Check for updates

SOFTWARE TOOL ARTICLE

RBP-Tar – a searchable database for experimental RBP binding sites [version 1; peer review: 2 approved with reservations]

Katarina Gresova[1,2], Tomas Racek[1,2], Vlastimil Martinek[1,2], David Cechak[1,2], Radka Svobodova[1,2], Panagiotis Alexiou [2-4]

[1]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic
[2]Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czech Republic
[3]Centre for Molecular Medicine & Biobanking, University of Malta, Msida, Malta
[4]Department of Applied Biomedical science, University of Malta, Msida, Malta

Abstract

Background: RNA-binding proteins (RBPs) play a critical role in regulating gene expression by binding to specific sites on RNA molecules. Identifying these binding sites is crucial for understanding the many functions of RBPs in cellular function, development and disease. Current experimental methods for identifying RBP binding sites, such as ultra-violet (UV) crosslinking and immunoprecipitation (CLIP), and especially the enhanced CLIP (eCLIP) protocol, were developed to identify authentic RBP binding sites experimentally.
Methods: To make this data more accessible to the scientific community, we have developed RBP-Tar (https://ncbr.muni.cz/RBP-Tar), a web server and database that utilises eCLIP data for 167 RBPs mapped on the human genome. The web server allows researchers to easily search and retrieve binding site information by genomic location and RBP name.
Use case: Researchers can produce lists of all known RBP binding sites on a gene of interest, or produce lists of binding sites for one RBP on different genomic loci.
Conclusions: Our future goal is to continue to populate the web server with additional experimental datasets from CLIP experiments as they become available and processed, making it an increasingly valuable resource for the scientific community.

Keywords
RNA Binding Proteins, CLIP, RBP, Web-server

Open Peer Review

Approval Status ❓ ❓

	1	2
version 1 27 Jun 2023	❓ view	❓ view

1. **Xiaoyong Pan** ID, Shanghai Jiao Tong University, Shanghai, China

2. **Clifford A Meyer**, Harvard T.H. Chan School of Public Health, Dana-Farber Cancer Institute, Boston, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Bioinformatics gateway.

Corresponding author: Panagiotis Alexiou (panagiotis.alexiou@um.edu.mt)

Author roles: Gresova K: Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Racek T**: Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Martinek V**: Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Cechak D**: Data Curation, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Svobodova R**: Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Alexiou P**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

How to cite this article: Gresova K, Racek T, Martinek V *et al.* **RBP-Tar – a searchable database for experimental RBP binding sites [version 1; peer review: 2 approved with reservations]** F1000Research 2023, **12**:755 https://doi.org/10.12688/f1000research.131014.1

First published: 27 Jun 2023, **12**:755 https://doi.org/10.12688/f1000research.131014.1

Introduction

RNA binding proteins (RBPs) are key players in a broad spectrum of RNA regulation, including all stages of the RNA lifecycle (Gerstberger *et al.* 2014; Gebauer *et al.* 2021). Eukaryotic genomes typically encode hundreds of RBPs. For example, over 1500 human RBPs involved in the maturation, transport, stability and translation of coding and non-coding RNA were recently characterised and manually curated (Gerstberger *et al.* 2014). Each RBP can typically target hundreds of RNAs in a complex coordinated fashion (Hogan *et al.* 2008). The general transcriptomic locations of thousands of RNA binding sites corresponding to hundreds of RBPs have been identified using a family of experimental techniques based on RBP CrossLinking, ImmunoPrecipitation and Sequencing (CLIP-Seq) (Chi *et al.* 2009; Licatalosi *et al.* 2008; Moore *et al.* 2014; Hafner *et al.* 2010). A thorough exploration of tens of RBPs binding characteristics in vitro has shown that RBPs can differentiate their binding sites with context preferences beyond narrowly defined binding sequence motif and secondary structure often involving complex binding configurations (Dominguez *et al.* 2018). Among the experimental techniques available to date, the enhanced CLIP (eCLIP) protocol (Van Nostrand *et al.* 2016) is particularly important, as it significantly reduces required amplification and increases specificity in identifying authentic binding sites. This improves the efficiency and accuracy of RBP binding site identification and allows for a deeper understanding of the role of RBPs in gene regulation.

