post-transcriptional regulation, including splicing, RNA localization, transportation, decay, and translation (Supplementary Fig. 1).

Our definition of the RBP regulome covers 52.6 Mbp of the human genome after duplicate and blacklist removal (Fig. 2A). This is 1.5 and 5.9 times the size the whole exome and lincRNAs, respectively. In addition, only 53.1% of the RBP regulome has transcription-level annotations, such as transcription binding sites, open chromatin regions, and enhancers (Supplementary Fig. 2). Unlike the transcription regulome, which has many distal elements, 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 conservation scores in the peak regions versus the non-peak regions in almost all annotation categories, providing additional evidence of regulatory roles of RBP peaks (Fig. 2C). In summary, the large size of the regulome, the limited overlap with previous annotations, and the elevated conservation scores highlight the necessity of our computational efforts to define 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 \cite{CADD, Funseq, GWAVA}. For example, GERP score maps the human genome to other species to identify nucleotide-level evolutional constraints[19,20]. We used the GERP score in our baseline RADAR framework to detect potentially deleterious mutations in the RBP regulome.

Because the enrichment of rare variants indicates a purifying selection in functional regions in human genomes[21-23], we also inferred conservation of RBP binding sites by integrating population-level polymorphism data from large cohorts (i.e. the 1,000 Genomes Project)[24,25]. GC percentage may confound such inference by introducing read coverage variations, which is a sensitive parameter in the downstream variant calling process[26,27]. Therefore, we calculated the fraction of rare variants, defined as those with derived allele frequencies (DAFs) less than 0.5%, within each RBP binding site and compared them with those from regions with similar GC content as the 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 2).

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[28]. Specifically, it demonstrates 12.7% and 10.3% more rare variants in coding and noncoding regions, respectively (adjusted P values are 1.89×10^{-9} and 2.85×10^{-118} for one-sided binomial tests)[29]. Hence, we used the enrichment of rare variants to infer the

selection pressure in RBP binding sites to weight the variants in such regulator regions (see methods for more details).

RNA secondary structures have been reported to affect every step of protein expression and RNA stability[30]. We incorporated structural features predicted by Evofold, which uses phylogenetic stochastic context-free grammars to identify functional RNAs in the human genome that are deeply conserved across species[31]. 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 score.

Incorporating network information

Highlighting variants in binding hubs

It has been reported that genes within network hubs demonstrate elevated cross-population conservations by demonstrating enrichment of rare variants—a sign of strong purifying selection[21,22,32]. We hypothesized that RBP binding hubs would show similar characteristics because once mutated larger regulation alterations may be introduced. To test this, we separated the regulome based on the number of associated RBPs. Most regulome regions (62%) were associated with only one RBP (Fig. 3B and Supplementary Fig. 4). As the number of RBPs increased, we observed a clear trend of larger rare variant enrichment. 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 3C). This observation supports our hypothesis that the RNA regulome hubs are under stronger selection pressure and, thus, should be given high priority when evaluating the functional impacts 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 differentially expressed genes and simultaneously disrupted the binding of the corresponding RBPs (schematic in Supplementary Fig. 5).

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 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 (details see 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 in to 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, element, and variant level information to up-weight any factors that are possibly associated with target disease.

Highlighting key regulators through expression profiles

Key regulators are often associated with disease progression, so variants that affect such regulation should be prioritized. Therefore, we used RADAR to find disease-specific expression profiles, RADAR tried to find such key regulators by combining the RBP regulatory network information with expressions. Specifically, we first constructed the RBP network from the eCLIP profiles and defined differentially expressed genes from disease and normal cell types. Then for each RBP, we quantified the regulation power for each RBP by association with aggregated disease-to-normal differential expressions from many samples. We applied this approach on 19 cancer types from TCGA and the regulation powers are given in Fig. 4. We found that among many of the RBPs with larger for each RBP power have been reported as cancer-associated genes (with * in Fig. 4B, Table S3). Interestingly, the regulatory power of two key genes 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 significant regulators in their corresponding cancer types by adding extra point to their baseline scores (details see methods). We can easily extend such analysis for other diseases by incorporating differential expression profiles from others cohorts such as GTEx[33,34].

