Nucleic Acids Research, 2024, **52**, e1
https://doi.org/10.1093/nar/gkad998
Advance access publication date: 14 November 2023
Methods

Improved discovery of RNA-binding protein binding sites in eCLIP data using DEWSeq

Thomas Schwarzl [1,†], Sudeep Sahadevan [1,†], Benjamin Lang [2,†], Milad Miladi [3], Rolf Backofen [3], Wolfgang Huber [1], Matthias W. Hentze [1,*] and Gian Gaetano Tartaglia [4,*]

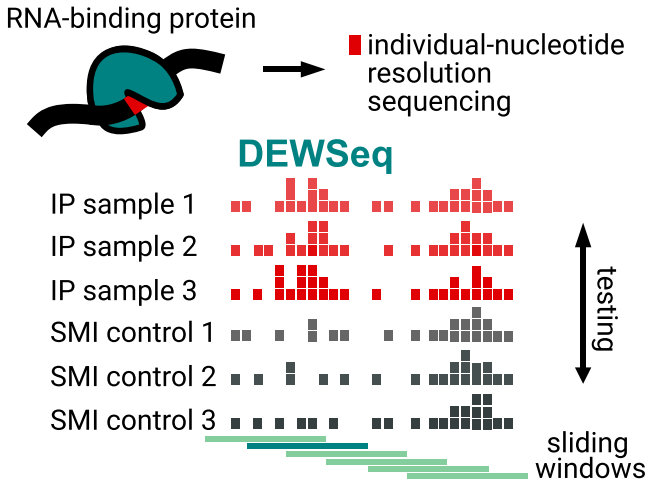
[1]European Molecular Biology Laboratory (EMBL), Meyerhofstraße 1, 69117 Heidelberg, Germany
[2]Department of Structural Biology and Center of Excellence for Data-Driven Discovery, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA
[3]Bioinformatics Group, Department of Computer Science, University of Freiburg, 79098 Freiburg im Breisgau, Germany
[4]Center for Life Nano & Neuroscience, Italian Institute of Technology, 00161 Rome, Italy and Department of Biology, Sapienza University of Rome, 00185 Rome, Italy

[*]To whom correspondence should be addressed. Tel: +39 010 2897 604; Email: gian.tartaglia@iit.it
Correspondence may also be addressed to Matthias W. Hentze. Tel: +49 6221 3878501; Email: hentze@embl.org
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Abstract

Enhanced crosslinking and immunoprecipitation (eCLIP) sequencing is a method for transcriptome-wide detection of binding sites of RNA-binding proteins (RBPs). However, identified crosslink sites can deviate from experimentally established functional elements of even well-studied RBPs. Current peak-calling strategies result in low replication and high false positive rates. Here, we present the R/Bioconductor package *DEWSeq* that makes use of replicate information and size-matched input controls. We benchmarked *DEWSeq* on 107 RBPs for which both eCLIP data and RNA sequence motifs are available and were able to more than double the number of motif-containing binding regions relative to standard eCLIP processing. The improvement not only relates to the number of binding sites (3.1-fold with known motifs for RBFOX2), but also their subcellular localization (1.9-fold of mitochondrial genes for FASTKD2) and structural targets (2.2-fold increase of stem–loop regions for SLBP. On several orthogonal CLIP-seq datasets, *DEWSeq* recovers a larger number of motif-containing binding sites (3.3-fold). *DEWSeq* is a well-documented R/Bioconductor package, scalable to adequate numbers of replicates, and tends to substantially increase the proportion and total number of RBP binding sites containing biologically relevant features.

Graphical abstract



Introduction

RNA-binding proteins play major roles in biological processes such as splicing (1), polyadenylation, nuclear export, subcellular localization, transcript stabilization and degradation as well as translation (2). In recent years, thousands of mammalian proteins have been found to bind to RNA (3,4), many of which have unknown RNA targets. To identify RNA sites bound by an RBP of interest, several related crosslinking and immunoprecipitation (CLIP) high-throughput sequencing methods have been developed (5). These methods exploit

Received: February 17, 2023. **Revised:** September 4, 2023. **Editorial Decision:** October 5, 2023. **Accepted:** October 18, 2023

the phenomenon that UV light induces covalent RNA-protein crosslinks between RNA nucleotides and protein amino acids in immediate contact with each other (6). Over the years, several variants of CLIP-based sequencing methods have been developed: HITS-CLIP (7) directly sequences the crosslinked RNA fragment, PAR-CLIP detects mutations induced at the crosslink site (8), the related methods iCLIP (9) and eCLIP (10), as well as further derivatives such as irCLIP (11), seCLIP (12), easyCLIP (13) and iCLIP2 (14) optimise the protocols for reverse transcription truncations for precise identification of RNA-protein crosslink sites.

