Transcriptome - wide bioinformatics analysis of the binding sites of RNA - binding proteins and their putative role in mendelian diseases

Phan Nguyen Anh Thu[1], Matteo Floris[2], Maria Laura Idda[3], Nguyen Hoang Bach[4*]
(1) Department of Physiology, University of Medicine and Pharmacy, Hue University
(2) Department of Science Biomedicine, Sassari University
(3) National Research Council - Institute of Genetic and Biomedical Research (CNR-IRGB)
(4) Department of Microbiology, University of Medicine and Pharmacy, Hue University

Abstract

Background: Post-transcriptional regulation is the control of gene expression at the RNA level. After produced, the stability and distribution of the different transcripts are regulated by means of RNA-binding proteins (RBPs). Mutations in RNA-binding proteins can cause Mendelian diseases - prominently neuro-muscular disorders and cancers. This study determines the interaction between RBPs and target-RNA complexes from public data of the ENCODE project and identifies mutations associated with Mendelian diseases that could disrupt the RBP-RNA interactions. **Materials and methods:** we performed a transcriptome - wide bioinformatics prediction of the binding sites of RBPs in the human transcriptome from public data of the ENCODE project. **Results:** The majority (54%) of pathogenic mutation putatively affecting the binding sites of RBPs are located in protein - coding genes and are mainly classified as loss - of - function mutations. Mutations located in the binding sites of RBPs related to RNA processing. For 13 diseases, Familial hypercholesterolemia is the most significant disease with about 40% of mutations in ClinVar database located into the binding sites of RBPs (p=2.3e-65), but congenital hypogonadotropic hypogonadism is the disease with the highest percentage of mutations affecting the binding sites of RBPs (98%, *p*=2.7e-25). The RBPs most involved in human Mendelian diseases by binding sites-disrupting mutations are YBX3, AQR and PRPF8. **Conclusions:** A large number of Mendelian diseases are potentially mediated by disease - causing variants that potentially disrupt the binding sites of RBPs. This will provide insight sharper on post - transcriptional mechanisms. Besides, it is useful to know the role of protein - RNA interactome networks in pathologies, thereby serving the treatment of diseases.

Keywords: *bioinformatics analysis, ENCODE project, ClinVar, RNA-binding proteins, Mendelian diseases.*

1. INTRODUCTION

Post-transcriptional regulation, also known as the control of gene expression at the RNA level, occurs between the transcription and translation of the gene [1]. It makes a significant contribution to the control of gene expression in all human tissues [2,3]. After being produced, the stability and distribution of the different transcripts are regulated by means of RNA - binding proteins (RBPs). RBPs are widely and abundantly produced in cells. They participate and coordinate crucially in maintaining the integrity of the genome and play a crucial and conserved role in gene regulation. RBPs have a wide range of functions, including regulating polyadenylation, splicing, translation, editing, and post-transcriptional regulation of mRNA stability, which ultimately affects the expression of every gene in the cell [4]. RBPs also contain regulatory regions that post-transcriptionally affect gene expression [5].

The role and process by which these proteins control gene expression is of great interest, and there is evidence of their involvement in a wide range of illnesses. Recent research has identified human cell in vivo mRNA interactions that are linked to more than 1.100 RBPs. Most RNAs interact with all proteins, and many proteins interact with several RNAs [6]. RNA - protein networks, which control gene expression at the RNA level, are formed as a result of the combinations of individual RNA - protein interactions [7]. Defects or deregulation of RNA - protein networks often cause disease. Cancers and Mendelian diseases, particularly neuro - muscular disorders can be brought on by mutations in RBPs[8–10]. In this work, we first determined the interaction between the RBPs and target-RNA complexes from public data of the ENCODE project (Encode Project Consortium, 2004) [11]. In particular, we identified disease mutations associated with Mendelian diseases that could disrupt the RBP-RNA interactions.

Corresponding author: Nguyen Hoang Bach, email: nhbach@huemed-univ.edu.vn
Recieved: 22/10/2022; Accepted: 28/11/2022; Published: 30/12/2022
DOI: 10.34071/jmp.2022.7.11

2. MATERIALS AND METHODS

Construction of RBP - RNA regulatory network as well as relationship between RBP mutations and Mendelian diseases

To identify RBP-RNA interactions, the full list of eCLIP binding assay was retrieved from the Encode website (https://www.encodeproject.org/eclip/) [12, 13]. The standard eCLIP pipeline has been described at the ENCODE project (https://www.encodeproject.org/pipelines/ENCPL357ADL/). In total, 225 eCLIP - seq datasets for 103 diverse RBPs in HepG2 cells, 120 in K562 cells and 2 in adrenal gland cells were collected. The final bam files were then processed with the PureClip pipeline with basic mode settings [14].

