RAD: A framework to annotate and prioritize post-transcriptional regulome variants for RNA binDing proteins

Abstract

====================== currently 146 words, limit 150 =======================

Dysregulation of RNA binding proteins (RBP) can cause numerous diseases, but how variants affect their regulome have been barely investigated due to the lack of annotations. The new release of ENCODE has substantially remedied this by performing large-scale eCLIP experiments. Here, we collected 112 RBP binding profiles from 318 ENCODE eCLIP experiments to accurately define the RBP regulome. We showed the majority of RBPs display significant enrichment of rare variants in their binding sites, suggesting extensive purifying selections. We further proposed an entropy based scoring framework, RAD, to investigate variant impact in RBP binding sites. Specifically, it first combines the regulator-, element-, and nucleotide-level effects to provide a baseline variant score, and also allows user-specific inputs to highlight disease- or tissue-specific variants. We demonstrate that RAD can successfully pinpoint disease-relevant germline and somatic variants missed by *state-of-the-art* scoring methods and provides complementary insights on post-transcriptional regulations.

1 Introduction and Background

Dysregulation of gene expression is a hallmark of many diseases, including cancer[1]. In recent years, the accumulation of functional characterization data on the transcription-level, such as transcription factor binding, chromatin accessibility, histone modification, and methylation, has brought great success to annotating and prioritizing deleterious variants. However, after (or simultaneously when) DNA are transcribed to premature RNAs, genes also experience a series of precise and delicately controlled processing, such as when they are converted to mature RNA, transported, translated, and then degraded in the cell. Alterations in any one of these steps may alter the final fate of gene products and result in abnormal phenotypes[2-4]. Despite its importance in regulation, the post-transcriptional regulome has been underdeveloped, partially due to its less systematic functional mapping as compared with the transcription-level regulome.

RNA binding proteins (RBPs) have been reported to play essential roles during both co- and post-transcriptional regulation[5-7]. They bind to thousands of genes in the cell through multiple processes, including splicing, cleavage and polyadenylation, RNA editing, localization, stability, and translation[8-12]. Recently, many efforts have been made to complete the annotation of the post- and co-transcriptional regulome by synthesizing public RBP binding profiles[13-16]. However, such data are usually from heterogeneous experiments with different profiling sensitivities, including HITS-CLIP, par-CLIP, and iCLIP experiments. Depending on the experiment, the number of peaks for the same protein can also differ by several orders of magnitude. Moreover, these databases may suffer from lack of quality control, e.g. lack of replicates. Since 2016, the ENCODE consortium started to release large-scale enhanced CLIP (eCLIP) experiments for hundreds of RBPs[17]. It provides high-quality RBP binding profiles with strict quality control and uniform peak calling to accurately catalogue the RBP regulome at the nucleotide resolution.

In this paper, we collected the full catalogue of 318 eCLIP experiments for 112 unique RBPs from ENCODE to construct a comprehensive RBP regulome for post-transcriptional regulation annotations. By combining polymorphism data from large sequencing cohorts, like the 1000 Genomes Project, we demonstrated that 88 and 94 percent of RBPs showed significant enrichment of rare variants in coding and noncoding regions respectively. It strongly indicates the purifying selection of the RBP regulome. Furthermore, we proposed a top-down scheme, named RAD (***R***N***A*** Bin***D***ing Protein regulome), to investigate the variant functional impact in such regions. RAD first combines the regulator, element, and nucleotide level effects to provide a baseline functional impact score for each variant. Then it allows tissue- or disease-specific inputs to incorporate features likes differential expression, somatic mutation landscape, and prior knowledge of genes to highlight relevant variants. By applying our scoring scheme on both somatic and germline variants from disease genomes, we demonstrate that RAD is able to pinpoint disease associated variants missed by other methods. Finally, we implemented our RAD annotation and prioritization framework into a software for community use (www.rad.gersteinlab.org).

2 Results

2.1 Establishing RBP regulomes by integration of 318 eCLIP experiments from ENCODE

Here we collected 318 eCLIP experiments for 112 distinct RBPs from ENCODE to fully explore the human RBP regulome. These RBPs are known to play various roles in post-transcriptional regulation, including splicing, RNA localization, transportation and decay, and translation. Many RBPs play more than one role in the cell (Supplementary Figure S1, JL).