Enhanced CLIP (eCLIP) introduced changes to the sequence library generation as well as a size-matched input (SMI) control to address background noise and false positives in CLIP data (10). The ENCODE Consortium has used eCLIP to generate the largest coherent public set of CLIP data, covering 150 RNA-binding proteins in two cell types (HepG2 and K562), processed with the computational peak-calling analysis pipeline *CLIPper* (10). Detected binding sites were compared against SMI controls individually for each replicate and extended by 50 nucleotides from their 5′ end for functional analyses, with the reasoning that the 5′ end of a peak represents the crosslink site (15). An analysis of these data shows the relatively low reproducibility of reported binding sites between replicates (Figure 1A). While RBPs such as the SBDS Ribosome Maturation Factor (SBDS), NOP2/Sun RNA Methyltransferase 2 (NSUN2), and Small RNA Binding Exonuclease Protection Factor La (SSB) show almost perfect reproduction of the respective binding sites, Transforming Growth Factor Beta Regulator 4 (TBRG4), Splicing Factor 3b Subunit 1 (SF3B1), and WD Repeat Domain 3 (WDR3) binding sites display high replicate to replicate variation. More recently, a '*CLIPper reproducible*' (*CLIPper*$_{Rep.}$) dataset was introduced, featuring only binding sites that were identical at the base level in both replicates (15). This approach greatly reduced the number of reported binding sites. However, it raises the question whether better data analysis approaches exist.

A particular challenge in the analysis of eCLIP data is that the measured crosslink peaks can be at an offset from the RBP's actual binding regions, which can result from an RBP's particular structure and physicochemical crosslinking behaviour. CLIP methods are often tested against classical RBPs such as the RNA Binding Fox-1 Homolog 2 (RBFOX2) or splicing factor Heterogeneous Nuclear Ribonucleoprotein C (hnRNPC) with well-known binding sites and RNA sequence motifs (10,13,16). RBFOX2's crosslink site pattern around discovered binding sites is strongest at its known RNA sequence motifs (Figure 1B). Similarly, hnRNPC contacts RNA in a motif- and position-dependent context, displaying bell shaped crosslink distribution around the expected binding site (17). Both cases support the use of traditional peak-callers. However, other RBPs display profound divergences between their biological binding sites and their crosslink behaviours in terms of positioning and shape of the truncation/crosslink sites (Figure 2 and Supplementary Figure S1a, b): For example, the stem–loop binding protein (SLBP), a protein which binds conserved 3′UTR stem–loop structures in histone genes (18), shows systematic crosslink site enrichment upstream of the actual stem–loop (Figure 1E). CSTF2, known to interact with the AAUAAA polyadenylation signal (19), conversely displays crosslink enrichment downstream of its binding motif (Figure 1D), while U2AF2 binds either directly at or downstream of its uridine/cytidine-rich motifs (20) (Figure 1E). Other RBPs such as HNRNPL (21) (Figure 1F), CPEB4 (22) (Figure 2 and Supplementary Figure S1a, b) or the non-classical RBP ENO1 show different crosslinking behaviour (10,23). HNRNPL crosslink sites are in fact depleted at its known sequence motif (Figure 1F and Supplementary Figure S1a, b). Generally, the crosslink sites are enriched at or in proximity of the binding motif (Figure 2 and Supplementary Figure S1a, b). The shift of crosslink site peaks was also evident in some iCLIP data: for eIF4A3, an exon junction complex subunit with well-known binding site locations, the bell-shaped crosslink site curve was shifted by >10 nt compared to other exon junction complex proteins (24,25). Without prior knowledge of an RBP's behaviour, such shifts in positioning and varying crosslink site behaviour are likely to lead to misinterpretation of the binding sites.

Given the varying crosslinking behaviour of each protein and the relatively low reproducibility of binding site detections in eCLIP experiments, we developed a method, *DEWSeq*, that allows accurate and robust identification of RBP interactions in eCLIP data by detecting regions enriched in crosslink sites compared to the control. *DEWSeq* is a sliding-window-based approach that uses single-nucleotide-precision information across multiple replicates and control experiments for significance testing. To test *DEWSeq* and to facilitate a comprehensive analysis of RBP-RNA interactions, we benchmarked it on the eCLIP data for RBPs provided by ENCODE. Notably, 107 out of 150 RBPs in the dataset have known experimentally determined RNA sequence motifs, and one, SLBP, is known to recognise a specific secondary structure (histone mRNA stem–loops). We used validated RNA motifs as a proxy for the biological relevance of a given binding site. This compilation represents, to the best of our knowledge, the most comprehensive eCLIP benchmark based on known sequence motifs to date. We show that RNA binding regions identified by *DEWSeq* show a consistent improvement in sensitivity as well as specificity relative to HITS-CLIP, iCLIP and PAR-CLIP experiments.

Materials and methods

eCLIP data

In order to compare the results from our newly developed *DEWSeq* package, we chose the eCLIP data published by the ENCODE Project (10). This dataset provides consistently produced 223 experimental studies, covering 150 RBPs in either one or both of the two human cell lines: HepG2 and K562. Each study in this dataset consists of two biological replicate IP samples and one size-matched input (SMI) control sample (26). For data reanalysis using *DEWSeq*, we downloaded alignment files (.bam) with reads mapped to the GRCh38 genome annotations from the ENCODE Project data portal (27). BAM file accessions and additional details are given in Supplementary Table S1, Sheet 1.

