UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE BIOLOGIA

FELIPE EDUARDO CIAMPONI

IDENTIFICATION OF BIOLOGICAL CHARACTERISTICS ASSOCIATED WITH RNA-BINDING PROTEINS (RBPs) TARGET SITES

IDENTIFICAÇÃO DE CARACTERÍSTICAS BIOLÓGICAS ASSOCIADAS A SÍTIOS-ALVO DE PROTEÍNAS DE LIGAÇÃO AO RNA (RBPs)

CAMPINAS

2018

FELIPE EDUARDO CIAMPONI

IDENTIFICATION OF BIOLOGICAL CHARACTERISTICS ASSOCIATED WITH RNA-BINDING PROTEINS (RBPs) TARGET SITES

IDENTIFICAÇÃO DE CARACTERÍSTICAS BIOLÓGICAS ASSOCIADAS A SÍTIOS-ALVO DE PROTEÍNAS DE LIGAÇÃO AO RNA (RBPs)

Dissertation presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Genetics and Molecular Biology in the field of Animal Genetics and Evolution

Dissertação apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do Título de Mestre em Genética e Biologia Molecular na área de Genética Animal e Evolução

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO FELIPE EDUARDO CIAMPONI E ORIENTADO PELA DRA. KATLIN BRAUER MASSIRER.

Orientador: Katlin Brauer Massirer

CAMPINAS

2018

Agência(s) de fomento e nº(s) de processo(s): FAPESP, 2012/00195-3; BNDES
ORCID: https://orcid.org/0000-0002-0076-882

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

C481i

Ciamponi, Felipe Eduardo, 1991-
Identification of biological characteristics associated with RNA-binding proteins (RBPs) target sites / Felipe Eduardo Ciamponi. – Campinas, SP : [s.n.], 2018.

Orientador: Katlin Brauer Massirer.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Bioinformática. 2. Proteínas de ligação ao RNA. 3. Transcriptoma. I. Massirer, Katlin Brauer, 1975-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Identificação de características biológicas associadas a sítios-alvo de proteínas de ligação ao RNA (RBPs)
Palavras-chave em inglês:
Bioinformatics
RNA-binding proteins
Transcriptome
Área de concentração: Genética Animal e Evolução
Titulação: Mestre em Genética e Biologia Molecular
Banca examinadora:
Katlin Brauer Massirer [Orientador]
Marcelo Alves da Silva Mori
Paulo Sergio Lopes de Oliveira
Data de defesa: 16-02-2018
Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, 16/02/2018.

COMISSÃO EXAMINADORA

Profa. Dra. Katlin Brauer Massirer

Prof. Dr. Marcelo Alves da Silva Mori

Prof. Dr. Paulo Sérgio Lopes de Oliveira

Os membros da Comissão Examinadora acima assinaram a Ata de defesa, que se encontra no processo de vida acadêmica do aluno.

Dedicatory

I dedicate this work to my family and friends, specially to my recently deceased grandfather, Durval Ciamponi, who was one of my greatest inspirations for becoming a scientist and always encouraged me to be the best that I could, both as a person and as professional. Wherever you are now, know that if I'm standing here today, it is in large part due to your guidance and support. Thank you.

Acknowledgments

I would like to thank my family and friend for all the support in this journey. Always by my side, crying and celebrating with me at every step. I would not have accomplished this work without your support and dedication. You helped me through a lot of bad times and also cheered with me at the good times. Thank you all for everything.

To my supervisor, Dra. Katlin Brauer Massirer, for the opportunity, patience and guidance during my 5 years in her lab.

To my lab colleagues, who became my second family during these years. With special thanks to Laura Alonso, Natacha Migita and Pedro Cruz who worked directly with me in projects presented here.

To Prof. Dr. Mário Henrique Bengtson and Prof. Dr. Marcelo Mendes Brandão, who provided valuable insights during several discussions.

To Prof. Dr. Marcelo Alves da Silva Mori, Dr. Michel Yamagishi and Prof. Dr. Henrique Marques-Souza for valuable feedback and ideas provided during my qualification exam.

I would like to extend a special thanks to my co-supervisor Michael Thomas Lovci, who was more than a scientific advisor. Mike was not only a friend but also my mentor during the first steps as a bioinformatician, teaching me how to write my first lines of code and how to “think big picture without losing focus on the small details”. This work would not exist without him.

Lastly, I would also like to thanks the funding agencies CAPES and FAPESP for the fellowships and financial support to carry out this study (process nº 2016/25521-1, Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP)).

Resumo

Grânulos de stress são agregados de proteínas e RNAs encontrados no citoplasma das células, em geral produzidos em resposta uma forma de stress (ex. Hipóxia, infecções virais, privação de nutrientes e choques térmicos). No entanto, diversas doenças neurodegenerativas, como esclerose lateral amiotrófica e Alzheimer, já foram associadas com inclusões patogênicas desses agregados que geram consequências nocivas para a célula. Dentre as proteínas presentes no granulo de stress o complexo G3BP1-CAPRIN1-USP10 é essencial para a condensação do granulo e sua associação com subunidades ribossomais, sendo que a expressão ectópica de CAPRIN1 é suficiente para induzir a formação desses agregados. Apesar dos mecanismos moleculares associados aos grânulos não estarem completamente elucidados, as principais funções propostas para estas estruturas são: a proteção dos RNAs de situações danosas, degradação de mRNAs alvo, seleção da tradução de mRNAs de resposta a stress e reprogramação da expressão gênica (transcriptoma).

Para caracterizar os mRNAs-alvo de Caprin-1 nos grânulos de stress, utilizamos da combinação de técnicas de CLIP-seq, RIP-seq e RNA-seq para determinar os sítios de ligação que Caprin-1 apresenta nos alvos, bem como quais classes funcionais de transcritos estão enriquecidas nesse tipo de amostra. Nossas análises revelaram que Caprin-1 apresenta uma preferência de ligação por regiões organizadas em forma de “stemloops” na estrutura secundária do mRNA. Adicionalmente, essas regiões estão enriquecidas em repetições “GG”, que podem sugerir a formação de estruturas secundárias do tipo G-quadruplex, sendo mais estáveis para o nosso modelo encontrado que os loops. Do ponto de vista funcional, encontramos que os alvos identificados pelo CLIP-seq estão enriquecidos em transcritos codantes para proteínas da ligação ao RNA, associadas principalmente com processos catabólicos e controle do ciclo celular. Adicionalmente, comparamos as categorias enriquecidas com alterações preditas nas vias metabólicas obtidas a partir do RNA-seq, encontrando 19 vias que estão simultaneamente alteradas após expressão ectópica de Caprin-1 e enriquecidas nos alvos ligados a proteína nos grânulos de stress. Por fim, identificamos que sítios de ligação associados a Caprin-1, encontramos que estes apresentam uma tendência a apresentarem mais sítios de

ligação microRNAs e estão associados também a outra proteina de ligação ao RNA chamada PUM2.

Em conjunto, esses dados nos permitiram coletar informações importantes em relação ao papel da Caprin-1 nos grânulos de stress, tanto no campo de seleção de alvos de ligação bem como nas alterações funcionais decorrentes da expressão ectópica desta proteína. Nossos achados corroboram, com novas abordagens, alguns dados já sugeridos anteriormente pela literatura bem como propondo novos mecanismos que não haviam sido descritos previamente para esse modelo.

Abstract

Stress granules are protein and RNA aggregates found in the cytoplasm of cells, in general are produced in response to a source of stress (ex. Hypoxia, viral infections, nutriente deprivation and heat shock). However, several neurodegenerative diseases, such as amyotrophic lateral sclerosis and Alzheimer's disease, have already been associated with pathogenic inclusions of these aggregates with harmful consequences for the cells. Amongst the proteins presente in the stress granule, the complex G3BP1-CAPRIN1-USP10 is essential for the condensation fo the granule and it's association with ribosomal subunits, with the ectopic expression of CAPRIN1 being suufficient to induce the formation of these aggregates. Although the molecular mechanisms associated with stress granules are not completely elucidated, the main functions postulated for these structures are: protection of RNAs from harmful situations, degradation of target mRNAs, selection of stress-response mRNAs for translation and reprogamming of overall gene expression.

In order to characterize the mRNA-targets of Caprin-1 in stress granules, we used a combination of eCLIP-seq, RIP-seq and RNA-seq techiniques to identify the RNA binding sites for Caprin-1, as well as identifying which functional classes of transcripts are enriched in these samples. Our analysis revealed that Caprin-1 posesses a preference for binding to stemloops RNA secondary structures, additionally these regions were also enriched in GG repeats, which might suggest the formation of secondary structures known as G-quadruplexes, which are more stable than the stemloop model. From the functional standpoint, we found that targets identified by eCLIP-seq are enriched in transcripts coding for other RNA-binding proteins, associated mostly with catabolic processes and cell cycle control. Additionally, we compared the enriched categories with predicted alterations in metabolic pathways obtained from RNA-seq data. Overall, we found 19 pathways which are simultaneously enriched in eCLIP-seq targets and had significant alterations in their activation after CAPRIN1 ectopic expression. Lastly, we identified that Caprin-1 binding are also enriched in target sites for microRNAs and are associated with PUM2, another RNA binding protein.

Taken together, our data allowed us to gather important information on the role of Caprin-1 in the stress granules, both for the selection of binding targets as well

as the functional alterations resulting from the ectopic expression of this protein. Our finds corroborate, with new approaches, data previously suggested in the literature as well as propose novel mechanisms previously unreported for this model.

Figure index

Appendix 1 (Article) - BioFeatureFinder: Flexible, unbiased analysis of biological characteristics associated with genomic regions

Table index

Appendix 1 (Article) - BioFeatureFinder: Flexible, unbiased analysis of biological characteristics associated with genomic regions

Abbreviations

ALS: Amyotrophic lateral sclerosis
aMI: Adjusted mutual information
AUC: Area under curve
BFF: BioFeatureFinder
CDF: Cumulative distribution function
CDS: Coding sequence
ChIP-seq: Chromatin Immunoprecipitation sequencing
CLIP-seq: Cross-linking Immunoprecipitation sequencing
CpG: 5'—C—phosphate—G—3'
DNA: Deoxyribonucleic acid
eCLIP-seq: enhanced Cross-link Immunoprecipitation sequencing
fE: fold Enrichment
FTD: Frontotemporal dementia
GFF: General feature format
GTF: General transfer format
IRES: Internal ribosomal entry site
KDE: Kernel density estimation
KST: Kolmogorv-Smirnov Test
MFE : Minimum free energy
MSE: Mean squared error
N.P.V.: Negative predictive value
P.P.V.: Positive predictive value
P-R AUC: Precision-Recall area under curve
pVal: p-value
RBP: RNA-binding protein
RNA: Ribonucleic acid
ROC AUC: Receiver operating characteristic area under curve
SG: Stress granule
SNP: Single-nucleotide polymorphism
St-GBCLF: Stochastic gradient boost classifier
UTR: Untranslated region

Index

Introduction

Stress granules (SGs) are cytoplasmic aggregates composed of protein and RNA complexes commonly found in cells exposed to stress conditions, such as heat, UV radiation exposure, presence of reactive oxygen species, starvation and hypoxia [1,2]. Amongst suggested mechanisms for assembly of these aggregates, one of the is the hyperphosphorylation of the translation factor eIF2α by kinases as PERK, HCN2, Z-DNAk, PKR and HRI. This phosphorylation cascade causes the blockage of the translation initiation machinery, dismantling the polysomes and recruiting RNAs and proteins to the cytoplasmic granules, one example is the reduction of assembly viability of the eIF2α-GTP-tRNAi^{MET} complex [1–3]. In healthy cells, specially neurons, stress granules assemble and disassemble regularly in response to temporary factors, however in some neurodegenerative diseases (ALS, Alzheimer's and FTD) occurs the appearance of pathogenic inclusions of these granules. These processes have already been associated with alterations in the FUS and TARDBP proteins, when they leave the cell nucleus and are included in the stress granules present in the cytoplasm. It has been suggested that, once in contact with these aggregates, the proteins can undergo modification processes, as phosphorylation, ubiquitination and partial proteolysis, by other proteins already present in the SGs and inhibit their return to the nucleus. The abnormal accumulation of these cytoplasmic aggregates, catalyzed by the stress granules, appear to be related to neurodegenerative process associated with neuronal diseases [4–7]. However, there is still a clear deficiency in the comprehension of the impact that the stress granules cause in the transcriptome of the cells. It is fundamental that we explore the molecular dynamics involved in the RNA regulatory process involved in these granules, which will also lead to better an understanding of the biological processes involved in neurodegenerative diseases associated with SG formation and aberrant behavior. Understanding the molecular regulatory involved in these granules is the first step in the development of new alternatives for treatment and prevention of diseases which are currently still untreatable or incurable, such as ALS and Alzheimer's.

Among the diverse proteins found in stress granules, the RNA binding protein Caprin-1 is part of the activation complex of the PKR kinase and, consequently, phosphorylation of the eIF2α, an essential protein in the assembly of the aggregate. In addition, it was shown that the ectopic expression of Caprin-1 is sufficient to induce

the formation of cytoplasmic stress granules. It was also demonstrated that Caprin-1 binding to the G3BP1 complex is responsible for promoting the assembly and condensation of the granules, with G3BP1 being capable of associating to the 40S subunit of the ribosome after the formation of the granules [8–12]. Caprin-1 is important protein for embryonic development in mice, capable of inducing the formation of stress granules upon it's overexpression [13] and is also related to cellular proliferation processes and some forms of cancer [14].

Initially it has been proposed that stress granules act as global repressor of cellular translation, although more recent works suggest that mRNA translation levels are more associated with the mRNP context it is inserted. However, other works revealed that RNAs which are preferentially transcribed during stress situations present internal ribosomal entry sites (IRES), with estimations suggesting that 10 to 15% of transcripts, from multiple cell lines, are capable of presenting these sites. With that in consideration, there is a possibility the existence of stress granules is associated with the creation of an optimal cellular condition for the expression of stress-response mRNAs [15–17]. Studies performed in yeast demonstrate that mutations which affect the assembly of stress granules lead to transcriptomic alterations that impair the stress response to glucose deprivation, generating a less effective response and also suggesting that the absence of stress granules can be lethal to cells exposed to stress conditions [18]. A second study, also in yeast, showed that 10 to 15% of transcriptional processes are regulated by stress conditions, with the storage capacity of stress granules and subsequent direction of transcripts to either translation or degradation is essential to generate phenotypical variability in a genetically identical population, increasing the ability to survive under stress conditions [19]. Lastly, a third study performed in human HEK293T cells, the same model used in this work, showed that proteins associated with stress granules (FUS, EWSR1 and TAF-15) can modulate the expression of target transcripts in stress conditions [20].

Among the diverse strategies used by the cells in order to alter their transcriptome and increase the chances of survival in stress conditions, the modulation of the splicing process of pre-mRNAs poses a role of significant importance. An example is the assembly inhibition of the U4-U5-U6 trisnRNP complex due to inactivation of the HSLF splicing factor, which occurs in heat stress [21,22]. In general, stress conditions are capable of modulate splicing event in a context-dependent manner, since signaling pathways and stress response modulate the activity of RNA-

binding proteins and promoting alterations in the pre-mRNA processing steps [23,24]. One such example is the alteration on the activity of the TIA-1 protein, which is present in stress granules. Not only this protein is bound to transcripts in the cytoplasmic granules, but it also has the function of regulating the splicing of FAS and FGFR2 genes. With alterations in TIA-1 being capable of leading to the production of the apoptotic isoform of the FAS gene, while TIA-1 depletion in HeLa cells leads to an increase in cell proliferation [25]. Another example of the action of stress conditions in splicing regulation is the hyperphosphorylation of the hnRNPA1 RNA-binding protein. This process induces the exit of hnRNPA1 from the nucleus to the cytoplasm, where it irreversibly aggregates in stress granules, leading to splicing alterations in multiple transcripts [17,26]. Another alteration found in the splicing process as part of stress response is the production of non-functional isoforms as a means of protection. As an example, there is the occurrence of multiple processes of “exon skipping” in the MDM2 gene in response to exposure to genotoxic agents. This process leads the transcript to the nonsense-mediated decay machinery and favors the p53 response [27].

In general, the alternative splicing process is fundamental for the response to environmental pressures, conferring the proteins the ability to modulate their domains, modifying their functions without radically altering their structure of global functions [28], or leading transcripts to degradation processes. Due to the vast amount of isoforms produced by eukaryotic genes, the application of computational methodologies is imperative in the search and characterization of these events in global scale [29]. Studies performed in plants have already shown the importance of a global analysis of the production of alternative isoform as a mechanism of stress response and the presence of cytoplasmic granules [30,31]. Additionally, studies from human cancer samples have also shown the existence of thousands of splicing alterations detected by high-throughput RNA sequencing, evidencing the capabilities of this technique in identifying such events and provide a better understanding the cellular processes associated to a phenotype [32].

Our study has an exploratory characteristic, with the objective in acting in two approaches related to the role of the Caprin-1 in RNA-regulatory processes under induction of stress granules. Our experimental model is cultured human HEK293T cells, which is one of the mostly widespread cellular models both in academic and industrial environments. Considering those observations, we wanted to understanding

of how these cells respond to stress conditions, especially in the case of transcriptomic alterations induced by the appearance of cytoplasmic stress granules.

Objectives

1. Characterize the RNAs associate with Caprin-1 in HEK293T cell cultures, under induction of stress granules, and identify the binding sites preferentially used in the target RNA to bind this RBP.

2. Identify functional groups (Gene ontology, metabolic pathways, protein domains and others) enriched in RNA targets associated with Caprin-1

3. Identify biological characteristics associated with Caprin-1 binding regions, comparing these results with other regions of the transcriptome

4. Identification of the set of splicing alterations in human HEK293T cell cultures under induction of stress granules via ectopic expression of Caprin-1

5. Analyze and characterize the splicing alterations found in functional categories

6. Evaluate the existence of new transcripts and/or alterations associated with stress conditions.

Material and Methods

Preparation of RNA-seq and eCLIP-seq libraries

The RNA libraries used in this project were prepared by the student Natacha Azussa Migita during her MSc project (projeto FAPESP 2014/20174-6) [33]. eCLIP-seq libraries were generated in accordance with the protocol described by Van Nostrand et al, 2016 [34] (Supplementary Protocol 1: eCLIP-seq Experimental Procedures).

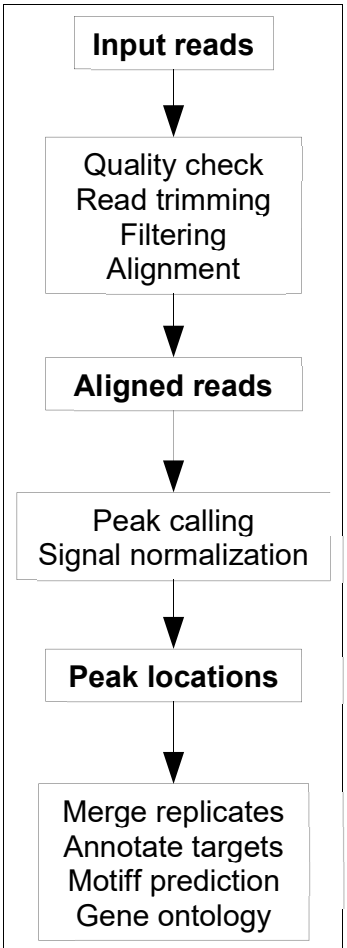

Figure 1: Summary of the computational pipeline

Computational pipeline for analysis of Caprin-1 eCLIP-seq and RNA-seq

The computational analyzes were performed in accordance with the protocol described by Van Nostrand et al., 2016 [34] (Supplementary Protocol 2: eCLIP-seq Processing Pipeline), that are summarized in the Figure 1. In a more detailed description, the steps performed were:

1. Assessment of the quality of the reads obtained from the sequencing platform Illumina HiSeq 3500 using FastQC [35] package.

2. Adapter removal, trimming and removal of low-quality and duplicated reads performed with the cutadapt [36] and trimmomatic [37] softwares.

3. Alignment of remaining reads to human GRCh37.p13 genome using STAR v. 2.4.0 [38] aligner.

4. Characterization of binding peaks, signal normalization, target identification and motif prediction with the packages: CLIPPER [39] and GraphProt [40].
5. Complementary analysis of data using StringDB v10.5 [41] (for enrichment of biological groups), SPIA v2.30 [42] (for pathway impact analysis) and microRNA

targets sites from microRNA.org 2010 [43]. These complementary analysis were done using Python 2.7 [44] e R 3.2.3 [45]

6. RNA-seq reads were aligned to the human GRCh37.p13 genome using STAR v.2.4.0, with the GENCODE [46] V19 comprehensive annotation as reference.

7. Differential gene expression was calculated using edgeR v3.12.1 [47], alternative splicing events at exon-level were identified using rMATS v3.2.5 [48], isoform quantification and characterization was done using stringitie v1.3.1 [49].

Unless specifically stated, all quality assessment, filtering, treatment, alignment and post-processing of the data was performed in accordance with ENCODE's guidelines and best practices for RNA-seq of human samples [50].

Results and Discussion

Quality assessment and alignment of the reads

The reads obtained from the 6 experimental libraries were processed for adapter removal and general quality controls (including removal of duplicated and repetitive reads, in the case of eCLIP-seq) was performed using FastQC, cutadapt and trimmomatic. The amount of aligned reads obtained from eCLIP-seq for Caprin-1 were approximately 9.8 times higher than those obtained for eCLIP-seq for GFP and approximately 1.3 times higher than the size-matched control from HEK293T cells (Table 1). In general, the reads presented a high sequencing quality after processing for removal of low-quality entries (Figure 2A), with the majority of reads achieving a score between 38 and 41 in Phred scale (Figure 2B). The resulting bam files, after alignment with the GRCh37.p13 human genome, were

Individual libraries	
Library	Mapped Reads
CAPRIN1.1	10856410
CAPRIN1.2	12893987
GFP.1	1695382
GFP.2	713459
HEK293T.1	7693677
HEK293T.2	9593415
Caprin1 eCLIP-seq vs others	
Comparison	Fold
CAP1 vs GFP	9,86
CAP1 vs HEK	1,37

Table 1: Number of mapped reads for eCLIP-seq libraries (upper) and fold ratios (lower)

concatenated by condition and then converted to bedGraph format and uploaded to a UCSC Genome Browser session [51], for visual exploration of the data.