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

RADAR reconsiders the functional impact difference among RBP peaks either by their associated genes or cohort-level mutational profiles. Users can input a prioritized gene list, such as well-documented risk genes for a disease of interest, and RADAR up-weights all the RBP peaks that are close to these genes. Genes that undergo significant expression or epigenetic changes are mostly cell type specific and be can be used to highlight more relevant variants.

In addition, RADAR can incorporate variant recurrence, which has been widely used to discover key disease regions, to reweight different RBP peaks. Regions with more mutations than expected are often considered to be disease-driving[35-37]. For example, we used RADAR to define a local background somatic mutation rate from a large cohort of cancer patients to evaluate the mutation burden in each RBP peak. Variants that are associated with burdened elements are given higher priority in the scoring scheme.

Prioritizing variants with a RADAR weighted scoring scheme

By integrating the universal and user-specific information described above, our entropy-based scoring scheme investigates the functional impacts of variants that are specific to post-transcriptional regulation (Fig. 1, Table 1, and Supplementary Fig. 5). First, RADAR adds up the (weighted) score of variants for all universal features, which include sequence and structure conservation, binding hubs, and motif disruptions. Then, depending on the user inputs, RADAR further up-weights variants with mutations in the key RBP binding sites, nearby genes of interest, or within regulatory elements.

Table 1. Features used by RADAR

Category	Feature	Source	Scoring Scheme
Universal	Selection pressure	eCLIP	Corrected-entropy
	Binding hotspots	eCLIP	Corrected-entropy
	RBP-gene association	shRNA RNA-seq	Entropy
	Motif disruption	Bind-n-Seq DREME	Corrected-entropy
	Structure sensitivity	Evofold	Entropy
	Conservation	Gerp	Entropy
User-specific	RBP regulatory power	Survival Expression	Entropy
	Key genes	Prior knowledge	Entropy
	Mutation Recurrence	Mutation profiles	Entropy

Applying RADAR to pathological germline variants

We calculated baseline RADAR scores on all pathological variants from HGMD. We used the 1,000 Genomes variants as the background to compare the distribution of scores. As expected, the HGMD variants scored significantly higher than somatic mutations (Supplementary Fig. 6). For example, the mean RADAR score for HGMD variants is 0.445, while it is only 0.044 for 1,000 genomes variants (P value <2.2e-16 for two sided Wilcoxon test). We further compared RADAR scores of HGMD variants with other methods (Supplementary Table 6). In total, 992 HGMD variants were identified by only our methods, 29.6% of which were noncoding variants located in a nearby intron, 5'UTR, or 3'UTR (and their extended regions). An example of such a variant is shown in Fig. 5; this variant is located 28 base pairs away from the acceptor site of exon 3 in TP53. eCLIP experiments showed strong binding evidence in seven RBPs, most of which are splicing factors. The co-binding of these above-mentioned splicing factors strongly indicate that this is a key splicing regulatory site. Specifically, the A to T mutation strongly disrupts the binding motif of SF3B4, increasing the possibility of splicing alteration effects. Our finding is not reflected in previous methods for variant prioritization.

Applying RADAR to somatic variants in cancer

We next aimed to leverage our 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[38,39], variants associated with these genes are likely to have the highest functional impact[21]. To test this hypothesis, we first linked each variant with a gene by the shortest distance according to the Gencode v19 annotation. We tested four cancer types, breast, liver, lung, and prostate cancer, and found in all cases that variants associated with cancer-related genes showed significant enrichment, with a larger RNA level functional impact (Supplementary Fig. 7). For example, we found a 3.27- and 3.36-fold increase in high-

impact variants at a threshold level of 2.5 and 3, respectively, in breast cancer patients ($P < 2.2e-16$, single-sided Wilcoxon).

In our second approach, we hypothesized that variant recurrence could be a sign of functionality and may indicate an association with cancer[21-23]. Thus, we compared the variants' score distribution from RNA binding peaks with or without recurrence. Specifically, we separated the peaks of variants from more than one sample from those that were mutated in only one sample, and then compared the percentage of high-impact scores. We found that in most cancer types, elements with recurrent variants were associated with a larger fraction of high-impact mutations. For example, in breast cancer recurrent elements demonstrated a 1.20, 1.55, and 1.77-fold enrichment of high-impact variants with RADAR greater than 1.5, 2.5, and 3.0, respectively, resulting in a P value of 1.71e-19 (one-sided Wilcoxon test).

