

JZ2JL&LL: use of colors

Blues text: words that I dislike but can not find a better phrase, please edit

JZ2JL: please add reference at places I marked as `\cite{JL_add}`, try Nature big papers and papers from NAR, RNA, Plos Computatiobal Biology in the first round as much as you can.

Comments: highlights yellow

Action terms: highlights green

PASPorT: A Post-transcriptional regulome Annotation, Scoring, and Prioritization Tool for RNA binding proteins

Abstract

===== v1 (199) words =====
Numerous RNA binding proteins (RBP) interact with single or double stranded RNAs to play critical post-transcriptional roles and their dysregulation has been reported to cause numerous diseases. However, the functional impacts of individual variants in RBP regulomes have been barely investigated, partially due to the lack of large-scale binding events profiling. The ENCODE Consortium has substantially remedied this by uniformly producing high quality eCLIP data on a variety of RBPs as part of the ENCODE data release. Here, we collected 112 RBP binding profiles from 318 eCLIP experiments from ENCODE to annotate the most comprehensive post-transcriptional regulome for RBPs. We discovered that around ~~88%~~ and 93.8 percent of the RBP binding sites show significant enrichment of rare variants, which support the claim that RBP binding sites have high functionality and purifying selection pressure. In addition, based on population-level polymorphism data we proposed an entropy based scoring system to evaluate the functional impact from mutations on a post-transcriptional level. It successfully pinpoints relevant germline and somatic mutations missed by state-of-the-art variant scoring methods. Furthermore, we further linked the RBPs to genes to build RNA regulatory networks and utilized a regression based scheme to prioritize key regulators, such as BCCIP, that drive disease specific gene expressions. Survival analysis confirmed the association of BCCIP regulation with patient survival in multiple cancer types. In summary, we released our tool PASPorT at passport.gersteinlab.org to annotate, score, and prioritize the post-transcriptional regulome in disease genomes, which provides complementary regulation insights from the RNA level.

(Annotate + Score + Prioritize)

1 Introduction and Background

Dysregulation of gene expression are hallmarks of many diseases, including cancer `\cite{19763153}`. In recent years, with the development of sequencing technologies, various types of assays have been used to systematically map the functional elements in human genome, such as transcriptional factor binding site, chromatin accessibility, histone modification, and methylation in numerous cell types `\cite{23221638,25762420,25693562}`. The accumulation of such data has led to a comprehensive catalogue of functional elements on the transcription-level and has brought great success to annotating variants and pinpointing deleterious mutations that are associated with diseases. However, after (or simultaneously while) DNA are transcribed to premature RNAs, genes experienced a series of precisely and delicately controlled processing, including being processed into mature RNA, transported, translated, and then degraded in the cell. Dysregulation of any one of the following steps may alter the final fate of gene products and

result in abnormal phenotypes \cite{23945584,25159663,15211354}. Despite its importance in regulation, the post-transcriptional regulome has been underdeveloped, partially due to its less systematic functional mapping as compared with the regulome on a DNA level.

RNA binding proteins (RBPs) have been reported to play essential roles during both co- and post-transcriptional regulation \cite{25365966,21822212,16769980,16769980}. They have been reported to bind to thousands of genes to govern the fate of transcribed genes in the cell through multiple processes, including splicing, cleavage and polyadenylation, RNA editing, localization, stability, and translation \cite{11994740,25112293,25201104,24916387,18342629}. Recently, many efforts have been made to complete these post- or co-transcriptional regulomes by synthesizing public RBP binding profiles \cite{24297251,25416797,22086949,28053162}. However, such data are usually from heterogeneous experiments with different profiling sensitivity, including HITS-CLIP, par-CLIP, and iCLIP experiments. Also, these databases may suffer from non-uniform quality control effects, e.g. lack of replicates in each experiment. Since 2016, the ENCODE community started to release large-scale RBP profiling data from enhanced CLIP (eCLIP) experiments \cite{27018577}. It not only provides improved specificity in defining binding sites for RBPs with single nucleotide resolution, but also uniformly processed data with strict quality control.

