Biochemical and Biophysical Research Communications 525 (2020) 53–56

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

SMAtool reveals sequences and structural principles of protein-RNA interaction

Pengcheng Du [1], Pengfei Cai [1], Beibei Huang, Chen Jiang, Quan Wu***, Bin Li*, Kun Qu**

The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, 230001, China

ARTICLE INFO

Article history:
Received 19 January 2020
Accepted 9 February 2020
Available online 16 February 2020

Keywords:
SMAtool
Structure-sequencing
RNA secondary structure preference
RBP regulation
Pipeline

ABSTRACT

Protein binding events on RNA are highly related to RNA secondary structure, which affects post-transcriptional regulation and translation. However, it remains challenging to describe the association between RNA secondary structure and protein binding events. Here, we present Structure Motif Analysis tool (SMAtool), a pipeline that integrates RNA secondary structure and protein binding site information to profile the binding structure preference of each protein. As an example of applying SMAtool, we extracted the RNA-structure and binding site information respectively from the DMS-seq and eCLIP-seq data of the K562 cell-line, and used SMAtool to analyze the structure motif of each RNA binding protein (RBP). This new approach provided results consistent with X-ray crystallography data from the protein data bank (PDB) database, demonstrating that it can help researchers investigate the structure preference of RBP, and understand the role of RNA secondary structure in gene expression. Availability and implementation: https://github.com/QuKunLab/SMAtool.

1. Introduction

The human transcriptome encodes a large amount of RNA binding proteins (RBPs) and each of them interact with RNAs with different affinities [1]. RBPs are essential in the post-transcriptional regulation of genes, affect signaling pathways, and function directly in disease formation and cell fate [2,3]. Therefore, the interaction between RBP and RNA has long been of great interest in a wide range of fields. In recent years, a variety of high-throughput sequencing methods have been developed for RNA binding domain (RBD), and a large amount of data has been generated. For example, cross-linking immunoprecipitation (CLIP) based on UV-crosslinking and immunoprecipitation is very effective for detecting the binding of proteins and RNA. Multiple versions of the CLIP method have been derived, such as single-nucleotide resolution individual-nucleotide resolution CLIP (iCLIP) [4], and the more powerful enhanced CLIP (eCLIP) [5] and photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) [6], which have realized the precise recognition of each protein's binding site on the target RNA molecule at single base resolution.

Previous studies have shown that in addition to the sequence of the binding site, the secondary structure of RNA also plays an important role in RBP binding. Taliaferro et al. combined MBNL1 and RBFOX2 binding motifs in intron flanking exons and found that sequence motifs with a high degree of structured RNA may inhibit protein binding [7]. This result shows that the interdependence between RNA structure and sequence is very important for protein-RNA interaction. At present, a variety of high-throughput sequencing technologies have been developed to detect RNA secondary structures. Single-base-resolution RNA secondary structure detection methods such as PARS, dimethyl sulfate sequencing (DMS-seq), and Structure-seq [8–10] have been applied in multiple species. PARS uses RNAse V1 and S1 to specifically detect the double-stranded and single-stranded sites of each RNA, and achieves efficient genome-wide structure detection based on deep sequencing. DMS can methylate unpaired adenine and cytosine residues, and is used in methods such as DMS-seq and Structure-seq to detect the structure of *in vivo* RNA.

Based on the methods described above, the volume of RNA structure sequencing data and RBP binding sequencing data have greatly increased, and multiple related databases are available, such as Encode, structure surfer, RNA Mapping DataBase (RMDB) and database of RNA interactions in post-transcriptional regulation

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: powerwu@mail.ustc.edu.cn (Q. Wu), libin03@ustc.edu.cn (B. Li), qukun@ustc.edu.cn (K. Qu).

[1] These authors contributed equally.

https://doi.org/10.1016/j.bbrc.2020.02.068
0006-291X/© 2020 Elsevier Inc. All rights reserved.

(doRiNA) [11–14]. The availability of these databases highlights the need for tools that integrate high-throughput data to study the interactions between RBP and RNA. Therefore, we have developed SMAtool, which analyzes the sequence and structural motifs enriched in the RBD to show the preference of RBP for the secondary structure of RNA.