To identify RBPs mutated in genetic disease, we crossed our RBPs with Mendelian diseases association data from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/). A public list of mutations involved in Mendelian diseases has been compiled from the ClinVar FTP repository (ClinVar version 13/01/2020). Only disease variants classified as "Pathogenic" and/or "Likely_pathogenic" were retained for this analysis. The Human Genome reference built here used in the context of this analysis is GRCh38.

Statistical analysis and network visualization

All statistical analyses were performed by R language. Enrichment analysis used to identify biological themes among genes that mutated the binding sites of RBPs has been performed with the R package ReactomePA [15]. A hypergeometric model has been used to assess whether the number of selected genes associated with Reactome pathways is larger than expected. The *p* values were calculated based on the hypergeometric model. A Fisher exact test statistic has been used to calculate the significance. To control the familywise error rate, we applied here the Bonferroni correction method [16].

3. RESULTS

The interaction between the RBPs and target - RNA complexes from public data of the ENCODE project

A total of *496,672 binding sites* were predicted by the PureClip pipeline. Only binding sites with PureClip score within the 4th quartile of the score distribution was retained for further analysis.

For RBP tested in more than 1 cell line, all the binding sites were merged into 1 single file. Individual crosslink sites with a distance below 8 bp were then merged into binding sites and given out in a separate BED6 file, available on demand.

The positions of the predicted the binding sites of RBPs (extended by 5 nucleotides in both directions) were then intersected with the position 80,902 ClinVar entries (release 13/01/2020, considering only variants classified as pathogenic, likely pathogenic, risk factor or affects). A total of 13,127 intersections were obtained, with 7,688 unique variants associated with 2,383 disorders in 6,100 unique binding sites. The majority (54%) of pathogenic mutation putatively affecting the binding sites of RBPs are located in Protein coding genes and are mainly classified as loss of function mutations (missense, frameshift, stop gain and splice – site variants) (Figures 1A and 1B).

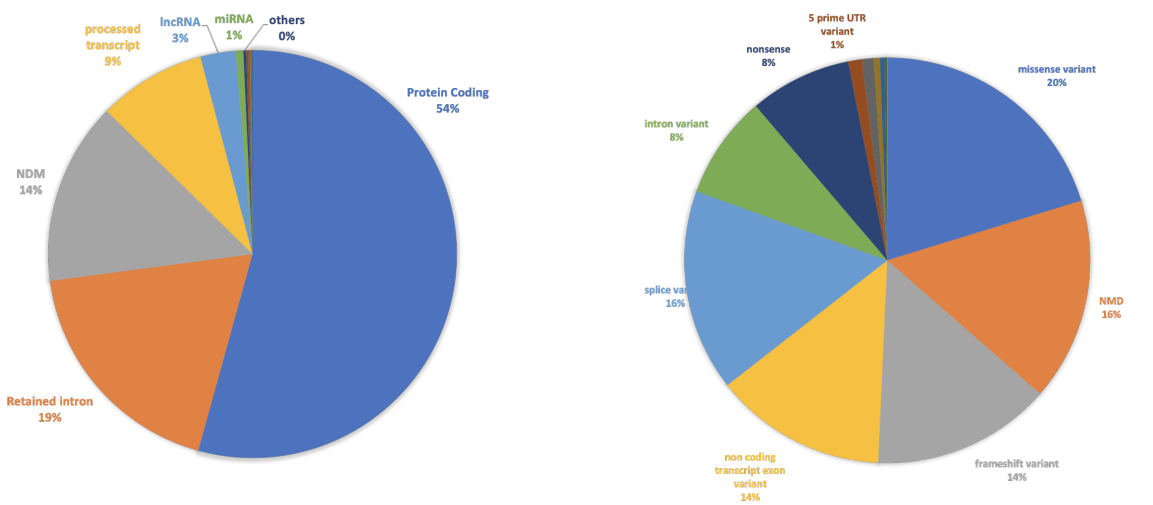

Figure 1. Functional consequences of mutation in functional classes of genes with the binding sites of RBPs. A. Functional consequences of mutations of the binding sites of RBPs.
B. Functional classes of genes with mutated the binding sites of RBPsing sites of RBPs