After collecting the binding sites, merging the overlaps, and removing the blacklist region, the overall RBP regulome covers 52.6 Mbp, which is around 1.6 percent on the genome (details see methods). It is roughly 1.5 times of the size of coding regions (35.3 Mbp), but less than that of the transcription factor binding sites (3693 Mbp) and open chromatin regions (434 Mbp). Among these 52.6 Mbp RBP binding sites, 54 percent is located within annotated regions, such as coding regions, 3’ or 5’ UTRs, and introns, or their immediate extended regions (details see methods). In total, 53.1 percent of the RBP regulome is overlapped by the DNA-level regulomes, including transcription binding sites, open chromatin regions, and enhancers (Supplementary Figure S2). The limited overlap between transcription and post-transcription regulomes highlights the immediate necessity of our resource and its role as a useful complement to the many existing DNA-level efforts to annotate the human genome.

In terms of the binding sites of each RBP, we found heterogeneous binding preferences of RBPs over different annotated regions. The distribution of these binding preferences is given in Fig 1 C. Several examples were given in Table 1 and supplementary Table 1. To identify RBPs that bind together to jointly carry out certain biological functions, we calculated the co-binding coefficient for each pair of RBPs and did a hierarchical clustering of the co-binding matrix (details see methods). Using a significance level threshold of 0.02 (details see methods), we found several groups of well-known regulatory partners with different binding preferences, as summarized in Fig 1D. The discovery of the co-binding of such functional relevant proteins at various regions indicates the high quality of our regulome.

2.2 The default baseline RAD score

Our RAD framework first assigns a baseline score in a tissue- and disease- independent manner by only integrating the RBP annotations and polymorphism data. It treats peaks of each RBP equally by the same regulator score inferred from the purifying selection pressure, and then highlights peaks surrounded by multiple RBPs. Finally, it pinpoints deleterious nucleotides by quantifying the motif disruptiveness of each variant. This baseline RAD score provides a general insight to variant impact on the RBP regulome and can be useful when no appropriate prior knowledge is given for a specific variant.

2.2.1 Regulator Score from purifying selection pressure for individual RBPs

Many literatures have pointed out that enrichment of rare variants indicates purifying selection in functional regions of human genomes[18-20]. Similar to the scheme used by Khurana et al, we inferred the purifying selection pressure on the binding sites of each RBP by integrating population-level polymorphism data from large cohorts, e.g. the 1000 Genomes Project[21]. However, in our analysis we found that GC percentage usually confounds inference of the purifying selection pressure (Figure S1). This is partially because GC bias causes read coverage variations, which is a sensitive parameter in the downstream variant calling process[22,23]. Hence, we first calculated the fraction of rare variants (derived allele frequency (DAF) less than 0.5%) within each RBP's binding site, 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. Similarly, in the noncoding regions of the binding sites, 93.8 percent of RBPs (105 out of 112) exhibit a enrichment of rare variants. This observation convincingly demonstrates the accuracy of our RBP regulome definition. (Supplementary Table 3).

Some well characterized disease-causing RNA binding proteins are among the top RBPs with larger difference of rare variants fraction when comparing to the GC-corrected background regions. For example, the well-known oncogene XRN2 was reported to bind to the 3' end of transcripts to degrade aberrantly transcribed isoforms[24]. It showed significant enrichment of rare variants in its binding sites. Specifically, XRN2 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)[25]. Another example is the core splicing factor HNRNPA1, defects of which are known to cause a variety of diseases including cancer[26]. According to the profile of eCLIP binding sites, it preferentially binds to noncoding regions and controls the recognition of splice sites, and also demonstrates noticeable enrichment of rare variants (8.25%, adjusted P value is $6.70*10^{-6}$ one sided binomial tests)[27,28].

We incorporated the purifying selection pressure of individual RBPs to provide a regulator score for the variants in its binding sites. Specifically, the entropy of variants was weighted by rare variant enrichment in the binding sites of RBP to represent the regulator score (Figure 2, see details in methods). If multiple RBPs have overlapping peaks on one targeted region, the maximum score was used.