CLIPper binding sites

In this manuscript, the results from the reanalysis of this ENCODE dataset were compared against two sets of binding site results: the original set called on individual IP samples with respect to SMI controls (10) (referred to as '*CLIPper original*' (*CLIPper*$_{Orig.}$), and the refined set of binding sites based on stringent thresholds, IDR analysis and exclude-list region removal (15) (referred to as '*CLIPper reproducible*'

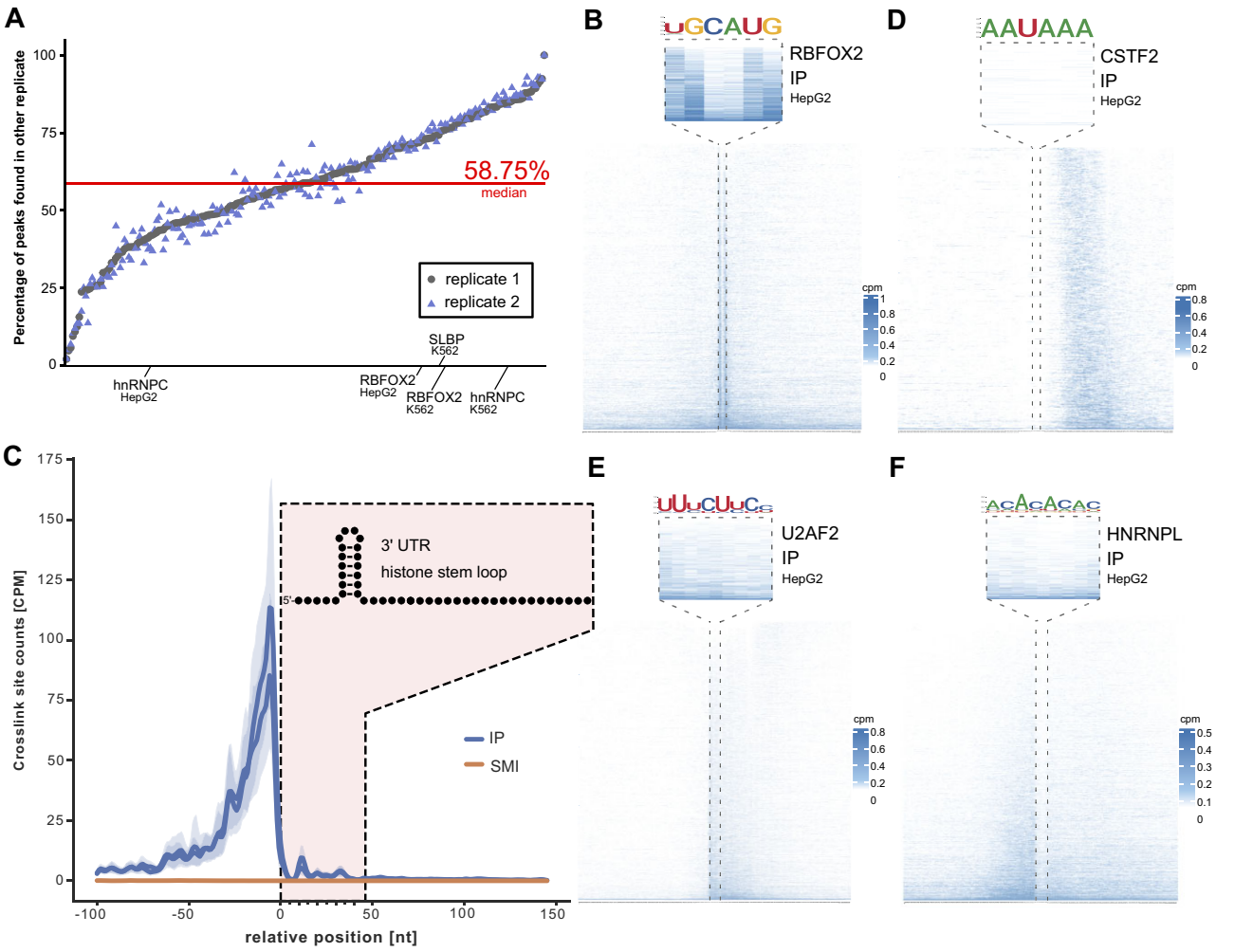


Figure 1. eCLIP crosslink sites around functional elements. (**A**) Reproducibility of binding sites between ENCODE eCLIP datasets replicate 1 and 2. A binding site is counted as reproducible if at least 1 nucleotide overlaps with a binding site called in the other replicate. (**B**) Example of RBFOX2 crosslink site distribution for ENCODE dataset ENCSR756CKJ (K562) and ENCSR987FTF (HepG2) relative to known RBFOX2 UGCAUG motifs. (**C**) Crosslink site distribution of SLBP eCLIP dataset on 34 histone genes relative to known histone mRNA 3′ UTR stem–loops (ENCODE eCLIP dataset ENCSR483NOP, K562 cell line). (**D**) CSTF2 crosslink site distribution for ENCODE eCLIP dataset ENCSR384MWO (HepG2 cell line) relative to known AAUAAA polyadenylation signals. (**E**) U2AF2 crosslink site distribution for ENCODE eCLIP dataset ENCSR202BFN (HepG2 cell line) relative to uridine/cytidine-rich motifs. (**F**) HNRNPL crosslink site distribution for ENCODE eCLIP dataset ENCSR724RDN (HepG2 cell line) relative to CA repeat motifs.