#!/*== JZ2MG: should we add so much result part in the introduction, sounds like a little repetitive with the ==*/

In this paper, we collected the full catalogue of eCLIP experiments from ENCODE to build a comprehensive RNA regulome to annotate the functional regions after the transcription process. It contains the binding profiles of 112 RBPs from 318 eCLIP experiments, providing an unprecedented opportunity for us to annotate variants. By combining eCLIP data with population genetics data from large sequencing cohorts, like the 1000 Genomes Project, we were able to propose a scoring system that considers the annotation, network, and motif disruption events to evaluate the functional consequence of mutations \cite{26432245,26432246,23128226,20981092}. By applying our scoring system on both somatic and germline variants from disease genomes, we demonstrate that our scoring scheme is able to pinpoint the disease associated variants. We further built up RNA regulatory networks by linking RBPs to their respective gene targets and quantify the regulatory potential on disease-specific gene expression patterns in multiple cancer types through a regression scheme. Application of our scheme on the full catalog of TCGA data indicates that novel RNA regulators, such as BCCIP, significantly drive tumor specific gene expressions in multiple cancer types. We implemented our Post-transcriptional regulome Annotation, Scoring, and PriORitization Tool for RNA binding proteins into a software for community use (passport.gersteinlab.org).

2 Results

2.1 Establishing RNA regulomes through integration of hundreds of eCLIP experiments from ENCODE

2.1.1 Summary of the RBP binding profiles

Here we collected 318 eCLIP experiments for 112 distinct RBPs from ENCODE to fully explore the human RNA 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 table 1, Figure S1, JL to prepare).

After collecting the binding sites, merging the overlaps, and removing the blacklist region, the overall RNA regulome by all RBPs covers 52.6 mbp, which is around 1.6 percent on the genome (details see methods). It is roughly comparable to the length of the 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 has allocated within annotated regions, such as coding regions, 3' or 5' UTRs, and introns, or their immediate extended regions (details see methods). In total, xxx percent of the RNA regulomes are overlapped by the DNA-level regulomes, including transcription binding sites, open chromatin regions, and enhancers (supplementary table 2, Figure S2, JL to

prepare). The limited overlapped between transcription and post-transcription regulomes highlights the immediate necessity of our resource to serve 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 are given in Figure 1 C. Specifically, RBPs like XXX preferentially bind to the coding regions, with xx percent of coding region peaks. On the other hand, RBPs like xxx and xxx preferentially bind to the noncoding regions (xxx and xxx percent respectively). Besides, we found that the cleavage and polyadenylation specific factors CPSF6 shows noticeably enriched binding events in the 3' UTR and extended regions (Figure S3 JL to prepare). Xxx percent of its peaks are located within such regions, confirming its documented role in cleavage site \cite{28253363}.

2.1.2 Co-binding analysis identified well-known collaborating RBPs with diverse functions

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 0.05, we found several pairs of well-known regulatory partners with different binding preferences. For example, the famous heterogeneous nuclear ribonucleoprotein (hnRNP) family protein HNRNPU and its paralog HNRNPU1 were found to bind together in the nearby intron region, probably regulating the pre-mRNA splicing process (Fig X C) \cite{26039991}. SF3A3 and SF3B4, which encode two units of splicing factor 3a protein, were also found to co-bind in the nearby intron region in our data (Fig X C) \cite{28348170,22314233}. The SR family protein U2AF1 and U2AF2 are found to co-bind near the intron/exon junctions to jointly control splicing events (Fig X C) \cite{22314233,25965565}. Two cleavage stimulation factor (CSTF) complex proteins, CSTF2 and CSTF2T, were found to bind near the 3' UTRs, and were reported to be associated with 3' end cleavage and polyadenylation of pre-mRNAs. Consistent with previous report, three functional similar genes FMR1, FXR1, and FXR2 were found to co-express, and shuttle between the nucleus and cytoplasm and associate with polyribosomes, predominantly with the 60S ribosomal subunit \cite{21912443,11545690}. The discovery of the co-binding of such functional relevant proteins at various regions indicates the high quality of our regulome.

2.2 De novo motif discoveries from eCLIP binding profiles

To identify the context effect of the binding events, we performed de novo motif discovery for all 112 binding proteins. We used DREME to search for enriched short sequencing using the default settings (details see methods). Some of our *de novo* motifs are highly consistent with previous work (supplementary table 3, JL to add). 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 \cite{24213538,25110022}. Prior literature spanning various experimental and computational methods confirmed our analysis. We also checked the binding preference of RBFOX2 with respect to exon/intron junctions. Consistent with previous reports, we found RBFOX2 to preferentially bind to regions within xxx bp of splice junctions, convincingly confirming its role in splicing regulation.