The availability of a large amount of data points produced from the same experimental technique can be very beneficial for applications such as machine learning, as it allows researchers to train and test models with more confidence. Here we present RBP-Tar (Tomáš Raček 2023), a centralised and searchable database of experimentally identified RBP binding sites that can significantly facilitate the study of the RBP mediated gene regulation. Using RBP-Tar, researchers can quickly and cleanly retrieve RBP binding sites constrained by both genomic location and associated RBP for hundreds of RBPs.

Methods
Implementation
Pipeline for reproducible data download and annotation

We have developed a reproducible and easy-to-use pipeline for downloading and annotating RBP eCLIP data from the ENCODE (Luo *et al.* 2020) database.

Metadata for eCLIP experiments, as well as additional files containing genomic coordinates are downloaded from the ENCODE database with the following parameters:

> *"status= released&*
> *internal_tags=ENCORE&*
> *assay_title=eCLIP&*
> *biosample_ontology.term_name=K562&*
> *biosample_ontology.term_name=HepG2&*
> *files.file_type=bed+narrowPeak&*
> *type=Experiment".*

Following the download, information about the chromosome, start position, end position and strand of reads are extracted for each RBP binding site. Binding sites are filtered by length, excluding ones shorter than 20 and longer than 100 nucleotides. As a last step, genomic sequences of binding sites are retrieved.

The described pipeline is implemented as a set of python scripts and is freely available at GitHub.

Using this pipeline, data for 167 RBPs on two cell lines (K562, HepG2) were downloaded. In total, 1.8 Gb of data, representing more than 25 million binding sites, were thus processed.

Web server

Here we present RBP-Tar, a web server that can access the above curated dataset of RBP binding sites (https://ncbr.muni.cz/RBP-Tar) and was built with Python (RRID:SCR_008394), the web development framework Flask, and a simple SQLite database (RRID:SCR_017672). The application's source codes can be found on the project's GitHub page, along with the requirements and instructions for the deployment if a user wants to run the application locally.

The web user interface allows searching and filtering based on the start and end position of the binding site, strand, chromosome, and protein name. It offers the download of the filtered data based on the search done by the user. Due to the size of the dataset, the view is limited to 10 000 results. However, the whole dataset can be conveniently downloaded as a gzipped CSV (579 MB) (here).

Operation

The RBP-Tar web server can take as input any of the following user-provided parameters: [Start min, Start max] denote the limits of the low genomic coordinate of the locus of interest. Similarly, [End min, End max] denote the limits of the high coordinate. [Strand] and [Chromosome] can be used to narrow down the search to only one strand and a specific chromosome. Using these combinations of parameters, a user can easily search for binding sites on their favourite gene, exon, or even a whole chromosome. The last parameter is [Protein name], which brings out a drop-down menu of all the RBPs in the database. If this parameter is not set, all RBPs are queried.

Results are shown as a table with the [Chromosome, Start, End, Strand] genomic location of the binding site, followed by the [Protein name] of the associated RBP and the [Sequence] of the binding site contains the genomic sequence of the binding site. Results can be seen online in a table format or downloaded as a CSV file with one button click. We expect most users to download the results and use them for further downstream analyses.

Use cases
Use case 1: all known RBP binding sites on the gene of interest
The first and potentially most common use case would be the query of all known RBP binding sites on a gene of interest. For example, we can query our web server with the coordinates of the Fused in Sarcoma (FUS) gene (chr16: 31180139-31191605, +) and leave the protein field empty. After this search, all 5 045 known RBP binding sites on this gene are returned and can be easily downloaded in a CSV file for further analysis (Figure 1).