A case study on breast cancer patients

We applied our method on a set of breast cancer somatic variants from 963 patients released by Alexandrov *et al*[40]. We used COSMIC genes and expression and mutational profiles as additional features. In total, we determined that around 3% of the 68,000 variants alter post-transcriptional regulation to some degree. Specifically, 169 out of the 501 highly ranked variants only reported by our tool were located in noncoding regions, with 15, 28, and 24 from nearby introns, the 5' UTR, and the 3' UTR, respectively (Supplementary Fig. 8). We found that variants in the intronic region usually bind within 30 bp of the splice sites and break the motifs of many splicing factor binding sites. For the 3' UTR regions, variants reported only by RADAR were within the binding peaks of cleavage stimulation factor binding sites, strongly indicative of a role in the polyadenylation of pre-mRNAs. The discovery of such meaningful results indicates the ability of RADAR to differentiate deleterious mutations that disrupt post-transcriptional regulation.

Discussion

In this study, we integrated the full catalog of eCLIP, Bind-n-Seq, and shRNA RNA-Seq experiments from ENCODE to build a RNA regulome for post-transcriptional regulation. Although DNA-level regulation takes up a larger part of the genome, our defined RBP regulome is larger than previously thought and covered as much as 56.2 Mbp of the genome (Fig. 2A). In fact, the regulome is larger than the size of whole exome and only showed limited overlap with previous transcription-level annotations (Supplementary Fig. 2.). Furthermore, we found that the RBP regulome demonstrated noticeably larger conservation in two areas: higher cross-species conservation across all annotation categories (Fig. 2C) and significant enrichment in rare variants for most RBPs (Fig. 3A). These two sources of evidence support the notion that the RBP regulome is under strong purifying selection and carries out important biological functions. In addition, these results signify the necessity of computational tools to annotate and prioritize variants in the RBP regulome, which are under investigated.

By integrating a variety of regulator-, element-, and nucleotide-level features, we propose an entropy-based scoring frame, RADAR, to investigate 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. Applying this score to HGMD pathological variants highlighted many candidates that were solely discovered by RADAR and provided detailed explanation of the underlying disease-causing mechanism (Fig. 5). In addition to the baseline score, RADAR also allows user-specific inputs such as prior gene knowledge, patient expression, and mutation and survival profiles for a re-weighting process to highlight relevant variants in a disease-specific manner. As an example, we performed a breast cancer variant prioritization and score re-weighting scheme with user inputs in the well-known

tumor suppressor gene TP53. Also results from somatic variants from several cancer types and showed that this scheme is able to identify relevant variants (Supplementary Fig. 7).

In summary, we have shown that RADAR is a useful tool for annotating and prioritizing post-transcriptional regulomes for RBPs, which has not been covered by most of the current variant functional impact interpretation tools. Our method also provides additional layers of information to current gene regulomes. Importantly, the RADAR scoring scheme can be used in conjunction with current transcriptional variant functional evaluation tools, such as Funseq, to evaluate variant impacts. Given the fast expanding collection of RBP binding profiles from additional cell types, we envision that RADAR 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 RBP 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 binding site locations containing blacklisted regions. We further separated the peaks into coding regions and the noncoding regions.

Cross-population conservation inference

We used germline variants from the 1,000 Genomes Project to infer the selection pressure of RBP binding sites by evaluating their enrichment of rare variants. Specifically, for any RBP i, suppose its binding peaks contain n_i^r rare variants (DAF<= 0.005) and n_i^c common variants. The percentage of rare variants is defined as

$$r_i = \frac{n_i^r}{n_i^r + n_i^c} \quad (1)$$

In order to correct for potential GC bias, we first binned the genome into 500 base pair bins and grouped them according to their GC percentage. For RBP i with GC percentage gc_i, we only selected bin groups with closest GC to calculate the background rare variant percentage $r_{bg}^{gc_i}$. As a result, the corrected enrichment of rare variants was defined as

$$\acute{r}_i^{gc} = r_i / r_{bg}^{gc_i} \quad (2)$$

Regions with $\acute{r}_i^{gc}$ larger than 1 suggested a purifying selection.