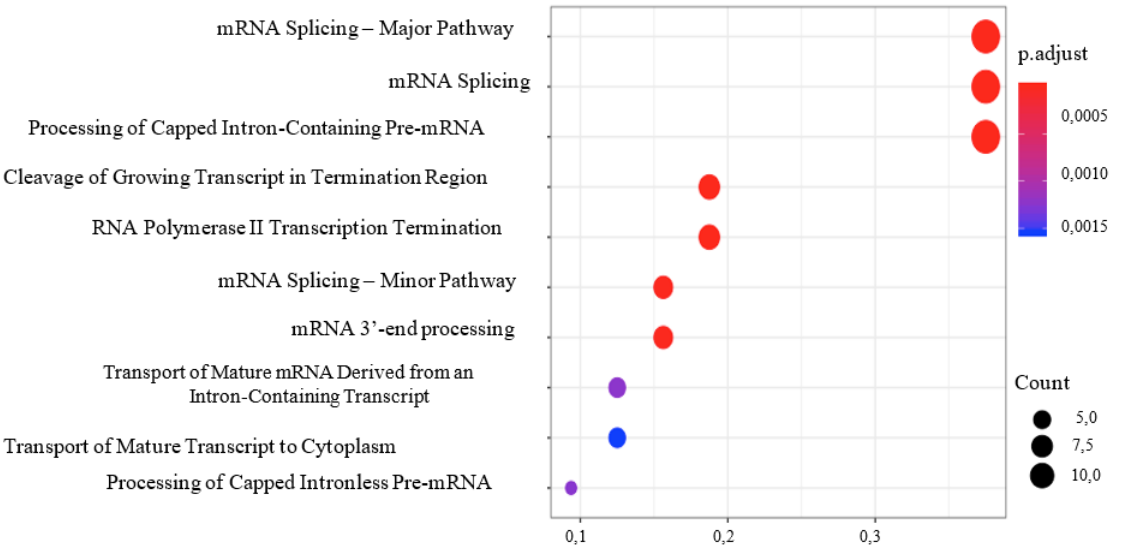

Figure 2. Plot with Enrichment analysis.

Enrichment analysis used to identify biological themes among genes that mutated the binding sites of RBPs (Figure. 2) reveal that most significantly represented Reactome pathways are those related to RNA processing, in particular maturation through splicing, capping and 3' end processing.

Disease mutations associated with Mendelian diseases that could disrupt the RBP - RNA interactions

For 13 diseases, there is a significant portion of disease - causing mutations that putatively disrupt. The binding sites of RBPs: familial hypercholesterolemia is the most significant disease, with about 40% of mutations in ClinVar database located into the binding sites of RBPs (p=2.3e-65), but Congenital hypogonadotropic hypogonadism is the disease with the highest percentage of mutations affecting the binding sites of RBPs (98%, p=2.7e-25) (Table 1, 2 and 3).

Table 1. The percentage of mutations affect the binding sites of RBPs: modified with *p* value calculation.

Disease	Mutations in binding sites	Total mutations	% of mutations in binding sites	*p* value
FH	579	1473	39.31	2.31843E-65
CHH	56	57	98.25	2.72081E-25
Hereditary cancer-predisposing syndrome	196	2127	9.21	4.63581E-18
HBOC	98	1198	8.18	1.88067E-13
ATS1	48	703	6.83	7.87654E-11
PKU	76	212	35.85	3.31945E-08
Inborn genetic diseases	91	854	10.66	7.01478E-06
VLCAD	32	80	40	4.22778E-05
VHL	38	109	34.86	0.000168184
PH1	53	171	30.99	0.000197395
FANCA	40	123	32.52	0.000480888
CDLS1	27	256	10.55	0.012383962

NPC1	29	107	27.1	0.032895705
NF1	121	808	14.98	0.122885423
HNPCC	126	807	15.61	0.257038843
Wilson disease	30	203	14.78	0.402741815
RSTS	38	190	20	0.433439617
NKH	28	180	15.56	0.580451727
KABUK1	34	186	18.28	0.792176888
PXE	51	292	17.47	0.981090427

Table 2. The percentage of mutations affect the binding sites of RBPs. Diseases with $p < 0.05$ sorted by percentage of mutations value in the binding sites