In fact, the gene we used encodes the RBP FUS, which plays important roles, among others, in neurodegeneration and cancer progression. We can use the filter [Protein Name] to identify the potential self-targeting of FUS on itself. Indeed, we can thus identify 72 potential FUS self-targeting binding sites identified *via* eCLIP (Grešová & Raček 2023) (Figure 2). Of course, the biological relevance of this type of finding is left to the users, as is the further validation of high-throughput derived RBP binding sites.
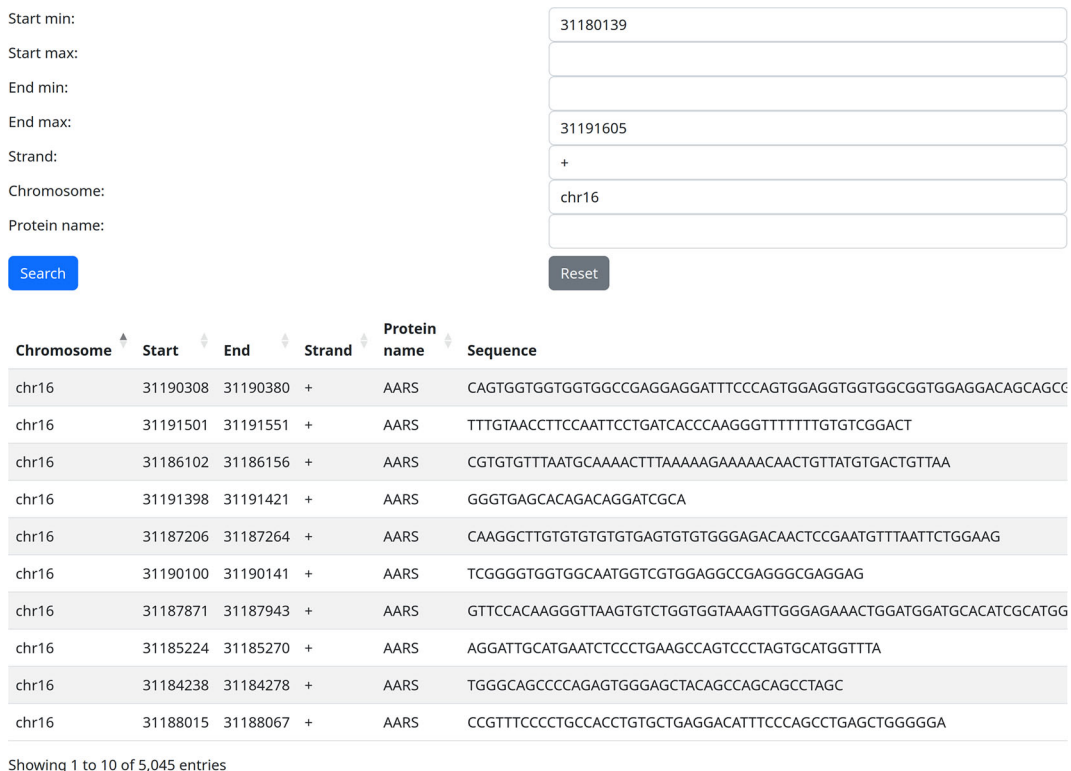

Figure 1. Extracting binding sites of all RNA-binding proteins on one locus (use case 1).

Start min:	31180139
Start max:	
End min:	
End max:	31191605
Strand:	+
Chromosome:	chr16
Protein name:	FUS

Search   Reset

Chromosome	Start	End	Strand	Protein name	Sequence
chr16	31186931	31186968	+	FUS	AGGTTTTGAGGTGTCCAGAACCACCTCCAGAAAGGGG
chr16	31187073	31187136	+	FUS	AGGGGTCTAGTAGGCCTTGGACTGGGCCGTTGCCACACCTGGCACTTAGTGACCATCATCATG
chr16	31187183	31187247	+	FUS	AGGATCCTTTTGATCGTTGTGTCCAAGGCTTGTGTGTGTGTGAGTGTGTGGGAGACAACTCCGA
chr16	31186986	31187023	+	FUS	GTTGCCTGGTGTGTGCTAACCTGGAGCAGGTAGGGGT
chr16	31187879	31187929	+	FUS	AGGGTTAAGTGTCTGGTGGTAAAGTTGGGAGAAACTGGATGGATGCACAT
chr16	31187996	31188057	+	FUS	TGTTCACTTTAATGGGTCTCCGTTTCCCCTGCCACCTGTGCTGAGGACATTTCCCAGCCTG
chr16	31187247	31187268	+	FUS	ATGTTTAATTCTGGAAGAGGG
chr16	31187041	31187073	+	FUS	TTTACCAAATGGGTTTGCCCCAGGTTAATAAG
chr16	31187929	31187981	+	FUS	CGCATGGCTGGTGGCGAGCCCATCTCTCTTCTCTCGGGTGAGAGAACCGGGC
chr16	31188136	31188179	+	FUS	AGAATTTTCTCCTCTGGGCAGGCGACCCAGGAAAGGGTTTGGA