#!/*==== JZ2DL: more to add in this section ====*

Furthermore, since the GCAUG motif plays an essential role in RBFOX2 binding and cross linking events, disruption of such canonical motifs usually significantly alters the splicing regulation process, and results in abnormal phenotypes \cite{24613350}. Hence, we investigated motif disruption events in both normal and tumor genomes. In normal genomes, we found that the two guanine positions are the top two most frequently disrupted sites. Furthermore, as compared to other RBPs, RBFOX2 undergoes significantly more severe motif disruption events. Similarly, in tumor genomes, the binding sites of RBFOX2 are more severely disrupted, which are possibly linked to its oncogenic role in multiple cancer types.

2.3 Extensive purifying selection in the RNA regulome

NOT COMMENT

2.3.1 Purifying selection pressure for individual RBPs

Lines of literatures have demonstrated that the enrichment of rare variants is a signature of functionality in human genomes \cite{25273974,24092746,24487276}. Similar to the scheme used by Khurana et al, we analyzed the purifying selection pressure by utilizing the full set of polymorphism data from the 1000 Genomes Project through a \cite{26432246} on our defined RNA regulomes. However, GC percentage is a well-known confounding factor of high throughput genome sequencing technology that affects the read depth at each loci, which further confounds the following variant calling process \cite{21328616,28049435}. If uncorrected, it will severely confound our purifying selection pressure inference. 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. In total, in coding regions, 99 out of 112 RBPs show elevated rare variant fraction compared to the background regions after GC correction, and 105 out of 112 were found in the noncoding regions. This observation convincingly demonstrates the premier quality of the binding sites from eCLIP experiments, indicating the functionality of our annotated regions. (supplementary table 4, JL 2 prepare).

Some well characterized disease-causing RNA binding proteins are among the top RBPs with larger difference of rare variants fraction when comparing to the background regions. For example, the well-known oncogene XRN2 was reported to bind to the 3' end of transcripts to degrade aberrantly transcribed isoforms \cite{19915612}. It showed significant enrichment of rare variants in its binding sites. Specifically, XRN2 demonstrates 12.7% and 10.3% more rare variants and ranks 2nd and 5th in coding and noncoding regions respectively (adjusted P values are 1.89×10^{-9} and 2.85×10^{-118} for one sided binomial tests) \cite{22522706}. Another example is the core splicing factor HNRNPA1, defects of which are known to cause a variety of diseases including cancer \cite{26151392}. 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) \cite{23169495,28394350}.

#!/*===== candidate example with good results but no bio story =====*/

For example, we found that the immunophilin protein family protein FKBP4, which plays a role in immunoregulation and basic cellular processes involving protein folding and trafficking, showed significant enrichment of rare variants in both coding and noncoding regions. Specifically, it showed 23.6% and 11.0% more rare variants in coding and noncoding regions respectively, resulting in an adjusted p value of 2.20×10^{-6} and 4.07×10^{-41} (one sided binomial test, ranked 1st and 3rd for coding and noncoding).

2.3.2 Increased purifying selection pressure in network hubs

It has been reported that genes within network hubs usually exhibit greater enrichment of rare variants—a sign of strong purifying selection pressure \cite{25273974,24092746,23505346}. Similarly, in the RNA regulome, we suspect that binding hot spots, where multiple RBPs preferentially bind, might demonstrate similar characteristics. Once mutated, these regions would result in dysregulation of gene networks. 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 of the regulome regions are associated with 1 RBP (xxx percent, Fig 3 A and Fig S5). However, as the number of RBPs increased, we observed an obvious trend of increasing rare variants (Fig xxx B), 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, and they exhibited 3 (JZ2JL: recalculate this number, use the normalized one) percent more rare variants compared to the whole genome. For regions that are surrounded by at least 10 RBPs, which are around 1 percent of the whole regulome, we observed up to 12 percent more rare variants (Fig xxx B). 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.