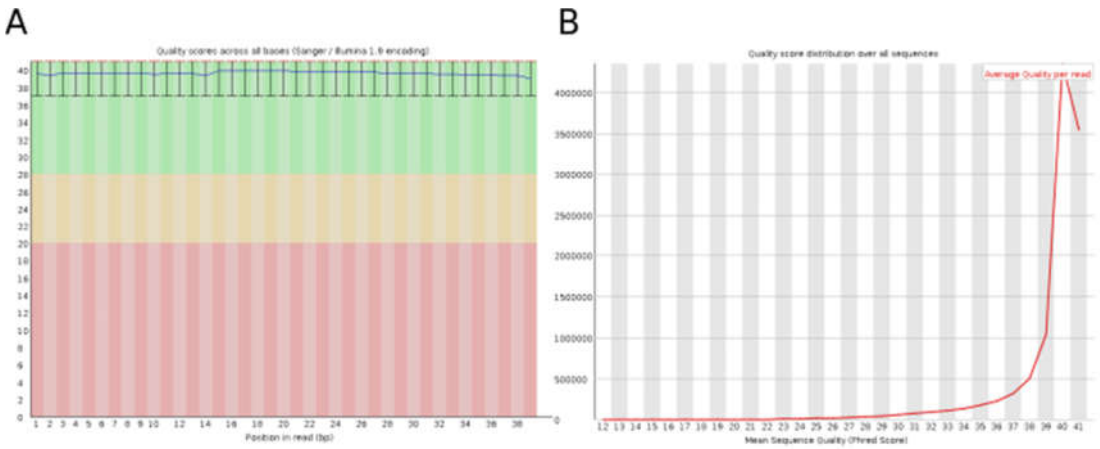

Figure 2: eCLIP-seq sequenced libraries show high quality scores per-base. (A) Histogram showing average per-base quality scores after removal of low-quality sequences in Caprin-1 eCLIP-seq library. The X-axis show the position of the base in each read and the Y-axis show the quality of the sequencing in Phred scale. The error bars show the standard deviation found for scores at each position. **(B)** Graph showing the distribution of average quality scores by read found in the Caprin-1 eCLIP-seq library. The X-axis show the quality score in Phred scale and the Y-axis show the number of reads found with that score.

Identification and filtering of RNA binding sites by peak calling approach.

After treatment of aligned reads, we used the CLIPper package [39] to identify the binding site peaks in each eCLIP-seq replicate separately, which represent potential binding sites for the Caprin-1 protein. In summary, the peaks were quantified those against background, normalizing peak signal strength when compared to GFP eCLIP-seq and size-matched input from HEK293T, in order to obtain both pValues and foldEnrichment (fE) values, as described by the supplementary protocol 2 in Van Nostrand et al, 2016 [34]. Both of these values (pValue and FE) were used to filter significant peaks, with cutoff values of pValue ≤ 10^{-5} and FE ≥ 8 used for significance threshold (Figure 3A). For the first replicate (R1) we identified 2836 significant peaks (~1.97% of total peaks) and for the second replicate (R2) we found 2426 significant peaks (~1.59% of total peaks) (Figure 3B). When analyzing the overlap between the peaks found in the two samples, from the total of significant peaks found in both samples (5262) we observed that 1470 were found in both replicates, 1366 only in R1 and 956 only in R2 (Figure 3C), these peaks were distributed in 1724 different genes. Although the raw number of significant peaks and their intersection might appear small,

it is consistent with previously published studies using this technique [34,52]. Along with the final alignment results (in bedGraph format), the peak positions were uploaded to a UCSC Genome Browser session in BED format, allowing for a visual representation and comparison of identified significant peaks with the genome coverage observed from the alignment (Figure 4).

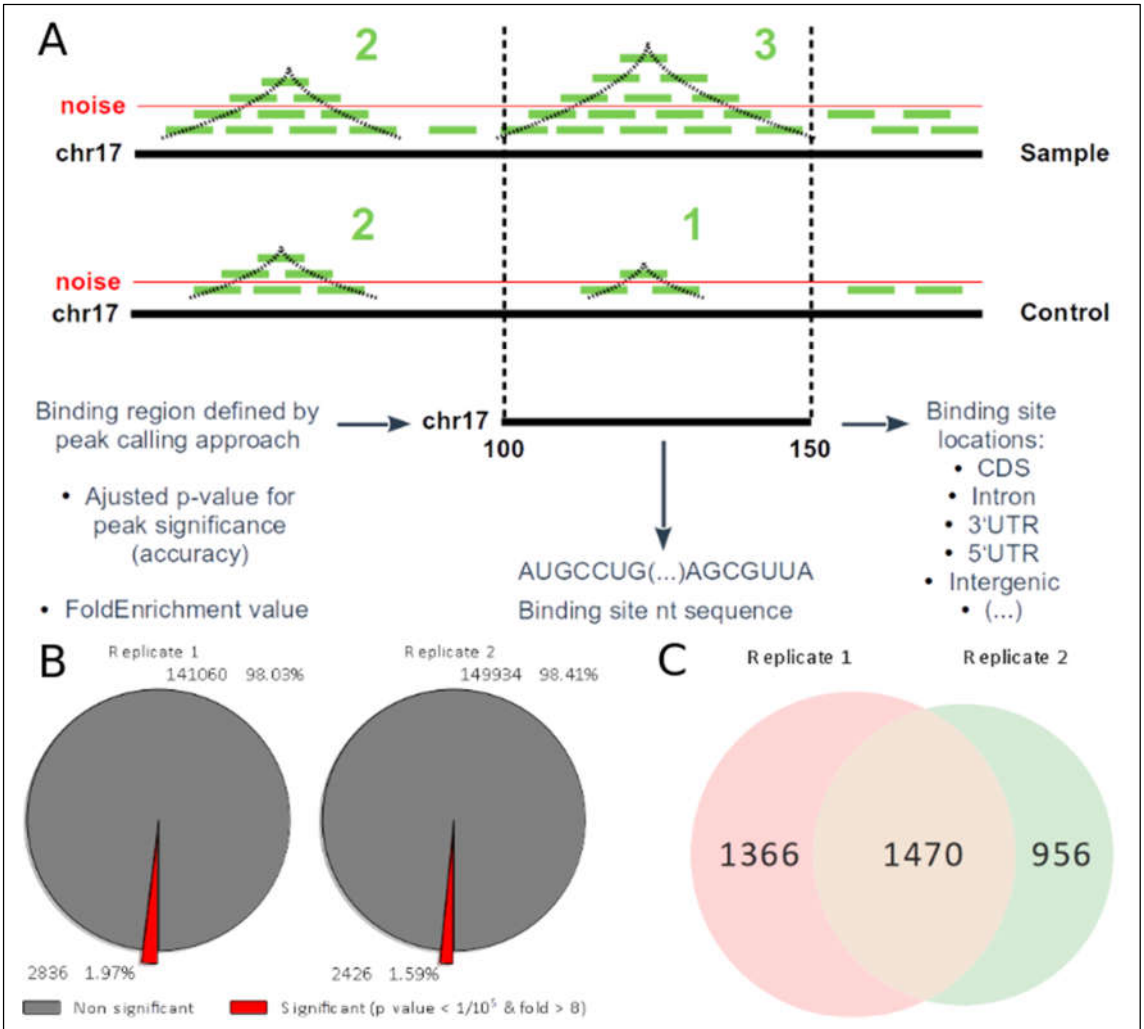

Figure 3: Prediction of Caprin-1 binding sites through significant peaks from eCLIP-seq. (A) Simplified schematic representation for the process of peak calling from reads aligned to a reference genome (chr17 chosen randomly for representation, not to be considered). In the image, the black bar represents a generic DNA fragmented, the red bar represents the background noise threshold for foldEnrichment, the dashed lines represent the theoretical extremities of the peaks and the green represent the short RNA reads. The numbers in green show the theoretical height of the peak, as an example for the read density in the region. **(B)** Pie chart representing the quantity of significant peaks (red slice) found in each replicate. For replicate 1 we found a total of 143896 peaks, with 2836 (~1.97% of the total) deemed significant. For replicate 2, we found 152360 peaks with 2426 (~1.59% of the total) labeled as significant. **(C)** Venn diagram showing the overlap between significant peaks found in both replicates, with 1470 peaks found in both replicates (minimum of 50% of bases), while 1366 peaks were identified only um replicate 1 and 956 only in replicate 2.

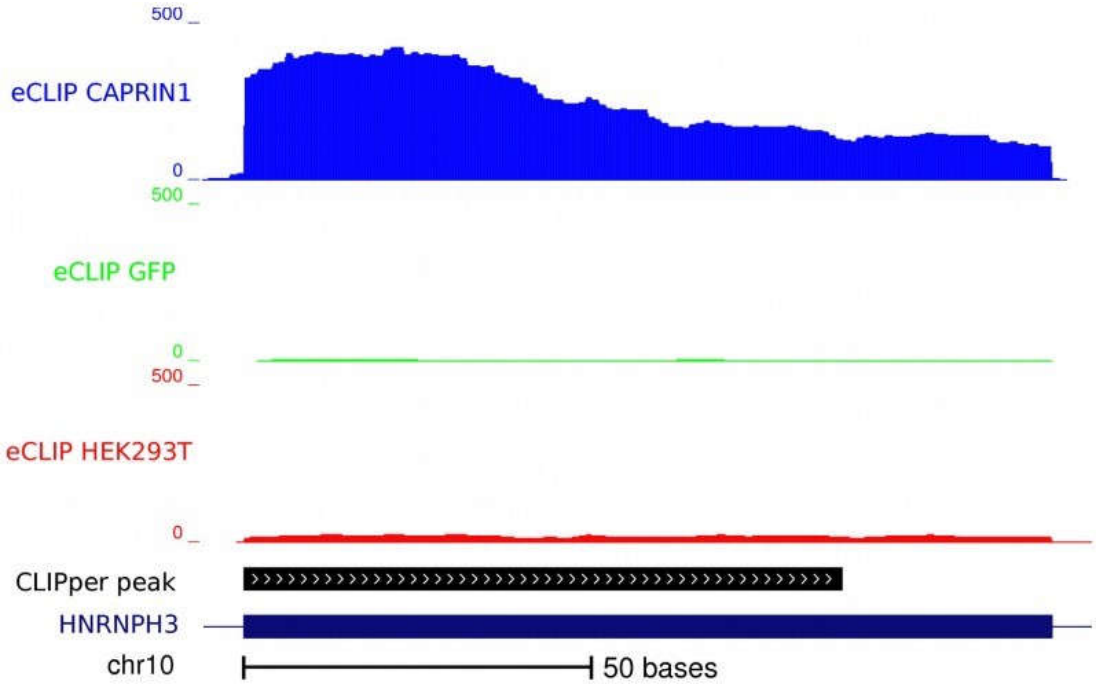

Figure 4: UCSC Genome Browser graphical visualization of genome coverage from alignment and significant peaks identified by CLIPper. The figure shows a region of the HNRNPH3 gene, in which we detected a significant peak in our analyses. The figure shows the bedGraph track containing the genome coverage for the experimental conditions in our experiment: Caprin-1 eCLIP (blue), GFP eCLIP (green), and size-matched input from HEK293T (red). In black is the BED track containing the genomic coordinates for the significant peaks.

Analysis and annotation of significant peaks

For the following analysis, the significant peaks found in each replicate were concatenated in a single BED file and processed with the script “clip_analysis.py”, which is part of the CLIPper package. Initially, the peaks that have any degree of overlap were aggregated in form of “clusters” (being considered a single region), and approximately 16.5% of peaks were aggregated in this manner. Both peaks and clusters were then characterized as their position in the transcript, distribution between mRNA and pre-mRNA, read identity from the central point and average length of the clusters/peaks. The results obtained revealed that Caprin-1 binds mostly to 50 nucleotide regions in exons of mRNAs, with a preference for coding regions (CDS), with 44.15% of clusters and peaks in this region, and 3’UTR with 50.8%. Other identified binding regions were: 5’UTR (2.2%), distal intron (more than 500nt in distance from exon, 2.02%) and proximal intron (less than 500nt of distance from exon, 0.78%) (Figure 5).

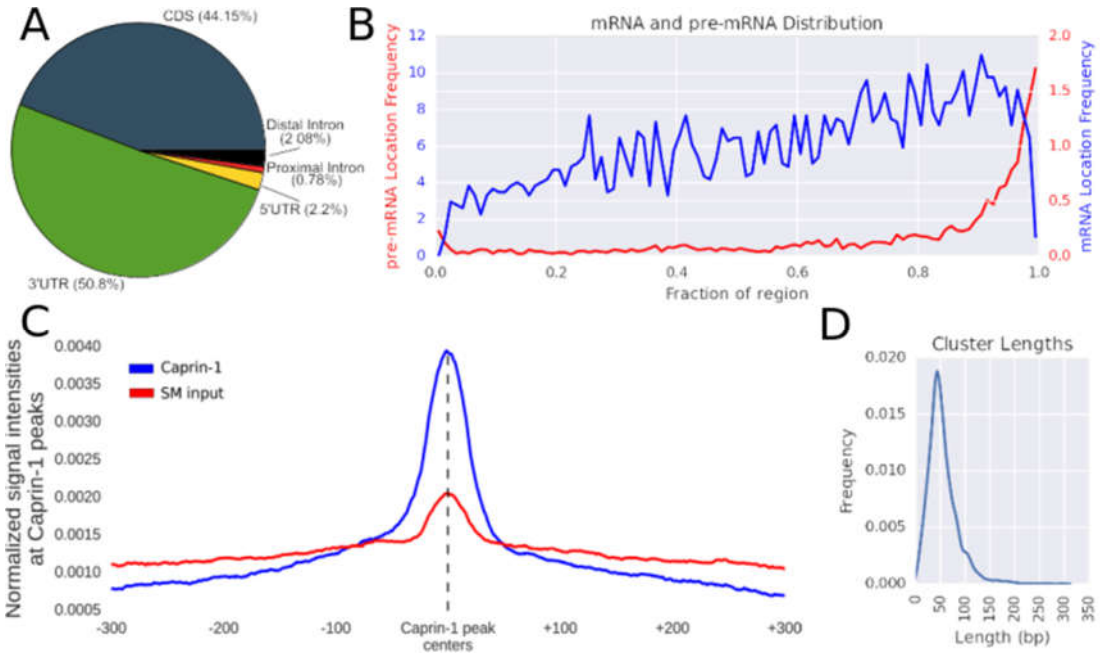

Figure 5: Caprin1 binds primarily to exons in the 3'UTR and CDS of mRNAs. (A) Pie chart showing the distribution of peaks within transcript regions. The two most abundant regions found in the analysis were 3'UTR, with 50.8% of peaks, and CD, with 44.15% of peaks. **(B)** Graph showing the relation between the fraction of regions found bound to Caprin-1 (X-axis) and the frequency of their localization in pre-mRNA (in red) and in mRNA (in blue). **(C)** Graph showing the normalized signal for read density obtained in identified peaks. The X-axis represents the distance (in nucleotides) from the central point of the peaks and the Y-axis shows the normalized signal intensity. The blue line represents the signal obtained from Caprin-1 eCLIP-seq and the red line from size-matched HEK293T control.

In order to evaluate the identified regions and verify/filter possible technical artifacts in a broader scope, we analyzed the foldEnrichment (fE) of the whole regions (CDS, 3'UTR, 5'UTR an intron) of the transcripts in which the clusters were identified and compare the results from both replicates. We found that the CDS and 3'UTR regions both show high correlations of fE values found in R1 and R2 (Pearson's R^2 of 0.83 and 0.80, respectively), as well as the majority of sample fractions showing positive fE values. The 5'UTR region, although it has the majority of its sample with positive fE, showed smaller correlation scores between replicates (Pearson's R^2 0.51), indicating a degree of discordance among the results obtained for R1 and R2. The intronic regions, however, showed high correlation between R1 and R2 (Pearson's R^2 0.86), but few regions with positive fE (Figure 6). These results suggest that the CDS and 3'UTRare the real binding regions of the protein, while the presence of peaks in the intronic and 5'UTR regions could be due to unspecific interactions, non-coding RNAs or other technical artifacts.

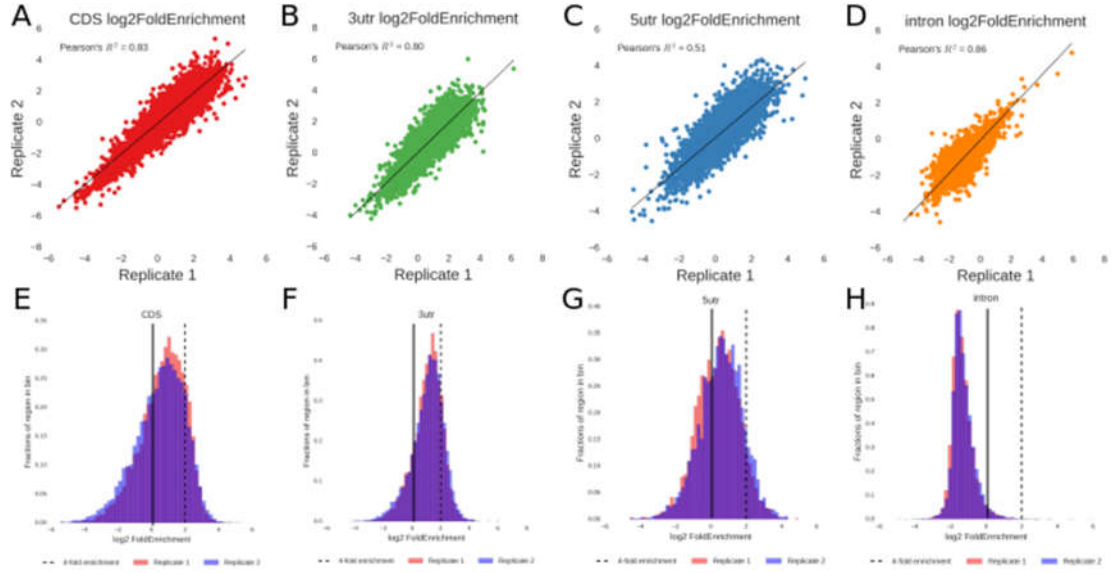

Figure 6: CDS and 3'UTR regions bound to Caprin-1 show high correlation between replicates and positive enrichment at region-level. (A-D) Scatter plots showing the foldEnrichment values for transcript regions found in Caprin-1 eCLIP-seq (CDS [A, red], 3UTR [B, green], 5'UTR [C, blue] and intron [D, orange]). The fE values for replicate 1 (R1) are on the X-axis and the fE values for replicate 2 (R2) are on the Y-axis. The solid line represents the trendline associated with the graph values. **(E-H)** Histograms showing the distribution of fE values in log2 scale (X-axis) for both replicates (R1 in red and R2 in blue) and the frequency which they were found (Y-axis). The solid lines represent the 0 log2 value, separating positive and negative values, while the dashed line represents the log2 = 2 threshold (fE >= 4).

In order to perform a functional analysis of Caprin-1 binding the targets, we used the STRINGdb dabatase to identify GO classes, metabolic pathways, protein domains and phenotypical associations which could be enriched in our dataset (Figure 7). We observed statistically significant results in all categories, however the most enriched class was poly(A) RNA-binding proteins (Gene Ontology – Molecular Function, GO:0044822; -log10(BH) > 50). It is also worth noting the presence of enrichment in classes related to cell cycle, viral carcinogenesis, miRNAs in cancer and retinoblastoma. Caprin-1 is a protein known to be related to cell cycle control in both humans and in *D. melanogaster* [53–56], additionally the association between Caprin-1 and the proliferation of certain types of tumors has already been suggested by previous works [8,14,57].

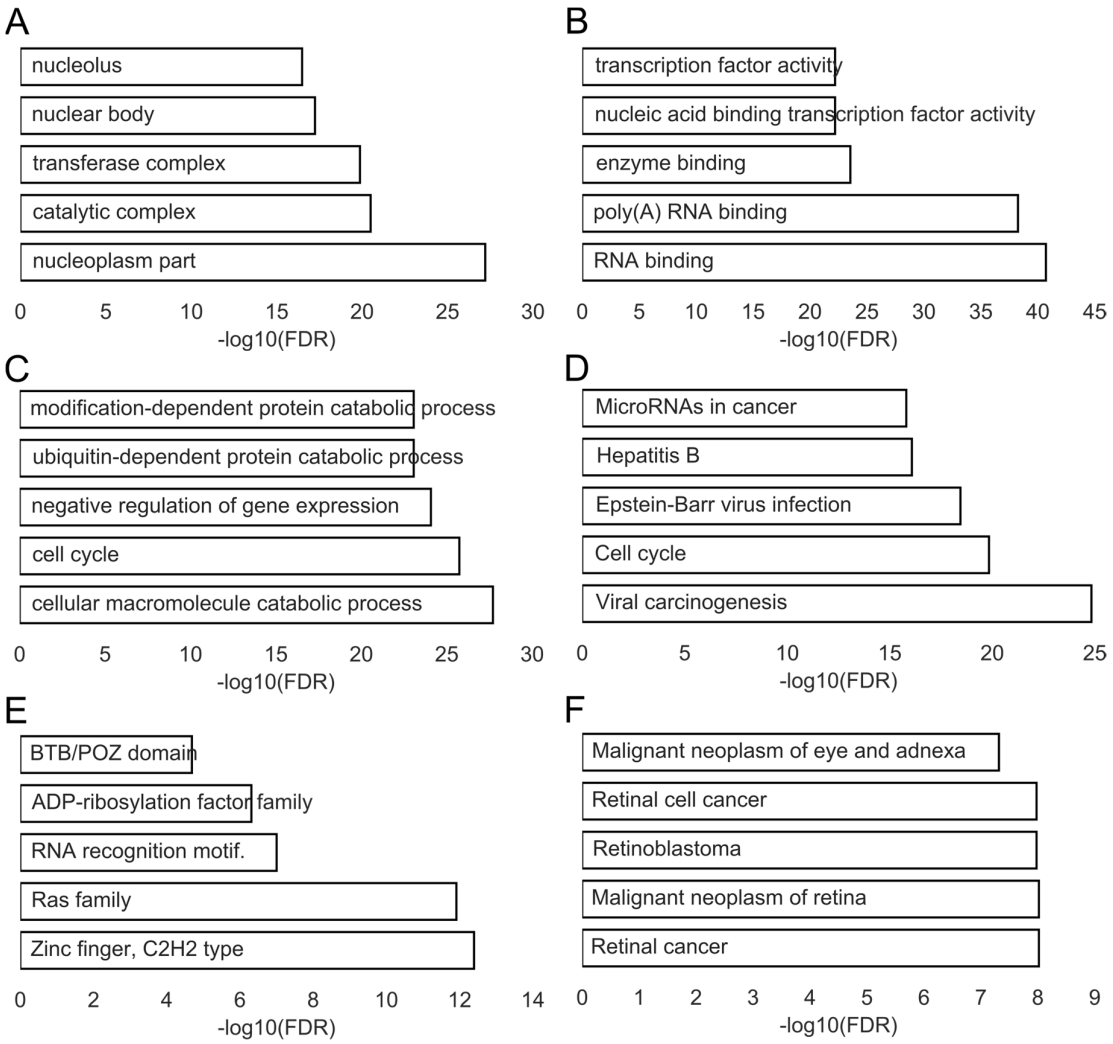

Figure 7: Enrichment analysis for biological categories of targets bound to Caprin-1. The horizontal bar charts correspond to the biologically functional categories used for enrichment analysis. The X-axis represents the p-value corrected by the Benjamini-Hochberg method (BH), in -log10 scale, and the Y-axis represent the 5 most enriched classes found in each category: **(A)** Celular component (GO); **(B)** Molecular function (GO)**; (C)** Biological process (GO); **(D)** Metabolic pathway (KEGG); **(E)** Protein domains (PFAM) and **(F)** Phenotypes (OMIM, ICD).

