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

Abstract

Dysregulation of RNA binding proteins (RBP) can cause numerous diseases but the effect of variants in their regulome has been barely investigated. Hence, we integrated 318 eCLIP, 76 RNA Bind-n-Seq, and 472 shRNA RNA-Seq experiments from the new release of the ENCODE project to deeply annotate the RBP regulome. First, we showed that around 90 percent of RBPs are enriched with rare variants in their binding sites, suggesting extensive purifying selections. We then proposed a variant impact scoring framework, RADAR, by combining RBP binding, structure, context, network, and conservation features with polymorphism data to provide a baseline impact score. Then we incorporated user-specific inputs, such as patient survival, expression, mutational profiles and prior knowledge of genes to reweight the variants to further highlight disease- and tissue-specific causal ones. Results on both germline and somatic variant datasets demonstrate that RADAR can successfully pinpoint disease-relevant variants and uncover the underlying regulation mechanism of post-transcriptional regulation.

1 Introduction and 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, after (or simultaneously while) DNA has been transcribed to premature RNAs, genes also experience a series delicately controlled processing, such as conversion to mature RNA, followed by transportation, translation, and then degradation in the cell. Dysregulation of any one of these steps may alter the final fate of gene products and result in abnormal phenotypes[4-6]. Despite its importance in regulation, such post-transcriptional regulome has been underdeveloped, partially due to its less systematic functional.

RNA binding proteins (RBPs) have been reported to play essential roles during both co- and post-transcriptional regulations[7-9]. They bind to thousands of genes in the cell through multiple processes, including splicing, cleavage and polyadenylation, RNA editing, localization, stability, and translation[10-14]. Recently, many efforts have been made to complete these post- or co-transcriptional regulomes by synthesizing public RBP binding profiles[15-18], which have greatly expanded our understanding of RBP regulation. Since 2016, the ENCODE consortium started to map the post-transcriptional regulome using various types of assays on matched cell types. First, ENCODE has released large-scale enhanced CLIP (eCLIP) experiments for hundreds of RBPs[19]. It 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. It also simultaneously performed expression quantification by RNA-Seq after knocking down various of RBPs. Also, ENCODE also performed a quantitative assessment of the context and structural binding specificity of many RBPs by Bind-n-Seq experiments \cite{24837674}.

JZ2MG: whenever I talked about conservation, I am thinking about cross species conservation, not from the rare DAF stuff. Please pay attention to the red text below. I found it is confusing actually.

In this paper, we collected the 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 1000 Genomes Project, we demonstrated that 88 and 94 percent of RBPs showed noticeably higher conservations respectively. This strongly indicates the purifying selection of the RBP regulome. Furthermore, we proposed a top-down 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, structure, context, network, and conservation 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 (Figure 1). By applying our scoring scheme on both somatic and germline variants from disease genomes, we demonstrate that RADAR is able to pinpoint disease-associated variants missed by other methods. Finally, we implemented the RADAR annotation and prioritization scheme into software for community use (radar.gersteinlab.org).

2 Results

2.1 Define the RBP regulome through eCLIP data

Here we first used the binding profiles of 112 distinct RBPs from ENCODE to fully explore the human RBP regulome (Table S1). Many of these RBPs are known to play key roles in post-transcriptional regulation, including splicing, RNA localization, transportation and decay, and translation (Fig. S1).

While under-investigated, our defined RBP regulome covers a decent part of the human genome (52.6Mbp after duplicate and blacklist removal, Fig. 2A). It is 1.5 and 5.9 times of the size the whole exome and lincRNAs respectively. Besides, only 53.1 percent of the RBP regulome is overlapped by the transcription-level annotations, including transcription binding sites, open chromatin regions, and enhancers (Fig. S2). Unlike the transcription regulome, which has many distal elements, 55.1 percent of the RBP regulome is located in the immediate neighborhood of the exome regions, such as coding exons, 3' or 5' UTRs, and nearby introns (Fig. 2B, details see methods). Furthermore, in almost all annotation categories, we observed significantly higher PhastCons scores in the peak regions vs. the non-peak regions, providing additional evidence of their regulatory roles (Fig. 2C). In summary, the large size of the regulome, the limited overlapping with previous annotations, and the elevated conservation scores underscore the immediate necessity of computational efforts to annotate and prioritize variants in the RBP regulome.