Showing 1 to 10 of 72 entries

Figure 2. Fused in Sarcoma (FUS) self-targeting identified by filtering by gene location and protein name.

Start min:	
Start max:	
End min:	
End max:	
Strand:	
Chromosome:	chr1
Protein name:	AARS

Search   Reset

Chromosome	Start	End	Strand	Protein name	Sequence
chr1	109100272	109100308	+	AARS	AGTGTGAGTGGACGCGTGAGTGTGTGAGTGTGCGCG
chr1	109100384	109100418	+	AARS	GGATCGTGACCCCAGCCCCGCCGGGCCGACCCGA
chr1	92841776	92841826	+	AARS	CCAAAATGTCCCTTGCTCAGAAGAAGGATCGGGTAGCTCAAAAGAAGGCA
chr1	75787915	75787952	+	AARS	CTGAATCTATTGCTGATGTGTAATAACACTTTAGCTC
chr1	44776516	44776556	+	AARS	GGCGTGGTTGTGGCCGTCTTGGTCACCTGTGTGCCACTTG
chr1	171541238	171541327	+	AARS	AAAGAAGACGACAGCGGGGTTCAGAGACTGACACAGACAGTGAAATTCATGAAAGTGCAAGTC
chr1	75787115	75787143	+	AARS	TTACTACTTTAGCTCTAGAATTACTCTG
chr1	75789485	75789547	+	AARS	TGTAATGGCATGTATTAGCTGAATCTAAAGTTGATGTGAGTTCTAAAATTACACTGAGACCT
chr1	45614292	45614360	+	AARS	GCAAGAAACAAAAGAAGCACAGCTTTATGCTGCCCAGGCACATCTTAAACTCGGAGAAGTTAGT
chr1	23694664	23694757	+	AARS	GGAGTATGAGTTAAGAAAAAACAACTTCTCAGATACTGGAAACTTTGGTTTTGGGATCCAGGAA

Showing 1 to 10 of 5,999 entries

Figure 3. Extracting all binding sites of a single protein on a single chromosome (use case 2).

Use case 2: training/testing dataset for one RBP

A potential user may want to develop an RBP binding site machine learning tool that would be able to predict binding sites based on a sequence. It is important to make sure that training and testing sets for their machine learning method are not overlapping. Using our web server, they can download all binding sites for a specific RBP, for example, on chromosome 1, and use them as a training set (Figure 3). Then do the same for chromosome 2 and use it as an independent testing set, thus ensuring that training and testing set do not overlap.

Conclusion

Recent advancements in experimental techniques, such as eCLIP, have greatly expanded our understanding of RBP binding preferences and their role in gene regulation. The development of centralised and searchable databases of experimentally identified RBP binding sites allows researchers to access and analyse the binding preferences of RBPs easily. This information can be used to identify known binding sites on genes of interest and aid in training machine-learning models for RBP binding site prediction. This paper presents RBP-Tar, a centralised web server that can retrieve RBP Target sites with location and RBP constraints. RBP-Tar has been designed to be easily accessible by non-experts. It is still confined to a single source of data, which is helpful for avoiding experimental design effects, but makes its scope limited. We plan to expand the web server with other sources of data, as well as ways for the user to be able to take into account provenance and experimental variation.