Considering that the molecular mechanism involving the biological role of Caprin-1 is still poorly understood, our results can contribute significantly to the understanding of how Caprin-1 relates to cell cycle control and its biological importance in the context of severe human diseases, such as cancer.

Analysis of alterations in metabolic pathways from RNA-seq data and comparison with enriched classes in eCLIP-seq

To better comprehend the impact of Caprin-1 in the cellular context involving stress granules, we focused on analyzing the eCLIP-seq peaks/clusters which were found in the 3'UTR and CDS regions, since they account for more than 95% of total peaks (Figura 8A). When analyzing the overlap between genes bound at the CDS and genes bound at the 3UTR, we observed a small degree of overlap between the groups (genes targeted at both their CDS and 3UTR regions), however the majority of genes are either bound by either their CDS or 3'UTR regions (Figure 8B). This suggests that there might be separate post-transcriptional regulatory mechanisms associated with Caprin-1 in the stress granules, since it has already been shown that binding sites in the CDS are capable of repressing the translation process [58], while binding sites in the 3'UTR region might be more associated with stability of the target mRNA [59]. Subsequently, we select the genes found in these regions and performed enrichment analysis for KEGG pathways within these two classes (3'UTR-bound and CDS-bound). Overall, were found a total of 73 enriched terms in CDS targets and 93 terms enriched for 3'UTR targets, with 55 terms being shared between the groups. This indicates that, although Caprin-1 binds to different sets of genes, there is a convergence of enriched KEGG pathways in eCLIP-seq targets. When comparing these enriched terms with affected metabolic pathways, obtained from differential gene expression data derived from RNA-seq libraries [33], we found 25 altered metabolic pathways, with 19 of those also present in the enriched groups found for eCLIP-seq targets, 13 being in the CDS/3'UTR shared targets and 6 for targets enriched only in the 3'UTR group (Figure 8C). The majority of altered pathways found are related to cell cycle control processes, immunological response and cancer, with 7 pathways being activated and 12 being repressed after ectopic expression of Caprin-1 and stress granule induction (Figure 8D). The pathway which showed the biggest alteration was small-cell lung cancer, enriched only for targets in the 3'UTR, while the most enriched pathways was microRNAs in cancer, which was enriched in targets in both CDS and 3'UTR.

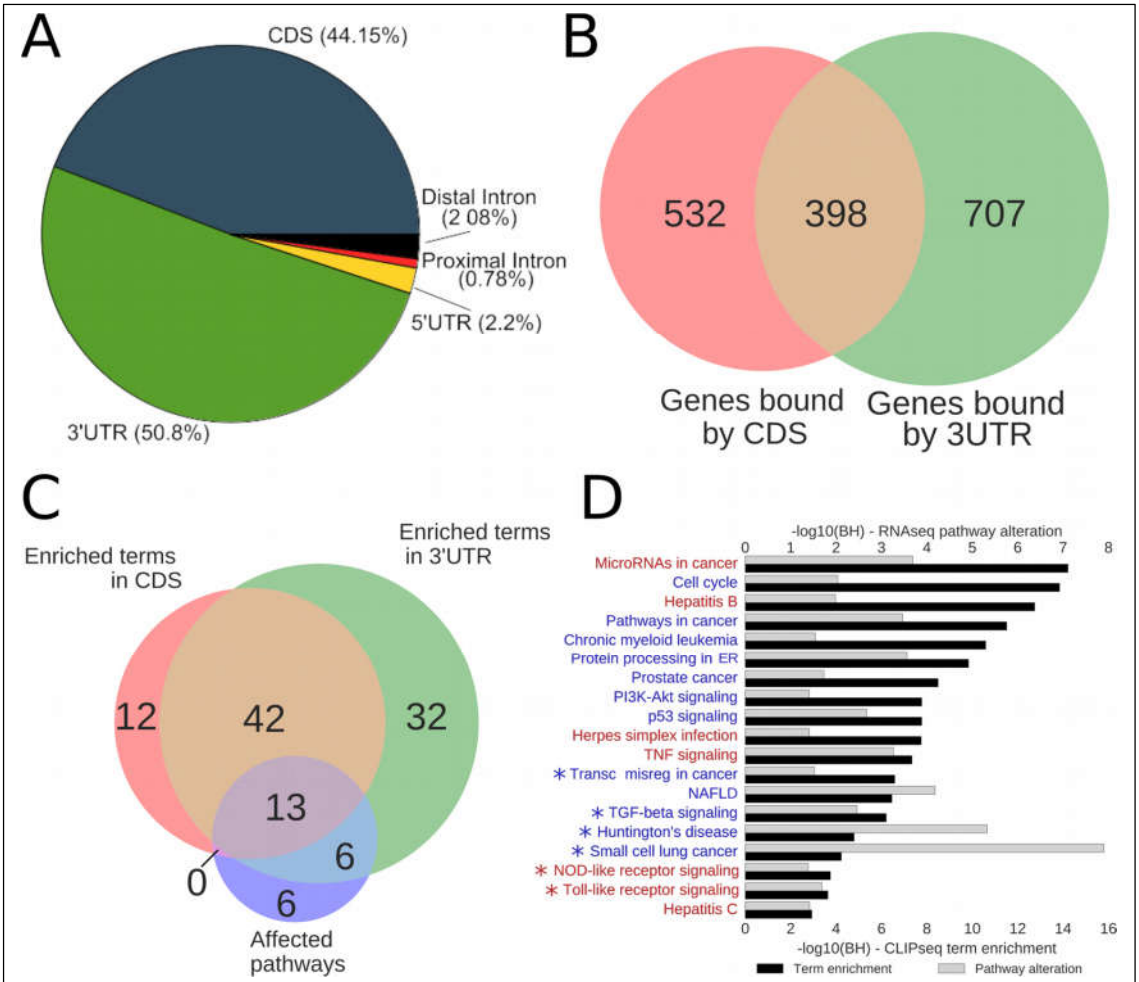

Figure 8: Caprin-1 binds to exons in the CDS and 3'UTR of targets associated with altered metabolic pathways associated with cell cycle control, immunological response and cancer **(A)** The majority of eCLIP-seq defined clusters (3998) are located in either the 3'UTR (2091, ~50.08%) or at the CDS (1765, ~ 44.15%) of target genes. **(B)** We identified 1105 genes which are bound to Caprin-1 by their 3'UTR and 930 genes that are bound by their CDS region. From the total of 1634 different genes, 398 (~24.30%) are bound by both CDS and 3'UTR regions. **(C)** Venn diagram showing the overlap between enrichment analysis for 3'UTR targets (green), CDS targets (red) and altered pathways (blue). We found a total of 105 KEGG pathways enriched, with 55 being shared by both 3'UTR and CDS targets, 12 enriched only in CDS targets and 38 enriched only in 3'UTR targets. We also identified 25 altered pathways, from those 13 were also enriched for targets in both CDS- and 3'UTR-bound groups and 6 were enriched only for targets bound only by their 3'UTR region. **(D)** Horizontal bar chart showing the 19 affected and enriched pathways found in (C). The black bars represent significance value (in -log10(BH)) for enrichment of these pathways in the 3'UTR group. The gray bars indicate the significance value (in -log10(BH)) for pathway alteration. The activated pathways are labeled in red, while the repressed pathways are labeled in blue, an asterisk is used to highlight the ones found enriched only in 3'UTR-bound targets.

Since one of the suggested roles for Caprin-1 is the stabilization of RNAs in the stress granules, which is more commonly attributed to binding in the 3'UTR regions, we selected the 6 affected pathways which are also enriched in 3'UTR targets: Small cell lung cancer, Huntington's disease, TGF-β signaling, transcriptional misregulation in cancer, NOD-like receptor signaling and Toll-like receptor signaling. We then filtered

the genes found in eCLIP-seq which were present in any of these pathways and filtered them by their expression levels in the RIP-seq (TPM ≥ 1) and differential expression in the RNA-seq (FDR ≤ 0.05 e log2(FoldChange) ≥ 0.5). Using these criteria, we narrowed down our list to 13 candidate genes which might be associated with the biological processes mentioned above (Table 2).

Gene	CLIP-seq clusters	RIP-seq TPM	RNA-seq log2(FC)
SESN1	7	52,63	0,94
GADD45A	4	123,75	0,64
PIK3R3	4	32,06	1,07
EP300	4	31,02	0,66
PHLPP2	3	24,98	0,73
CREBBP	3	35,91	0,74
CASP7	3	24,65	0,61
RBL1	3	54,04	0,50
BBC3	2	13,30	0,60
SMURF1	1	6,75	0,65
NFKBIA	1	14,53	0,73
FADD	1	14,53	0,67
CDKN1A	1	32,97	0,66

Table 2: Candidate genes associated with Caprin-1. The table shows the 13 selected genes from the combined analysis of eCLP-, RNA- and RIP-seq .

Prediction of binding site motif and RNA-recognition sequences

Once the peak positions were accurately placed in regions on the genome, it was possible to use this information to extract the nucleotide sequence of the region. These sequences were then used to predict binding site motifs, which are over-represented sequences of “k” nucleotides (called k-mers) in the sample which might be associated with the RNA-recognition sequences of the protein. Due to the fact that Caprin-1 presents clusters of 50 nucleotides in length both in the CDS and 3’UTR regions, these might contain both sequence-specific information as well as RNA secondary structure information. Therefore, we initially searched for a wider range of k-mers, including repetitions of: 6, 8, 10, 12, 14, 16, 18 and 20 nucleotides. As background for our analysis, we used randomized sequences drawn from the same regions without overlap with Caprin-1 binding sites. The CLIPper package uses as basis the algorithm Hypergeometric Optimization of Motif EnRichment (HOMER, [60]) for motif prediction.

The results from our analysis show that the most enriched motif, present in 69.99% of Caprin-1 peaks, is a sequence of 14 nucleotides rich in GG repeats, with the most common repetitions is the 3-mer CGG and the consensus sequence is "GCGGCGGCGGCGGC" (Figure 9A). Being a relatively long sequence for simple sequence-specific RNA recognition, we considered the possibility of this particular sequence assembly itself in a stable RNA secondary structure. For this, we used the RNAfold tool from the Vienna RNA package, using as basis the consensus sequence indentified by CLIPper. We considered both simple base-pairing as well as G-quadruplexes (due to the GG repetitions present), when using the RNAfold algorithm. Our results for the analysis of RNA secondary structure, based on the minimum free energy model (MFE), show that the formation of G-quadruplex structures is more stable for this particular motif. The secondary structure formed by simple base pairing achieved a MFE of -5 kcal/mol, while the G-quadruplex structure showed a MFE of -18 kcal/mol, which is 3.6 times lower (Figure 9B). As an additional analysis, we also calculated the occurrence of non-overlapping G-quaruplex in both Caprin-1 binding sites and background regions using QGRS Mapper, requiring at least 2 tetrads and a maximum of 1 mismatch for formation of the G-quadruplex. We found that Caprin-1 bound regions have statistically significant higher occurrences of G-quadruplexes (o-value < 0.05, T-test for population mean vs sample) when compared to randomized background regions, the 5'UTR showed the highest frequency of G-quadruplex with an average of 1.2 G-quadruplex per binding site, while remaining regions exhibited smaller frequencies (~0.54 by CDS site and ~0.52 by 3'UTR site). If we considered an average for all sites (independent of region), we obtain an average of 0.57 G-quadruplex per binding site (Figure 9C) It is important to note that although the association between Caprin-1 and G-quadruplex structure is novel, there are other RBPs from the same protein family [61] (RG/RGG RNA-binding proteins) which have already been shown to bind to these types of structures [62–64]. Additionally, there are other publications associating this type of RNA structure to translational regulation [65] and neurodegenerative diseases such as ALS or FTD [66], both characteristics have also been linked to biological processes involving stress granules.

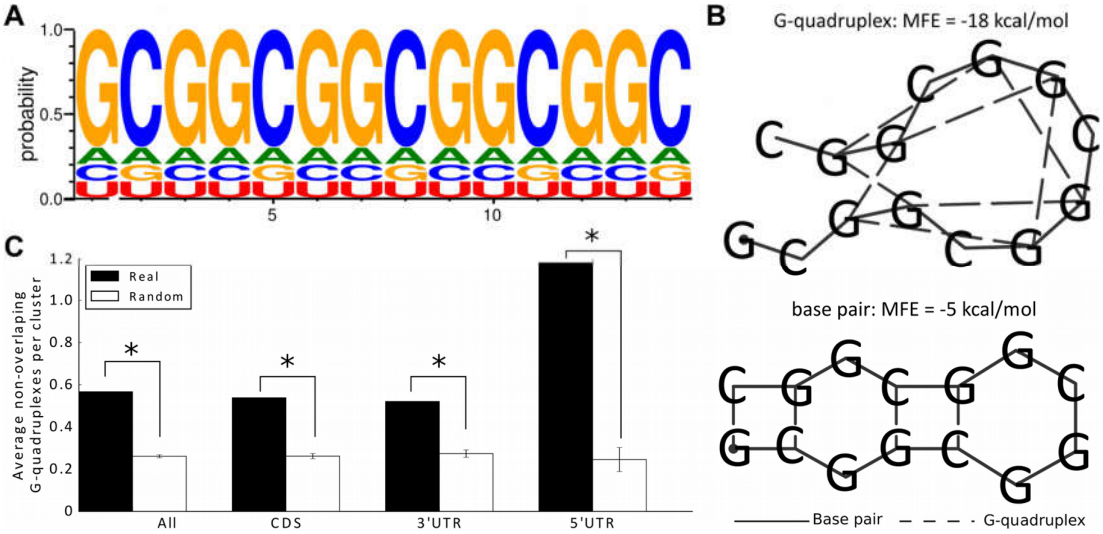

Figure 9: Caprin-1 binds to G-quadruplex regions formed by CGG repeats (A) Predicted binding motif by CLIPper shows that 69.99% of binding sites are enriched with a 14 nucleotide sequence rich in CGG repeats. In the graph, the X-axis represents the position of each nucleotide and the Y-axis is the probability of occurrancce of each base in that position. **(B)** The predicted binding motif assembles in a G-quadruplex secondary RNA structure via two towards created by the Gs in positions [3,6,9,12] and [4,7,10,13] with a MFE of -18 kcal/mol. In contrast, the same motif when assembled in a simpler base-pairing model achieves only a MFE value of -5 kcal/mol. **(C)** Bar chart showing the enrichment analysis for the distribution of G-quadruplex structures shows a significant increase in this type of structures in Caprin-1 binding sites (real, in black) when compared to randomized background regions (random, in white). The X-axis shows each region and the Y-axis the average number of non-overlapping G-quadruplex, asterisks indicate regions with statistically significant differences (All, p-value = 5.16e-16, foldEnrichment [fE] = 2.17), CDS (p-value = 1.61e-13, fE = 2.06), 3'UTR (p-value = 1.14e-11, fE = 1.90) e 5'UTR (p-value = 2.43e-12, fE = 4.81).

In order to analyze with more detail the structural relationships found previously in Carpin-1 eCLIP peaks, we used the GraphProt [40] algorithm to evaluate the binding preferences of the protein in relation to its binding nucleotide sequence and secondary RNA structure. This software uses a different approach than CLIPper, using graphs created from the eCLIP defined RNA-binding regions (CLIP peaks + 150 flanking nucleotides at both directions) and comparing against randomized unbound regions. This graph model is then used to evaluate binding preferences, both to nucleotide sequences and RNA structure, and also to predict the binding potential of the RBP to target transcripts.

For Caprin-1 binding sites, this analysis did not reveal a specific nucleotide K-mer (unlike CLIPper), but the GraphProt model found an enrichment for degenerate sequences rich in GG repetitions (Figure 10A). When evaluating the structural composition of the predicted model, it is possible to observe that Caprin-1 binding regions are associated with structured RNA regions in the form of "stem loops", binding mostly to regions near the boundary between the stem and the hairpin portions of this

secondary RNA structure (Figure 10B). Among the identified nucleotide sequences, the most common K-mer (a 14nt repetition of GGU) also shows increased stability when associated in a G-quadruplex secondary RNA structure, with a minimum free energy (MFE) value of -9.69 kcal/mol (Figure 10C)

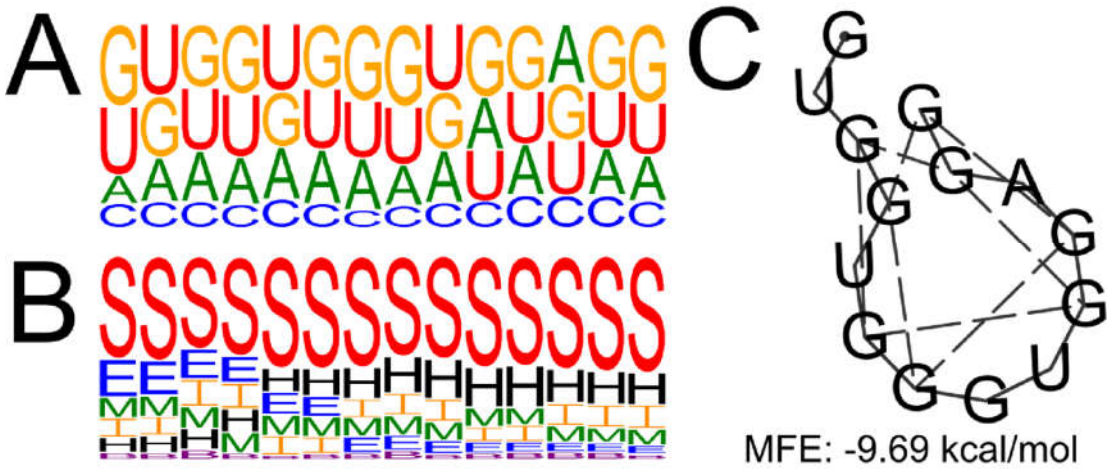

Figure 10: Caprin-1 binds to structured mRNA regions. (A) Nucleotide binding preference predicted by the computational model for Caprin-1 eCLIP-seq. **(B)** RNA structure binding preference predicted by the computational model for Caprin-1 eCLIP-seq. S = stem; H = hairpin loop; E = external region; I = internal loop; M = multiloop; B = bulge loop. **(C)** The most common binding motif predicted assembles in a G-quadruplex secondary RNA structure via 2 tetrads created by Gs in the positions [3,6,10,13] and [4,7,11,14], with an MFE (Minnimum Free Energy) of -9.69 kcal/mol.

Comparing these results with previously published studies [40], which also performed analysis of Caprin-1 binding sites from another dataset, we observed that there is a clear difference in the predicted nucleotide sequence, however the structural model prediction is also composed of stem loop regions associated with hairpins. This comparison increases the credibility of our findings, with Caprin-1 being a protein that recognizes primarily a particular type of RNA secondary structure.

Characterization of microRNA target sites inside Caprin-1 eCLIP-seq peaks

In addition to the main structural motif described above, we also found several other secondary motifs, which are shorter (between 6 and 8 nt) and less frequent (enrichment values vary between 46.65% and 21.02%) in Caprin-1 binding sites. However, these secondary motifs correspond do possible binding sites for microRNAs with similarity scores for miRNA target sequences above 0.7 (Figure 11). Although the regulatory mechanisms of miRNAs in coding sequences is still a underexplored and highly controversial field, there is evidence in the literature that

show the action of these regulatory RNAs in transcript coding regions [67–70], with a particular work suggesting that they are actively responsible for translation inhibition of the target transcript [71] and another publication showing post-transcriptional regulation of the BRAC1 gene by action of microRNA group miR-15/107 in its coding sequence [72].

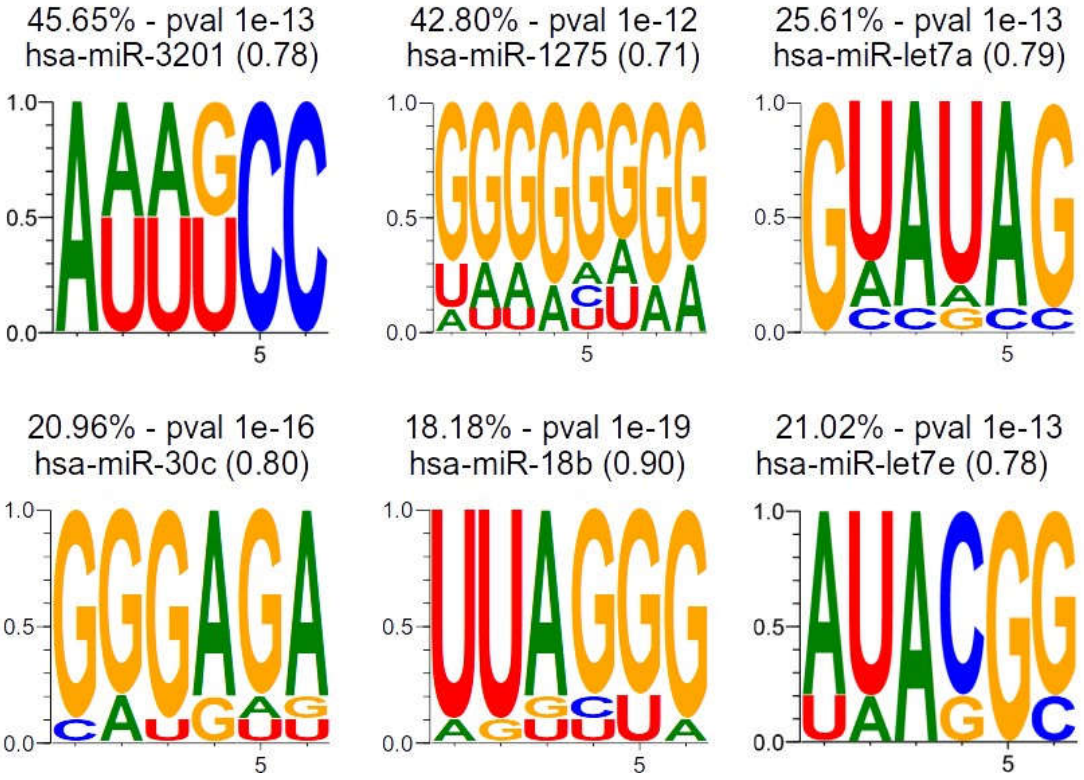

Figure 11: Secondary binding motifs for Caprin-1 show similarities with miRNA target sites. The figure above shows secondary binding site motifs identified in Caprin-1 binding sites. The percentage values indicate the frequency of occurrence of the motif and their associated p-value for enrichment. Below, there is the microRNA identification and the similarity score, in parenthesis, found for the extracted motif. In the logos, the X-axis represents the nucleotide position in the sequence and the Y-axis represents the probability of occurrence of the base in that position.