(*CLIPper*$_{Rep.}$). Files providing these *CLIPper* results were also downloaded from the ENCODE Project data portal in narrowPeak BED format.

Preprocessing of eCLIP data with *htseq-clip*

We have developed a custom Python package called *htseq-clip* (28) to count and extract crosslink sites from sequencing alignment files. *htseq-clip* is designed to preprocess alignment files from CLIP experiments and to generate a count matrix of crosslink sites that can be used in further downstream analysis. The required inputs for *htseq-clip* are: gene annotation in GFF format and alignment files (.bam format, coordinate-sorted and indexed).

htseq-clip flattens the input gene annotation file and creates sliding windows by splitting each individual gene annotation feature into a series of overlapping (sliding) windows, where length and overlap (slide) are user supplied parameters. In subsequent steps, *htseq-clip* filters and extracts the crosslink sites based on user supplied experiment specifications and computes the number of crosslink sites per sliding window. In the final step, crosslink counts for multiple samples from the same experiment are summarised into a crosslink site count matrix file, which will be used as input for further downstream analysis. In this study, we used windows of size 50, 75 and 100 base pairs and slides of size 5 and 20 base pairs. The IP and SMI samples in each study were processed and concatenated into a crosslink site count matrix which was used as input for downstream analysis using *DEWSeq*.

Calling differentially enriched regions for eCLIP data with *DEWSeq*

We developed the R/Bioconductor package *DEWSeq* for analysing high-throughput single-nucleotide resolution data. Crosslink site count matrix from *htseq-clip* is tested for differential enrichment in IP samples in comparison with SMI samples. For statistical testing, *DEWSeq* utilises *DESeq2* (29), a well-established R/Bioconductor package primarily used for the analysis of differentially expressed genes in RNA-seq data. After *DESeq2* initial pruning, normalization and dispersion estimation, *DEWSeq* uses a custom