2.2 Universal features used for baseline RADAR score

2.2.1 Inference of purifying selection in RBP binding sites

Many researchers have pointed out that enrichment of rare variants indicates purifying selection in functional regions in human genomes[34-36]. Hence, aside from the cross-species conservations mentioned above, we also inferred the purifying selection on the RBP binding sites by integrating population-level polymorphism data from large cohorts, e.g. the 1000 Genomes Project cite {26432245, 23128226}. GC percentage may confound such inference because it causes read coverage variations, a sensitive parameter in the downstream variant calling process[37,38]. Hence, we first calculated the fraction of rare variants (derived allele frequency (DAF) less than 0.5%) within each RBP's binding sites, and compared it with those from regions with similar GC content as a background (see details in methods). In total, 88.4 percent of the RBPs (99 out of 112) show elevated rare variant fraction in coding regions compared to those of the background regions after GC correction (Fig. 3A). Similarly, in the noncoding part of the binding sites, 93.8 percent of RBPs (105 out of 112) exhibit an enrichment of rare variants. This observation convincingly demonstrates the accuracy of our RBP regulome definition. (Table S2).

Some well-known disease-causing RNA binding proteins demonstrate the largest rare DAF enrichment. 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[39]. Specifically, it demonstrates 12.7% and 10.3% more rare variants in coding and noncoding regions (adjusted P values are 1.89×10^{-9} and 2.85×10^{-118} for one-sided binomial

tests)[40]. Hence, we used the enrichment of rare variants as a feature to infer the selection pressure in RBP binding sites to weight the variants in such regulator regions (details see methods).

2.2.2 Highlight variants in RBP binding hubs

It has been reported that genes within network hubs usually exhibit larger enrichment of rare variants—a sign of strong purifying selection[34,35,44]. Similarly, we suspect that RBP binding hubs may demonstrate similar characteristics because once mutated larger regulation alterations may be introduced. To test this hypothesis, we separated the regulome based on the number of associated RBPs. The majority of the regulome regions (62 percent) are associated with only 1 RBP (Fig. 3B and Fig. S4). As the number of RBPs increased, we observed a clear trend of larger rare variant enrichment. For instance, in the noncoding regions, regions with at least 5 and 10 RBPs exhibited 2.2 and 13.4 percent more rare variants (top 5 and 1 percent, Fig 3C). This observation convincingly supports our hypothesis that the RNA regulome hubs are under stronger selection pressure, and hence should be given high priority when evaluating the functional impacts of mutations.

2.2.3 Motif analysis for nucleotide impact

Mutations that change the RBP binding affinity may alter RBP regulation via loss-of-function effect. To quantify such impacts, we used the difference of position weight matrix (PWM) scores of the ~~mutant~~ allele against the reference allele. RADAR consists of two sources of motifs. First, it has been reported that many of the RBPs' binding events *in vivo* can be captured by binding preferences *in vitro*. Hence, we used motifs reported by RNA Bind-n-Seq experiments from ENCODE. Then, we also used the same scheme on the *de novo* motifs discovered directly from the binding peaks using the default settings in DREME (details see methods). For each variant, the highest motif score from the above two sources is used to quantify the nucleotide effect.

2.2.4 Structure and context conservations

RNA secondary structures have been reported to affect every step of the protein expression and RNA stability \cite{24821474}. We incorporated structure features predicted by Evofold, which uses phylogenetic stochastic context-free grammars to identify functional RNAs encoded in the human genome that are deeply conserved across species \cite{16628248}. We found after intersecting with conserved structure regions defined by Evofold, the RBP binding sites demonstrate significantly higher conversation, and hence we used the Evofold regions a feature in the baseline score.

Besides, cross-species sequence conservations have also been widely used as an important feature to discover regions with biological functions. For example, The Genomic Evolutionary Rate Profiling (GERP) score was developed to identify nucleotide level evolutional constraints by mapping human genome to other species. We used GERP score in our baseline RADAR framework to detect potentially deleterious mutations in the RBP regulome.

2.2.6 Highlight differentially expressed genes after RBP knockdown

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 the RBP binding and hence affect gene expressions in a similar way. Therefore, we extracted the differentially expressed genes from such experiments and up-weight all variants that are located near such genes and at the same time disrupt the binding of corresponding RBPs (schematic in Fig. S5).

2.3 User-specific features to reweight variant impact

2.3.1 Use expression profiles to prioritize key regulators

In many diseases, several key regulators can be associated with disease progression and variants that affect such regulation should be prioritized. Therefore, given additional 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 differential expressions from disease and normal cell types. Then for each RBP, we quantified its regulation power by association with aggregated disease-to-normal differential expressions from many sampls. 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. We can easily extend such analysis for other diseases by incorporating normal expression profiles from others cohorts such as GTEx \cite{JZ2JL: GTEx nature marker papers}.