Data availability

Source data

All data used in RBP-Tar has been downloaded from ENCODE and Ensembl projects in April 2022. RBP eCLIP metadata were downloaded from https://www.encodeproject.org/metadata/?status=released&internal_tags=ENCORE &assay_title=eCLIP&biosample_ontology.term_name=K562&biosample_ontology.term_name=HepG2&files.file_type= bed+narrowPeak&type=Experiment&files.analyses.status=released&files.preferred_default=true. List of all downloaded files containing genomic coordinates can be found here: [https://github.com/ML-Bioinfo-CEITEC/rbp_encode_eclip/ blob/main/csv/coord_links.txt] (hundreds of filenames). Reference genome was downloaded from http://ftp.ensembl. org/pub/release-97/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.toplevel.fa.gz

Columns 'chrom', 'chromStart', 'chromEnd' and 'strand' from downloaded files containing genomic coordinates can be found here: https://github.com/ML-Bioinfo-CEITEC/rbp_encode_eclip/tree/main/csv. The whole dataset (curated and including sequence) can be downloaded from https://rbp-tar.ncbr.muni.cz/ as a gzipped CSV (579 MB) (https://rbp-tar. ncbr.muni.cz/download_all).

Software availability

RBP-Tar

Software: https://ncbr.muni.cz/RBP-Tar

Source code: https://github.com/sb-ncbr/RBP-Tar/tree/v1.0.0

Archived source code at time of publication: https://doi.org/10.5281/zenodo.7807678 (Raček 2023)

License: MIT

eCLIP RBP

Source code: https://github.com/ML-Bioinfo-CEITEC/rbp_encode_eclip

Archived source code at the time of publication: https://doi.org/10.5281/zenodo.7802803 (Grešová & Raček 2023)

License: Apache 2.0

References

Chi SW, Zang JB, Mele A, *et al*.: **Argonaute HITS-CLIP Decodes microRNA–mRNA Interaction Maps.** *Nature.* 2009; **460**: 479–486.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Dominguez D, Freese P, Alexis MS, *et al*.: **Sequence, Structure, and Context Preferences of Human RNA Binding Proteins.** *Mol. Cell* 2018;

70(5): 854–67.e9.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Gerstberger S, Hafner M, Tuschl T: **A Census of Human RNA-Binding Proteins.** *Nat. Rev. Genet.* 2014; **15**(12): 829–845.
PubMed Abstract | **Publisher Full Text**

Gebauer F, Schwarzl T, Valcárcel J, *et al.*: **A-binding proteins in human genetic disease.** *Nat. Rev. Genet.* 2021; **22**: 185–198.
Publisher Full Text

Grešová K, Raček T: **ML-Bioinfo-CEITEC/rbp_encode_eclip: Publication without data (v1.0.1). [Software].** *Zenodo.* 2023.
Publisher Full Text

Hafner M, Landthaler M, Burger L, *et al.*: **Transcriptome-Wide Identification of RNA-Binding Protein and microRNA Target Sites by PAR-CLIP.** *Cell.* 2010; **141**(1): 129–141.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Hogan DJ, Riordan DP, Gerber AP, *et al.*: **Diverse RNA-Binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System.** *PLoS Biol.* 2008; **6**(10): e255.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Licatalosi DD, Mele A, Fak JJ, *et al.*: **HITS-CLIP Yields Genome-Wide Insights into Brain Alternative RNA Processing.** *Nature* 2008; **456**(7221): 464–469.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Luo Y, Hitz BC, Gabdank I, *et al.*: **New developments on the Encyclopedia of DNA Elements (ENCODE) data portal.** *Nucleic Acids Res.* 2020 Jan 8; **48**(D1): D882–D889.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Moore MJ, Zhang C, Gantman EC, *et al.*: **Mapping Argonaute and Conventional RNA-Binding Protein Interactions with RNA at Single-Nucleotide Resolution Using HITS-CLIP and CIMS Analysis.** *Nat. Protoc.* 2014; **9**(2): 263–293.
Publisher Full Text