In order to further explore this possibility, we downloaded the miRNA target site predictions from microrna.org [43], which uses a combination of miRanda [73] and mirSVR [74] algorithms to identify both canonical and non-canonical binding sites for microRNAs in the human transcriptome. We extracted all target sites for the miRNAs identified in Figure 11 (hsa-miR-3201, hsa-miR-1275, hsa-miR-let7a/e, hsa-miR-30c and has-miR-18b) and compared their target genes with both Caprin-1 eCLIP-seq targets and also differentially expressed genes drawn from RNA-seq after ectopic expression of Caprin-1. Overall, our results identified 411 target genes that are, at the same time, targeted by at least 1 of these microRNAs, bound to Caprin-1 by either

CDS or 3'UTR and significantly upregulated in the RNA-seq (FDR ≤ 0.05 e log2(FoldChange) ≥ 0.5). In contrast, we identified only 21 significantly downregulated genes (FDR ≤ 0.05 e log2(FoldChange) ≤ -0.5) which are at the same targeted by Caprin-1 and the microRNAs (Figure 12). We then selected all genes that were differentially expressed and targeted by microRNAs, then used a chi-squared test to compare the results for targets up/downregulated and bound/unbound to Caprin-1 (Table 3). Our results showed a statistically significant relationship between the variables (x^2 = 161.40; pvalue = 9,15E-35), with unbound targets skewed to downregulation (Observed: 533, Expected: 427.44), which is consistent with known microRNA regulatory mechanisms [75–77]. However, our results also show that microRNA target genes that are bound to Caprin-1 are skewed towards upregulation (Observed: 411; Expected: 305.44), which could indicate a protective effect of Caprin-1. We theorize that Caprin-1 binding mRNAs blocks the access of these microRNAs to their target sites which, in turn, leads to and increased abundance of these transcripts, possibly increasing their half-life and/or overall stability.

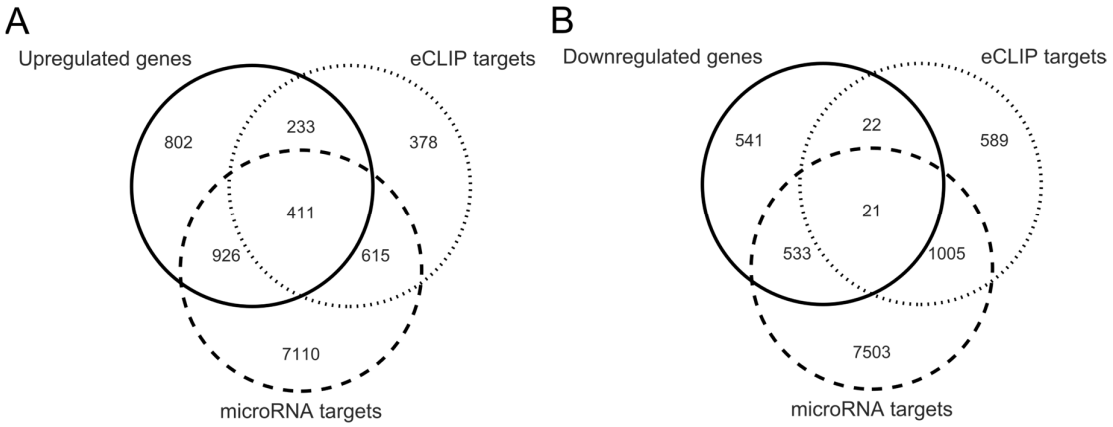

Figure 12: microRNA targets bound to Caprin-1 are upregulated. 3-way Venn diagrams showing the overlaps between genes bound to Caprin-1 (dotted circle), targeted by at least one of the microRNAs (dashed circle) and differentially expressed (solid circle) after Caprin-1 ectopic expression. The comparison shows both upregulated **(A)** and downregulated genes **(B)**.

microRNA targets	Upregulated	Downregulated
eCLIP targets	411 (305.44) [36.48]	21 (126.56) [88.05]
Not eCLIP targets	926 (1031.56) [10.8]	533 (427.44) [26.07]

Table 3: microRNA targets bound to Caprin-1 are more likely to be upregulated. Table displaying the number of observed targets in each class, the expected result (in parenthesis) and the associated x^2 value for each class (in brackets).

Identification of biological characteristics associated with exons bound to Caprin-1

In order to further explore the characteristics of regions bound to Caprin-1, we searched biological characteristics (such as: gc content, length, protein-binding sites, microRNA sites, CpG islands, SNPs, methylation marks and others) associated with the exons bound do Caprin-1 that could explain differences between these regions and the remaining exons encountered in the transcriptome. In order to perform this analysis, we developed an algorithm based on machine learning and big data approaches that is capable of analyzing sets of genomic/transcriptomic regions (in this case, exons bound to Caprin-1) and estimate which features are most important for separating those from the other parts of the genome/transcriptome (see Part 1: BioFeatureFinder: Flexible, unbiased analysis of biological characteristics associated with genomic regions). For this analysis, the final datamatrix was comprised of, in total (input and background), 20971 lines (exons) e 724 columns (features). For all biological features we applied a Kolmogorov-Smirnov test and filtered a total of 88 statistically significant features (q-value ≤ 0.05). Amongst the most significant features for exons bound to Caprin-1, we identified that the length of the exons, the presence of microRNA target sites, binding sites for PUM2 RNA-binding protein and exon conservation among primates are the most important features (Figure 13A). Due to the large number of background regions and the biological variability encountered in exons, we used 500 classification runs to calculate the final average scores for feature importance and their associated standard deviations (STD). The final quality scores for our classifier were: accuracy (86.9%, STD=1.0), positive predictive value (PPV, 90.5%, STD=1.3), negative predictive value (NPV, 83.3%, STD=1.3), sensibility (84.4%,STD=1.5), specificity (89.7%, STD=1.4) and area under curves (AUCs) for ROC (0.938, STD=0.007) and Precision-Recall (0.927, STD=0.011) (Figures 13B-D).

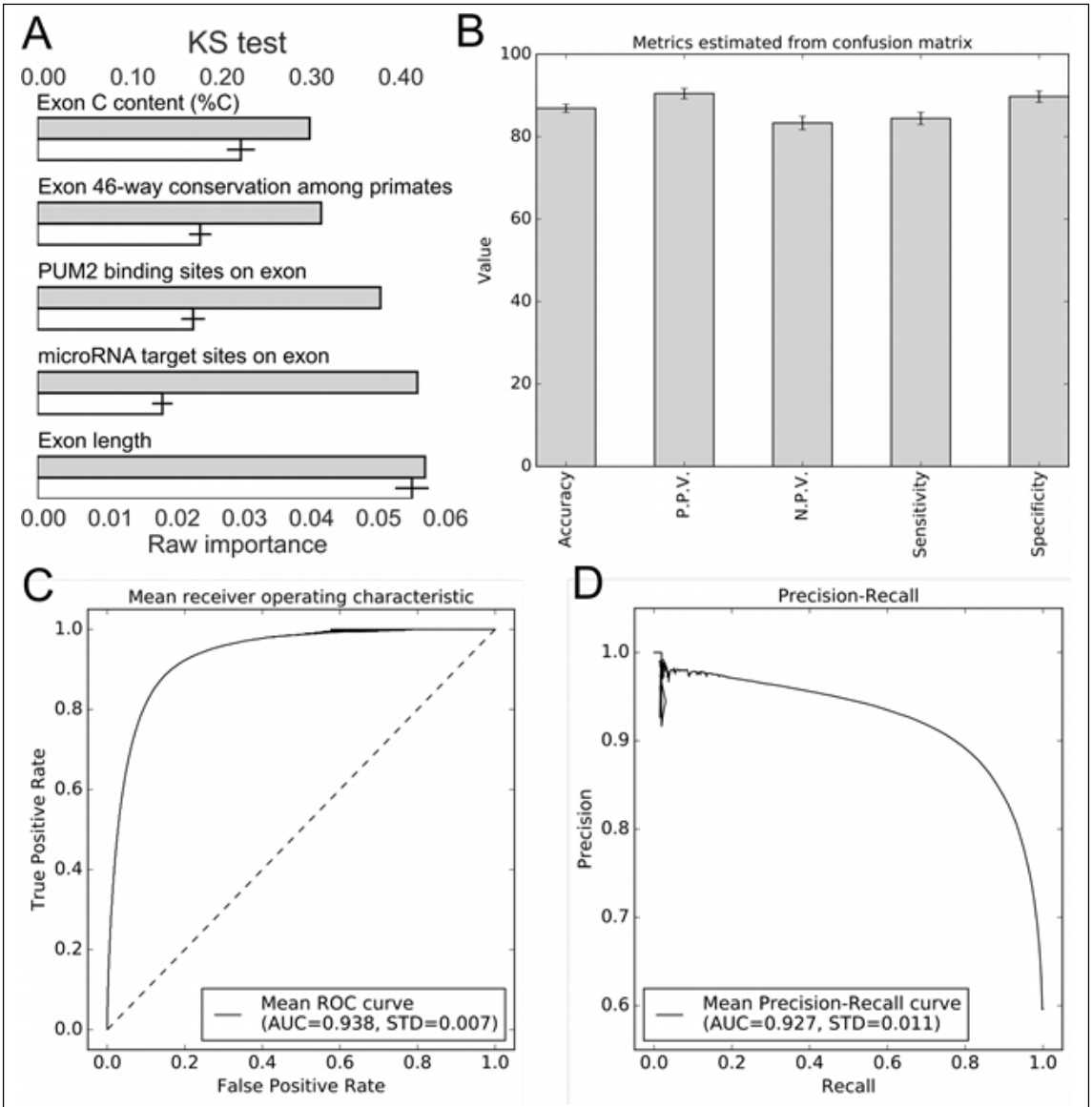

Figure 13: Length, conservation, microRNA and PUM2 binding sites are the most important features of exons bound to Caprin-1. (A) Horizontal barchart showing the 5 highest socring features for classification of Caprin-1 exons. The gray bar representes the values obtained for the Kolmogorov-Smirnov test and the white bars represent the variable importance values for each feature. **(B)** Bar charts for mean classifier performance scores for RBFOX2 eCLIP sites (P.P.V.: Positive predictive value, N.P.V.: Negative predictive value). **(C-D)** Graphs showing the ROC (C) and Precision-Recall (D) curves obtained for the classifier performance, the mean values for área under curve (AUC) and their associated standard deviation (STD) are represented in the legends.

When analyzing the differences found in the curves for the cumulative distribution functions of the exons bound to Caprin-1 and comparing with background unbound exons, it was possible to observe that exons associated with Caprin-1 have a tendency to be longer (Figure 14A), more conserved among primates (Figure 14B) and with a higher number of microRNA target sites (Figure 14C). This is in accordance to our previous findings regarding the occurrence of microRNA target sites in Caprin-

1 binding sites and the enrichment of upregulated microRNA targets that are bound to Caprin-1. Additionally, we also identified that exons bound to Caprin-1 have a high occurrence of binding sites for PUM2, another RNA-binding protein, with approximately 50% of Caprin-1 exons also having at least 1 binding site for this RBP (Figure 14D). This last result is particularly interesting due to the fact that PUM2 has been associated with cell cycle control and translational regulation [78–81], which are biological functions also associated with Caprin-1.

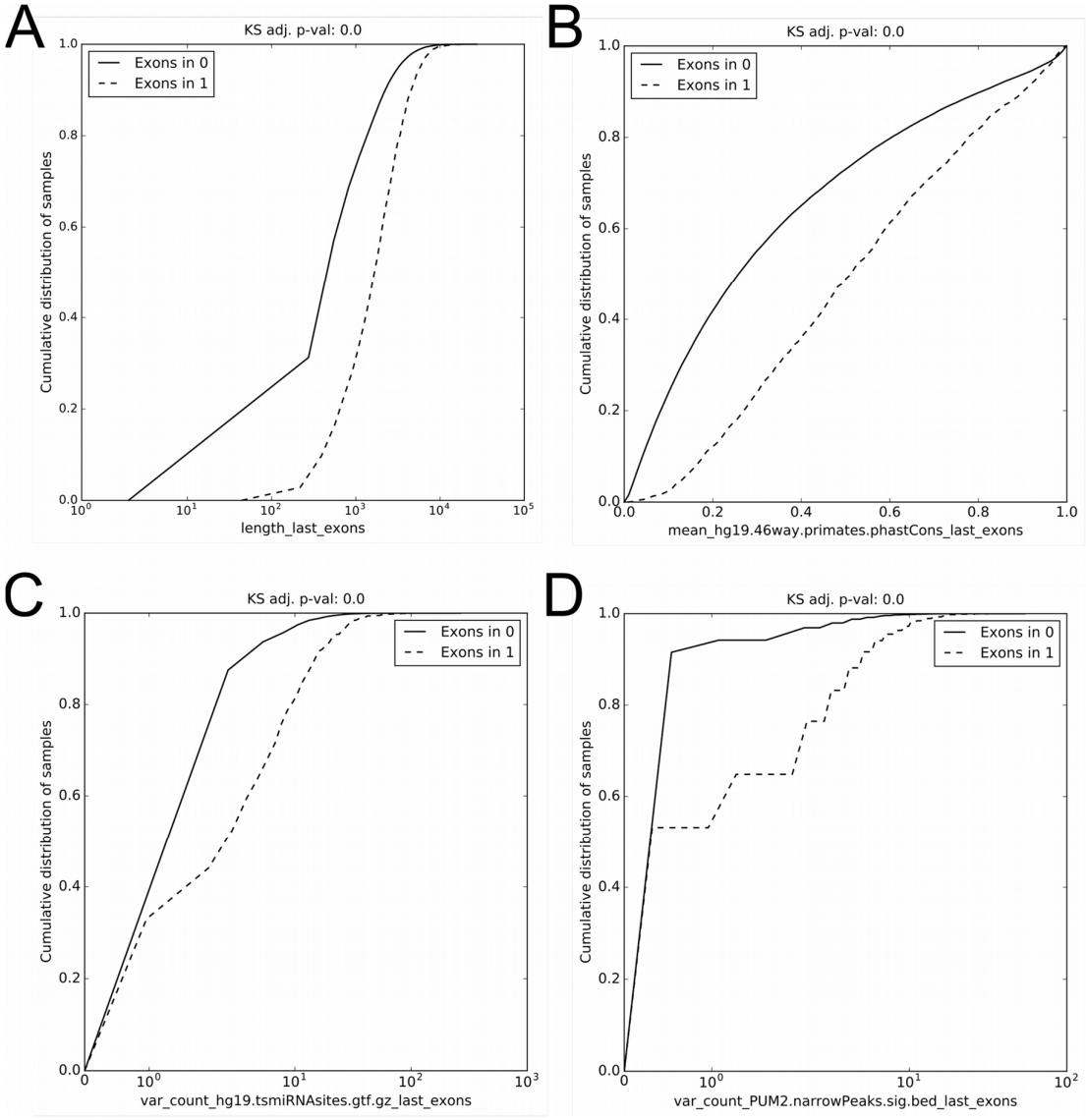

Figure 14: Exons bound to Caprin-1 are longer, highly conserved and possess higher number of microRNA and PUM2 target sites. (A-D) Cumulative distribution function graphs comparing the exons bound to Caprin-1 (group 1, dashed line), with exons not bound to Caprin-1 (group 0, solid line). The cumulative distribution is represented in the Y-axis and the values for each characteristic is represented in the X-axis. The graphs show the comparisons for exon length (A), conservation among primates (B), microRNA target sites (C) and PUM2 binding sites (D).

However, it did not escape our attention the possible correlation between the exon length and the higher occurrence of microRNA target sites (i.e. longer exons can have a tendency to accumulate more microRNA target sites). In order to assess this possibility, we used linear regression with the values encountered for these two characteristics in the exons bound to Caprin-1, obtaining a R^2 value of 0.25 (Figure 15). We also performed pairwise correlation assessments for the 10 most important features, however no combination achieved a R^2 score higher than 0.2.

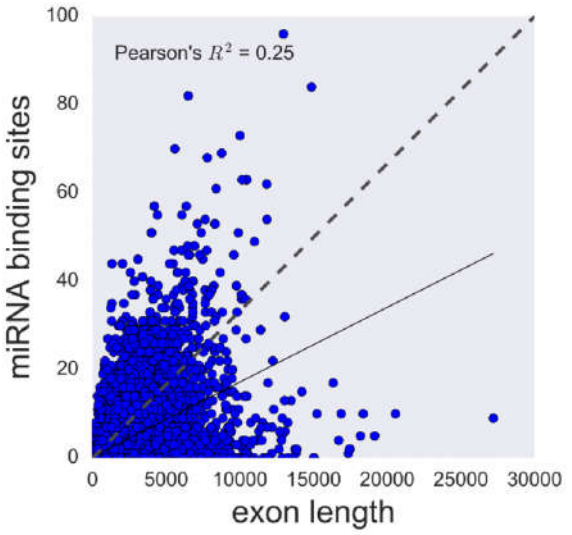

Figure 15: Exons bound to Caprin-1 do not have direct correlation between their length and the number of microRNA target sites. Scatter plot showing the correlation between exon length (X-axis) and the number of microRNA target sites (Y-axis). Each blue dot on the graph represents one exon bound to Caprin-1, with the diagonal (dashed) and trend (solid) lines also represented. The Pearson's R^2 value for correlation is 0.25.

Characterization of splicing events after Caprin-1 ectopic expression

Data was obtained from the RNAseq libraries generated during Natacha Azussa Migita MSc project (FAPESP project 2014/20174-6) [33]. In summary, the experiment is comprised 2 experimental conditions, with a total of 6 samples: 3 controls and 3 ectopic expression of Caprin-1, upon induction of stress granules. In total, these libraries offer a coverage of 1,90E+4 gigabases (190.091.292 million reads of 100 bases). These reads were aligned to the reference genome GRCh38.p5 (hg38) and

the comprehensive annotation, both obtained from *Gencode Release 24* [46], with the STAR v2.5.1 [38] aligner with the options:

–outSAMstrandField intronMotif

--outFilterIntronMotifs RemoveNoncanonical

--twopassMode Basic

The resulting alignments were processed by the rMATS v3.2.5 [48] software for the identification of splicing events at exon-level. Our results indicate the presence of 1058 differential splicing events between the Caprin-1 e GFP conditions (Table 4), with the most common class being the exon-skipping type of splicing event (652 events identified), but we also identified a high incidence of intron retention events (191). Other events were also identified, albeit in smaller number. The same alignments were also used as input for the StringTie [49,82], which allows the reconstruction of potential transcripts existing in the analyzed libraries. For this step, we used the reference guided assembly approach (-G option), using as basis the comprehensive annotation from Gencode Release 24. We then proceeded to compare the resulting transcripts from the StringTie assembly with the reference annotation using the *gffcompare* algorithm. Our results indicate that, from the 199,169 transcripts existing in the reference, we were able to recover 199.103 with our assembly using “exact match”. Additionally, we also identified 12,484 potential novel isoforms and 875 unknown transcripts (Table 5), both which could represent novel isoform/transcripts specific to stress response and/or stress granule induction. Manual assessment via visualization on UCSC’s GenomeBrowser revealed that some of those isoforms showed and elongation in their 5’UTR (Figure 16), which can be related to the formation of structures of the IRES (internal ribosome entry site) which would affect the translational regulation of the

AS Event	Number
Skipped exon	652
Mutually exclusive exon	55
Alternative 3’ splice site	86
Alternative 5’ splice site	74
Retained intron	191

Table 4. Number of splicing events identified by rMATS The number of alternative splicing events at exon-level identified by rMATS divided by class

Transcript Class	Number
Complete match	199103
Potentially novel isoform	12484
Inronic transfrag	985
Unkown trascript	875
Polymerase fragment	416
Generic exonic overlap	135
Pre-mRNA fragments	91
Exonic overlap on oposite strand	44
Intronic overlap on oposite strand	8

Table 5: Total number of transcripts assembled by StringTie divided by class.

transcript within the stress granules, as has been previously demonstrated in the literature [15].

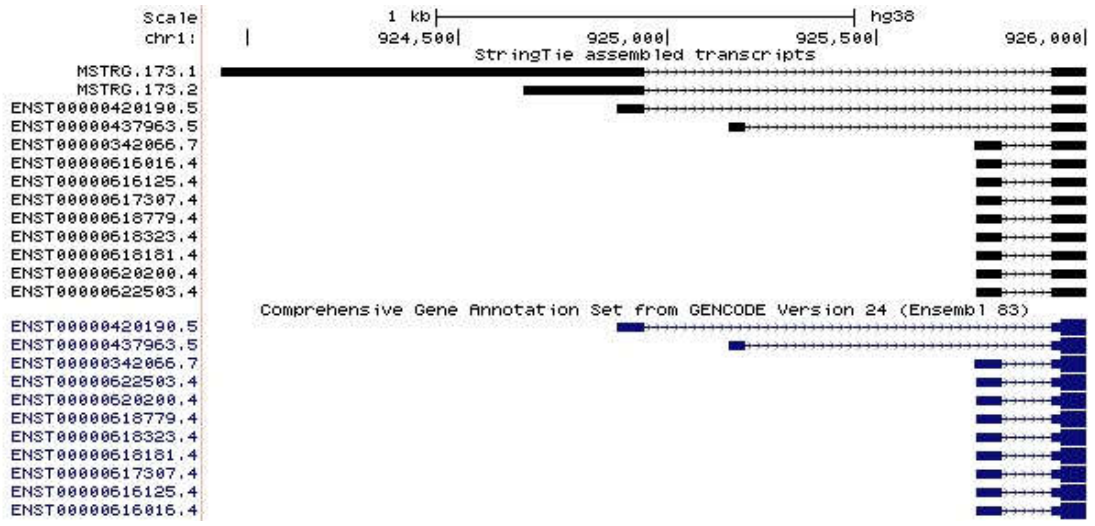

Figure 16: Comparison between the transcripts assembled by StringTie and the Gencode V24 comprehensive annotation. The image represents a portion of the 5'UTR region of the SAMD11 gene. At the top, in black, are the transcripts assembled by StringTie from the Caprin-1 and GFP libraries. At the bottom, in blue, are the transcripts annotated by the Gencode V24 comprehensive gene annotation. All the known isoforms were found by StringTie (with the corresponding ENST annotation), with the addition of 2 novel isoforms (MSTRG.173.1 and MSTRG.173.2), both of which show an elongated 5'UTR.