2.2.2 Element score from increased purifying selection pressure in RBP binding hotspots

It has been reported that genes within network hubs usually exhibit larger enrichment of rare variants—a sign of strong purifying selection pressure[18,19,29]. Similarly, in the RBP regulome, we suspect that binding hot spots, where multiple RBPs preferentially bind together, might demonstrate similar characteristics because once mutated these regions might have a larger chance of dysregulation of genes. To test this hypothesis, we separated the whole genome regions into different groups based on the number of RBPs that bind to each region. Due to the specificity of the RBPs, the majority (62 percent) of the regulome regions are associated with 1 RBP (Figure 3 and Supplementary Figure S4). However, we observed an obvious trend of increasing rare variants in regions where the number of RBPs increased(Fig XXX), similar to the trend observed for the gene level. For instance, in the noncoding regions, around 5 percent of the regulome is surrounded with at least 5 RBPs, exhibiting 2.2 percent more rare variants compared to the whole genome average. For regions that are surrounded by at least 10 RBPs, which are around 1 percent of the whole regulome, we observed up to 13.4 percent more rare variants (Fig XXX). This observation significantly supports our hypothesis that the RNA regulome hubs are under stronger selection pressure, and should be given high priority when evaluating the functional impacts of mutations.

Hence to quantify the element-wise (binding peaks) impact difference, we further organize the RBP regulome according to number of binding RBPs. Then regions with top 5 and 1 percent of RBPs were defined as the hot and ultra-hot region for coding and noncoding regions respectively. Similar to the regulator score, the rare variant enrichment weighted entropy value was used to weight element-wise hot and ultra-hot regions.

2.2.3 Nucleotide score from motif disruptive events

Within each element, we further differentiate each nucleotide by investigating the motif disruptive events. Loss-of-function mutations in the RBP binding peaks are more likely to alter the RBP binding events and thus cause deleterious impacts. However, due to the lack of golden standard binding motif analysis for RBPs, we performed *de novo* motif discovery for all 112 binding proteins by searching for enriched short sequences by DREME (details see methods). Many of our *de novo* discovered motifs are highly consistent with previous work (Supplementary Table 3). For example, the key splicing regulator for alternative exons in the central neural system, RBFOX2, was reported to consistently bind to a canonical sequence GCAUG[30,31]. Prior literature spanning various experimental and computational methods confirmed our analysis. Our RAD framework defined a Dscore to quantify the variant effect to RBP binding by calculating PWM P-value changes (details see methods). For each PWM altering event (Dscore greater than 3), we assigned a semi-continuous nucleotide score by counting the weighted entropy of all mutations with equal or larger Dscore.

2.3 Tissue- and disease-specific RAD score by integrating user inputs

In order to identify deleterious variants specific for diseases, our RAD scheme is able to incorporate user specific inputs at all regulator-, element-, and nucleotide- levels to increase specificity.

2.3.1 Prioritizing key regulators by incorporating gene expressions profiles

Despite the difference of purifying selection pressure, RBPs also plays key roles during post-transcriptional regulation in modulating tissue- or disease-specific expression profiles. For example, recent studies suggest that RBPs control the expression of key genes in cancer that are associated with key pathways including cell proliferation, apoptosis, and angiogenesis[1-4,32]. Therefore, we can further quantify the impact of variants associated with the regulatory potential of RBPs by including higher regulator scores for variants located in binding sites of RBPs with statistically significant regulatory potential. For example, we first built up RBP-to-gene networks directly from the binding profiles. Then RAD took either user-specific or a pre-processed expression profiles from TCGA to calculate the cancer-specific associations between each RBP and their targets (see details in methods). If a variant was found to be associated with one of such associated RBPs, an extra entropy-based regulator score was added to highlight the regulator effect.

2.3.2 Highlighting elements by gene linkage and variant recurrence

Since the majority of peaks are in gene-proximal regions, we first link each peak to genes by its shortest distance in the genome. Within the peaks of a specific RBP, we can further elevate those that are known to be associated with a disease of interest. For example, when studying diabetes, relevant genes discovered by previous GWAS studies can be used as an input to elevate the element score. Another example is in cancer studies, as many cancers have associated genes, such as COSMIC and Vogelstein cancer genes[33,34], that can be used to increase the element-wise scores.

Besides, variant recurrence is another hallmark of disease association when evaluating functional impact. Our RAD framework allows user-specific somatic variant inputs by providing extra recurrence scores to elements with more than expected somatic variants. To handle the heterogeneity in the mutation landscape, we first calculated the local mutation rate within each 1Mb bins and then used them as a background to search for enrichment of variants in RBP peaks by a binomial test (see details in method). All variants within peaks with more-than-expected variants are given additions to the element score.

2.3.3 Incorporating user-specific nucleotide scores

To accommodate studies with specific aims, RAD also allows user-specific nucleotide level scores to pinpoint variants has more focused functional impacts. For example, if users are interested in cross-species analysis, various types of conservation scores, such as Gerp, Phastcon, and Phylop, can be incorporated in RAD.