2. Materials and methods

2.1. RBP binding sites identification

SMATool accepts aligned eCLIP data (*.bam files) as input files for the analysis of binding sites information. The pipeline calculates cumulative RT-count (reverse transcription stop count) by locating the last base of each read on the reference, and normalizes counts by the total count of each protein. Replicates of each protein are automatically merged in this program to minimize the bias of experimental deviation. Since eCLIP signals are usually discrete, the normalized counts per base are smoothed by a 5 bp width window with a 2 bp step size, and SMATool calls peaks in each region with a remarkably smoothed signal per window. Then the fold change of raw signal over mock data and enrichment P-value are calculated to filter significant peaks for each protein on its binding RNAs. All filtered peaks are summarized by their location and normalized summit counts as binding sites information and the data are prepared for the subsequent procedures.

2.2. RNA structure detection

RNA Structure Framework [13,15] is highly recommended for processing secondary structure information, and is compatible with experimental raw data from PARS, DMS-seq and Structure-seq. Per base structure annotation is achieved through forgi [16], a Python library used to analyze the tertiary structure of RNA secondary structure elements. In this paper, *in vivo* DMS-seq data for the K562 cell line were adopted to distinguish specific structural traits for RNAs in different cell types. The command lines and details about alignment, read counting, normalization and structure reconstruction are described in detail in the SMAtool manual (https://github.com/QuKunLab/SMAtool). The structure information for all probed transcripts (*.db files) is required to identify the secondary structure around each binding site (45 bp range around each site). By combining binding events information from eCLIP analysis and corresponding structure analysis by PARS and DMS-seq, a table file is generated to depict the sequence and structure information of the region around all binding sites of each protein.

2.3. Motif searching

To reveal the structure bias for RBPs, SMATool employs MEME Suit to search a 25 bp structure motif on the neighboring regions of all binding sites [17]. Some proteins have more than 10,000 binding sites, requiring an excessive amount of computer time for the motif searching process. Therefore, for these proteins with too many binding sites, we set cut-off value so that only the top 4000 most eCLIP-signal significant binding sites were chosen for *de novo* structure-motif searching, then we applied FIMO to scan all binding sequences to find all sites that highly matched (P-value less than 10^{-10}) the searched motif. In logo-profile of each structure motif, we used symbols “s”, “h”, “m”, and “i” (i.e. alphabet) to label per base stem, hairpin, multi-loop and interior structures, respectively. After determining structure motifs for a protein, SMATool employs MEME again to search for a 5–6 bp sequence motif on all relevant binding sites of each structure motif. All motif information is reported in html format, and precise values of per base alphabet (structure or sequence) ratio are listed in a table (txt file) for each protein/motif.

3. Results

3.1. Integration of RBP binding sites and RNA secondary structure

SMAtool combines RBP binding sites and per base structure information, and systematically reveals the protein-RNA affiliations and their relevant secondary structure (Fig. 1). Among all binding sites of a structure motif, the sequence motif can be searched to figure out whether the nucleotide base bias and structure bias are related for a corresponding protein. Finally, a figure is displayed to show the RNA secondary structure preference of an RBP on the corresponding structural motifs.

3.2. K562 cell-line analysis with SMAtool

As an example of the utility of SMAtool, a test run was performed on publicly available datasets of the K562 cell-line. We processed eCLIP data for RBPs and DMS data in the K562 cell line through the SMAtool pipeline and obtained per base structure information of structure motifs, then performed 25 bp distributions of four different structure/loop types (stem, hairpin, interior and multi-loop) by clustering motifs with their significant on-site structures. In the K562 cell line, 18,790 transcripts were probed by DMS-seq. We selected RBPs with X-ray results of RNA binding sites in the RSCB PDB, a member of Worldwide PDB(wwPDB), as targets. The distributions of sequence motif sites based on structure motifs for these RBP were calculated (Fig. 2A). We performed a sequence motif search on the most significant structural motifs bound by each RBP, and selected sequence motifs that matched the sequence of the RNA binding site provided by the PDB database, and these sequence motifs were significantly enriched (sFig. 1). X-ray crystallography from the RSCB PDB database, showing the structure of these proteins and the 3D geometry of their binding domains (Fig. 2B), confirmed that sequence-motifs and structure bias identified in our results are consistent with the actual RBP binding sites. These several RBPs all show an obvious preference for RNA secondary structure. Among them, SRSF1, PUM2, and GEMIN5 are enriched at the multi-loop structure, while TROVE2 tends to bind to the RNA of the stem structure. These results validate that SMAtool is capable of revealing RBP’s preference for RNA secondary structures by integrating high-throughput data.