Conclusions

In summary, our results show that Caprin-1 is an RNA-binding protein that has a preference for binding to exonic regions of processed messenger RNAs (mRNAs) located primarily in the coding sequence (CDS) and 3' untranslated (3'UTR) portions of the transcript. Furthermore, Caprin-1 target genes in stress granules are enriched in transcripts coding for other RBPs which are, in turn, associated with protein complexes responsible for regulation metabolic pathways and control of cellular cycle. In addition, we also identified key genes targeted by Caprin-1 that are located within major metabolic pathways (such as cell cycle control, cancer -related pathways and immunological response pathways) that are either activated or repressed after induction of stress granules by ectopic expression of Caprin-1.

Our results for the prediction of the binding motif revealed that Caprin-1 has a preference for binding to structured RNA regions rich in GG repeats, which are capable of assembling in secondary RNA structures called G-quadruplex. We also identified that Caprin-1 binding sites are enriched in target sites for a group of 6

microRNAs, which at the same time are bound to Caprin-1 and are upregulated in the RNA-seq. When taken together, these results suggest that Caprin-1 binds to specific RNA structures in its targets, blocking the access of the microRNA to mRNA target site and, therefore, promoting an increase in the overall mRNA abundance and stability (Figure 17). However, our results cannot define the actual effects of Caprin-1 blockage of the microRNA target sites, requiring further experiments and more in-depth exploration of this possible regulatory mechanism in order to fully comprehend the biological process by which Caprin-1 regulate its target mRNAs.

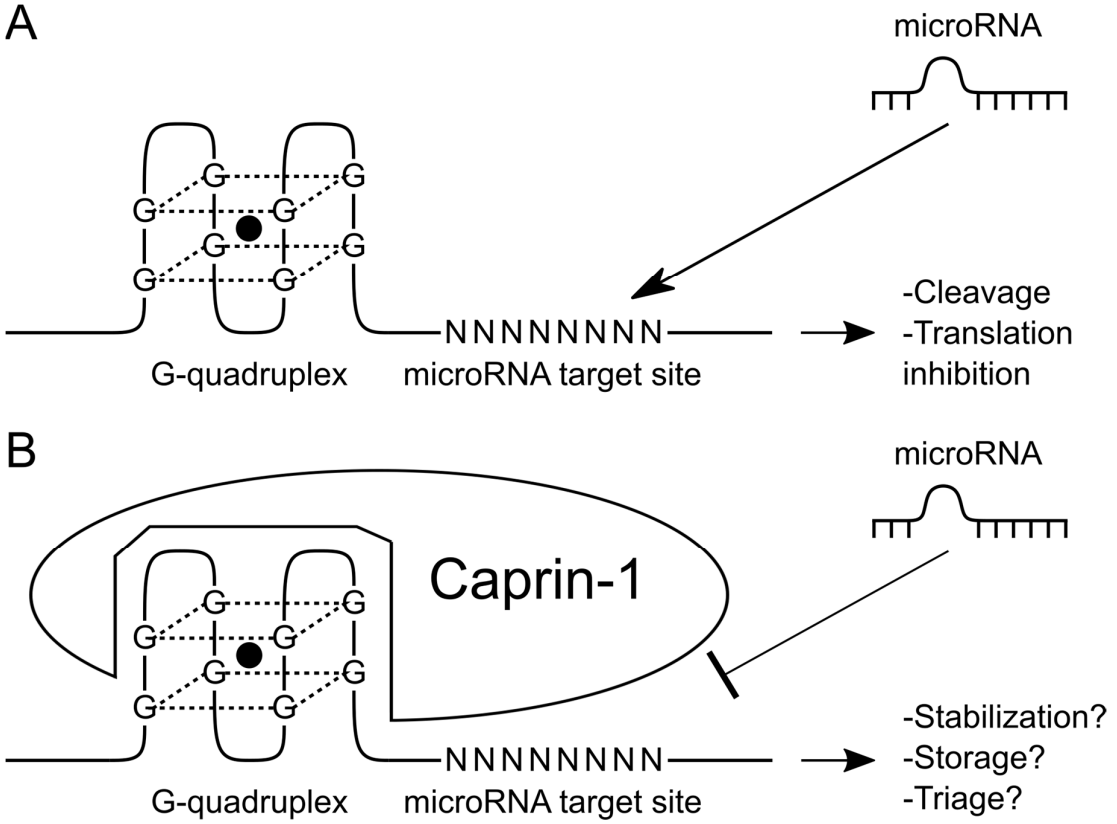

Figure 17: Proposed mechanism for Caprin-1 RNA-regulatory mechanism. Schematic drawing of the Caprin-1 regulatory mechanism for mRNAs that can be inferred from the results presented in the current study. **(A)** In physiological conditions, the microRNAs target the intended transcripts and promote mRNA cleavage and/or translational inhibition. **(B)** When Caprin-1 is present, however, it can recognize specific RNA secondary structures (G-quadruplexes) in the target transcripts and binds to its target, blocking the access of the microRNA to the target site and increasing the overall abundance of the mRNA. This could lead to either increased stability of the target, a long-term storage of the mRNA in the stress granule or a step in the transcript triage and production of stress-response proteins.

References

1. Anderson P, Kedersha N. Stress granules: the Tao of RNA triage. Trends Biochem. Sci. 2008. p. 141–50.

2. Anderson P, Kedersha N. RNA granules. J. Cell Biol. 2006. p. 803–8.

3. Kedersha NL, Gupta M, Li W, Miller I, Anderson P. RNA-binding proteins TIA-1 and TIAR link the phosphorylation of eIF-2?? to the assembly of mammalian stress granules. J. Cell Biol. 1999;147:1431–41.

4. Li YR, King OD, Shorter J, Gitler AD. Stress granules as crucibles of ALS pathogenesis. J. Cell Biol. 2013. p. 361–72.

5. Wolozin B. Regulated protein aggregation: stress granules and neurodegeneration. Mol. Neurodegener. [Internet]. 2012;7:56. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3519755&tool=pmcentrez&rendertype=abstract

6. Bentmann E, Haass C, Dormann D. Stress Granules in Neurodegeneration - Lessons learnt from TDP-43 and FUS. FEBS J. [Internet]. 2013;1–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23587065

7. Liu-Yesucevitz L, Bilgutay A, Zhang YJ, Vanderwyde T, Citro A, Mehta T, et al. Tar DNA binding protein-43 (TDP-43) associates with stress granules: Analysis of cultured cells and pathological brain tissue. PLoS One. 2010;5.

8. Sabile AA, Arlt MJE, Muff R, Husmann K, Hess D, Bertz J, et al. Caprin-1, a novel Cyr61-interacting protein, promotes osteosarcoma tumor growth and lung metastasis in mice. Biochim. Biophys. Acta - Mol. Basis Dis. 2013;1832:1173–82.

9. Kedersha N, Panas MD, Achorn CA, Lyons S, Tisdale S, Hickman T, et al. G3BP-Caprin1-USP10 complexes mediate stress granule condensation and associate with 40S subunits. J. Cell Biol. 2016;212:845–60.

10. Bidet K, Dadlani D, Garcia-Blanco MA. G3BP1, G3BP2 and CAPRIN1 Are Required for Translation of Interferon Stimulated mRNAs and Are Targeted by a

Dengue Virus Non-coding RNA. PLoS Pathog. 2014;10.

11. Solomon S, Xu Y, Wang B, David MD, Schubert P, Kennedy D, et al. Distinct structural features of caprin-1 mediate its interaction with G3BP-1 and its induction of phosphorylation of eukaryotic translation initiation factor 2alpha, entry to cytoplasmic stress granules, and selective interaction with a subset of mRNAs. Mol. Cell. Biol. [Internet]. 2007;27:2324–42. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1820512&tool=pmcentrez&rendertype=abstract

12. Reineke LC, Kedersha N, Langereis MA, van Kuppeveld FJM, Lloyd RE. Stress granules regulate double-stranded RNA-dependent protein kinase activation through a complex containing G3BP1 and Caprin1. MBio. 2015;6.

13. Shiina N, Yamaguchi K, Tokunaga M. RNG105 deficiency impairs the dendritic localization of mRNAs for Na+/K+ ATPase subunit isoforms and leads to the degeneration of neuronal networks. J. Neurosci. 2010;30:12816–30.

14. Gong B, Hu H, Chen J, Cao S, Yu J, Xue J, et al. Caprin-1 is a novel microRNA-223 target for regulating the proliferation and invasion of human breast cancer cells. Biomed. Pharmacother. 2013;67:629–36.

15. Buchan JR, Parker R. Eukaryotic Stress Granules: The Ins and Outs of Translation. Mol. Cell. 2009. p. 932–41.

16. Spriggs K a, Stoneley M, Bushell M, Willis AE. Re-programming of translation following cell stress allows IRES-mediated translation to predominate. Biol. Cell. 2008;100:27–38.

17. Guil S, Long JC, Cáceres JF. hnRNP A1 relocalization to the stress granules reflects a role in the stress response. Mol. Cell. Biol. [Internet]. 2006;26:5744–58. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1592774&tool=pmcentrez&rendertype=abstract

18. Yang X, Shen Y, Garre E, Hao X, Krumlinde D, Cvijović M, et al. Stress Granule-Defective Mutants Deregulate Stress Responsive Transcripts. PLoS Genet. 2014;10.

19. Lavut A, Raveh D. Sequestration of highly expressed mrnas in cytoplasmic granules, p-bodies, and stress granules enhances cell viability. PLoS Genet. 2012;8.

20. Blechingberg J, Luo Y, Bolund L, Damgaard CK, Nielsen AL. Gene Expression Responses to FUS, EWS, and TAF15 Reduction and Stress Granule Sequestration Analyses Identifies FET-Protein Non-Redundant Functions. PLoS One. 2012;7.

21. Biamonti G, Caceres JF. Cellular stress and RNA splicing. Trends Biochem. Sci. 2009. p. 146–53.

22. Metz A, Soret J, Vourc'h C, Tazi J, Jolly C. A key role for stress-induced satellite III transcripts in the relocalization of splicing factors into nuclear stress granules. J. Cell Sci. [Internet]. 2004;117:4551–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15331664

23. Albert H, Battaglia E, Monteiro C, Bagrel D. Genotoxic stress modulates CDC25C phosphatase alternative splicing in human breast cancer cell lines. Mol. Oncol. 2012;6:542–52.

24. Fu XD, Ares Jr. M. Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. [Internet]. 2014;15:689–701. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25112293%5Cnhttp://www.nature.com/nrg/journal/v15/n10/pdf/nrg3778.pdf

25. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: Pathways and programs unhinged. Genes Dev. 2010. p. 2343–64.

26. Van Oordt WVDH, Diaz-Meco MT, Lozano J, Krainer AR, Moscat J, Cáceres JF. The MKK(3/6)-p38-signaling cascade alters the subcellular distribution of hnRNP A1 and modulates alternative splicing regulation. J. Cell Biol. 2000;149:307–16.

27. Dutertre M, Sanchez G, Barbier J, Corcos L, Auboeuf D. The emerging role of pre-messenger RNA splicing in stress responses: Sending alternative messages and silent messengers. RNA Biol. 2011;8:740–7.

28. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, et al. Function of alternative splicing. Gene. 2013. p. 1–30.

29. Black DL. Protein Diversity from Alternative Splicing. Cell [Internet]. 2000;103:367–70. Available from: http://www.sciencedirect.com/science/article/pii/S0092867400001288

30. Goyal E, Amit SK, Singh RS, Mahato AK, Chand S, Kanika K. Transcriptome profiling of the salt-stress response in Triticum aestivum cv. Kharchia Local. Sci. Rep. [Internet]. Nature Publishing Group; 2016;6:27752. Available from: http://www.nature.com/articles/srep27752

31. Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L. Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. BMC Genomics [Internet]. 2014;15:431. Available from: http://www.biomedcentral.com/1471-2164/15/431

32. Eswaran J, Horvath A, Godbole S, Reddy SD, Mudvari P, Ohshiro K, et al. RNA sequencing of cancer reveals novel splicing alterations. Sci. Rep. [Internet]. 2013;3:1689. Available from: http://www.nature.com/articles/srep01689%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3631769&tool=pmcentrez&rendertype=abstract

33. Migita NA. Perfil de RNAs ligados a Proteína Caprin-1. 2016.

34. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods [Internet]. 2016;13:1–9. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3810%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27018577

35. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Available from: citeulike-article-id:11583827

36. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal [Internet]. 2011;17:10. Available from: http://journal.embnet.org/index.php/embnetjournal/article/view/200/479

37. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

38. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

39. Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat. Struct. Mol. Biol. 2013;20:1434–42.

40. Maticzka D, Lange SJ, Costa F, Backofen R. GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biol. [Internet]. 2014;15:R17. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053806&tool=pmcentrez&rendertype=abstract

41. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45:D362–8.

42. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009;25:75–82.

43. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: Targets and expression. Nucleic Acids Res. 2008;36.

44. Python Software Foundation. Python Language Reference, version 2.7 [Internet]. Python Softw. Found. 2013. Available from: http://www.python.org

45. R Core Team. R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria [Internet]. 2014;2014. Available from: http://www.r-project.org

46. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 2012;22:1760–74.

47. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for

differential expression analysis of digital gene expression data. Bioinformatics [Internet]. 2010;26:139–40. Available from: http://dx.doi.org/10.1093/bioinformatics/btp616

48. Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. [Internet]. 2014;111:E5593–601. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1419161111

49. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. [Internet]. 2015;33:290–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25690850%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4643835

50. Encode. ENCODE Guidelines and Best Practices for RNA-Seq: Revised December 2016. 2016;1–5. Available from: https://www.encodeproject.org/documents/cede0cbe-d324-4ce7-ace4-f0c3eddf5972/@@download/attachment/ENCODE Best Practices for RNA_v2.pdf

51. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002. p. 996–1006.

52. Conway AE, Van Nostrand EL, Pratt GA, Aigner S, Wilbert ML, Sundararaman B, et al. Enhanced CLIP Uncovers IMP Protein-RNA Targets in Human Pluripotent Stem Cells Important for Cell Adhesion and Survival. Cell Rep. The Authors; 2016;15:666–79.

53. Papoulas O, Monzo KF, Cantin GT, Ruse C, Yates JR, Ryu YH, et al. dFMRP and Caprin, translational regulators of synaptic plasticity, control the cell cycle at the Drosophila mid-blastula transition. Development. 2010;137:4201–9.

54. Kaddar T, Rouault J-P, Chien WW, Chebel A, Gadoux M, Salles G, et al. Two new miR-16 targets: caprin-1 and HMGA1, proteins implicated in cell proliferation. Biol. Cell. 2009;101:511–24.

55. Xiao H, Zeng J, Li H, Chen K, Yu G, Hu J, et al. MiR-1 downregulation correlates

with poor survival in clear cell renal cell carcinoma where it interferes with cell cycle regulation and metastasis. Oncotarget [Internet]. 2015;6:13201–15. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4537008&tool=pmcentrez&rendertype=abstract

56. Wang B, David MD, Schrader JW. Absence of caprin-1 results in defects in cellular proliferation. J. Immunol. 2005;175:4274–82.

57. Qiu Y-Q, Yang C-W, Lee Y-Z, Yang R-B, Lee C-H, Hsu H-Y, et al. Targeting a ribonucleoprotein complex containing the caprin-1 protein and the c-Myc mRNA suppresses tumor growth in mice: an identification of a novel oncotarget. Oncotarget. 2015;6:2148–63.

58. Brümmer A, Kishore S, Subasic D, Hengartner M, Zavolan M. Modeling the binding specificity of the RNA-binding protein GLD-1 suggests a function of coding region-located sites in translational repression. RNA [Internet]. 2013;19:1317–26. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3854522&tool=pmcentrez&rendertype=abstract

59. Pulcrano G, Leonardo R, Piscopo M, Nargi E, Locascio A, Aniello F, et al. PLAUF binding to the 3'UTR of the H3.3 histone transcript affects mRNA stability. Gene. 2007;406:124–33.

60. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell. 2010;38:576–89.

61. Thandapani P, O'Connor TR, Bailey TL, Richard S. Defining the RGG/RG Motif. Mol. Cell. 2013;50:613–23.

62. Siomi MC, Zhang Y, Siomi H, Dreyfuss G. Specific sequences in the fragile X syndrome protein FMR1 and the FXR proteins mediate their binding to 60S ribosomal subunits and the interactions among them. Mol. Cell. Biol. [Internet]. 1996;16:3825–32. Available from: http://mcb.asm.org/lookup/doi/10.1128/MCB.16.7.3825

63. Suhl JA, Chopra P, Anderson BR, Bassell GJ, Warren ST. Analysis of FMRP mRNA target datasets reveals highly associated mRNAs mediated by G-quadruplex structures formed via clustered WGGA sequences. Hum. Mol. Genet. 2014;23:5479–91.

64. Vasilyev N, Polonskaia A, Darnell JC, Darnell RB, Patel DJ, Serganov A. Crystal structure reveals specific recognition of a G-quadruplex RNA by a β-turn in the RGG motif of FMRP. Proc. Natl. Acad. Sci. U. S. A. [Internet]. 2015;112:E5391-400. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26374839%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4593078

65. Murat P, Zhong J, Lekieffre L, Cowieson NP, Clancy JL, Preiss T, et al. G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. Nat. Chem. Biol. [Internet]. 2014;10:358–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24633353

66. Fratta P, Mizielinska S, Nicoll AJ, Zloh M, Fisher EMC, Parkinson G, et al. C9orf72 hexanucleotide repeat associated with amyotrophic lateral sclerosis and frontotemporal dementia forms RNA G-quadruplexes. Sci. Rep. [Internet]. 2012;2:1016. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3527825&tool=pmcentrez&rendertype=abstract%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/23264878%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3527825

67. Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature [Internet]. 2008;455:1124–8. Available from: http://www.nature.com/doifinder/10.1038/nature07299%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/18806776

68. Fang Z, Rajewsky N. The impact of miRNA target sites in coding sequences and in 3???UTRs. PLoS One. 2011;6.

69. Schnall-Levin M, Zhao Y, Perrimon N, Berger B. Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3’UTRs. Proc. Natl. Acad. Sci.

U. S. A. 2010;107:15751–6.

70. Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. Functional microRNA targets in protein coding sequences. Bioinformatics [Internet]. 2012;28:771–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22285563

71. Hausser J, Syed AP, Bilen B, Zavolan M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. Genome Res. 2013;23:604–15.

72. Quann K, Jing Y, Rigoutsos I. Post-transcriptional regulation of BRCA1 through its coding sequence by the miR-15/107 group of miRNAs. Front. Genet. 2015;6.

73. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. [Internet]. 2003;5:R1. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-5-1-r1

74. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. 2010;11.

75. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature [Internet]. 2005;433:769–73. Available from: http://www.nature.com/doifinder/10.1038/nature03315

76. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature. 2010;466:835–40.

77. Ruike Y, Ichimura A, Tsuchiya S, Shimizu K, Kunimoto R, Okuno Y, et al. Global correlation analysis for micro-RNA and mRNA expression profiles in human cell lines. J. Hum. Genet. 2008;53:515–23.

78. Zhao H, Cui J, Wang Y, Liu X, Zhao D, Duan J. Spatial-temporal expression of pum1 and pum2 in medaka Oryzias latipes. J. Fish Biol. 2012;80:100–9.

79. Huang YH, Wu CC, Chou CK, Huang CYF. A translational regulator, PUM2, promotes both protein stability and kinase activity of aurora-A. PLoS One. 2011;6.

80. Vessey JP, Schoderboeck L, Gingl E, Luzi E, Riefler J, Di Leva F, et al. Mammalian Pumilio 2 regulates dendrite morphogenesis and synaptic function. Proc. Natl. Acad. Sci. U. S. A. [Internet]. 2010;107:3222–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2840302&tool=pmcentrez&rendertype=abstract

81. Spassov DS, Jurecic R. Mouse Pum1 and Pum2 genes, members of the Pumilio family of RNA-binding proteins, show differential expression in fetal and adult hematopoietic stem cells and progenitors. Blood Cells, Mol. Dis. 2003;30:55–69.

82. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. [Internet]. 2016;11:1650–67. Available from: http://dx.doi.org/10.1038/nprot.2016.095%5Cnhttp://10.1038/nprot.2016.095%5Cnhttp://www.nature.com/nprot/journal/v11/n9/abs/nprot.2016.095.html#supplementary-information

83. Reuter JA, Spacek D V., Snyder MP. High-Throughput Sequencing Technologies. Mol. Cell. 2015. p. 586–97.

84. Park PJ. ChIP-seq: Advantages and challenges of a maturing technology. Nat. Rev. Genet. 2009. p. 669–80.

85. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP Identifies Nova-Regulated RNA Networks in the Brain. Science (80-.). 2003;302:1212–5.

86. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 2008;40:1413–5.

87. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14.

88. Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations. Bioessays [Internet]. 2010;32:524–36. Available from: http://dx.doi.org/10.1002/bies.200900181%5Cnhttp://onlinelibrary.wiley.com/store/10.

1002/bies.200900181/asset/524_ftp.pdf?v=1&t=hz4sgjb3&s=d7669c48bf7843ed9e0715d320188c85879897ca

89. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods. 2015;12:115–21.

90. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. [Internet]. 2016;17:13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26813401%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4728800

91. Uhl M, Houwaart T, Corrado G, Wright PR, Backofen R. Computational analysis of CLIP-seq data. Methods. 2017. p. 60–72.

92. Liu Q, Zhong X, Madison BB, Rustgi AK, Shyr Y. Assessing Computational Steps for CLIP-Seq Data Analysis. Biomed Res. Int. 2015;2015.

93. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. Brief. Bioinform. [Internet]. 2016;bbv110. Available from: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbv110

94. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput. Biol. 2013;9.

95. Zhang MQ. Statistical features of human exons and their flanking regions. Hum. Mol. Genet. 1998;7:919–32.

96. Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. Structural and functional features of eukaryotic mRNA untranslated regions. Gene. 2001. p. 73–81.

97. Zhang MQ. Computational prediction of eukaryotic protein-coding genes. Nat. Rev. Genet. 2002. p. 698–709.

98. Larsen F, Gundersen G, Lopez R, Prydz H. CpG islands as gene markers in the human genome. Genomics. 1992;13:1095–107.

99. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.

100. Sethupathy P, Collins FS. MicroRNA target site polymorphisms and human disease. Trends Genet. 2008;24:489–97.

101. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. 2012;13.

102. Schübeler D. Function and information content of DNA methylation. Nature. 2015. p. 321–6.

103. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, Sklar P, et al. Identifying relationships among genomic disease regions: Predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 2009;5.

104. Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. 2003;4:R13.

105. 1000 Genomes Project Consortium T 1000 GP, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature [Internet]. 2012;491:56–65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23128226%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3498066

106. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science (80-.). [Internet]. 2007;318:420–6. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1149504

107. Stein L. Genome annotation: From sequence to biology. Nat. Rev. Genet. 2001. p. 493–503.

108. Liu C, Che D, Liu X, Song Y. Applications of machine learning in genomics and

systems biology. Comput. Math. Methods Med. 2013;2013.

109. Libbrecht MW, Noble WS. Machine learning in genetics and genomics. 2017;16:321–32.

110. Zhang YQ, Rajapakse JC. Machine Learning in Bioinformatics. Mach. Learn. Bioinforma. 2008.

111. Singireddy S, Alkhateeb A, Rezaeian I, Rueda L, Cavallo-Medved D, Porter L. Identifying differentially expressed transcripts associated with prostate cancer progression using RNA-Seq and machine learning techniques. 2015 IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. [Internet]. 2015. p. 1–5. Available from: http://ieeexplore.ieee.org/document/7300302/

112. Xue B, Oldfield CJ, Dunker AK, Uversky VN. CDF it all: Consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. FEBS Lett. 2009;583:1469–74.

113. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44:D710–6.

114. Encode Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2013;489:57–74.

115. Wei C, Xiao R, Chen L, Cui H, Zhou Y, Xue Y, et al. RBFox2 Binds Nascent RNA to Globally Regulate Polycomb Complex 2 Targeting in Mammalian Genomes. Mol. Cell. 2016;62:875–89.

116. Jangi M, Boutz PL, Paul P, Sharp PA. Rbfox2 controls autoregulation in RNA-binding protein networks. Genes Dev. 2014;28:637–51.

117. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

118. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: Enabling browsing of large distributed datasets. Bioinformatics. 2010;26:2204–7.

119. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology

Open Software Suite. Trends Genet. [Internet]. 2000;16:276–7. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0168952500020242

120. Tafer H, Höner zu Siederdissen C, Stadler PF, Bernhart SH, Hofacker IL, Lorenz R, et al. ViennaRNA Package 2.0. Algorithms Mol. Biol. 2011;6:26.

121. Kikin O, D'Antonio L, Bagga PS. QGRS Mapper: A web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res. 2006;34.

122. Oliphant TE. SciPy: Open source scientific tools for Python. Comput. Sci. Eng. [Internet]. 2007;9:10–20. Available from: http://www.scipy.org/

123. Weston J, Mukherjee S, Chapelle O, Pontil M. Feature selection for SVMs. Nips [Internet]. 2000;13:668–74. Available from: http://www.ee.columbia.edu/~sfchang/course/svia-F04/slides/feature_selection_for_SVMs.pdf

124. Ivanov A, Riccardi G. Kolmogorov-Smirnov test for feature selection in emotion recognition from speech. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2012. p. 5125–8.

125. Subrahmanyam K, Sankar NS, Baggam SP. A Modified KS-test for Feature Selection. IOSR J. Comput. Eng. 2013;13:73–9.

126. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. [Internet]. 2012;12:2825–30. Available from: http://dl.acm.org/citation.cfm?id=2078195%5Cnhttp://arxiv.org/abs/1201.0490

127. Blagus R, Lusa L. Boosting for High-Dimensional Two-Class Prediction. BMC Bioinformatics [Internet]. 2015;16:300. Available from: http://www.biomedcentral.com/1471-2105/16/300

128. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 2012. p. 463–84.

129. Schapire RE. A brief introduction to boosting. IJCAI Int. Jt. Conf. Artif. Intell.

1999. p. 1401–6.

130. Hastie T, Tibshirani R, Friedman J. Relative Importance of Predictor Variables. Elem. Stat. Learn. Elem. Stat. Learn. Mining, Inference, Predict. Second Ed. [Internet]. 2001. p. 367–9. Available from: http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf%255Cnhttp://www-stat.stanford.edu/~tibs/book/preface.ps

131. Longadge R, Dongre SS, Malik L. Class imbalance problem in data mining: review. Int. J. Comput. Sci. Netw. 2013;2:83–7.

132. Chan CY, Carmack CS, Long DD, Maliyekkel A, Shao Y, Roninson IB, et al. A structural interpretation of the effect of GC-content on efficiency of RNA interference. BMC Bioinformatics. 2009;10:1–7.

133. Zhang J, Kuo CCJ, Chen L. GC content around splice sites affects splicing through pre-mRNA secondary structures. BMC Genomics [Internet]. 2011;12:90. Available from: http://www.biomedcentral.com/1471-2164/12/90

134. Li X, Kazan H, Lipshitz HD, Morris QD. Finding the target sites of RNA-binding proteins. Wiley Interdiscip. Rev. RNA. 2014. p. 111–30.

135. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. RNAcontext: A new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Comput. Biol. 2010;6:28.

136. Marcel V, Tran PLT, Sagne C, Martel-Planche G, Vaslin L, Teulade-Fichou MP, et al. G-quadruplex structures in TP53 intron 3: Role in alternative splicing and in production of p53 mRNA isoforms. Carcinogenesis. 2011;32:271–8.

137. Ribeiro MM, Teixeira GS, Martins L, Marques MR, de Souza AP, Line SRP. G-quadruplex formation enhances splicing efficiency of PAX9 intron 1. Hum. Genet. 2014;134:37–44.

138. Gazzara MR, Mallory MJ, Roytenberg R, Lindberg J, Jha A, Lynch KW, et al. Ancient antagonism between CELF and RBFOX families tunes mRNA splicing outcomes. Genome Res. [Internet]. 2017;gr.220517.117. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/28512194%5Cnhttp://genome.cshlp.org/lookup/doi/10.1101/gr.220517.117

139. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45:D369–79.

140. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell. 2015;162:425–40.

141. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the human interactome defines protein communities and disease networks. Nature. 2017;545:505–9.

142. Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual JF, et al. A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration. Cell. 2006;125:801–14.

143. Hegele A, Kamburov A, Grossmann A, Sourlis C, Wowro S, Weimann M, et al. Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome. Mol. Cell. 2012;45:567–80.

144. Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, et al. Panorama of ancient metazoan macromolecular complexes. Nature. 2015;525:339–44.

145. Shao C, Yang B, Wu T, Huang J, Tang P, Zhou Y, et al. Mechanisms for U2AF to define 3’ splice sites and regulate alternative splicing in the human genome. Nat. Struct. Mol. Biol. 2014;21:997–1005.

146. Tanackovic G, Krämer A. Human splicing factor SF3a, but not SF1, is essential for pre-mRNA splicing in vivo. Mol. Biol. Cell [Internet]. 2005;16:1366–77. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=551499&tool=pmcentrez&rendertype=abstract

147. Wickramasinghe VO, Gonzàlez-Porta M, Perera D, Bartolozzi AR, Sibley CR,

Hallegger M, et al. Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5’ splice site strength. Genome Biol. 2015;16.

148. Ascano M, Mukherjee N, Bandaru P, Miller JB, Nusbaum JD, Corcoran DL, et al. FMRP targets distinct mRNA sequence elements to regulate protein expression. Nature. 2012;492:382–6.

149. Lu G, Hall TMT. Alternate modes of cognate RNA recognition by human PUMILIO proteins. Structure. 2011;19:361–7.

150. Matoulkova E, Michalova E, Vojtesek B, Hrstka R. The role of the 3′ untranslated region in post-transcriptional regulation of protein expression in mammalian cells. RNA Biol. 2012. p. 563–76.

151. Tollervey JR, Curk T, Rogelj B, Briese M, Cereda M, Kayikci M, et al. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nat. Neurosci. [Internet]. Nature Publishing Group; 2011 [cited 2014 Mar 22];14:452–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3108889&tool=pmcentrez&rendertype=abstract

152. Lou T-F, Weidmann CA, Killingsworth J, Tanaka Hall TM, Goldstrohm AC, Campbell ZT. Integrated analysis of RNA-binding protein complexes using in vitro selection and high-throughput sequencing and sequence specificity landscapes (SEQRS). Methods [Internet]. Elsevier Inc.; 2017;118–119:171–81. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1046202316303383

153. Campbell ZT, Wickens M. Probing RNA-protein networks: Biochemistry meets genomics. Trends Biochem. Sci. [Internet]. Elsevier Ltd; 2015;40:157–64. Available from: http://dx.doi.org/10.1016/j.tibs.2015.01.003

154. Colombrita C, Onesto E, Megiorni F, Pizzuti A, Baralle FE, Buratti E, et al. TDP-43 and FUS RNA-binding proteins bind distinct sets of cytoplasmic messenger RNAs and differently regulate their post-transcriptional fate in motoneuron-like cells. J. Biol. Chem. 2012;287:15635–47.

155. Galarneau A, Richard S. Target RNA motif and target mRNAs of the Quaking STAR protein. Nat. Struct. Mol. Biol. 2005;12:691–8.

156. Pérez I, Lin CH, McAfee JG, Patton JG. Mutation of PTB binding sites causes misregulation of alternative 3' splice site selection in vivo. RNA [Internet]. 1997;3:764–78. Available from: http://rnajournal.cshlp.org/content/3/7/764.abstract

157. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. 2010;17:909–15.

158. Swanson MS, Dreyfuss G. Classification and purification of proteins of heterogeneous nuclear ribonucleoprotein particles by RNA-binding specificities. Mol. Cell. Biol. [Internet]. 1988;8:2237–41. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=363409&tool=pmcentrez&rendertype=abstract

159. Galarneau A, Richard S. The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs. BMC Mol. Biol. 2009;10.

160. Dember LM, Kim ND, Liu KQ, Anderson P. Individual RNA recognition motifs of TIA-1 and TIAR have different RNA binding specificities. J. Biol. Chem. 1996;271:2783–8.

161. Takahama K, Kino K, Arai S, Kurokawa R, Oyoshi T. Identification of Ewing's sarcoma protein as a G-quadruplex DNA- and RNA-binding protein. FEBS J. 2011;278:988–98.

162. Skourti-Stathaki K, Proudfoot NJ, Gromak N. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. Mol. Cell. 2011;42:794–805.

163. Stefanovic S, DeMarco BA, Underwood A, Williams KR, Bassell GJ, Mihailescu MR. Fragile X mental retardation protein interactions with a G quadruplex structure in the 3'-untranslated region of NR2B mRNA. Mol. Biosyst. [Internet]. 2015;11:3222–30. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4643373&tool=pmcentrez

&rendertype=abstract

164. Cammas A, Millevoi S. RNA G-quadruplexes: emerging mechanisms in disease. Nucleic Acids Res. 2017;45:1584–95.

165. BioFeatureFindder [Internet]. GitHub. Available from: https://github.com/kbmlab/BioFeatureFinder

Appendix 1 (Article) - BioFeatureFinder: Flexible, unbiased analysis of biological characteristics associated with genomic regions

Abstract

BioFeatureFinder (BFF) interrogates interesting genomic landmarks (including alternatively spliced exons, DNA/RNA-binding protein binding sites, and gene/transcript functional elements) to identify distinguishing biological features (nucleotide content, conservation, k-mers, secondary structure, protein binding sites and others). BFF uses a flexible underlying model, combining classical statistical tests with big data machine learning strategies, that uses thousands of biological characteristics (features) to interpret category labels in genomic ranges or numerical scales from genome graphs. Our results show that BFF provides a reliable analysis platform for large-scale datasets, capable of recovering several well-known features from the literature for RNA-binding proteins as well as uncovering novel associations for 112 eCLIP-seq datasets. BioFeatureFinder is freely available at https://github.com/kbmlab/BioFeatureFinder/.

Background

The emergence of high-throughput sequencing technologies led to an increase in the magnitude of datasets available for researchers, and multiple types of analysis were built based on these technologies [83]. These strategies can be applied to identify protein binding sites (ex. ChIP [84]/CLIP-seq [85]), alternative splicing (AS) events [86], differentially expressed genes [87], detection of SNPs [88] and a multitude of other applications [89,90], resulting in large sets of genomic coordinates (ex. Binding sites, AS exons, polymorphisms). This type of result is particularly challenging to interpret from a biological perspective and is time-consuming. Several approaches have been used for characterization of sets of genomic coordinates and for the identification of enriched characteristics (features) in these datasets, especially when dealing with results of ChIP-seq or CLIP-seq experiments [90–94]. However, several of the most commonly used tools in these analyses focus on particular aspect of the target regions in their process, such as structure models or sequence motifs [40,91,93]. Although these tools provide valuable insight in which characteristics are enriched in

the genomic regions associated with these datasets, there is a clear deficiency in computational tools capable of performing more comprehensive analyses and integrate multiple types of sources of variation.

From previous studies, specific features were shown as contributors: GC content [95–97], nucleotide composition [95], length [95,97], CpG islands [97,98], conservation [99], microRNA [70,100] and protein-binding [101] target sites, methylation sites [102], single nucleotide polymorphisms (SNPs) [103], microssatellite regions [104] and also the aforementioned sequence motifs and structural characteristics of these regions. However, due to the biological variability occurring within the genome sequences [105–107], we must also consider that sets of genomic regions are also composed of a heterogeneous population of sequences, each with its unique profile of characteristics. While not all these characteristics are enriched and/or important for creating a profile for the whole region groups, it is possible that a combination of many factors is responsible for separating groups of genomic regions and/or determining the binding of a protein to that particular region.

In addition to the identification of enriched features in a set of genomic coordinates, BioFeatureFinder can help discover how these characteristics interact with each other. This is useful to create an accurate map of which feature are more important for explaining differences between the input genomic regions and the remaining regions of the genome. In order to achieve that, we applied machine-learning strategies, which are already widely used in other transcriptomic, genomics and system biology studies [101,108–111]. Instead of analyzing the genomic regions as individual data-points, we analyzed the cumulative distribution functions that are drawn from the population of regions for each of the features described above, then proceed to apply binary classification algorithms in order to identify which characteristics are more important for group separation. This type of strategy has already been applied in other studies [112], but it is applied for the first time in the context of classification of feature associated with groups of genomic regions in this scale. Furthermore, we also aimed at developing a widely flexible tool that can use data available from multiple public databases such as UCSC GenomeBroswer [51], Ensembl [113], GENCODE [46], ENCODE [114] and others, as well as capable of performing in an unsupervised and unbiased way.

Results and Discussion

We present BioFeatureFinder (BFF), a flexible and unbiased algorithm for discovery of distinguishing biological features associated with groups of genomic regions. We define “biological features” as the set of characteristics that can be used to distinguish regions from other sections of the genome. These features can include, but are not limited to: nucleotide content, length, conservation, k-mer occurance, presence of SNPs, protein-binding sites, microRNA target sites, methylation sites, microsatellite, CpG islands, repeating elements, protein domains and others.

Our algorithm is capable of analyzing the distribution of values for each of those features in a set of genomic regions of interest, then the algorithm compares this distribution with a randomized background in order to identify which features represent the most distinguishing characteristics associated with the input dataset and ranks them by importance values. This tool can be used as an important information source for scientists, using the data provided to generate new and more accurate hypothesis and also guide wet-lab experiments more efficiently. Also, BFF can be used in large-scale computational projects, being capable of analyzing hundreds of datasets with ease and produce consistent results.

For the first time, we apply big data strategies in an unbiased way, effectively reducing observer bias, to take advantage of the large amounts of data produced by high-throughput experiments, such as CLIP/CHIP/RNA/DNA-seq, and data deposited on publicly available databases (UCSC GenomeBroswer [51], Ensembl [113], GENCODE [46], ENCODE [114] and others) to extract a set of significant informations from genomic regions in order to uncover latent relationships inside datasets. First, we present the framework used by BFF in its analytic process, with an overview of the input data types, workflow and output. Second, as a control, we applied BFF to the RBP (RNA-binding protein) RBFOX2 eCLIP-seq (enhanced crosslink immunoprecipitation RNA-sequencig) data, since this protein is widely studied and it’s binding sites are well-characterized in the literature [34,39,115,116]. Finally, to showcase potential applications of the algorithm, we analyzed 112 eCLIP datasets obtained from human cell lines, which are available from ENCODE database, identifying biological features associated with binding sites of all RBPs and their respective importance scores.

The BioFeatureFinder workflow

BioFeatureFinder focuses on flexibility, consistency and scalability. It is python-based with scalable multi-thread capabilities, being memory-friendly and compatible with most commonly used UNIX-based systems (such as CentOS, Ubuntu, openSUSE and macOS) it can be used with a wide-range of hardware, ranging from notebooks to HPC clusters. In Figure 1 we show a schematic representation of the BFF workflow. The kinds of inputs required for using the algorithm are: a set of BED coordinates with genomic regions of interest (ex. CLIP/CHIP-seq binding sites, promoter regions for differentially expressed genes, splice sites for alternatively spliced exons/introns and others), compatible fasta files with sequences (ex. Reference genomeranscriptome), for increased accuracy it is also possible to use a GTF/GFF file with regions annotations (exons, introns, CDS, UTR and others). Optionally, to increase the number of features analyzed by using BED files with genomic regions of biological features (ex. microRNA sites, methylation sites, CpG islands, protein binding sites, SNPs and mutations, repeating elements and multiple bigWig files with phastCons scores for multiple alignment.

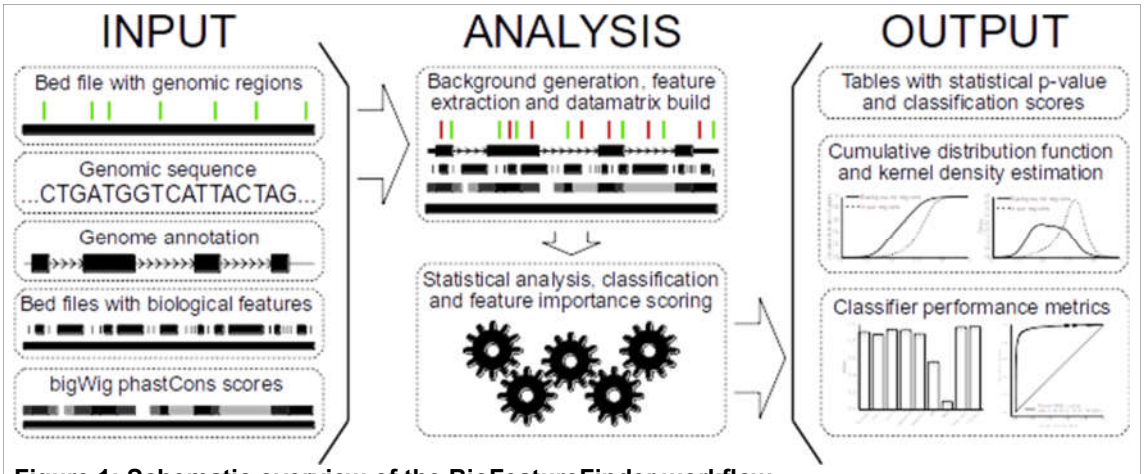

Figure 1: Schematic overview of the BioFeatureFinder workflow.

The analytic process of the algorithm is divided in 2 sub-sections: Build /datamatrix and Analyze features. Building the datamatrix starts with selecting an appropriate background for comparison of the input regions of interest, which is obtained using the *shuffle* function of bedtools. Although not required, the usage of a reference annotation improves accuracy of the algorithm by guiding the included/excluded background regions. The total number of background regions (B) is

proportional to the number of regions in the input list (I) of bed coordinates, which can be represented by the following formula: $B = I * N$, where N is an integer variable that can be set as one of the options (default = 3, i.e. The number of background regions is 3 times the number of input regions). These two sets of regions are then used to produce a datamatrix. Each bed entry in the regions is converted into a line in the matrix, and each feature corresponds to a column. Every feature is represented as numeric value, which can be a continuous, discrete or boolean variable. For obtaining these values, BFF uses multiple freely available tools such as bedtools [117], for nucleotide content and counting intersections with features in bed format, bigWigAverageOverBed [118], for extraction of conservation scores, EMBOSS wordcount [119] for k-mer counting, Vienna's RNAfold [120] for RNA secondary structure MFE (minimum free energy) values and QRGS Mapper [121] for G-quadruplex scoring. Designed with a modular concept, new functions and sources of data can be easily added by researchers to answer project-specific questions.

Once the matrix is created, the algorithm applies a two-step analysis for identifying important features in the dataset. The first step is to analyze each feature in the matrix with a two-sample Kolmogorov-Smirnov test (KST) implementation by SciPy [122], comparing the distribution of values of the regions of interest (group 1) with background regions (group 0). Aside from a statistical tool for identification of significant features, it's also possible to use KST as a tool for feature selection, extracting statistically significant features which are correlated with differences between the groups, a strategy which has been shown to improve classification performance on high-dimensional data [123–125]. As an additional benefit, filtering the features by KST p-values also reduces the size of the datamatrix used in the following classification step, which can be helpful in reducing both computational time and resources used in the analyses. The second step involves the usage of a Stochastic GradientBoost Classifier (St-GBCLF) from Scikit-learn [126], which is capable of naturally handling mixed datatypes, is fairly robust to outliers and possesses reliable predicitive power. Additionally, this method has been shown to be preferable for high-dimensional two-class prediction [127,128]. Also, as other ensemble methods, St-GBCLF is less likely to suffer from overfitting [129]. This stage will use the feature values extracted from the matrix (which can be filtered, or not, by KST) for each group (0, background, and 1, input) and calculate feature importance, a score which measures how valuable each feature was in the decision-making process of the trees

[130]. Higher importance values indicate that the feature considered to make key decisions and, therefore, can be inferred to have more biological significance. In order to address the issue of class imbalance problems that are inherent to these types of analysis [128,131], our algorithm draws a random sample from the background (group 0) that is the same size as the input dataset (group 1), increasing the overall accuracy of the classifier. However, in order to address the biological variability, our algorithm performs multiple classification runs, with each time drawing a new sample of background regions. The final classification score, then, is calculated as an average of importance values obtained in each classification run.

Both classification and statistical results are compiled into a table, which allows easy interpretation. Additionally, graphical representations of each feature are outputted in both cumulative distribution function (CDF) and kernel density estimation (KDE) plots. This allows visualization of the distributions found in the input and background, and leads to conclusions on how the distribution is shifted from the reference. Additionally, classifier importance and KS test values are output in barcharts. Lastly, the classifier performance is measured by several parameters: accuracy, sensitivity, sensibility, positive predictive value, negative predictive value, adjusted mutual information (aMI), mean squared error (MSE) values and receiver operating characteristic (ROC) and precision-recall (P-R) area under curve. These metrics are outputted in both table (with scores) and graphical (barcharts and curves) formats. Taken together, these outputs can be used for exploration of the data and identification of features which can be of significance in a biological context.