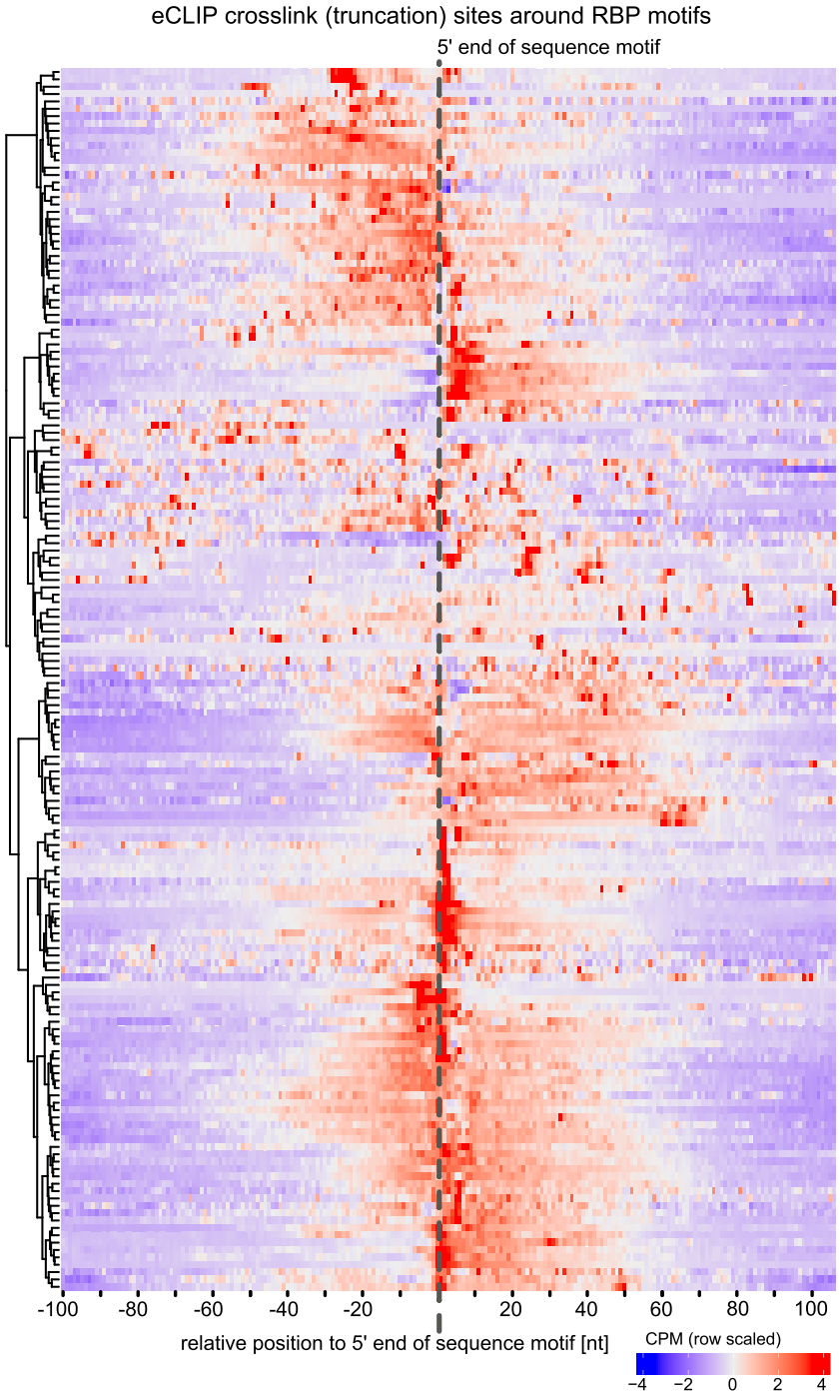


Figure 2. eCLIP crosslink sites around known motifs. Heatmap of ENCODE eCLIP crosslink (truncation) sites around experimentally derived RNA sequence motifs for 107 RBPs. Each row displays a motif for an RNA-binding protein per eCLIP dataset. Rows were clustered with WARD2 to group similar binding patterns.

one-tailed test for detecting significant crosslink regions enriched in IP samples over SMI, followed by two multiple hypothesis correction steps. In the first optional step, dependencies between overlapping windows are corrected using Bonferroni correction, as the adjacent sliding windows share crosslink site count information. In the second step, all windows are corrected for False Discovery Rates (FDR) at the genome level using either Benjamini-Hochberg (BH) method or independent hypothesis weighting (IHW) (30). Finally, all enriched windows passing user specified enrichment thresholds are merged into regions. Users can sort this result table using a combination of $\log_2$ fold change and P_{adj}-value in order to get a list of top enriched regions. *DEWSeq* is available as an open-source R/Bioconductor package (https://bioconductor.org/packages/release/bioc/vignettes/DEWSeq/inst/doc/DEWSeq.html). A sample pipeline for the analysis of eCLIP/iCLIP data using *htseq-clip* and *DEWSeq* is also available (31).

In this study, to test for the impact of *DESeq2* and *DEWSeq* parameters on the final results, we ran *DEWSeq* with the following set of parameters on all ENCODE studies:

1. Dispersion estimation: Using *DESeq2* default 'parametric' dispersion estimation (29) or a custom function to decide the best fit (either 'parametric' or 'local' from *DESeq2*). Referred to either as 'parametric' or 'auto' throughout the rest of this manuscript.
2. Choice of statistical test: Either Wald test or Likelihood Ratio Test (LRT) from *DESeq2*. Referred to either as 'Wald' or 'LRT' throughout the rest of this manuscript.
3. Influence of dependent windows: An optional step to either correct for dependencies between overlapping windows using Bonferroni correction or to use no dependency correction. Referred to either as 'Bonferroni' or 'no correction' throughout the rest of this manuscript.
4. Choice of FDR correction methods: Either using the Benjamini–Hochberg (BH) method for FDR correction or using IHW for FDR correction. Referred to either as 'BH' or 'IHW' throughout the rest of this manuscript.

All the enriched regions resulting from these parameter combinations were filtered using the following thresholds: $\log_2$ fold change ≥ 0.5 and P_{adj} value ≤ 0.1. Parameter combinations were benchmarked for robustness and parameters 100 nt window size, 5 nt step size, LRT testing, 'no correction', IHW, and 'auto' fit were selected. For the rest of this manuscript, these results will be referred to as *'DEWSeq binding sites'*. Supplementary Figure S7 shows a comparison of various parameter combinations tested for *DEWSeq*.

Gene annotations

We used all primary assembly gene models from GENCODE release 27 (GRCh38.p10) to map reported RBP binding sites to genes.

Reference set of known RBP motifs

In this analysis, we used the presence of an RNA sequence motif in a binding region as a proxy for the biological relevance of the binding region. We used motifs from the catRAPID omics v2.0 RBP motifs database, where the authors collected and curated motifs from a comprehensive set of sources including ATtRACT, cisBP-RNA, mCrossBase, oRNAment and RBPmap (32). Supplementary Table S1, Sheet 2 shows the total number of RBPs per motif length in the data source and the number of RBPs in common with the ENCODE dataset. We post-processed this set of motifs by removing any peripheral positions with information content ≤ 0.1 and rounding down base probabilities ≤ 0.025 using the R/Bioconductor package universalmotif (33). We then selected motifs with length ≥ 6 nt to reduce the probability of random occurrences. This selected set of motifs comprised 604 motifs from 258 RBPs, 107 of which had ENCODE eCLIP data available. Using motifs from this common set of RBPs, our final benchmark set contained 322 motifs for 107 RBPs (Supplementary File 1).

Discovery and comparison of known motifs sites

We predicted the positions of human RNA-binding protein motifs within eCLIP binding regions using version 5.4.0 of FIMO (34) from the MEME Suite (35) and the comprehensive benchmark set of RBP motifs described above. For *CLIPper*, eCLIP binding regions were extended 50 nt upstream of their 5′ end, as previously described (15,36). Enriched regions from *DEWSeq* were analysed without any extension. We used a near-equiprobable background sequence model calculated across GENCODE 27 transcripts using fasta-get-markov (-norc) from the MEME Suite. We filtered FIMO results using a *p*-value cutoff ≤ 0.001. We consciously avoided the use of *q*-values to filter the results due to the variation in the number of binding sites between different RBPs, which would penalise experiments that succeeded at identifying a larger number of binding sites, and reasoned that *p*-values would offer more comparable results across experiments.