2.3.2 Prior knowledge of regulators, target genes, and variants

Variant Prioritization can be improved if prior knowledge of the regulators, target genes, and variants can be appropriately incorporated. For example, the splicing factor QKI has been reported by many reports to be associated with different cancer types and it is reasonable to further prioritize key variants that affect QKI regulation for cancer \cite{JZ2JL27841882 and others}. Besides, many databases have enumerated hundreds of cancer-associated genes that are known to play critical roles in cancer. Cell proliferation and DNA repair-related genes are also important for cancer research. For other diseases, many GWAS studies have pointed out many risk genes. Furthermore, genes undergo significant expression or epigenetic changes are mostly cell-type-specific and be can be used to highlight more relevant variants. Hence, our RADAR framework allows users to input their disease-specific genes of interest to prune the candidate variant list further. Another example useful prior knowledge is variant recurrence, which has been widely used to discover key regions in disease. Regions with more than expected mutations are often considered as disease driving events \cite{mutsigCV, larva, moat}. Taken the cancer variants as an example. Given the somatic mutation from a large cohort of patients, we can first define a local background mutation rate to evaluate the mutation burden in each RBP peaks. Variants that are associated with burdened elements are given higher priority in the prioritization scheme.

2.4 RADAR weighted scoring scheme to prioritize variants

By integrating the universal and user-specific information mentioned above, we proposed an entropy-based scoring scheme to investigate the functional impacts of variants specific to post-transcriptional regulation (Fig. 1 and Fig. S5). First, we added up the (weighted) entropy score of variants for all universal features, which include rare variant enrichment, binding hotspot, structure, motif, and conservations. Then depending on the user inputs, we further up weight mutations that fall into the key RBP binding sites, nearby genes of interest, or within elements with more than expected variants.

Table 1. Features used by RADAR

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

2.5 Application on pathological germline variants

We calculated baseline RADAR score on all pathological variants from HGMD. We used the 1000 genomes (1kg) variants, as the background to compare the distribution of scores. As expected, the HGMD variants are scored

significantly higher than somatic mutations (Fig. S6). For example, the mean RADAR score for HGMD variants is 0.445, while it is only 0.044 for 1kg variants (P value <2.2e-16 for two sided Wilcoxon test). We further compared RADAR scores of HGMD variants with other methods (Table S6). Specifically, we found 992 HGMD variants that are explained by only our methods and 29.6% of them are noncoding variants that are located in the nearby intron, 5'UTR, and 3'UTR (and their extended regions). An example of such variants is given in Fig 5. It is located 28 bp away from the acceptor site of exon 3 in TP53. eCLIP experiments showed strong binding evidence in 7 RBPs, most of which are splicing factors. The co-binding of these above mentioned splicing factors strongly indicate this is key splicing regulatory site. Specifically, this 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.

2.6 Application on somatic variants in cancer

2.6.1 Somatic variants associated with COSMIC genes and recurrence

We applied our scheme to evaluate the deleteriousness of somatic variants from public datasets. Due to the lack of golden standard, we evaluate our results from two aspects. First, hundreds of cancer-associated gene are known to play essential roles through various pathways[47,48]. Hence, in general, variants associated with these genes are supposed to have a higher functional impact compared to others[34]. To test this hypothesis, we first associated each variant with a gene by the shortest distance according to Gencode v19 annotation. We found that in all four cancer types we tested, including the breast, liver, lung, and prostate cancer, variants associated with cancer associated genes showed significantly enrichment in variants with larger RNA level functional impact (Fig. S7). 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). This pattern is consistent in all four cancer types we investigated (Fig S7).

In addition, because variant recurrence is considered a sign of functionality and may indicate association with cancer[34-36], we also compared the variants' score distribution from RNA binding peaks with or without recurrence. Specifically, we separated the peaks with variants from more than one sample from those that are mutated in only one sample and compared the percentage of higher impact scores. We found that in most cancer types, elements with recurrent variants are associated with a larger fraction of high impact mutations. For example, in Breast cancer, recurrent elements demonstrated a factor of 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 at 1.71e-19 from one-sided Wilcoxon test.

2.6.2 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*[50] and used COSMIC genes, expression and mutational profiles as additional features. In total around 3 percent out of the 68k variants was evaluated to alter post-transcriptional regulations to some degree. Specifically, 169 out of the 501 highly ranked variants only reported by our tool are located in the noncoding regions, with 15, 28, and 24 are from nearby introns, 5' UTR and 3' UTR regions, respectively (Fig. S8). For the intronic one, we find that such variants 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 are 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 regulations.