2.4 Entropy based scoring scheme helps to evaluate somatic variant effects

[[JZ2everyone: we never mentioned the annotation, hotness here, ppl will be confused, think about other names for this 3 groups]]

By integrating the information gained from ~~our~~ annotations, designation of hot spots, and motif analysis of RBP binding profiles, we proposed an entropy based scoring scheme to investigate the functional impacts of somatic variants specific to post transcriptional regulation (Fig xxx A). First, we combined both entropy and purifying selection pressure in the binding peaks of each RBP to evaluate an individual annotation score. For each position, our scoring system selected the highest annotation score for all possible RBPs with peaks overlapping this position. Then, we quantified the RBP hot spots in the target region, and assigned higher hotness scores accordingly. Finally, we considered potential changes in binding motifs defined by changes in PWM score. We added up these three scores to evaluate the RNA regulatory potential for individual variants (Figure xxx).

2.4.1 Somatic variants associated with COSMIC genes

We applied our scheme to evaluate the deleteriousness of somatic variants from public datasets. Due to the lack of an experimentally validated golden standard, we evaluate our results from two aspects. First, due to the efforts of the cancer community, hundreds of well-known genes have documented cancer associations from multiple aspects \cite{18428421,25355519}. Such genes play essential roles through various pathways \cite{15286780,26781813}. Hence, in general, variants associated with these genes are supposed to have a higher functional impact compared to others \cite{25273974}. To test this hypothesis, we first associated each variant with a gene by the shortest distance according to Genecode 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 xxx). For example, in Breast cancer, 16,861 out of 668,286 somatic variants from xxx 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 5, JZ 2 prepare).

2.4.2 Somatic variants associated with recurrence

In addition, because variant recurrence is considered a sign of functionality and may indicate association with cancer \cite{25273974,24092746,24487276}, 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.

2.4.3 A case study on breast cancer patients

Currently, several variant scoring tools utilized different schemes to evaluate the functional consequences of mutations with different emphasis. For example, Gerp score profiles the evolution rate over the genome by comparing sequence similarity across species to infer purifying selection pressure \cite{21152010}. 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 the transcriptional regulatory annotations \cite{25273974,24092746,24487276}. Our tool provides a different perspective on variant interpretation. As a comparison, we applied our method on a set of breast cancer somatic variants from xxx patients released by xxx \cite{24657537}.

#!/*===== JZ2MG: is it good to show the following highlighted comparison? =====*/

In total around 3 percent out of the 68k variants was evaluated to alter post-transcriptional regulations to some degree. We first calculated the spearman rank correlation of the scores from these tools. RSCORE showed 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 the highest impact variants. First, we selected 2906 impactful variants 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.

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. S X). 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 RSCORE 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 RSCORE to differentiate deleterious mutations that disrupt post-transcriptional regulations.

2.5 Scoring on pathological germline variants

We applied our method on pathological variants from HGMD. 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 XX). 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-pruned features like Gerp score, 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 PASPorT 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 top 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 zoomed in to an intronic variant of TP53 as an example. It has a high RSCORE at 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 XX). 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 (Dscore = xx), increasing the possibility of splicing alteration effects. Our finding is not reflected in previous methods for variant prioritization.

2.6 Combining RNA regulatory networks help to prioritize key regulators for cancer genomes

In order to measure the regulatory activity of RBPs, we first constructed the networks that associate with each individual RBP (4.7). From here, we run our logistic regression for each RBP and each of the 23 cancer types from TCGA, incorporating information from tissue specific gene expressions (TCGA) as well as nodes in the gene network. From the logistic regression, we obtain a value (in the form of the regression coefficient) that can be interpreted as the regulatory activity of an RBP. In general, the regulatory activity across RBPs within a tissue type varies, including both positive and negative regulatory RBPs. However, some cancer types, such as READ and COAD demonstrate predominantly RBPs with positive regulatory activities while the reverse is true in CHOL (Fig XX). Besides the heterogenous nature of RBP regulatory behavior within a cancer type, there exists differing RBP regulatory activities across different cancer types, further supporting the idea that cancer tumors are heterogenous in nature \cite{16077728}.