Analysis of RBFOX2 eCLIP dataset

In order to evaluate the performance of our algorithm, we analyzed the binding sites in the group of targets RNAs for the RNA binding protein RBFOX2, available from eCLIP experiments deposited in ENCODE database. We found 922 statistically significant biological features by the KS test, which were used in the classification step. The classification algorithm achieved a satisfactory performance, with an overall 91% mean accuracy score, 88% positive predictive value (P.P.V.), 93% negative predictive value (N.P.V.), 93% sensitivity, and 89% specificity. Overall this means that, in average, our classifier provides accurate predictions 9 out of 10 times. Also, we've obtained average scores for adjusted mutual information (aMI) and mean

squared error (MSE) of 0.56 and 0.09 respectively. Both receiver operating characteristic (ROC) and Precision-Recall (P-R) area under curve (AUC) were measured at 0.97 (Figure 2A, Additional File 1). Among all statically significant features, our approach identified 16 features which had a relative importance score of at least 10% (i.e. 1/10 of the importance score of the highest scoring feature), containing both known features from the literature, such as conservation of the binding site and an enrichment for the GCAUG k-mer, which is the known binding motif for this protein, but we also identified novel features, such as higher GC-content of binding sites, lower MFE for RNA secondary structure, higher G-quadruplex score and overlap of binding sites with RPS5 and other RBPs (Figure 2B, Additional File 2).

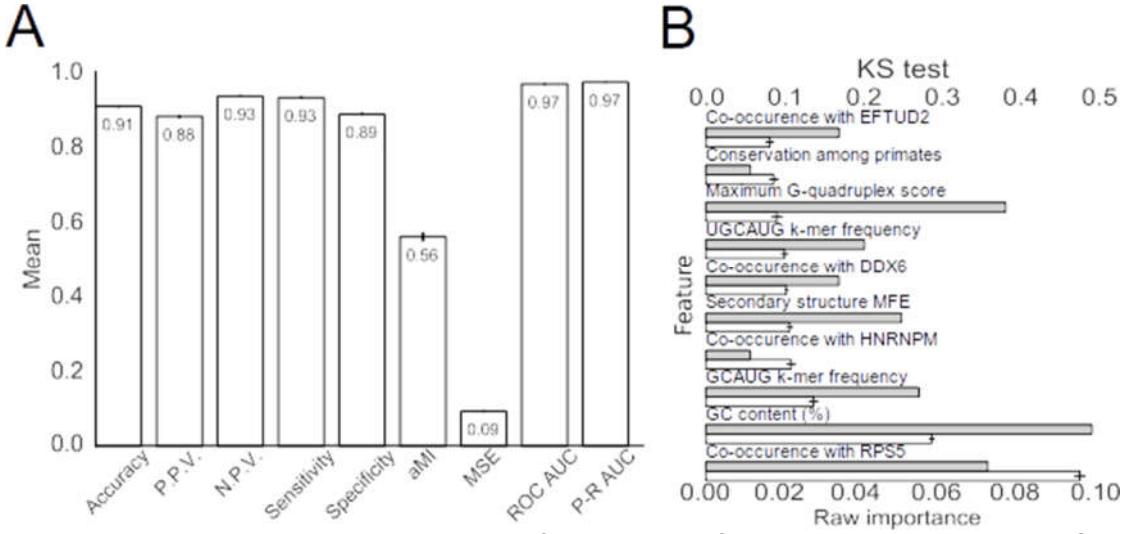

Figure 2: BioFeatureFinder accurately identifies biological features associated with RBFOX2 binding sites. A. Bar charts for mean classifier performance scores for RBFOX2 eCLIP sites (P.P.V.: Positive predictive value, N.P.V.: Negative predictive value, aMI: Adjusted mutual information, MSE: Mean squared error, ROC AUC: Receiver operating characteristic area under curve, P-R AUC: Precision-Recall area under curve).; **B.** Horizontal bar chart showing the importance score (white) and Kolmogorov-Smirnov test value (grey) for the top 10 features associated with RBFOX2 sites. The black bars represent the standard deviation found in each scoring parameter.

We identified the enrichment of the GCAUG 5-mer as one of the major features that characterized the RBFOX2 eCLIP binding sites by both KS test (p-value < 0.001) and variable importance in classification. Our analysis indicated that 32.83% of binding sites identified in RBFOX2 eCLIP contained at least 1 repetition of the GCAUG motif, while only 5.69% randomized background regions exhibited at least 1 instance of this motif (Figure 3A). Additionally, we also found significant enrichment for the UGCAUG 6-mer, which occurred in 21.87% of binding sites in contrast with 1.67% of background regions (Figure 3B). Both of these results are consistent with RBFOX2

nucleotide sequence motif enrichment and occurrence in binding sites [39,116], indicating that our algorithm successfully recovered known features.

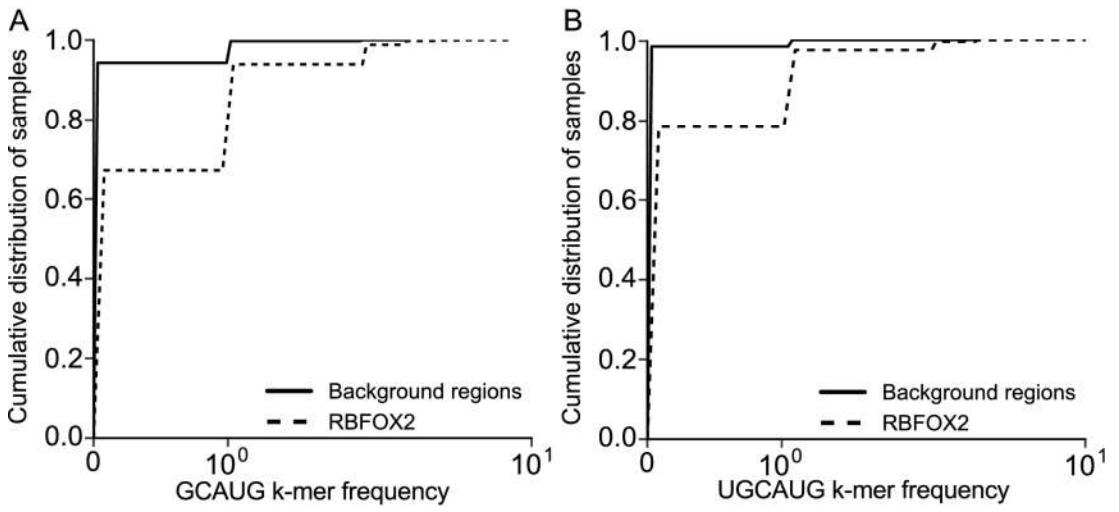

Figure 3: RBFOX2 binding sites are enriched for the (U)GCAUG motif. A-B. Cumulative distribution function curves for GCAUG (A) and UGCAUG (B) k-mer sequences. The Y-axis shows the cumulative distribution of samples and the X-axis indicates the number of occurrences for each k-mer. The solid lines represent randomized background regions while dashed lines represent RBFOX2 binding sites.

Interestingly, we identified several major components associated with RNA secondary structure among the important features for RBFOX2 binding sites. GC content had the second highest importance value from all features, with RBFOX2 binding sites exhibiting a higher distribution of GC than randomized background regions, with the majority of binding sites having a range of 50% to 80% GC content in their sequences while background regions were more evenly distributed between 20% to 60% (Figure 4A). It is known that RNA regions with higher GC content tend to have a more stable secondary RNA structure than regions with lower GC content [132], also GC content has already been associated with alterations in splicing patterns by affecting pre-mRNA secondary structure [133], which is a known mechanism for RBFOX2 splicing regulation [39]. Although the importance of RNA secondary structure as a guiding factor for RBP binding has already been determined by previous studies [40,134,135], our finding shows this association occurs in RBFOX2, since we identified that minimum free energy (MFE) for RNA folding is a major feature for distinguishing the protein binding sites from randomized background. We identified that 70.31% of binding sites had a MFE lower than 0, indicating the possible existence of a localized secondary structure, while only 47.52% of background regions exhibited similar

behavior (Figure 4B). Additionally, we also identified that the presence of G-quadruplexes, a specific type of secondary structure, appears to be enriched in this protein's binding sites. Our analysis indicates that 53.29% of RBFOX2 binding sites had a positive score for their presence, in contrast with only 15.10% of background regions (Figure 4C).

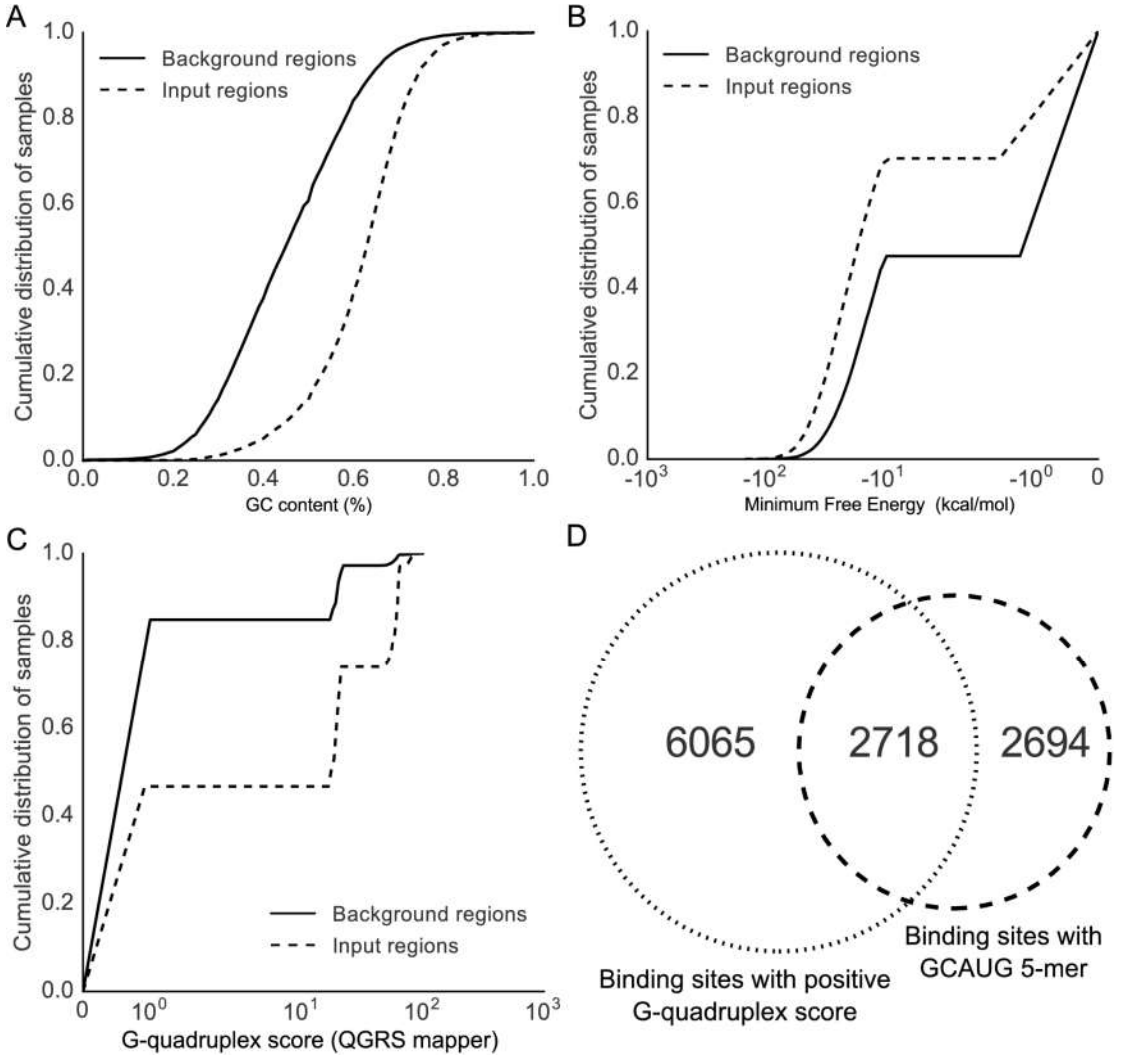

Figure 4: RBFOX2 binds preferentially to structured RNA regions enriched in GC content. A-C. Cumulative distribution function curves for GC content (A), Minnimum Free Energy (MFE, B) and maximum G-quadruplex score (C). The Y-axis shows the cumulative distribution of samples and the X-axis indicates the values obtained for each feature. The solid lines represent randomized background regions while dashed lines represent RBFOX2 binding sites. **D**. 2-way Venn diagram showing the overlap between the number of peaks identified with the GCAUG k-mer and the ones with positive G-quadruplex score.

This result is particularly interesting because, although this feature was not previously associated with RBFOX2, it can be supported extensively by literature evidence. First, RBFOX2 has been described as a member of the RG/RGG family of

RNA-binding proteins, with arginine-glycine rich regions, known as RGG-box, which is responsible for RNA recognition [61]. Second, other proteins from this family have been shown to bind to RNA G-quadruplex by their RG/RGG regions [64]. Third, the existence of G-quadruplexes in intronic regions can have impacts on alternative splicing regulation [136,137]. Taken together, these results indicate that secondary RNA structure may play a bigger role in RBFOX2 targeting for binding sites than previously assumed, combining with the existence of the GCAUG motif for increased accuracy in target selection. This is further evidenced by the fact that 50.22% of binding sites containing GCAUG are also positive for the presence of G-quadruplexes (Figure 4D).

Lastly, the most important feature had not been shown previously and represents to overlap of RBFOX2 binding sites with RPS5 binding sites. We identified that 35.93% of RBFOX2 peaks had at least 1 nucleotide position in common with RPS5 binding sites, which is significantly higher than the value of obtained for randomized background regions that scored less than 0.01% of overlap (Figure 5A). Although this association is novel, it can also be observed in another study which showed that CELF2-repressed exons were not only enriched for RBFOX2 in their downstream intron but also for RPS5 in HepG2 cell lines [138]. We also identified several other RPBs which had significant overlap with RBFOX2, including known splicing regulators and/or components of the spliceosome such as HNRNPM, EFTUD2, PRPF8, QKI, HNRNPK and PCBP2 (Additional File 2). Using data available from BioGrid 3.4 [139] and STRINGdb 10.5 [41], we found that these targets were associated to RBFOX2 by a curated protein-protein interaction network (Figure 5B), with RBFOX2 directly interacting with QKI [140–143] and HNRNPK [143], the latter, in turn, has been shown to interact with RPS5 [144]. Furthermore, when analyzing the eCLIP data for RPS5 with our algorithm we identified that 67.17% of binding sites showing overlap with RBFOX2 binding sites, 29.5% of them had at least 1 repetition for the GCAUG motif and 21.14% at least 1 instance of the UGCAUG motif, however we found little indication of RPS5 having preferences for binding to any type of secondary RNA structure (Additional Files 2, 3 and 4).

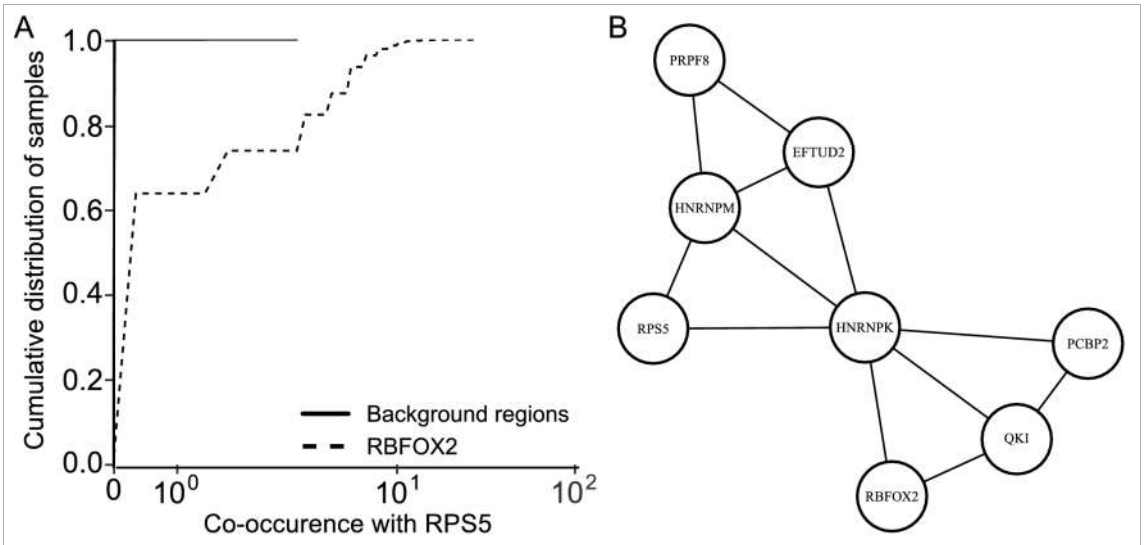

Figure 5: Overlapping RBPs identified as important features are connected to RBFOX2 via a protein-protein interaction (PPI) network. A. Cumulative distribution function curves for RPS5 overlapping binding sites with RBFOX2. The Y-axis shows the cumulative distribution of samples and the X-axis indicates the number of occurrences for each overlap. The solid lines represent randomized background regions while dashed lines represent RBFOX2 binding sites. **B.** PPI network created based on interactions drawn from BioGrid 3.4 and StringDB 10.5. Each node represents a different RBP and the lines represent known protein-protein interactions between them.

Taken together, our results indicate the existence of a combinatorial mechanism of both RNA structure and nucleotide sequence to direct binding specificity to either RBFOX2 or RPS5, especially since the former is a RBP associated with regulation of alternative splicing and the latter is a component of the small subunit of the ribosome associated with translational regulation. However, our analysis is limited to identifying the similar and diverging characteristics of their binding sites. For a more complete understanding of the relationship between RBFOX2 and RPS5, further experiments would be required to determine if they indeed have different molecular mechanisms and the differences on the target sequence is what determine the binding specificity or if there are novel biological functions for either RBFOX2 or RPS5 yet to be discovered.

Identification of important features for 112 RNA-binding proteins binding sites from ENCODE

To showcase potential applications of BioFeatureFinder in high-throughput studies, we applied our algorithm to 112 eCLIP-seq datasets available at ENCODE. First, we identified the preferential binding regions for each RNA Binding Protein (RBP)

in the dataset (Figure 6A, Additional File 6), with our results indicating that 59.8% of the proteins analyzed had preferential binding to the intronic regions. The second most frequent region was 3'UTR (15.2%), followed by CDS (14.3%), 3' splice site (7.1%), 5' splice site (1.8%) and 5'UTR (1.8%). Among the identified preferential regions for the RBPs, some were already known from the literature (such as U2AF1 [145], U2AF2 [145], SF3A3 [146], PRPF8 [147], FMR1 [148], PUM2 [149], TIA1 [150], TARDBP [151] and RBFOX2 [39]), which demonstrates that our algorithm correctly identified their binding region preferences. We used this information to generate the appropriate background for each RBP. Overall, our algorithm performed consistently with an average accuracy of 0.9 and average ROC and Precision-Recall AUCs of 0.95. The biggest variance encountered was with the aMI (adjusted mutual information) scores, with an average of 0.57 and standard deviation of 0.16 (Figure 6B, Additional File 2). We also observed a strong correlation (Pearson's $R^2 \geq 0.95$) between aMI scores and Accuracy (Figure 7A) and MSE (Mean Squared Error, Figure 7B), indicating that RBPs with higher aMI scores tend to reach a higher degree of resolution of the binding site features. This can be inferred to be a consequence of the binding characteristics of the RBPs, with some proteins, as TARDBP, possessing a strict set of characteristics that guide their binding to specific targets, while other RBPs, as SF3B1, appear to have a higher degree of flexibility in their binding target selection (Additional File 2).

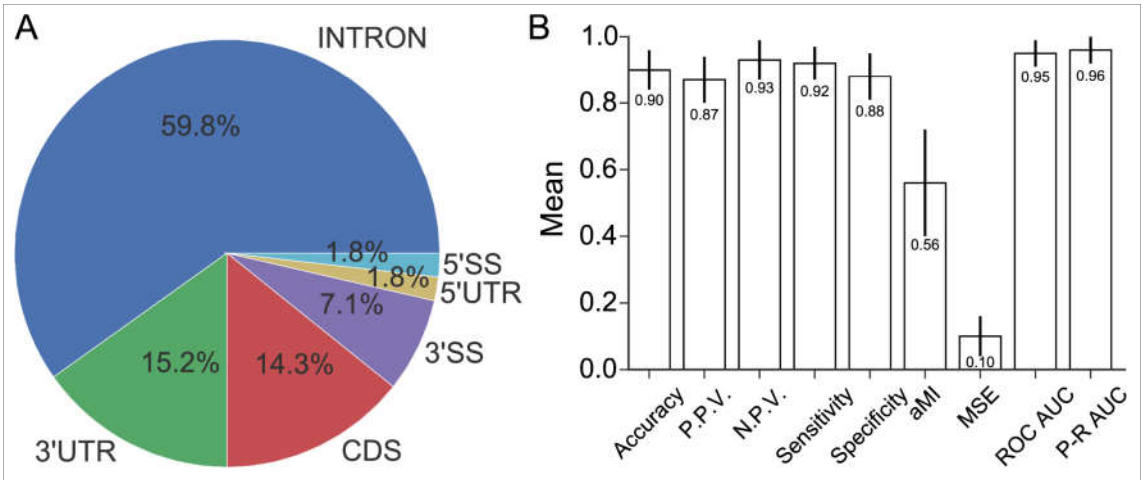

Figure 6: BioFeatureFinder performs consistently and accurately for 112 RBPs that bind to multiple transcript regions. A. Pie chart showing the percentage of RBPs that had preferential binding to each transcriptomic region. Each slice of the chart corresponds to a different region (Intron, 3'UTR, CDS, 3'SpliceSite, 5'UTR, 5'SpliceSite) and the percentages correspond to the number of RBPs which had higher number of binding sites to that region. **B.** Bar charts for mean classifier performance scores for 112 eCLIP sites (P.P.V.: Positive predictive value, N.P.V.: Negative predictive value, aMI: Adjusted mutual information, MSE: Mean squared error, ROC AUC: Receiver operating characteristic area under curve, P-R AUC: Precision-Recall area under curve).The black bars represent the standard deviation found in each scoring parameter.