Disease	Mutations in binding sites	Total mutations	Mutations in binding sites (%)	*p* value
CHH	56	57	98.25	2.72081E-25
VLCAD	32	80	40	4.22778E-05
FH	579	1473	39.31	2.31843E-65
PKU	76	212	35.85	3.31945E-08
VHL	38	109	34.86	0.000168184
FANCA	40	123	32.52	0.000480888
PH1	53	171	30.99	0.000197395
NPC1	29	107	27.1	0.032895705
Inborn genetic diseases	91	854	10.66	7.01478E-06
CDLS1	27	256	10.55	0.012383962
Hereditary cancer-predisposing syndrome	196	2127	9.21	4.63581E-18
HBOC	98	1198	8.18	1.88067E-13
ATS1	48	703	6.83	7.87654E-11

Table 3. Diseases-causing mutations in the binding sites of RBPs.

Disease	RBPs
CHH	AATF, AGGF1, AKAP1, AQR, BCLAF1, CSTF2T, CSTF2, DROSHA, EFTUD2, EIF3D, FAM120A, FASTKD2, FXR2, G3BP1, GEMIN5, GRWD1, HLTF, HNRNPL, HNRNPM, IGF2BP1, IGF2BP2, IGF2BP3, KHSRP, LARP7, LSM11, NONO, PABPN1, PCBP2, PRPF4, PRPF8, RBFOX2, RBM15, RPS3, SDAD1, SND1, SRSF1, SSB, SUGP2, U2AF1, U2AF2, UCHL5, YBX3, ZNF622, ZNF800, ZRANB2
VLCAD	AQR, BCCIP, BCLAF1, BUD13, DGCR8, EIF3H, FMR1, G3BP1, GRWD1, LIN28B, PPIG, PRPF4, PRPF8, RBM15, SF3B4, SRSF1, SRSF7, SRSF9, U2AF1, U2AF2, UCHL5, YBX3

FH	AQR, BCLAF1, BUD13, CPEB4, DDX6, FXR2, G3BP1, GPKOW, GRWD1, HLTF, HNRNPA1, IGF2BP1, IGF2BP2, IGF2BP3, LIN28B, LSM11, NKRF, PPIG, PRPF8, RBM15, SF3B4, SND1, SUB1, SUPV3L1, U2AF2, UCHL5, XRN2, YBX3, ZC3H11A, ZNF622, ZNF800
PKU	AQR, G3BP1, GRWD1, LIN28B, HLTF, NCBP2, PPIG, PRPF8, SRSF1, U2AF2, UCHL5
VHL	AQR, GRWD1, PRPF8, YBX3
FANCA	AQR, BCLAF1, DDX55, KHSRP, LSM11, PPIG, PRPF4, PRPF8, RBM15, SSB, UCHL5, YBX3, ZNF622
PH1	AQR, BCLAF1, LSM11, GRWD1, PPIG, PRPF4, PRPF8, UCHL5, ZNF800
NCP1	AQR, BUD13, GRWD1, LIN28B, LSM11, PRPF8, RBM15, SND1, U2AF2, UCHL5, YBX3
Inborn genetic diseases	ABCF1, AKAP1, APOBEC3C, AQR, BCLAF1, BUD13, CPEB4, DDX3X, DDX55, DKC1, EIF3H, EIF4G2, FMR1, FXR1, FXR2, GRWD1, HLTF, HNRNPU, IGF2BP1, IGF2BP2, IGF2BP3, KHSRP, LARP4, LIN28B, LSM11, METAP2, PPIG, PRPF4, PRPF8, RBM15, SF3B4, SLTM, SND1, SRSF1, SRSF7, SRSF9, SUB1, TIA1, U2AF1, U2AF2, UCHL5, UTP3, YBX3, ZC3H11A, ZNF622
CDLS1	AQR, BCLAF1, FXR2, IGF2BP1, IGF2BP2, U2AF2, UCHL5, YBX3, ZNF622
Hereditary cancer-predisposing syndrome	AQR, BCLAF1, BUD13, DDX3X, EIF3H, FXR1, FXR2, GPKOW, GRWD1, HLTF, HNRNPM, HNRNPU, IGF2BP2, IGF2BP3, KHSRP, LIN28B, PPIG, PRPF8, RBM15, RBM5, SF3B4, SND1, SRSF1, SSB, SUB1, TIA1, U2AF1, U2AF2, UCHL5, UTP3, XPO5, XRN2, YBX3, ZC3H11A, ZC3H8
HBOC	AQR, GRWD1, HLTF, LIN28B, PRPF8, YBX3, ZC3H11A, ZC3H8
ATS1	PPIG, PRPF4, PRPF8, SND1, U2AF1, U2AF2, YBX3