2.4 Performance of RAD on pathological germline variants

We applied RAD on pathological variants from HGMD. Since the HGMD database contains heterogeneous diseases, we only used the baseline RAD score to compare variants. For a fair comparison, we used the somatic variants, which are mainly composed of passenger mutations in cancer patients, as a rough background to compare the distribution of scores. As expected, the HGMD variants are scored significantly higher than somatic mutations (Fig XXX). For example, the mean Rscore for HGMD variants is 0.445, while it is only 0.044 for somatic variants (P value <2.2e-16 for two sided Wilcoxon test). Note that unlike the CADD score, which takes coding-focused features like Gerp, SIFT, and PolyPhen-2 scores as inputs of the training process, our score is purely based on the binding profiles from eCLIP experiments. This sharp discrepancy of pathological and background mutations demonstrates the ability of RAD to pinpoint functional variants. We further scrutinized HGMD variants that have been missed by other methods. Specifically, we found 992 HGMD variants that are highly ranked in our methods, but are not within the top-scoring list of CADD, Funseq, and Gerp score results (details in methods). 29.6% of them are noncoding variants that are located in the nearby intron, 5'UTR, and 3'UTR (and their extended regions). We focus on an intronic variant of TP53 as an example. This particular variant has a high RCORE of 3.43 (top 0.1 percent in all HGMD variants), but a moderate FunSeq score(0.999) and low CADD and Gerp score (3.316 and 0, respectively). Specifically, it is located 28 bp away from the acceptor site of exon 3 in TP53. eCLIP experiments showed strong binding evidence in 7 RBPs, including BUD13, EFTUD2, PRPF8, SF3A3, SF3B4, SMNDC1, and XRN2 (Fig XXX). The co-binding of these above mentioned splicing factors strongly indicate that this region corresponds to a key splicing regulatory site. Specifically, this A to T mutation strongly disrupts the binding motif of SF3B4 (JL2DL, need to find a dscore for this. Dscore = xx), increasing the possibility of splicing alteration effects. Our finding is not reflected in previous methods for variant prioritization.

2.5 Performance of RAD on somatic variants in cancer

2.5.1 Somatic variants associated with COSMIC genes

We applied our scheme to evaluate the deleterious effect of somatic variants from public datasets. Due to the lack of an experimentally validated golden standard, we evaluate our results from two perspectives. First, due to the efforts of the cancer community, hundreds of well-known genes have documented cancer associations from multiple aspects[33,35]. Such genes play essential roles through various pathways[34,36]. Hence, in general, variants associated with these genes are supposed to have a higher functional impact compared to others[18]. 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 (Cosmic Gene Census, CGC) showed significantly enrichment in variants with larger RNA level functional impact (Fig XXX). For example, in Breast cancer, 16,861 out of 668,286 somatic variants from 963 breast cancer patients were found to be associated with 567 CGC genes. 2.88 percent of them have a RSCORE greater than 1.5, while only 0.84 percent of the non-CGC related genes have an RSCORE greater than 1.5. Similarly, we found a 3.27 and 3.36 fold increase in high impact variants at a threshold level of 2.5 and 3 respectively. The P value for single sided Wilcoxon test is less than $2.2e^{-16}$. This pattern is consistent in all four cancer types we investigated (Supplementary Table 3, JZ).

2.5.2 Somatic variants associated with recurrence

In addition, because regions of variant recurrence can correspond to biological function and may indicate association with cancer[18-20], 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 RSCORE greater than 1.5, 2.5, and 3.0 respectively, resulting in a P value at $1.71e^{-9}$ from one-sided Wilcoxon test.

2.5.3 A case study on breast cancer patients

Currently, several variant scoring tools utilize different schemes to evaluate the functional consequences of mutations, resulting in varying perspectives of significance. For example, Gerp score profiles the evolution rate over the genome by comparing sequence similarity across species to infer purifying selection pressure[37]. However, it might under-weigh the newly evolved human specific functional regions. CADD and Funseq scores combine effects of various annotations to evaluate the deleteriousness of mutations, but they are focused more on transcriptional regulatory annotations[18-20]. Our tool provides a different perspective on variant interpretation, focusing on post-transcriptional regulation. As a comparison, we applied our method on a set of breast cancer somatic variants from 963 patients released by Alexandrov *et al*[38].