3 Discussion

In this paper, we integrated the full catalog of eCLIP, Bind-n-Seq, and shRNA RNA-Seq experiments from ENCODE to build the 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 covers as large as 56.2Mbp of the genome. It is larger than the size of whole exome and only showed some overlap with previous transcription-level annotations. Furthermore, we found that the RBP regulome demonstrates noticeably larger conservation from two aspects. It does not only show higher cross-species conservation under all annotation categories and but also

demonstrate significant enrichment in rare variants for most RBPs. These two sources of evidence consistently showed that the RBP regulome is under strong purifying selection and carrying out important biological functions. It signifies the necessity of computational tools to annotate and prioritize variants in the RBP regulome, which was previously under investigated.

By integrating a variety of regulator, element, and nucleotide level features, we proposed an entropy based scoring frame RADAR to investigate impacts of somatic and rare germline variants. RADAR contains two parts in the variant prioritization framework. First, by incorporating eCLIP, Bind-n-Seq, shRNA RNA-seq experiments with conservation and structure features, we built a pre-defined data context to quantify the baseline variant impact score. It is suitable for multiple disease analysis or cases where no other prior information can be used. We applied this score on HGMD pathological variants and highlighted many candidates that are solely discovered by RADAR with detailed explanation of the underlying disease cause mechanism. On top of the baseline score, RADAR also allow user-specific inputs such as prior regulator/target gene/variant knowledge and patient expression/profiles for a re-weighting process to highlight relevant variants in a disease specific manner. We showed an example of breast cancer variant prioritization and score re-weighting scheme by user inputs in the well-known tumor suppressor gene TP53. Also results from somatic variants from several cancer types show that RADAR can identify relevant variants.

In summary, we believe that RADAR can serve as a useful tool to annotate and prioritize the post-transcriptional regulomesfor RBPs, which has not been covered by most of the current variant functional impact interpretation tools. It is also able to provide additional layers of information to current gene regulomes. More importantly, the RADAR scoring scheme can be used in conjunction with some of the current transcriptional variant functional evaluation tools, such as Funseq, to add independent information to jointly evaluate variant impacts. With the fast expanding collection of binding profiles of more RBPs from more cell types, we envision that it can more extensively tackle the functional consequence of mutations from both somatic and germline genomes.

4 Methods

4.1 eCLIP Data Processing and Quality Control

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 cell lines, 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 per ENCODE 3 uniform data processing pipeline. The eCLIP peak calling method and processing pipeline were developed by the laboratory of Gene Yeo at the University of California, San Diego (https://github.com/YeoLab/clipper, CLIP-seq cluster-identification algorithm on PMID: 24213538). For each peak, the enrichment significance was calculated against a paired input, and we filtered those peaks with a significance flag of 1000. We ultimately used the recommended cutoff of the significance, which was -log10(P-value) >= 3 and log2(fold_enrichment) >= 3.

4.2 Annotation

RBPs bind along the genome in a variety of contexts. Using eCLIP data, we can synthesize a genomic landscape of where RBPs bind. Raw peak signals from eCLIP data are translated into binding sites, using a peak caller specialized for eCLIP data. Generally, these RBPs having binding sites that correspond to about 150 bp, with many RBPs having well over 10,000 binding sites. Binding site locations containing blacklisted regions are removed. These include regions on the genome with low sequencing depth or coverage. Despite filtering these blacklisted regions, over 99% of the binding locations are preserved. While the total number of base pairs corresponding to binding sites translates to a large number, compared to the scale of the genome it is still minute. Therefore, we annotate the genome, indicating at each position the set of RBPs that bind. This annotation set is known as the contextual annotations.

In addition to contextually annotating the genome with the preferential binding of RBPs, we also include a functional annotation – whether a specific position falls in the coding or noncoding region of the genome. The coding region consists of only the exons of protein coding genes. The noncoding region is further divided into 3'UTR, 5'UTR, 3'UTR extended, 5'UTR extended, and nearby intron regions. Coding and UTR annotations are retrieved from Gencode and UCSC, respectively. 3'UTR and 5'UTR extended regions consist of the 1000 base pairs downstream of the 3'UTR and 5'UTR regions, respectively. The nearby intron regions consist of the 100bp regions adjacent to each exon. While each of these region types are generally distinct, overlap is a possibility. Therefore, a hierarchy of which annotation takes precedence when annotation types overlap is established, from highest priority to lowest: coding, 3'UTR, 5'UTR, 3'UTR extended, 5'UTR extended, and nearby intron. Regions of the genome not classified by these annotations are labeled as "other" and may refer to other noncoding elements or blacklisted elements.