The RBP with the highest percentage of binding sites with disease causing mutations are PABPN1 (poly(A) binding protein nuclear 1, a member of a larger family of poly(A)-binding proteins in the human genome) and SND1 (staphylococcal nuclease and tudor domain containing 1, a main component of RISC complex with an important role in miRNA function) and SRSF1 (Serine and Arginine Rich Splicing Factor 1, an essential sequence specific splicing factor involved in pre-mRNA splicing.) (Table 4, 5).

Table 4. Relationship between RBPs, number of mutations in binding sites and mutated binding sites. RBPs with the highest number of mutated binding sites

RBP	Total mutations in binding sites	Total binding sites	% of sites with mutations
YBX3	2458	36449	6.74
AQR	2394	40011	5.98
PRPF8	1049	16111	6.51
GRWD1	830	11969	6.93
RBM15	828	17643	4.69
SND1	698	4743	14.72
LIN28B	615	9658	6.37
UCHL5	547	10282	5.32
U2AF2	362	11477	3.15
BCLAF1	233	2811	8.29
IGF2BP1	217	3139	6.91

IGF2BP2	210	3277	6.41
PPIG	203	2693	7.54
IGF2BP3	193	3480	5.55
SRSF1	174	1492	11.66
U2AF1	172	3881	4.43
BUD13	171	5315	3.22
SF3B4	147	6178	2.38
FXR2	146	2167	6.74
LSM11	123	2825	4.35

Table 5. Relationship between RBPs, number of mutations in binding sites and mutated binding sites. RBPs with the largest percentage of mutations in the binding sites.

RBP	Total mutations in binding sites	Total binding sites	% of sites with mutations
PABPN1	76	284	26.76
SND1	698	4743	14.72
SRSF1	174	1492	11.66
DDX51	15	154	9.74
SSB	103	1099	9.37
BCLAF1	233	2811	8.29
NOL12	17	209	8.13
G3BP1	95	1190	7.98
ABCF1	11	141	7.8
PPIG	203	2693	7.54
XRN2	73	994	7.34
GRWD1	830	11969	6.93
IGF2BP1	217	3139	6.91
EIF3D	93	1369	6.79
YBX3	2458	36449	6.74
FXR2	146	2167	6.74
PRPF8	1049	16111	6.51
HNRNPUL1	14	216	6.48
SUB1	72	1119	6.43
IGF2BP2	210	3277	6.41

Overall, the RBP most involved in human Mendelian diseases by binding sites-disrupting mutations are YBX3 (Y-Box Binding Protein 3, a RNA-binding protein that regulates distinct sets of mRNAs by discrete mechanisms, including mRNA abundance), AQR (Aquarius Intron-Binding Spliceosomal Factor, a component of the spliceosome) and PRPF8 (pre-mRNA Processing Factor 8, another component of mammalian spliceosome) (Table 4, 5).

4. DISCUSSION

The interaction between the RBP and target - RNA complexes from public data of the ENCODE project

The research described above is an important step in clarifying the functions of RBPs in post-transcriptional gene regulation. The ability of eCLIP-seq to reveal the recognition code of RBPs and their binding locations is perhaps what makes these experiments so significant. Understanding how RNA - binding proteins positively or negatively influence post - transcriptional processes like alternative splicing requires a thorough analysis of protein - RNA interactions. Integration of binding - site data with functional genomic techniques has the potential to show the overall structure of post - transcriptional regulation networks in mammalian cells as future eCLIP - seq investigations expand the database of known protein - RNA interactions [17,18].

Disease mutations associated with Mendelian diseases that could disrupt the RBP - RNA interactions.