4. Discussion

By using SMATool combined with RBP binding sites and experimental data of RNA structure, we provide a convenient and reliable direction for secondary structure substitution analysis. SMATool facilitates studies of the interdependence between sequence motifs and structural motifs of RBP. Recently, several improved high-throughput sequencing methods for RNA structure sequencing have been proposed.

For example, nextPARS enables higher throughput and sample multiplexing while achieving comparable accuracy to previous implementations [18]. Structure-seq2 increases sensitivity by at least 4-fold and improves data quality compared to previous versions [19]. This will make it easier for us to obtain reliable genome-wide RNA secondary structure information. The analysis of RNA secondary structure preference for RBPs should take advantage of these techniques to obtain more reliable analysis results instead of relying solely on prediction based on the binding sites sequence.

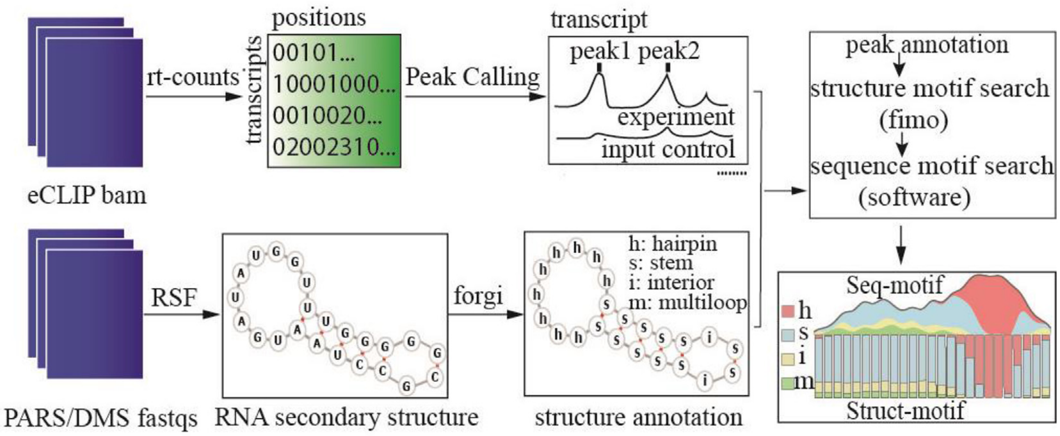

Fig. 1. Schematic representation of SMAtool with steps and workflow.