4.3 Regulatory Network Construction

In order to construct a regulatory network of protein coding genes associated with a given RBP, we first identify which annotation is associated with which protein coding gene. The network we construct is undirected between protein coding genes and consists of a set of genes that a given RBP interacts with. To determine which genes the RBP interacts with, all binding sites of the RBP are intersected with all annotations (4.2). With the additional information of the associated gene given the annotation, we compile a list of all protein coding genes associated with the RBP. A unique list is determined and such a set of genes is determined to be the network of genes associated with that RBP. This is performed across each RBP in order to obtain a set of genes associated with each RBP. Furthermore, since the annotations we use consist of multiple categories, we are able to form networks specific to certain annotations. For example, while the regulatory network of one RBP could be viewed as a group of all protein coding associated genes where an RBP binds in the exon, UTR, or intronic region of those genes, another network of only the RBPs binding in 3UTR associated regions can be established, which is used in our survival analysis (4.9.2).

4.4 Inference of selection pressure from population genetics data

4.4.1 Using rare derived allele frequency as a metric for selection pressure

In order to infer the selection pressure of a given region on the genome, we make use of germline variants from the 1000 Genomes Project. These germline variants consist of both common and rare variants. These variants are then classified into coding and noncoding variants based on our annotations derived from Gencode. Coding variants fall in regions annotated as coding, while noncoding variants fall in regions annotated as noncoding Section (4.2). Noncoding variants are not further classified into noncoding element subgroups in order to maintain a large sample size of variants for optimal statistical power in inferring negative selection pressure. The metric we use to represent negative selection pressure is the rare derived allele frequency (rare DAF). For a given region, i, containing rare variants r_i, and common variants c_i, the rare DAF is defined to be

$$Rare\ DAF = \frac{r_i}{r_i + c_i}$$

Since we have further categorized both rare and common variants as coding and noncoding, we can obtain a coding and noncoding rare DAF for a given region as well. Finally, we take the rare DAF value and divide it by the GC content corrected genome average (Section 4.4.2) in order to obtain a ratio. Regions with rare DAF ratios larger than 1 suggest an associated selection pressure higher than the expected selection pressure of regions on the genome of similar GC percent.

4.4.2 Rare DAF is confounded by GC content

Although negative selection pressure can be inferred from metrics such as rare DAF, it is not always accurate. In particular, the rare DAF of a region is severely confounded by its GC content. In order to correct for this bias, we first bin the genome into 500 base pair bins. Next, we estimate the average GC content within these 500 base pair bins,

which can range from 0% to 100%. We then group bins with similar GC content. Specifically, we establish 40 groups, using 2 percent intervals from 20 to 80 percent GC. Bins containing 0-20 and 80-100 percent GC content are ignored due to limited observations in these groups. For each of the 40 groups of 2% GC intervals, we associate a set of 500 base pair bins. Each of these sets are taken together to form a region, i, and the rare DAF is calculated. For each of the 40 regions, i, we obtain a rare DAF value, forming a discrete relationship between rare DAF and GC content. Using these discrete points, we fit a Gaussian kernel smoother with bandwidth of 10, resulting in a smoothed function between rare DAF and GC. This function serves as a way to estimate the genomic rare DAF given the GC content.

4.4.3 Negative selection pressure of RBP specific binding sites

We directly apply the method of determining a corrected rare DAF ratio to binding regions for a given RBP. The GC content of all binding sites for an RBP is estimated (from a genomic bigwig file), and using the derived smooth function between rare DAF and GC, a coding and noncoding rare DAF ratio is determined. For any given RBP a rare DAF ratio is used to measure the relative selection pressure of an RBP.

4.5 RBP network hubs

A natural extension to annotating locations based on the set of RBPs that preferentially bind, is to include the annotation of how many RBPs bind. The value associated with the number of RBPs that bind to a position is termed the "hub size". We hypothesize that the hub size of a region and the selection pressure of the region share a positive relationship. To determine the actual relationship, we annotate the genome with its hub size on a base pair resolution. For both noncoding and coding regions, we estimate the selection pressure using rare DAF ratio from germline variants within all regions showing equal to or more extreme hub size for a given hub size. The rare DAF ratio is found by taking the rare DAF and dividing by the corrected rare DAF, derived from evaluating the GC for regions with the same hub size and predicting the genomic rare DAF average (4.4.2). We show a cumulative relationship between rare DAF and hotness, with a generally increasing trend. When the hotness increases past 10 however, the lack of observations results in difficulty in producing a reliable rare DAF. Therefore, we cutoff the measure of rare DAF at a maximum hub size of 10, corresponding to the top 1% of the data. Furthermore, regions with hub size less than 5 and 6 respectively for the noncoding and coding regions, are deemed to have insignificant hub size, and are automatically given a 0 value for the hub score. The resulting discrete function is smoothed from hub size of 5 or 6 to 10, for coding and noncoding respectively. The function steps from 0 (from hotness of 1 to 4) to the rare DAF ratio at 5 or 6, and also maintains a constant rare DAF ratio for hotness values over 10, with the assumption that network hubs of size 10 or greater are extremely rare and therefore difficult to infer the selection pressure in these sparse regions.