RBP Network information

We first separated the RBP regions into coding and noncoding regions and then calculated the number of RBPs binding in each nucleotide. Then we grouped the RBP regulome by the number of binding RBPs and calculated the GC corrected enrichment of rare variants for each group in coding and noncoding regions separately by equation (2). We found that the conservation level of the binding sites increases as the number of binding RBPs grows.

Specifically, we selected regions with at least the top 5 and 1 percent of binding RBPs as the hot and ultra-hot regions to give extra score in our RADAR framework.

Motif Analysis

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

$$Dscore = motifscore_{ref} - motifscore_{alt} = -10 \times \log_{10}\left(\frac{p_{ref}}{p_{alt}}\right) \quad (3)$$

To quantify a motif breaking event, we require that the p value for the reference allele is at least 5e-4(JZ2JL: correct?). 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 utilized an *in vitro* RNA binding assay, RNA Bind-N-Seq (RBNS, Lambert et al. Molecular cell 54.5 (2014): 887-900 JZ2JL: fix), to characterize sequence and structural specificities of RBPs. For the RBNS motifs, we selected sequences based on an enrichment Z-score cutoff of 3 (JZ2JL: check the details of their paper). For each variant that affected multiple RBP binding profiles, we determined the max score.

Baseline Variant Scoring

We considered 6 different features in the RADAR baseline scoring scheme (details in Table S10). For any continuous feature i listed in Table S10, we first calculated the percent of SNPs from 1,000 Genomes that are falling into this feature category and denote it as f. Then the entropy related weight is used to evaluate the functional impact.

$$S_i = 1 + f_i \times \log(f_i) + (1 - f_i) \times \log(1 - f_i) \quad (4)$$

Specifically, for the motif analysis, for any motif breaking event with D-score $\acute{d}$, f_i should be the percent of SNPs with equal or larger D-scores among all the 1,000 Genomes SNPs.

$$f_i = \frac{\#(d \geq \acute{d})}{n} \quad (5)$$

For any continuous feature that used the weighted entropy value, we further used the GC corrected rare variant enrichment $\acute{f}_i^{gc}$ of the RBP as a coefficient to reweight the score S. Finally, the overall baseline RADAR score is $S = \sum_i S_i$

RBP Regulatory Power from Linear Regression

RADAR allows inputs in addition to the pre-built context used to calculate the baseline 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 xxx cancer patients of 24 types from TCGA. To get a robust differential expression analysis, we excluded 5 cancer types that have less than 10 normal expression profiles and used DESeq to find tumor-to-normal differentially expressed genes (corrected p-val from DESeq <0.05). Let y_i^k represent the differential expression status of gene i of the k^{th} cancer type.

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. For any RBP, $x_i = 1$ if it regulate gene i. $x_i = 0$ otherwise.

We inferred the regulatory power of each RBP by through a regression approach of the above differential expression and RBP network connectivity as $\vec{\boldsymbol{y}} = \beta_0 + \beta_1 \vec{\boldsymbol{x}}$. The associated p-value is also an indication of the statistical significance that such a regulatory potential exists. The full table of regulatory power in all

19 cancer types were given in Table S7 and Figure S9. Interestingly, we found that for the RBPs with larger regulatory power are those tends to be known to associate with cancer, as listed in Table S8.

For RBPs with high regulator powers, we also performed a patient-wise regulatory 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). The full list of RBPs that are associated with patient survival are given in Table S9.

Recurrence in Somatic Mutations

We prioritized variants that fell in elements (RBP binding sites) that were statistically enriched for somatic mutations. In order to do this, we first binned the genome using 1Mbp windows, and counted the number of somatic mutations in each window. This provided us a with a local mutation rate. Then, for each RBP binding site, 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 higher priority via the entropy scoring scheme.

Resource and Software Accessibility

We have made this RNA variant prioritization tool available as an open-source Python source at radar.gersteinlab.org. The website contains details on usage, examples, resources, and dependencies. We recommend a system with 10gb of RAM to avoid slowed performance for variant sets with a sample size below 1 million. We also provided a genome-wide pre-built RADAR 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. In addition, we released the RBP-gene regulatory network at radar.gersteinlab.org.

References

1 Croce, C. M. Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* **10**, 704-714, doi:10.1038/nrg2634 (2009).

2 Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L. & Almouzni, G. Epigenomics: Roadmap for regulation. *Nature* **518**, 314-316, doi:10.1038/518314a (2015).