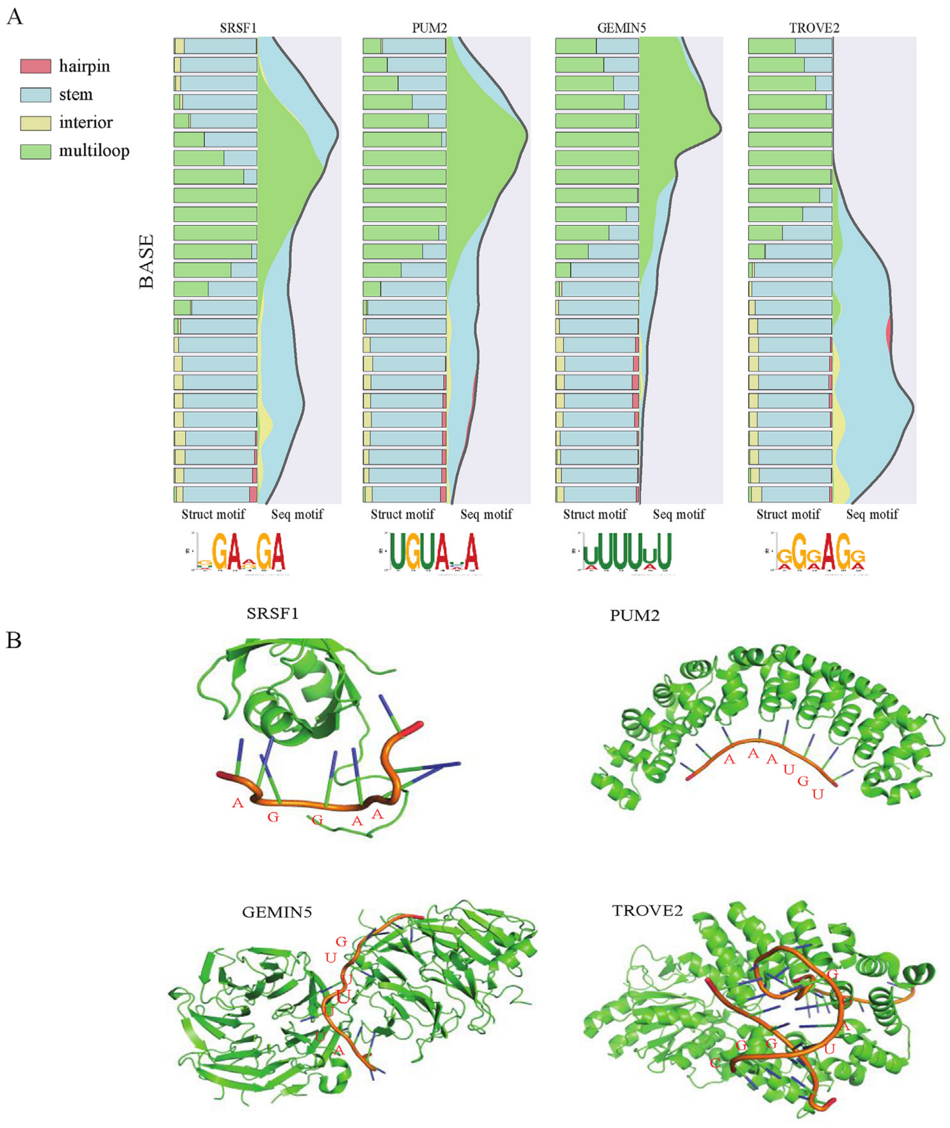

Fig. 2. Structure motifs of proteins in the K562 cell line and distribution of sequence motifs within structure motifs. (A): Structure motifs and representative sequence motifs for SRSF1, PUM2, GEMIN5, and TROVE2 in the K562 cell line. The left bar-plot profiles the per-base loop probability for its structure motif, and the right filled curve shows the distribution and loop-proportion of its sequence motif. (B): Protein 3D geometry structures and RNA-binding domains probed by X-ray crystallography experiments, from the RSCB PDB database, for SRSF1, PUM2, GEMIN5, and TROVE2.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key R&D Program of China 2017YFA0102903 (to K.Q.), the National Natural Science Foundation of China grant (91940306, 91640113, 31771428, 31970858 to K. Q., 81572255 to Q. W.), the Fundamental Research Funds for the Central Universities (to K.Q.). It was also supported by Anhui Provincial Natural Science Foundation grant BJ2070000097 (to B.L.).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbrc.2020.02.068.

References

[1] V. Anantharaman, E.V. Koonin, L. Aravind, Comparative genomics and evolution of proteins involved in RNA metabolism, Nucleic Acids Res. 30 (2002) 1427–1464, https://doi.org/10.1093/nar/30.7.1427.

[2] K.E. Lukong, K.W. Chang, E.W. Khandjian, S. Richard, RNA-binding proteins in human genetic disease, Trends Genet. 24 (2008) 416–425, https://doi.org/10.1016/j.tig.2008.05.004.

[3] J. Ye, R. Blelloch, Regulation of pluripotency by RNA binding proteins, Cell Stem Cell 15 (2014) 271–280, https://doi.org/10.1016/j.stem.2014.08.010.

[4] J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D.J. Turner, N.M. Luscombe, J. Ule, iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution, Nat. Struct. Mol. Biol. 17 (2010), https://doi.org/10.1038/nsmb.1838. 909-U166.

[5] E.L. Van Nostrand, G.A. Pratt, A.A. Shishkin, C. Gelboin-Burkhart, M.Y. Fang, B. Sundararaman, S.M. Blue, T.B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, G.W. Yeo, Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP), Nat. Methods 13 (2016) 508–514, https://doi.org/10.1038/nmeth.3810.