4.6 Motif analysis

4.6.1 De novo discovery

RBP motifs were found using DREME software (Version 4.12.0, http://meme-suite.org/tools/dreme, Timothy L. Bailey, "DREME: Motif discovery in transcription factor ChIP-seq data", Bioinformatics, 27(12):1653-1659, 2011.). De novo motif was called on a collection of significant eCLIP peaks.

4.6.2 Bind-n-seq motif processing

In addition to ENCODE eCLIP dataset, we utilized in vitro RNA binding assay, RNA Bind-N-Seq (RBNS, Lambert et al. Molecular cell 54.5 (2014): 887-900.), to characterize sequence and structural specificities of RNA-binding proteins. We used RBNS motifs from 78 human RBPs
(doi: http://dx.doi.org/10.1101/201996) to prioritize germline and somatic variants that could potentially disrupt RNA-binding domain (RBD). In summary, RBNS motifs were called based on the enrichment Z-score cutoff of 3. Some

RBPs had up to 4 motifs and they ranged from 5-mer to 9-mers. There were 17 RBPs that overlapped with eCLIP RBPs, and all RBNS motifs were treated independently from eCLIP-based de novo motifs.

Overlapping RBPs between eCLIP and RBNS (n=17):
EIF4G2, EWSR1, FUBP3, HNRNPC, HNRNPK, IGF2BP1, IGF2BP2, KHSRP, PCBP2, RBFOX2, RBM22, SFPQ, SRSF9, TAF15, TARDBP, TIA1, TRA2A

4.6.2 Evaluating Motif Disruption with MotifTools

To evaluate the functional importance of RNA-binding sites, we surveyed mutational impact on RBP motifs. We called potential RBP motifs on high-confidence RBP peaks and evaluated motif disruption power of each variant using a germline variant set (1000 Genomes Project, a somatic variant set (30 types of cancer somatic SNVs, Alexandrov et al., Nature 2013), and HGMD (version 2015 *** please confirm the version ***). Motif breaking power, which we labeled as D-score (D stands for disruptive-ness or deleterious-ness), was evaluated using MotifTools (https://github.com/hoondy/MotifTools). D-score was calculated based on the difference between sequence specificities of reference to alternative sequence.

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

We only considered positive D-scores, which denote a variant that decreases the likelihood that a TF will bind the motif (motif-break), and ignored negative D-scores where a variant that increases the likelihood that a TF to bind the motif (motif-gain). For assessing D-score, uniform nucleotide background was assumed, and the p-value threshold of $5e^{-2}$ was used. For each variant that affected multiple RBP binding profiles, the max score was taken.

4.7 Baseline Variant Scoring

The three features that are considered in our baseline scoring scheme are the selection pressure, network hubs, and motif disruption. In the first two scenarios, we calculate a value of, p, corresponding to the number of germline variants falling in a given annotation divided by the total number of germline variants. For each of the three cases, the annotation is different. In the case for selection pressure, we naturally use an annotation equivalent to the binding sites from the eCLIP profiles. Each set of annotation is unique for each RBP, resulting in 112 different annotations and p. In the scenario of network hubs, a p is calculated from the annotations corresponding to the number of RBPs binding at each base pair along the genome. For example, the annotation associated with a network hub of 1 corresponds to all the positions on the genome that only have 1 RBP binding. For each possible value for the network hub, a value for p is determined. We use an entropy based scoring scheme, with input p, and output between 0 and 1. The following scheme is used

$$Entropy\ score = 1 + p \log p + (1-p)\log(1-p)$$

In the case of the selection pressure score, we then reweight the entropy score by the rare DAF (Section 4.4.1) by multiplying the rare DAF ratio, which takes into the GC bias associated with rare DAF, and the entropy to give a final selection pressure score. When variants fall into RBP binding sites, the max selection pressure score is given to the variant's RADAR score.