In total, around 3 percent of the 68k variants were predicted to alter post-transcriptional regulations to some degree. We first calculated the spearman rank correlation of the scores from these tools. RAD score showed the highest rank based correlation with Gerp score (0.32), and moderate correlation with CADD score (0.17), while almost no correlation with Funseq score. The relatively higher correlation between RSCORE and Gerp Score is probably due to the majority of the RNA regulome is after, or at least near simultaneous with, transcription, where the conservations scores are usually higher than the rest of the genomes. However, RSCORE uses nearly orthogonal features with Funseq during the scoring process, resulting in larger discrepancy. We further compared these methods by focusing only on variants with the highest impact. Of these variants, we selected 2906 of them by merging the top 0.1 percent of the highly scored variants for each method, and then checked whether these variants are listed among the top 1 percent of variants in each method. As expected, due to the different emphasis of each method, 2106 genes are only reported by 1 methods (72.5%). RSCORE, Funseq, Gerp, and CADD score reported 501, 630, 491, and 484 unique variants respectively.

We also find that 169 out of the 501 highly ranked variants only reported by our tool are located in the noncoding region, with 15, 28, and 24 from nearby introns, 5' UTR, and 3' UTR regions, respectively (Fig XXX). In the intronic region, 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 RAD score are within the binding peaks of Cleavage Stimulation Factor binding sites, which is strongly indicative of a role in the polyadenylation of pre-mRNAs. The discovery of biologically relevant results showcases the ability of the RAD score to classify deleterious mutations that disrupt post-transcriptional regulations.

3 Discussion

Despite its extensive involvement in various diseases, variant impact on post-transcriptional regulation has not been systematically investigated partially due to the lack of annotations. In this paper, we first build the RBP regulome for post-transcriptional regulations by collecting the full catalogue of RBP binding profiles from ENCODE. We found that 88.3 and 93.8 percent of RBPs have shown significant enrichment in rare variants in coding and noncoding regions respectively, indicating a strong purifying selection in such functional regions. Some well-known post-transcriptional regulators, such as XRN2 and HNRNPA1, which are actively involved in RNA splicing processes demonstrate the strongest purifying selection. We also defined the hot and ultra-hot binding regions that are surrounded by many RBPs. Population level polymorphism analysis have shown that such regions display even greater enrichment of rare variants and thus undergo stronger selection pressure. We further performed *de novo* motif discovery from the binding peaks of each RBP and quantified the binding affinity changes due to individual mutations. Such hierarchical post-transcriptional annotations for RBPs provides alternative insights of gene regulations and new opportunities for variant impact interpretation and prioritizations.

Based on our RBP regulome, we proposed a framework called RAD to evaluate the variants effect during post-transcriptional regulation in a step-wise manner. As compared with other variant scoring tools, RAD considers three distinct characteristics to highlight pathological variants. First, it focused on a specific type of regulation that is missed by most of other tools. Previous methods either rely of comparative genomics from multiple species or use machine learning schemes that heavily rely on transcriptional level regulation. RAD focuses on eCLIP binding profiles for RBPs that are only mediocrely covered by previous annotations (around 50%). This is confirmed by a large number of variants that can only be highlighted by RAD. Second, RAD does not require any training set, which might be biased due to our limited mutation dataset with phenotypic consequence. Results showed that our RAD method demonstrates higher scores in variants associated with cancer associated genes (like COSMIC) or binding peak elements with recurrent mutations, demonstrating its ability to pinpoint disease associated variants in multiple cancer types. Application of RAD on pathological germline variants from HGMD showed that our method could identify up to thousands of disease-causing variants that are missed by other methods. Third, RAD allows user-specific inputs such as expression, mutation, and conservation profiles to further quantify the impact of disease specific variants in different disease and tissue contexts. Finally, RAD not only provides scores for variants, but also includes detailed annotations, such as motif breaking events, to explain the mechanism of high-scoring variants.

In summary, we believe that RAD can serve as a useful tool to annotate and prioritize the post-transcriptional regulomes for RBPs, which has not been covered by most of the current variant interpretation tools. It is also able to provide additional information on top of current gene regulomes. More importantly, its scoring can be immediately compared and added on to some of the current transcriptional variant function evaluation tools, such as Funseq2, to add independent information to jointly evaluate variant impacts. With the quick expansion of binding profiles of more RBPs from more cell types, we envision that RAD 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

RNA binding proteins 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 or […]. 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 Inference of negative selection pressure from population genetics data

4.3.1 Using rare derived allele frequency as a metric for negative selection pressure

It is useful to understand the negative selection pressure associated with particular regions or locations of the genome. In order to infer the negative selection, 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. 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 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 = 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 4.3.2) in order to obtain a ratio. Regions with rare DAF ratios larger than 1 suggest an above average negative selection pressure.