De novo motif discovery

The *de novo* motif discovery pipeline includes a *de novo* motif discovery step and a motif refinement step. In the first step, motifs are predicted from the input set of peaks/region FASTA sequences using the STREME algorithm (37) from the MEME Suite (version 5.4.1). The motif position weight matrices (PWMs) were generated by STREME and then post-processed by removing any peripheral positions with information content ≤ 0.1 and rounding down base probabilities ≤ 0.025 using the R/Bioconductor package universalmotif (33). These trimmed motif PWMs are then refined using BaMMmotif2 (38) and subjected to another round of trimming as described above. Finally, the FIMO tool from the MEME suite was used to scan the input FASTA sequences using the refined PWMs as described above, but using default parameter values.

Secondary structure analysis

This analysis was restricted to enriched regions from SLBP, as it was the only RBP in the ENCODE data with well characterised RNA secondary structure binding targets (histone 3′ UTR stem–loops), to the best of our knowledge. In this study, we used the *cmsearch* tool from Infernal suite version 1.1.4 (39) and covariance models from the Rfam database (40) to scan for secondary structures in enriched regions from both $CLIPper_{Rep.}$ and *DEWSeq* results. In the first step, we used *mergeBed* from the bedtools suite (41) to merge overlapping regions or regions separated by a maximum of 10 nucleotides in $CLIPper_{Rep.}$ SLBP results. In the case of the *DEWSeq* SLBP regions, no such merging step was necessary. In the next step, we sequentially extended the length of both *DEWSeq* and $CLIPper_{Rep.}$ regions by up to 300 nt (50, 100, 150, 200, 300) using the *slopBed* tool from the bedtools suite (41). With this extension step, we aimed to detect histone stem–loop structures that are found in close proximity to either $CLIPper_{Rep.}$ or *DEWSeq* enriched regions, and to assess the gain in number of stem–loop structures identified with each extension. The FASTA sequences extracted (using *getfasta* from bedtools) from the original set of regions and extended regions were subjected to a profile-based search with the histone 3′ UTR stem–loop family (Rfam ID: RF00032) covariance model and using the Infernal *cmsearch* tool. In this step, *cmsearch* sequence-based pre-filtering heuristics were turned off and an *E*-value threshold ≥ 5.0 was used to identify hits. The *cmsearch* output table was processed with the Bash *awk* command to obtain the genomic location of the model hits and extract unique hits based on the genomic coordinates.

The reference set of histone mRNA 3′ UTR stem–loop regions were retrieved from the Rfam database (Rfam ID: RF00032).

Comparison to orthogonal datasets

We retrieved iCLIP (16), HITS-CLIP (7) and PAR-CLIP (8) interaction datasets from the POSTAR2 database (previously known as CLIPdb) (42). For each CLIP experiment type, POSTAR2 provided peak calling results from *Piranha* (43), as well as the more specialised *CIMS* (crosslink-induced mutation sites) (44) and *PARalyzer* data analysis pipelines (45). We used POSTAR2's standard thresholds such as: p-value <0.01 for *Piranha*, score <0.01 for *CIMS*, and score >0.5 for *PARalyzer*. Additionally, we also obtained a cohesive PAR-CLIP dataset from the DoRiNA database (46), and used a minimum conversion specificity score of 5 to filter binding regions, which resulted in a similar number of regions as for $CLIPper_{Rep.}$. We performed motif calling on these enriched regions using the reference motif set and methodology described above and compared the fraction of unique motifs present in these results to that of the *DEWSeq* and $CLIPper_{Rep.}$ results.

Comparison to HyperTRIBE and STAMP data

We retrieved TARDBP (TDP-43) single nucleotide edit sites in human (HEK293T cell line) identified using the HyperTRIBE and STAMP methodologies (47). We used the *intersect* tool from the bedtools suite to find the overlaps between single nucleotide edit sites and enriched regions and motifs within those regions from *DEWSeq*, $CLIPper_{Orig.}$ and $CLIPper_{Rep.}$ results. To account for overlapping motif positions in these results we merged overlapping motifs based on their chromosomal coordinates. To check for the presence of edit sites in proximity to the motif positions, edit sites were extended ± 50 nt using *slopbed* from the bedtools suite and the overlaps between these extended edit sites and motif positions were calculated again using *intersect* from bedtools.

Comparison to RNA interference (RNAi) data

RNAi data for 125 RBPs were obtained from the ENCODE Project's experiment matrix by searching for shRNA knockdown RNA-Seq experiments in human HepG2 or K562 cells that targeted RNA-binding proteins for which eCLIP data were available (15). A matched control experiment assigned to each RBP was obtained by querying the ENCODE Project's API for shRNA gene silencing series. We then obtained TSV-format gene-level expression data mapped to GRCh38 (GENCODE release 29) for both the knockdown and control experiments. For each RBP, we calculated the knockdown effect (difference) on individual genes by subtracting the control's value for a given gene from that of the corresponding shRNA knockdown. We normalised these difference values by subtracting their mean and dividing by their standard deviation. Any gene with a knockdown effect more than one standard deviation below the mean was considered to be a knockdown hit. The set of knockdown hits for each RBP was then compared with its corresponding set of eCLIP target genes using the Jaccard index.