Nostrand V, Eric L, Pratt GA, *et al.*: **Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (eCLIP).** *Nat. Methods* 2016; **13**(6): 508–514.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Raček T: **RBP-Tar (v1.0.0). [Software].** *Zenodo.* 2023.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ❓ ❓

Version 1

Reviewer Report 18 December 2023

https://doi.org/10.5256/f1000research.143817.r224195

❓ **Clifford A Meyer**

Harvard T.H. Chan School of Public Health, Dana-Farber Cancer Institute, Boston, USA

This paper describes a database of RNA binding protein binding sites based on 2 ENCODE cell lines. The website allows uses to search for data by querying a gene or genomic interval.

The scope of the database is very limited. It is not clear why the data is restricted to eCLIP data and not CLIP data in general. Although the authors make the argument that eCLIP data is of good quality, quality will vary from sample to sample and there are likely to be high quality CLIP data sets as well as low quality eCLIP ones.

It appears that the processed peaks are downloaded from ENCODE, but the methods used by ENCODE to identify these peaks are not described, and statistics such as p-values, false discovery rates and enrichment relative to background are not provided. Quality controls are also not provided, so it is not possible to know if the experiment worked.

As processed RBP data is downloaded from ENCODE this resource does not add much to what is already available on the ENCODE website. A resource that processed the raw read data from other data resources, such as GEO, would be more valuable. There are over 2,000 eCLIP samples and over 6,000 CLIP samples in GEO.

The search features are useful, although it is hard to interpret the results. Some ranking of the results from most significant to least is needed. It would also be helpful to enable filtering based on some binding sites statistics.

Overall, the resource is limited in terms of the data included and the website features. It would be helpful to include quality control metrics to let users know which samples are more informative.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenetics, Chromatin analysis, Gene regulation, Single cell

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 18 December 2023

https://doi.org/10.5256/f1000research.143817.r224199

? **Xiaoyong Pan** 🆔

Shanghai Jiao Tong University, Shanghai, Shanghai, China

1. In the paper, the author claim they download narrowPeak metadata from https://www.encodeproject.org/metadata/?status=released&internal_tags=ENCORE&assay_title=eCLIP&b , and the link summary is provided in the file https://github.com/ML-Bioinfo-CEITEC/rbp_encode_eclip/blob/main/csv/coord_links.txt. In the ENCODE dataset, the majority of RNA-binding proteins (RBPs) have biological replicate(s) or technical replicate(s), with each replicate having an associated narrowPeak file. Additionally, there is another merged peak file containing reproducible peaks as determined by entropy-ordered peaks between two replicates. However, it is observed that only one file per RBP from K562 or HepG2, which is both the merged peak file. For eCLIP experiment on K562 against QKI, I just find ENCFF190XSX that is from replicate 1. A careful examination of the data is recommended to ensure accuracy and completeness.

2. After narrowPeak download, the author filter the peak by length, excluding ones shorter than 20 and longer than 100 nucleotides. It's important to note that must short (rarely more

than 4–6-nucleotide-long) sequence elements within RNA targets that are recognized and bound by RNA-binding proteins (Gerstberger, S et al. (2014 [1]). In my perspective, employing a filter based on fold change and p-value, in accordance with the eCLIP-seq Processing Pipeline of ENCODE (https://www.encodeproject.org/documents/739ca190-8d43-4a68-90ce-1a0ddfffc6fd/@@download/attachment/eCLIP_analysisSOP_v2.2.pdf) is powerful.

3. If this job is for the benefit of biologists, why not use transcripts as a reference? Because RBP binds to transcripts. And it's too simplistic for systems biologists, the author can refer to the POSTAR3 data.

References
1. Gerstberger S, Hafner M, Tuschl T: A census of human RNA-binding proteins.*Nat Rev Genet*. 2014; **15** (12): 829-45 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: The authors need solve some issues before published

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias

- You can publish traditional articles, null/negative results, case reports, data notes and more

- The peer review process is transparent and collaborative

- Your article is indexed in PubMed after passing peer review

- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research