While the regulatory activity of many RBPs are not well defined, we do find some examples that are corroborated by biological examples in literature. QKI is an RBP that is known to regulate pre-mRNA splicing, and serves as a repressor with BRCA2 in breast cancer. Its repressive behavior is reflected in the results of our logistic regression, demonstrating an average coefficient for QKI of -0.046. In breast cancer, the coefficient is -0.123, suggesting a strong

repressive nature. For comparison, the strongest repressive regulatory activity determined by our model corresponds to a coefficient of -0.160, which happens to correspond to QKI as well, in thymoma.

3 Discussion

In this paper, we collected the full catalogue of RBP binding profiles from ENCODE to build the RNA regulome for post-transcriptional regulations. We found that 88.3 and 93.8 percent of RBPs have shown significant enrichment in rare variants after removing the GC content effect in coding and noncoding regions respectively, strongly indicating the involvement of gene regulations of the binding regions. Some well-known RNA regulators, such as XRN2 and HNRNPA1, which are actively involved in RNA splicing process have shown the strongest purifying selection pressure in both coding and noncoding regions. We also found that some RNA regulatory hotspots, which are surrounded by many RBPs, have shown elevated 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 reported the binding motifs. Many such motifs match reports from literature. We evaluated the motif breaking events from both germline and somatic mutations and identified several RBPs that are more significantly disrupted in specific cancer types. Interestingly, we found that binding sites of the splicing factor RBFOX2 are significantly more disrupted in tumor cells, confirming its oncogenic role from lines of literatures.

By integrating the RBP related annotation, binding hot spots, and motif alteration events, we proposed an entropy based scoring system to score both somatic and germline variants. Different from most previous scoring methods, which either rely on comparative genomics from multiple species or use machine learning schemes that heavily rely on transcriptional level regulations, our method used only pre-built data contexts based on eCLIP binding profiles and population level polymorphisms from 1000 Genomes Project. It does not only focus on a new angle of regulations events that have been missed by many current methods, but also does not require any training set, which might be biased due to our limited mutation dataset with phenotypic consequence. Results showed that our PASPorT method demonstrate higher scores in variants associated with cancer associated genes (like COSMIC) or elements with recurrent mutations, demonstrating its ability to pinpoint disease associated variants in multiple cancer types. Application of PASPorT 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. In addition, PASPorT provides detailed annotations, such as motif breaking events, to explain the mechanism of highly ranked variants.

We further built up the RNA regulatory network and proposed a regression based scheme to identify key RNA regulators that drives tumor specific expression patterns. Applying this method to a set of xxx expression profiles from TCGA, we identified several RBPs such as PUM2 and BCCIP to be significantly associated with tumor specific expressions. As a validation, the regulatory potential of BCCIP has been found to be associated with patient survival in multiple cancer types.

In summary, we believe that PASPorT can serve as a useful tool to annotate, score, and prioritize the post-transcriptional regulomes for RBPs, which has not been covered by most of the current variant functional impact interpretation tools. It is also able to provide additional information on top of the 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 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 $-\log_{10}(P\text{-value}) \geq 3$ and $\log_2(\text{fold_enrichment}) \geq 3$.

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

$$\text{Rare DAF} = r_i / (r_i + c_i)$$

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 `pvcust` 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.

{List options used, etc.}

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.

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

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,

DISC

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.

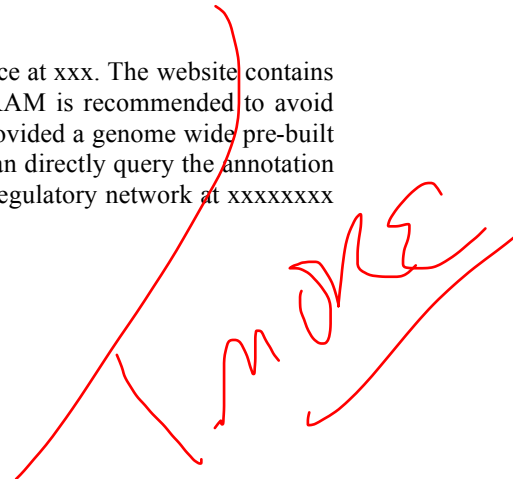
4.8.2 Survival analysis

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 PASPort score for every basepair on the genome (hg19 version of genome). Users can directly query the annotation and functional impact score from xxxxxxxx (link). We also released the RBP-gene regulatory network at xxxxxxxx (link).

References



more