4.3.2 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.

4.3.3 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.

4.4 Co-binding and Hotness (need to brainstorm another title)

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 "hotness". Regions with more RBPs binding are deemed to be more "hot" than locations with fewer RBPs binding. We hypothesize that the hotness of a region and the selection pressure of the region demonstrate a positive relationship. To determine the actual relationship, we annotate the genome with hotness 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 hotness for any given hotness. 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 hotness and predicting the genomic rare DAF average (4.3.2). 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 hotness of 10, corresponding to the top 1% of the data. Furthermore, regions with hotness less than 5% of the data, equal to a hotness of less than 5, are deemed to not be hot, and are automatically given a 0 value in rare DAF ratio. The resulting discrete function is smoothed from hotness of 5 to 10. The function steps from 0 (from hotness of 1 to

4) to the rare DAF ratio at 5, and also maintains a constant rare DAF ratio for hotness values over 10 by rounding them down to 10.

Many RBPs bind in similar locations across the genome, and this is measured by their co-binding percent. The co-binding between two RBPs, A and B, is defined to be the maximum ratio between the peaks that intersect between A and B and the total number of peaks for A or B. Intersection is defined for greater than or equal to one base pair. Here, the maximum is taken in order to allow for a symmetric matrix in plotting a co-binding heatmap, resulting in only a unique possible result for clustering RBPs by similarity of co-binding. Using the co-binding ratio values between pairwise RBPs, a symmetric matrix is constructed and clustering is performed. The R function pvrect in package pvclust is used for clustering with an alpha value of 0.02 instead of 0.05 in order to avoid clusters with large numbers of RBPs (>6). The resulting clusters of RBPs with significance were found to follow patterns of functional co-binding found in literature.

4.5 Motif analysis

4.5.1 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.

4.5.2 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.

$$\textbf{D-score} = \textbf{motif-score}_{\textbf{ref}} - \textbf{motif-score}_{\textbf{alt}} = -10 \cdot \log_{10}\left(\frac{p\text{-}value_{ref}}{p\text{-}value_{alt}}\right)$$

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 were ***averaged***(we need to decide if we average or max) over all D-scores.

4.6 Variant Scoring

4.7 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.

4.8 RNA Binding Protein Prioritization

4.8.1 Logistic regression and regulation potential (add the DEseq analysis, have the software version clearly labeled)

To prioritize the RBPs we use a logistic regression approach. Our goal is to assess the regulatory potential (positive or negative) that the RBPs have on their respective gene associated targets. For each RBP we perform a logistic regression to evaluate the individual regulatory potential on a set of its target genes. Our explanatory variable, y, in the logistic regression consists of a vector of 1s and 0s with vector length equal to the number of protein coding genes, xxx. For each gene, the corresponding position in the vector y 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 x variable consists of a vector of protein coding gene differential expressions. We determine these differential gene expression values for 24 different cancer types, allowing us to obtain 24 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.

4.8.2 Survival analysis

We also perform a patient wise regulatory potential logistic 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. Survival curves are plotted, with 95% confidence intervals.

4.9 Resource and software accessibility

This RNA variant prioritization tool is made available as an open source python source at xxx. 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 RAD score for every basepair on the genome (hg19 version of genome). Users can directly query the annotation and functional impact score from rad.gersteinlab.org (link). We also released the RBP-gene regulatory network at rad.gersteinlab.org (link).

5 Acknowledgement

This work was supported by the National Institutes of Health [5U41HG007000-04].

6 Author Contributions

MG and JZ designed the whole framework. JZ, JL, and JJ implemented the software. JL processed the expression data from TCGA. DL, JL, and SKL did the motif analysis. LL, JZ, and JL processed the somatic and germline variants. MG, JZ and JL wrote the manuscript.

7. Materials and Correspondence

The software and a pre-built baseline score can be downloaded in rad.gersteinlab.org. Correspondence for questions: pi@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 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).

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

20 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).

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

23 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).

24 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).

25 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).

26 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).

27 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).

28 Pala, M. *et al.* Population- and individual-specific regulatory variation in Sardinia. *Nat Genet* **49**, 700-707, doi:10.1038/ng.3840 (2017).

29 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).

30 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).

31 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).

32 Wurth, L. Versatility of RNA-Binding Proteins in Cancer. *Comp Funct Genomics* **2012**, 178525, doi:10.1155/2012/178525 (2012).

33 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).

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

35 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).

36 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).

37 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).

38 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).