3 Yang, G., Lu, X. & Yuan, L. LncRNA: a link between RNA and cancer. *Biochim Biophys Acta* **1839**, 1097-1109, doi:10.1016/j.bbagrm.2014.08.012 (2014).

4 Schmitt, A. M. & Chang, H. Y. Gene regulation: Long RNAs wire up cancer growth. *Nature* **500**, 536-537, doi:10.1038/nature12548 (2013).

5 Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829-845, doi:10.1038/nrg3813 (2014).

6 van Kouwenhove, M., Kedde, M. & Agami, R. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nat Rev Cancer* **11**, 644-656, doi:10.1038/nrc3107 (2011).

7 Swinburne, I. A., Meyer, C. A., Liu, X. S., Silver, P. A. & Brodsky, A. S. Genomic localization of RNA binding proteins reveals links between pre-mRNA processing and transcription. *Genome Res* **16**, 912-921, doi:10.1101/gr.5211806 (2006).

8 Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol* **3**, 195-205, doi:10.1038/nrm760 (2002).

9 Fu, X. D. & Ares, M., Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15**, 689-701, doi:10.1038/nrg3778 (2014).

10 Zheng, D. & Tian, B. RNA-binding proteins in regulation of alternative cleavage and polyadenylation. *Adv Exp Med Biol* **825**, 97-127, doi:10.1007/978-1-4939-1221-6_3 (2014).

11 Fossat, N. *et al.* C to U RNA editing mediated by APOBEC1 requires RNA-binding protein RBM47. *EMBO Rep* **15**, 903-910, doi:10.15252/embr.201438450 (2014).

12 Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**, 1977-1986, doi:10.1016/j.febslet.2008.03.004 (2008).

13 Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* **42**, D92-97, doi:10.1093/nar/gkt1248 (2014).

14 Blin, K. *et al.* DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* **43**, D160-167, doi:10.1093/nar/gku1180 (2015).

15 Anders, G. *et al.* doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* **40**, D180-186, doi:10.1093/nar/gkr1007 (2012).

16 Hu, B., Yang, Y. T., Huang, Y., Zhu, Y. & Lu, Z. J. POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res* **45**, D104-D114, doi:10.1093/nar/gkw888 (2017).

17 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514, doi:10.1038/nmeth.3810 (2016).

18 Lambert, N. *et al.* RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell* **54**, 887-900, doi:10.1016/j.molcel.2014.04.016 (2014).

19 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913, doi:10.1101/gr.3577405 (2005).

20

20 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

21 Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480, doi:10.1186/s13059-014-0480-5 (2014).

22 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).

23 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

24 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

25 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

26 Garner, C. Confounded by sequencing depth in association studies of rare alleles. *Genet Epidemiol* **35**, 261-268, doi:10.1002/gepi.20574 (2011).

27 Xu, C., Nezami Ranjbar, M. R., 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* **18**, 5, doi:10.1186/s12864-016-3425-4 (2017).

28 Lu, Y. *et al.* Genetic variants cis-regulating Xrn2 expression contribute to the risk of spontaneous lung tumor. *Oncogene* **29**, 1041-1049, doi:10.1038/onc.2009.396 (2010).

29 Davidson, L., Kerr, A. & West, S. Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. *EMBO J* **31**, 2566-2578, doi:10.1038/emboj.2012.101 (2012).

30 Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* **15**, 469-479, doi:10.1038/nrg3681 (2014).

31 Pedersen, J. S. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**, e33, doi:10.1371/journal.pcbi.0020033 (2006).

32 Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* **9**, e1002886, doi:10.1371/journal.pcbi.1002886 (2013).

33 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).

34 Consortium, G. T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

35 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

36 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* **43**, 8123-8134, doi:10.1093/nar/gkv803 (2015).

37 Lochovsky, L., Zhang, J. & Gerstein, M. MOAT: Efficient Detection of Highly Mutated Regions with the Mutations Overburdening Annotations Tool. *Bioinformatics*, doi:10.1093/bioinformatics/btx700 (2017).

38 Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat Med* **10**, 789-799, doi:10.1038/nm1087 (2004).

39 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108, doi:10.1038/nrg.2015.17 (2016).

40 Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* **24**, 52-60, doi:10.1016/j.gde.2013.11.014 (2014).

Acknowledgements

Author Contributions

Competing Financial Interests

Figure Legends and Tables