We found that genes with binding site mutations were likely to be bound by more RBPs. In total, the 13 Mendelian diseases are linked to disease-causing mutations that putatively disrupt the binding sites of RBPs, with a spectrum of pathologies including neuropathies, muscular atrophies, sensorial disorders, and cancer [19]. Similar symptoms and disorders could be caused by any protein in this protein - RNA interactome network malfunctioning. Additionally, we noticed that the mutations found were anticipated to show the percentage of mutation value in the binding region. These findings suggested that genetic mutations in the binding sites of RBPs play important functions. Our functional enrichment analysis revealed that the mutant target genes were considerably enriched in biological pathways, which allowed us to further study the role of the altered target genes. When considered as a whole, our study highlights the crucial functions that mutations in the binding sites of RBPs play. It is now necessary to assess the effects of mutation-mediated perturbations in the context of protein - RNA interactome networks.

5. CONCLUSION

Bioinformatics analysis performed in our study aim to perform the characterization of the binding sites of these RNA - binding proteins in the human transcriptome and to assess the putative role of RNA - binding protein in Mendelian diseases; our results suggest that a large number of Mendelian diseases are potentially mediated by disease - causing variants that potentially disrupt the binding sites of RBPs. This will provide insight sharper on post - transcriptional mechanisms. Besides, understanding the normal functions of RBPs throughout times of physiological change, such as during development, can reveal significant elements of function that are directly related to the pathogenic mechanisms and effects of disease. Thereby serving the treatment of diseases.

REFERENCES

[1] Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 2008;582:1977–86. https://doi.org/10.1016/j.febslet.2008.03.004.

[2] Franks A, Airoldi E, Slavov N. Post-transcriptional regulation across human tissues. PLoS Comput Biol 2017;13:e1005535–e1005535. https://doi.org/10.1371/journal.pcbi.1005535.

[3] Zhao BS, Roundtree IA, He C. Post-transcriptional gene regulation by mRNA modifications. Nat Rev Mol Cell Biol 2017;18:31–42. https://doi.org/10.1038/nrm.2016.132.

[4] Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet 2014;15:829–45. https://doi.org/10.1038/nrg3813.

[5] Brinegar AE, Cooper TA. Roles for RNA-binding proteins in development and disease. Brain Res 2016;1647:1–8. https://doi.org/10.1016/j.brainres.2016.02.050.

[6] Jankowsky E, Harris ME. Specificity and nonspecificity in RNA-protein interactions. Nat Rev Mol Cell Biol 2015;16:533–44. https://doi.org/10.1038/nrm4032.

[7] Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 2010;11:75–87. https://doi.org/10.1038/nrg2673.

[8] Castello A, Fischer B, Hentze MW, Preiss T. RNA-binding proteins in Mendelian disease. Trends in Genetics 2013;29:318–27. https://doi.org/10.1016/j.tig.2013.01.004.

[9] Nussbacher JK, Batra R, Lagier-Tourenne C, Yeo GW. RNA-binding proteins in neurodegeneration: Seq and you shall receive. Trends Neurosci 2015;38:226–36. https://doi.org/10.1016/j.tins.2015.02.003.

[10] Kapeli K, Yeo GW. Genome-wide approaches to

dissect the roles of RNA binding proteins in translational control: implications for neurological diseases. Front Neurosci 2012;6:144. https://doi.org/10.3389/fnins.2012.00144.

[11] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science (1979) 2004;306:636–40. https://doi.org/10.1126/science.1105136.

[12] van Nostrand EL, Kim SK. Integrative analysis of *C. elegans* modENCODE ChIP-seq data sets to infer gene regulatory interactions. Genome Res 2013;23:941–53. https://doi.org/10.1101/gr.152876.112.

[13] van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, et al. A large-scale binding and functional map of human RNA-binding proteins. Nature 2020;583:711–9. https://doi.org/10.1038/s41586-020-2077-3.

[14] Krakau S, Richard H, Marsico A. PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. Genome Biol 2017;18:240. https://doi.org/10.1186/s13059-017-1364-2.

[15] Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst 2016;12:477–9. https://doi.org/10.1039/c5mb00663e.

[16] Haynes W. Bonferroni Correction. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H, editors. Encyclopedia of Systems Biology, New York, NY: Springer New York; 2013, p. 154. https://doi.org/10.1007/978-1-4419-9863-7_1213.

[17] Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, et al. An RNA map predicting Nova-dependent splicing regulation. Nature 2006;444:580–6. https://doi.org/10.1038/nature05304.

[18] Wang Z, Burge CB. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. RNA 2008;14:802–13. https://doi.org/10.1261/rna.876308.

[19] Lukong KE, Chang K, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. Trends in Genetics 2008;24:416–25. https://doi.org/10.1016/j.tig.2008.05.004.