For network hubs, a similar manner is used to determine a variants network hub score. First, the variant is intersected with an annotation describing the number of RBPs that bind the position. This will give us the value of how many RBPs bind. The entropy value associated with this number of RBPs that bind (via the pre-calculated value of p) is multiplied by the corresponding rare DAF ratio (Section 4.4.1) in order to reweight the final network hub score of a variant.

When a variant disrupts a motif, we use a slightly adjusted entropy scheme to score the variant. We first determine a function between entropy score and motif D-score (Section 4.6.2) using the following formula:

$$M_e^D = 1 + p^{\geq D} \log p^{\geq D} + (1 - p^{\geq D}) \log(1 - p^{\geq D})$$

Here M_e^D is the entropy score associated for a variant that breaks a motif with D-score, D. The $p^{\geq D}$ describes the number of germline variants breaking an RBP motif with greater or equal to Dscore D, divided by the total number of germline variants. Therefore, we are able to form a function between D-score and entropy score. However, in order to make the scheme more computationally efficient, we discretize the motif function between M_e^D and D-score by using D-scores at a step size of 0.5. When a variant is given to be scored, its motif D-score is calculated, and the nearest value from the discretized function for that associated motif is used to calculate the entropy score. Finally, the entropy score here is multiplied by the rare DAF ratio of the RBP whose motif is broken.

4.8.1 Differentially expressed genes are important to the RBP regulome

RADAR allows for additional inputs in addition to the pre-built context used to calculate the baseline variant score. Differentially expressed genes are important to the RBP regulome because they add a layer of information to how variants could possibly affect functionality of RBPs. One way of approaching the feature of differentially expressed genes is to consider a dataset such as DESeq, which calculates the log fold change of tumor normal expression profiles for different cacner types. We use an entropy scoring scheme similar to the one used in the baseline scoring scheme to additionally prioritize those variants that fall in a region associated with a differentially expressed gene (corrected p-val from DESeq <0.05). Also, we consider the differential expression derived from shRNA RNA-seq experiments. If a motif of a certain RBP is broken by a variant, and the associated gene of that variant is found to be significantly differentially expressed after knockdown of that RBP (log fold change >2.5), then the variant is given extra prioroity based on the entropy scoring scheme.

4.9 RNA Binding Protein Prioritization

4.9.1 Linear regression and regulatory power

To prioritize the RBPs we use a linear regression approach. Our goal is to assess the regulatory power (positive or negative) that the RBPs have on their respective gene associated targets. For each RBP we perform a linear regression to evaluate the individual regulatory potential on a set of its target genes. Our x variable in the linear regression consists of a vector of 1s and 0s with vector length equal to the number of protein coding genes. For each gene, the corresponding position in the vector x is equal to 0 if that gene is not in the regulatory network, and 1 if it is. This vector is rather sparse, containing many more 0s than 1s. The y variable consists of a vector of protein coding gene differential expressions. We determine these differential gene expression values for 17 different cancer types, allowing us to obtain 17 different regulatory potentials, depending on tissue type. Expression data is downloaded from TCGA Data portal. The count data from RNA-Seq is used in the analysis. The goal in differential expression is to allow for the detection of an extreme value for positive or negative coefficient in the logistic regression in order to indicate upregulation or downregulation, respectively. To calculate the differential expression, DESeq2 (R Bioconductor package DESeq2 v3.5) is used, due to its flexibility in allowing varying numbers of tumor and normal samples. All cancer and normal samples are merged into categories of cancer and tumor, respectively, to determine an appropriate differential expression. Therefore, each RBP network for each cancer type satisfies a logistic regression, and the regulatory potential is inferred from the value of the coefficient. The associated p-value is also an indication of the statistical significance that such a regulatory potential exists.

4.9.2 Validation of RBPs through survival analysis

We also perform a patient wise regulatory potential linear regression, where the differential expression is determined as the individual expression fold change from a population mean. Each individual for a given cancer type is given a regulatory potential for each RBP, allowing for the regulatory potential of certain RBPs to serve as a prognosis marker. For each patient, the matching clinical XML data files are parsed for survival time. Patients who are alive use the

number of days since the last follow-up as a censored measure of survival time, but these patients are censored, so as to not contribute to an incorrect survival probability. Survival curves are plotted, with 95% confidence intervals.

4.10 Recurrence in somatic mutations

We prioritize variants that fall in elements (RBP binding sites) that are statistically enriched for somatic mutations. In order to do this, we first bin the genome using 1Mbp windows, and count the number of somatic mutations in each window. This provides us a with a local mutation rate. Then, for each RBP binding site, the number of somatic mutations are counted, and compared to the nearest local 1Mbp context using a one-sided binomial test. If the specific RBP binding site is enriched for somatic mutations, the variant falling in that specific RBP binding site is given higher priority via the entropy scoring scheme.