[6] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano Jr., A.C. Jungkamp, M. Munschauer, A. Ulrich, G.S. Wardle, S. Dewell, M. Zavolan, T. Tuschl, Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP, Cell 141 (2010) 129–141, https://doi.org/10.1016/j.cell.2010.03.009.

[7] J.M. Taliaferro, N.J. Lambert, P.H. Sudmant, D. Dominguez, J.J. Merkin, M.S. Alexis, C.A. Bazile, C.B. Burge, RNA sequence context effects measured in vitro predict in vivo protein binding and regulation, Mol. Cell 64 (2016) 294–306, https://doi.org/10.1016/j.molcel.2016.08.035.

[8] M. Kertesz, Y. Wan, E. Mazor, J.L. Rinn, R.C. Nutter, H.Y. Chang, E. Segal, Genome-wide measurement of RNA secondary structure in yeast, Nature 467 (2010) 103–107, https://doi.org/10.1038/nature09322.

[9] S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, J.S. Weissman, Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo, Nature 505 (2014) 701–705, https://doi.org/10.1038/nature12894.

[10] Y. Ding, Y. Tang, C.K. Kwok, Y. Zhang, P.C. Bevilacqua, S.M. Assmann, In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features, Nature 505 (2014) 696–700, https://doi.org/10.1038/nature12756.

[11] N.D. Berkowitz, I.M. Silverman, D.M. Childress, H. Kazan, L.S. Wang, B.D. Gregory, A comprehensive database of high-throughput sequencing-based RNA secondary structure probing data (Structure Surfer), BMC Bioinf. 17 (2016) 215, https://doi.org/10.1186/s12859-016-1071-0.

[12] P. Cordero, J.B. Lucks, R. Das, An RNA Mapping DataBase for curating RNA structure mapping experiments, Bioinformatics 28 (2012) 3006–3008, https://doi.org/10.1093/bioinformatics/bts554.

[13] J.D. Yesselman, S. Tian, X. Liu, L. Shi, J.B. Li, R. Das, Updates to the RNA mapping database (RMDB), version 2, Nucleic Acids Res. 46 (2018) D375–D379, https://doi.org/10.1093/nar/gkx873.

[14] G. Anders, S.D. Mackowiak, M. Jens, J. Maaskola, A. Kuntzagk, N. Rajewsky, M. Landthaler, C. Dieterich, doRiNA: a database of RNA interactions in post-transcriptional regulation, Nucleic Acids Res. 40 (2012) D180–D186, https://doi.org/10.1093/nar/gkr1007.

[15] D. Incarnato, F. Neri, F. Anselmi, S. Oliviero, RNA structure framework: automated transcriptome-wide reconstruction of RNA secondary structures from high-throughput structure probing data, Bioinformatics 32 (2016) 459–461, https://doi.org/10.1093/bioinformatics/btv571.

[16] B.C. Thiel, I.K. Beckmann, P. Kerpedjiev, I.L. Hofacker, 3D Based on 2D: Calculating Helix Angles and Stacking Patterns Using Forgi 2.0, an RNA Python Library Centered on Secondary Structure Elements, 2019. F1000Research 8.

[17] T.L. Bailey, M. Boden, F.A. Buske, M. Frith, C.E. Grant, L. Clementi, J. Ren, W.W. Li, W.S. Noble, Meme suite: tools for motif discovery and searching, Nucleic Acids Res. 37 (2009) W202–W208, https://doi.org/10.1093/nar/gkp335.

[18] E. Saus, J.R. Willis, L.P. Pryszcz, A. Hafez, C. Llorens, H. Himmelbauer, T. Gabaldon, nextPARS: parallel probing of RNA structures in Illumina, RNA 24 (2018) 609–619, https://doi.org/10.1261/rna.063073.117.

[19] L.E. Ritchey, Z. Su, Y. Tang, D.C. Tack, S.M. Assmann, P.C. Bevilacqua, Structure-seq2: sensitive and accurate genome-wide profiling of RNA structure in vivo, Nucleic Acids Res. 45 (2017), https://doi.org/10.1093/nar/gkx533. ARTN e135.