Comparing gene region and gene type enrichment using OLOGRAM

To determine whether *DEWSeq* or *CLIPper* results were biased towards gene regions (5′ UTR, exon, 3′ UTR) or gene types (e.g. protein-coding RNAs, non-coding RNAs, mtRNAs, …) we performed gene region and gene type enrichment analysis in both results using OLOGRAM (48). To avoid ambiguities in gene region annotation in genes with multiple transcripts, we selected the transcript with the highest abundance in each gene as the representative transcript. For this purpose we used rRNA-depleted total RNA-seq data from the HepG2 (ENCODE accession: ENCFF533XPJ, ENCFF321JIT) and K562 (ENCODE accession: ENCFF286GLL, ENCFF986DBN) cell lines available from the ENCODE Consortium. After removing lowly expressed transcripts with a TPM value ≤ 1, the datasets were merged and for each gene, the transcript with the highest abundance was selected as the representative transcript. We selected 15,274 transcripts as candidates and extracted region and type annotation for these selected transcripts. In the next step, we used this gene annotation data and enriched regions from *DEWSeq* and $CLIPper_{Rep.}$ results to assess the significance of overlaps between enriched regions in either of the two sets and the gene region or gene type annotations.

Comparing FIMO motifs and *de novo* motifs using OLOGRAM

OLOGRAM (48) methodology was used to compare the motif positions obtained from scanning enriched regions/peaks using catRAPID omics v2.0 motifs and those from our *de novo* motif prediction pipeline for each study.

Results

We developed *DEWSeq* as a new R/Bioconductor statistical analysis package for the robust detection of RBP binding regions from i/eCLIP datasets. The *DEWSeq* workflow starts from the output of an accompanying Python package for post-processing i/eCLIP alignment files, *htseq-clip* (28), which extracts crosslink site counts at single-nucleotide positions adjacent to the end of reads, flattens annotation of multiple transcripts and uses sliding windows to count and aggregate crosslink sites. *DEWSeq* performs one-tailed significance testing using *DESeq2* (29), result summarization and binding site visualization (Supplementary Figure S2). Similar to the *csaw* package (49) for ChIP-seq data, *DEWSeq* incorporates biological variation with significance testing, which reduces the false discovery rate (49), with the difference that *DEWSeq* is tailored to single-nucleotide position data. *DEWSeq* directly uses SMI samples quantitatively as controls for determining significance, while peak calling methods such as *CLIPper* may only use a rank-based measure to filter out irreproducible peaks in a post-processing step (50). 223 ENCODE eCLIP datasets covering 150 RBPs cell types HepG2 and K562 were processed and analysed with *htseq-clip*/*DEWSeq*.

Sequence motif-based evaluation strategy

We compared *DEWSeq*'s results to the *CLIPper* method used by the ENCODE Project, which uses peak-calling on two individual replicates compared against a single SMI control. We extended each peak 50 nt in the 5′ direction, as first introduced by the authors specifically for motif-based analyses (15), which is referred to as *CLIPper original* ($CLIPper_{Orig.}$) in our study. This dataset was further improved on by the authors to produce $CLIPper_{Rep.}$, which is the subset of $CLIPper_{Orig.}$ peaks that are reproducible at the nucleotide level across

both replicates (10,15). An overview of the RNA sequence motif-based benchmarking strategy we adopted in our study is shown in Figure 3A.

As a proxy for the likely biological relevance of the identified binding sites, we obtained known experimentally determined RNA sequence motifs of length 6 nucleotides or longer from catRAPID omics v2.0 (32) (Supplementary File 1). This curated motif dataset covered 107 of the 150 RBPs for which ENCODE eCLIP data were available.

To identify positions of known motifs within binding regions identified by *CLIPper* and *DEWSeq* from ENCODE datasets, we used FIMO from the MEME suite of motif analysis software (34). To compare these results to orthogonal datasets beyond eCLIP, we obtained iCLIP, HITS-CLIP, PAR-CLIP binding sites determined using different peak callers from the POSTAR2 database (51) as well as a separate dataset of author-processed PAR-CLIP binding sites (46), and scanned for motifs using FIMO. For each method, we then estimated the accuracy of binding site detection by calculating the proportion of reported binding sites that contained at least one expected sequence motif for the RBP of interest.