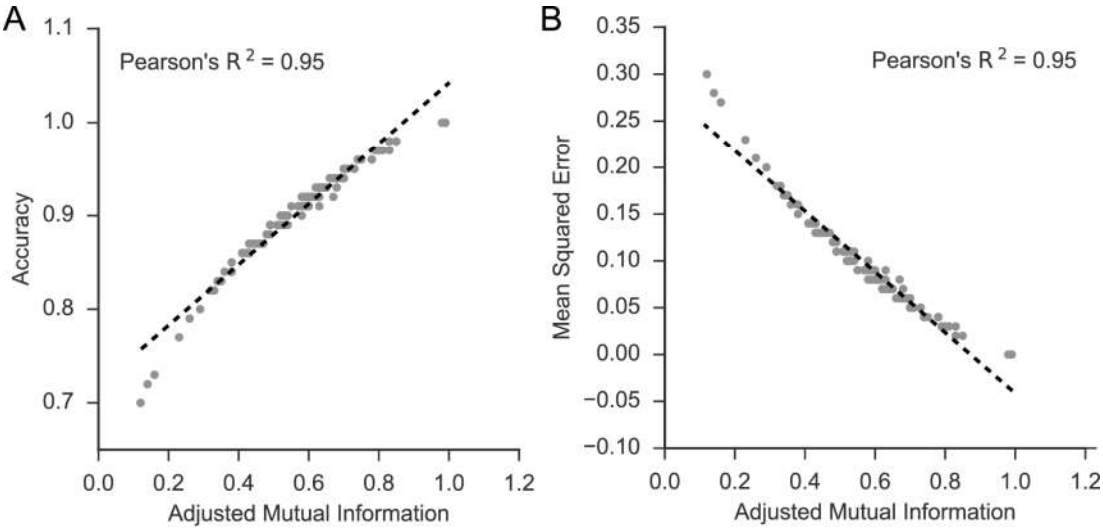

Figure 7: Classifier overall accuracy significantly improves for RBPs with strict set of characteristics defining their binding sites. A-B. Scatter plot showing the relationship observed between aMI scores (X-axis) and overall Accuracy (A) and MSE (B, Y-axis). In both cases, a high degree of correlation was identified by Pearson's R^2 (>= 0.95).

Overall, we identified 3 major classes of features that were important for determination of binding site selection for this group of RBPs: K-mer enrichment (motifs), existence of secondary RNA structure and overlap with other RBPs (Figure 8A, Additional File 7). All the 112 RBPs have at least one of these as an important feature for classification of their binding site with 10% or more relative importance. Furthermore, we also identified that the majority of RBPs (56.25%) have a combination of these three factors as important features for determination of binding site specificity. We identified that 109 (out of 112) RBPs showed some degree of overlap with at least 1 other RBP, which reflects the characteristic of RBPs working in protein complexes to perform biological functions [152,153]. We recovered information for known protein complexes, such as FMR1-FXR1-FXR2 [62,141] (Figure 8B), identifying that 68.32% of FXR1 binding sites overlap with FXR2 binding sites and 58.30% overlap with FMR1 binding sites. Interestingly the reciprocal did not hold true, with only 19.3% of FMR1 and 21.64% of FXR2 binding sites having overlap with FXR1 (See Additional File 4), which might reflect the molecular dynamics involved in the formation of the complex [62]. In addition, we also identified overlaps in the binding sites of RBPs without any previous association reported, such as the the case of AGGF1 which had 40.20% and 40.32% of overlap with TNRC6A and GTF2F1, respectively (See Additional File 4). While this information may indicate that they are only binding to the same targets in

similar positions, it could also suggest the existence of some biological relationship between these proteins which is yet to be uncovered.

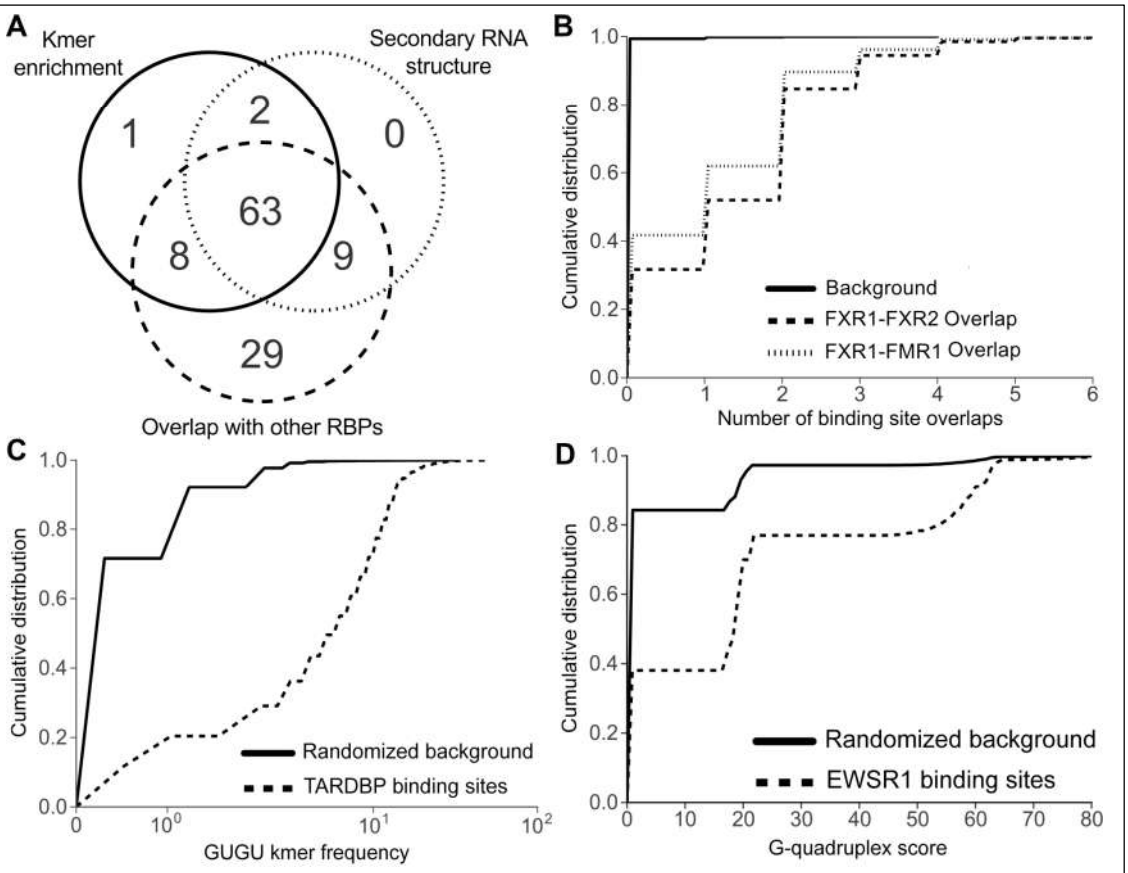

Figure 8: RNA-target selection by RNA-binding proteins is a multi-factorial biological process requiring cis- and trans-regulatory factors. (A) 3-way Venn diagram showing the overlap between RBPs identified with at least 1 K-mer enrichment (solid line), secondary RNA structure (dotted) and overlap of the binding site with other RBPs (dashed) as an important feature for characterization and group classification of their binding sites. **B-D.** Cumulative distribution function curves for FXR1 binding site overlaps with FMR1 and FXR2 (B), GUGU k-mer enrichment in TARDBP binding sites (C) and EWSR1 binding sites maximum G-quadruplex score (D). The Y-axis shows the cumulative distribution of samples and the X-axis indicates the values obtained for each feature. The solid lines represent randomized background regions while dashed lines represent RBPs binding sites.

Analysis of K-mer enrichment revealed that 74 RBPs had at least 1 K-mer with 10% or more relative importance (See Additional File 7), although that is an expressive number (66.07%) it also shows that the existence of a nucleotide sequence is not a requirement for directing the binding of an RBP to its target. We managed to recover several well-known examples from the literature, such as TARDBP's GUGU repeats [154] which are present in 88.31% of binding sites (Figure 8C). Other known examples include: QKI [155] (ACUAA in 56.72% and UAAC in 67.93% of binding sites), PUM2 [75] (UGUA in 72.82% of binding sites), PTBP1 [156] (UCUU, 80.20%), HNRNPC [157] (UUUU, 62.80%), HNRNPK [158] (CCCC, 87.01%), KHDRBS1 [159]

RBP	Motif	Binding sites (%)	Background (%)	Difference
FKBP4	GGGG	72.19	16.89	55.30
DDX42	GGGG	70.55	17.61	52.94
NKRF	GGGG	71.31	19.36	51.95
XRN2	GGGG	70.04	18.37	51.67
TRA2A	GAAGA	60.51	11.25	49.26
DDX59	CCCC	66.73	19.12	47.61
SRSF7	GUGUG	53.35	6.95	46.40
PCBP2	CCCU	72.58	26.28	46.30
GRSF1	GGGG	64.55	18.44	46.11
EIF4G2	GUGUG	52.05	6.73	45.32
SRSF9	GGAG	74.40	31.73	42.67
SERBP1	CGCC	54.26	11.74	42.52
KHSRP	UUGU	68.85	26.43	42.42
SLTM	GGGC	62.28	20.34	41.94
FUBP3	UUGU	66.28	24.52	41.76
AKAP8L	GGGG	59.15	18.42	40.73
GEMIN5	GCCG	47.84	8.72	39.12
FASTKD2	GGGG	54.44	16.66	37.78
DKC1	GUGUG	41.66	4.32	37.34
SRSF1	GGAG	69.00	31.84	37.16
TAF15	GAGG	63.55	28.59	34.96
DDX3X	GCGG	54.13	22.04	32.09
HNRNPM	UGUG	56.56	24.54	32.02
HNRNPA1	UUAG	49.01	17.26	31.75
SUPV3L1	GGGG	45.77	14.67	31.10
ZRANB2	GGUG	50.04	19.95	30.09
CPSF6	GAAGA	41.02	11.68	29.34
AARS	GGGG	44.32	15.54	28.78
RPS11	GCGG	31.60	3.16	28.44
U2AF2	UUUC	60.54	32.92	27.62
HNRNPU	GGGG	44.03	16.45	27.58
AGGF1	CACAC	32.67	5.73	26.94
DDX6	GGGG	41.61	16.99	24.62
HLTF	GAAA	50.00	25.64	24.36
SAFB2	GAAG	50.00	26.87	23.13
SFPQ	UGUG	50.94	28.25	22.69
CDC40	GGGG	40.11	17.55	22.56
SUGP2	UCUU	48.95	27.03	21.92
SUB1	UGUG	43.66	22.33	21.33
DGCR8	GGGG	31.91	10.76	21.15
XRCC6	CUGG	50.21	29.39	20.82
PRPF8	AGGU	46.67	26.34	20.33
SLBP	GAGC	32.06	11.85	20.21
LSM11	GCUG	43.18	23.89	19.29

Table1. Nucleotide motifs identified by BioFeatureFinder for 48 RNA-binding proteins. RBPs which had nucleotide sequences (motifs) identified as important features were analyzed for percentage of binding sites (BS) which had the identified motif in comparison with amount sampled from background (BG). The differences (Diff) in percentage percentage points (Diff = BS% – BG%) are also represented.

(UAAA, 80.08%) and TIA1 [160] (UUUU, 41.78%). Also, our algorithm identified motifs for other 47 RBPs, which were found in ~30% of binding sites and had at least 15% difference when compared to the background (Table 1, See Additional File 3).

Lastly, our algorithm also identified 74 RBPs which had secondary RNA structure (either by lower Minimum Free Energy, MFE, calculated by Vienna's RNAfold, or by a higher G-quadruplex score, from QGRS Mapper) as an important feature for classification (See Additional File 7). As an example, our algorithm identified an increased G-quadruplex score for EWSR1 binding sites, which is a known RBP that binds to these types of RNA structures [161], with 61.95% of the sites exhibiting a positive score while only 15.79% of background regions showed the same behavior (Figure 8D). Another RBP we identified as a binding to secondary RNA structure is XRN2, which had 86.12% of binding sites possessing an MFE score lower than 0, while only 51.09% of background regions had values lower than 0. This particular RBP has been shown to bind to R-loop structures formed by G-rich pause sites associated with transcription termination [162], which is also in accordance to our findings for motif enrichment, as we found that XRN2 had 70.04% of binding sites containing a GGGG 4-mer, while only 18.37% of background regions had the same 4-mer (Table 1). Other known examples from the literature we recovered include: FMR1, also known to bind to G-quadruplexes [64,163] and DDX3X, DDX6, DDX24 and DHX30, which are RNA-helicases. In addition, we also identified 34 RBPs which had positive G-quadruplexes and 15 percentage points or more of difference when comparing against the background. From those, 20 RBPs also exhibited an enrichment of GG repeats in their binding sites (Figure 9, Table 1, Additional File 5), which is a known characteristic for these structures [164], they are: AARS, AKAP8L, CDC40, DDX3X, DDX42, DDX6, DGCR8, FASTKD2, FKBP4, GRSF1, HNRNP, NKRF, SLTM, SRSF1, SRSF9, SUPV3L1, TAF15, XRCC6, XRN2, ZRANB2.

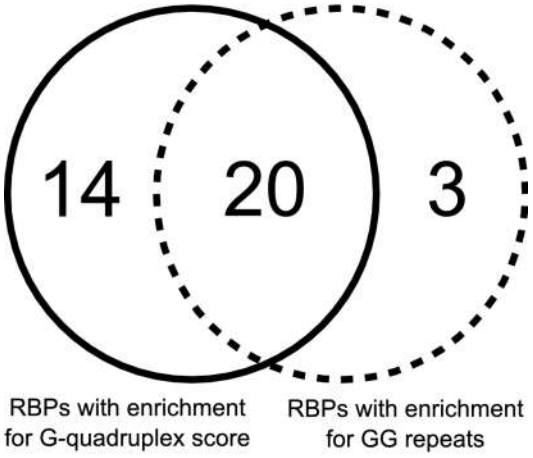

Figure 9: RBPs with enrichment for GG repeats in their motifs also have higher G-quadruplex scores. 2-way Venn diagram showing the overlap between the number of RBPs identified with the GG repeats in their enriched K-mers (dashed) and the RBPs with positive maximum G-quadruplex score (solid) as important features for group classification.

Conclusion

Our results show that BioFeatureFinder represents an accurate, flexible and reliable analysis platform for large-scale datasets, while at the same time providing a way to control observer bias and uncover latent relationships in biological datasets. By considering each genomic landmark as a separate data point in a distribution, we developed a novel implementation, combining both statistical analysis and big-data machine learning approaches, to provide accurate representations of differences in sets of genomic regions and identify which characteristics contribute more for separating these groups. As demonstrated by our analysis of the RBFOX2 dataset, our algorithm managed to recover multiple characteristics known from the literature, including nucleotide sequences for binding motifs and infer protein-protein interaction from overlaps between binding sites. In addition, we also uncovered new associations that might link RBFOX2 to targeting specific RNA secondary structures to increase RBFOX2 binding specificity, a hypothesis that is strengthened by inferences from the literature from multiple sources.

Furthermore, our analysis of 112 RNA-binding proteins CLIP-seq data from ENCODE also recovered several well-known features from the literature major characteristics that influence the targeting of these proteins. The results for this dataset indicate that RNA-target selection by RNA-binding proteins as a multi-factorial mechanism, demanding the existence of both cis- and trans-regulatory factors for increasing RBP affinity to target site. Amongst those features there are important factors as: existence of a particular set of nucleotide sequences (binding motif), accessibility of the target site via RNA secondary structure and also the neighboring RBP context (i.e. other proteins binding to neighboring/same region) contribute to determining if a particular RBP will bind to its target site. Additionally, our results suggest that the binding of RBPs to their targets is heavily dependent of the cellular context, with some proteins relying on fewer features for directing their binding specificity (i.e. the presence of a sequence is enough for recognition by the RBP), while other proteins require a more complex targeting context with multiple features involved in the binding of the RBP (i.e. requiring a specific sequence, accessory proteins nearby and a specific RNA structure). Taken together, our results not only deepen the knowledge of how these proteins select their targets in a broader scenario, but also

demonstrate how our approach can be applied to large-scale datasets from high-throughput experiments with a high degree of reproducibility.

Although the present study focused on the applications of BioFeatureFinder for RNA-binding proteins, our algorithm can be applied for any type of genomic landmark. Some examples of regions that could be analyzed when using BFF include: splicing sites for alternatively spliced exons (or whole exons), target sites for microRNAs, binding sites for DNA-binding proteins (for example, ChIP-seq data), promoter regions from differentially expressed genes, microsatellite/genomic markers, SNPs, whole transcript regions (5'UTR, 3'UTR and CDS) and any type of dataset that could be converted into a BED format.

Materials and methods

Extraction of information on biological features and RNA-binding proteins binding sites

We downloaded tracks from biological features associated with genomic features from UCSC Genome Browser [51] for human genome hg19, downloading tracks for conservation scores (phastCon scores in bigWig format), benign and pathological CNVs, common and flagged SNPs, TS microRNA target sites, CpG islands, layered H3K4Me1/3 and H3K27Ac and microsatellites. Additionally, we obtained data for 112 RNA-binding proteins (RBP) available from ENCODE [114] eCLIP experiments (Additional File 8: Table S1), downloading the bed files containing the narrowPeaks obtained for hg19 (Additional File 8: Table S2). Additionally, our algorithm is integrated with BedTools [117] (intersect, getfasta and nuc functions), UCSC's bigWigAverageOverBed [118], EMBOSS wordcount [119], Vienna's RNAfold [120] and QRGS Mapper [121]. Otherwise stated, all tracks were either downloaded or converted into BED format. We used GENCODE's [46] GRCh37.p13 as reference genomic sequence along with release 19 of the comprehensive annotation.

Preferential region identification and background selection

To identify preferential binding regions for each dataset analyzed, we separated the transcripts in 6 major regions: 5'UTR, 3'UTR, CDS, introns, 5' splice

sites and 3' splice sites. We then used bedtools intersect to count how many occurrences of RBP binding sites appeared in each of these regions. These values were normalized by Z-score and the highest scoring was selected as the preferential binding region. Randomized background was generated using bedtools shuffle (with -excl, -incl, -chrom and -noOverlapping options), using the GTF reference containing the preferential binding region (or regions) to guide the selection of regions, excluding overlaps with input regions (binding sites) and other randomized background regions. In cases of RBPs which had low difference in binding site z-scores (less than 10%) in their highest scoring regions, we selected the 2 highest scoring regions and used both as references for background generation. For each RBP dataset, we generate a randomized background regions 3 times the number of input regions (binding sites).

Assembly of a data matrix with biological features

The genomic regions (binding sites and background) and their associated biological features are converted in a numerical matrix, where each line is one region and each column is one of the biological features associated with that position. To convert biological features in numerical data, we used a combination of multiple freely available software. For most features we use BedTools intersect (-s and -c options) to count the number of occurrences of that feature in the corresponding region. For obtaining nucleotide sequence information we used a combination of BedTools getfasta and nucBed (both with -s option). For conservation score we used the tool bigWigAverageOverBed to obtain the average conservation score of covered bases in the region. For k-mer analysis, we used EMBOSS wordcount for counting the number of occurrences of each 4-mer, 5-mer and 6-mer in each region. For RNA structure analysis, we used both Vienna's RNAfold to calculate the lowest possible MFE (with -g option) and QGRS Mapper to calculate the maximum non-overlapping G-quadruplex score for each region. All operations were performed taking strandedness into consideration.

Group selection and statistical analysis of features

For the analysis process, input and background regions are separated in groups (1 and 0, respectively). This is done by using the unique identifier created during the datamatrix assembly and stored as the “name” field in the BED file generated. For all features in the matrix, we performed Kolmogorov-Smirnov comparing the cumulative distribution function of the input regions (group 1) with the background (group 0) to filter out the non-significant features between the groups. Features with a q-value ≤ 0.05, adjusted by Bonferroni, were selected for further analysis using the classification algorithm. This filtering aims to provide significantly different features between the two groups. Additionally, to minimize the noise introduced by non-significant features, while at the same time reducing the computational time required for the classification step and the overall required time for the analysis.

Classifier and feature importance estimation

To evaluate the importance of each feature’s ability in separating the genomic region groups we chose to use a stochastic gradient boost classifier python implementation from Scikit-learn [126]. The classifier was used with the following parameters: number of estimators = ‘1000’; learning rate = ‘0.01’; max depth = ‘8’; loss = ‘deviance’; max features = ‘sqrt’; minimum number of samples in each leaf = ‘0.001’; minimum number of samples to split = ‘0.01’; random state = ‘1’; subsample = ‘0.8’. Importance values for each feature are calculated at every run, with the final value representing the mean scores and their corresponding standard deviation. The same scoring strategy is employed for relative importance score (percentage in relation to the most important feature), accuracy, positive predictive value, negative predictive value, sensitivity, sensibility, ROC and Precision-Recall AUCs.

Availability

The BioFeatureFinder software is available for download at GitHub [165] and is also included as Additional file 9 for archival purposes.

Supplementary material

Additional File 1: Deviance, ROC and PR curves for RBFOX2

Additional File 2: Classifier Metrics and Results for all 112 RBPs

Additional File 3: K-mer enrichment scores and percentages

Additional File 4: Overlap percentages

Additional File 5: Structure percentages

Additional File 6: Heatmap with Z-score for binding region preference

Additional File 7: RBP classification for Kmer/Struct/Overlap and combinations

Additional File 8: List of accession numbers for RBPs from ENCODE

Additional File 9: BioFeatureFinder v1.0 algorithm

All files are available at:
https://drive.google.com/open?id=1lL6KC6BMwUN2xfP9XVZfgOf20-97jEAa

Annexes

COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br

UNICAMP

IB

DECLARAÇÃO

Em observância ao **§5º do Artigo 1º da Informação CCPG-UNICAMP/001/15**, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Dissertação de Mestrado, intitulada "***Identification of biological characteristics associated with RNAbinding proteins (RBPs) target sites***", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: _____
Nome do(a) aluno(a): Felipe Eduardo Ciamponi

Assinatura: _____
Nome do(a) orientador(a): Katlin Brauer Massirer

Dra. Katlin B. Massirer
Matrícula: 30083-1
CBMEG - UNICAMP

Data: 08 de Junho de 2018

Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Identification of biological characteristics associated with RNAbinding proteins (RBPs) target sites**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 08 de Junho de 2018

Assinatura : _____

Nome do(a) autor(a): **Felipe Eduardo Ciamponi**

RG n.° 503202836

Assinatura : _____

Nome do(a) orientador(a): **Katlin Brauer Massirer**

RG n.° 9006210398

Profa. Dra. Katlin B. Massirer
Matrícula: 30083-1
CBMEG - UNICAMP