4.11 Resource and software accessibility

This RNA variant prioritization tool is made available as an open source python source at radar.gersteinlab.org. The website contains details on usage, examples, resources, and dependencies. A system with 10gb of RAM is recommended to avoid slowed performance for variant sets with sample size less than 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 (link). We also released the RBP-gene regulatory network at radar.gersteinlab.org (link).

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 Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827-841, doi:10.1093/nar/gks1284 (2013).

3 Pazin, M. J. Using the ENCODE Resource for Functional Annotation of Genetic Variants. *Cold Spring Harb Protoc* **2015**, 522-536, doi:10.1101/pdb.top084988 (2015).

4 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).

5 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).

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

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

8 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).

9 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).
10 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).
11 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).
12 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).
13 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).
14 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).
15 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).
16 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).
17 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).
18 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).
19 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).
20 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
21 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).
22 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).
23 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).
24 Sasado, T., Kondoh, H., Furutani-Seiki, M. & Naruse, K. Mutation in cpsf6/CFIm68 (Cleavage and Polyadenylation Specificity Factor Subunit 6) causes short 3'UTRs and disturbs gene expression in developing embryos, as revealed by an analysis of primordial germ cell migration using the medaka mutant naruto. *PLoS One* **12**, e0172467, doi:10.1371/journal.pone.0172467 (2017).
25 Ye, J. *et al.* hnRNP U protein is required for normal pre-mRNA splicing and postnatal heart development and function. *Proc Natl Acad Sci U S A* **112**, E3020-3029, doi:10.1073/pnas.1508461112 (2015).

26 van Roon, A. M. *et al.* Crystal structure of U2 snRNP SF3b components: Hsh49p in complex with Cus1p-binding domain. *RNA* **23**, 968-981, doi:10.1261/rna.059378.116 (2017).
27 Lin, P. C. & Xu, R. M. Structure and assembly of the SF3a splicing factor complex of U2 snRNP. *EMBO J* **31**, 1579-1590, doi:10.1038/emboj.2012.7 (2012).
28 Obeng, E. A. & Ebert, B. L. Charting the "Splice" Routes to MDS. *Cancer Cell* **27**, 607-609, doi:10.1016/j.ccell.2015.04.016 (2015).
29 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* **32**, 135-162 (2011).
30 Crawford, D. C., Acuna, J. M. & Sherman, S. L. FMR1 and the fragile X syndrome: human genome epidemiology review. *Genet Med* **3**, 359-371, doi:10.109700125817-200109000-00006 (2001).
31 Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**, 1434-1442, doi:10.1038/nsmb.2699 (2013).
32 Arya, A. D., Wilson, D. I., Baralle, D. & Raponi, M. RBFOX2 protein domains and cellular activities. *Biochem Soc Trans* **42**, 1180-1183, doi:10.1042/BST20140050 (2014).
33 Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep* **6**, 1139-1152, doi:10.1016/j.celrep.2014.02.005 (2014).
34 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).
35 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).
36 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).
37 Garner, C. Confounded by sequencing depth in association studies of rare alleles. *Genet Epidemiol* **35**, 261-268, doi:10.1002/gepi.20574 (2011).
38 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).
39 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).
40 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).
41 Loh, T. J. *et al.* CD44 alternative splicing and hnRNP A1 expression are associated with the metastasis of breast cancer. *Oncol Rep* **34**, 1231-1238, doi:10.3892/or.2015.4110 (2015).
42 Piton, A. *et al.* Analysis of the effects of rare variants on splicing identifies alterations in GABAA receptor genes in autism spectrum disorder individuals. *Eur J Hum Genet* **21**, 749-756, doi:10.1038/ejhg.2012.243 (2013).
43 Pala, M. *et al.* Population- and individual-specific regulatory variation in Sardinia. *Nat Genet* **49**, 700-707, doi:10.1038/ng.3840 (2017).

44 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).
45 Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10 11, doi:10.1002/0471142905.hg1011s57 (2008).
46 Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-811, doi:10.1093/nar/gku1075 (2015).
47 Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat Med* **10**, 789-799, doi:10.1038/nm1087 (2004).
48 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).
49 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).
50 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).
51 Singh, R. & Valcarcel, J. Building specificity with nonspecific RNA-binding proteins. *Nat Struct Mol Biol* **12**, 645-653, doi:10.1038/nsmb961 (2005).