Performance comparison between *DEWSeq* and *CLIPper*

Slightly over half (51.8%) of the $CLIPper_{Orig.}$ binding sites contained a known sequence motif for the RBP under investigation (median across RBPs, cell types and replicates, Figure 3B, top). $CLIPper_{Rep.}$ slightly increased the motif-containing binding site fraction to 55.0%, but at the expense of reducing the total number of motif-containing regions identified from 1366 to 1021 median binding sites per RBP and cell type (Figure 3B, bottom). Conversely, while *DEWSeq* binding sites showed a further increased motif-containing rate (58.9%), *DEWSeq* also notably identified a total number of motif-containing binding sites (median 2,137) that was markedly higher than both $CLIPper_{Rep.}$ and the less stringent $CLIPper_{Orig.}$ set (a 2.25-fold and 1.81-fold median improvement, respectively). Complete results from this analysis are provided in Supplementary Table S2. Thus, *DEWSeq* achieves an increase in the number of detected binding sites without reducing the proportion of motif-containing sites, without any apparent systemic bias towards gene regions or gene types (Supplementary Figure S3). We thus posit that the detection quality is on median at least as good as that of the $CLIPper_{Rep.}$ approach, and at the same time, the detection rate has been improved by at least two-fold. At the gene level, *DEWSeq* consequently increases the discovery of RBP-RNA interactions from a median of 760 to 1,181 genes per RBP and cell type (Figure 3C). In order to check whether the motifs retrieved from FIMO scans could also be retrieved using *de novo* motif detection algorithms, we performed a *de novo* motif scan on *DEWSeq*, $CLIPper_{Orig.}$ and $CLIPper_{Rep.}$ results as mentioned in the Methods section and compared those motifs to the motifs retrieved from FIMO scans per RBP per cell-line using OLOGRAM. The results from this comparison show that at a P_{adj} threshold $\leq$0.1, out of 164 overlap comparisons between FIMO motif scans and *de novo* motif predictions, *DEWSeq* results have 7 non-significant overlaps, compared to 6 non-significant overlaps out of 165 comparisons for $CLIPper_{Rep.}$ results. For $CLIPper_{Orig.}$ results, the corresponding figures were 4 non-significant overlaps out of 165 comparisons for IP1 and 4 non-significant overlaps out of 166 comparisons for IP2. *DEWSeq* results have a slightly higher number (7) of non-significant overlaps compared to $CLIPper_{Rep.}$ results (6) and $CLIPper_{Orig.}$ results (4). These results indicate that across all the methods tested, there is a high degree of overlap between FIMO motif scans and *de novo* motif prediction results, albeit showing minor differences/non-significant overlaps for certain RBPs (Supplementary Table S3).

Motif exclusiveness analysis

Following the motif-containing binding site analysis, we assessed the proportion of binding sites that *DEWSeq* and $CLIPper_{Rep.}$ have in common and also the proportion of binding sites that are exclusively found by either one of the respective methods.

Here we chose $CLIPper_{Rep.}$ as the best performing representative of all *CLIPper* results (data for all runs are provided in Supplementary Table S4, Sheet 1), and focussed on the binding sites with known motifs. In this analysis, all detected motif-containing binding sites were classified based on whether they were detected by both methods (labelled 'Both'), or exclusively by only one of the methods (indicated either as $CLIPper_{Rep.}$ or *DEWSeq*, respectively). For each RBP, the percentage of motif-containing peaks in each group was calculated (Figure 3D, E). A substantial fraction of motif-containing binding regions was detected exclusively by *DEWSeq* (median across RBP–cell type experiments: 62.9%). *DEWSeq* and $CLIPper_{Rep.}$ showed good agreement on binding sites for some RBPs such as EFTUD2 (47.4% in K562 and 37.2% in HepG2 cells), while for other RBPs such as AKAP1 the agreement was high in one cell type (K562, 34.7%), but not in the other (HepG2, 11.4%) (Figure 3D). Overall, the median fraction of motif-containing binding regions exclusively detected by *DEWSeq* (62.9%) greatly exceeded the fraction agreed on by both methods (28.0%) and those found exclusively by $CLIPper_{Rep.}$ (4.5%) (Figure 3D, E).

Results for specific RBPs with well-defined biological roles

RBFOX2

RBFOX2 (RNA Binding Fox-1 Homolog 2) is an alternative splicing regulator that binds to UGCAUG motifs (52). It is regularly used for benchmarking in CLIP manuscripts (10). Here, we compared the number of RBFOX2 binding regions containing the UGCAUG motif reported by $CLIPper_{Rep.}$, *DEWSeq* or both. Figure 4A shows the total number of regions reported and the number of regions including the UGCAUG motif. The fraction of motif-containing regions per cell line (HepG2 and K562) are similar for both $CLIPper_{Rep.}$ and *DEWSeq* results: 61.2% and 60.2% in HepG2 cell lines and 44.9% and 46.7% in K562 cell lines. However, a striking difference can be seen for the number of motif-containing regions reported by *DEWSeq* as compared to $CLIPper_{Rep.}$. In the HepG2 cell line, $CLIPper_{Rep.}$ reported 3,223 UGCAUG motif-containing regions compared to 8,410 motif-containing regions for *DEWSeq*, and in K562 cells $CLIPper_{Rep.}$ reported 1,204 motif-containing regions in comparison to 4,257 from *DEWSeq*, indicating a substantial improvement in sensitivity when using *DEWSeq* (Supplementary Table S5, Sheet 1). Supplementary Figure S4a–c shows genomic tracks for 3 RBFOX2 targets with RBFOX2 HepG2 crosslink sites, *DEWSeq* and $CLIPper_{Rep.}$ enriched regions and motif positions identified in *DEWSeq* regions, visualised using JBrowse 2 (53).

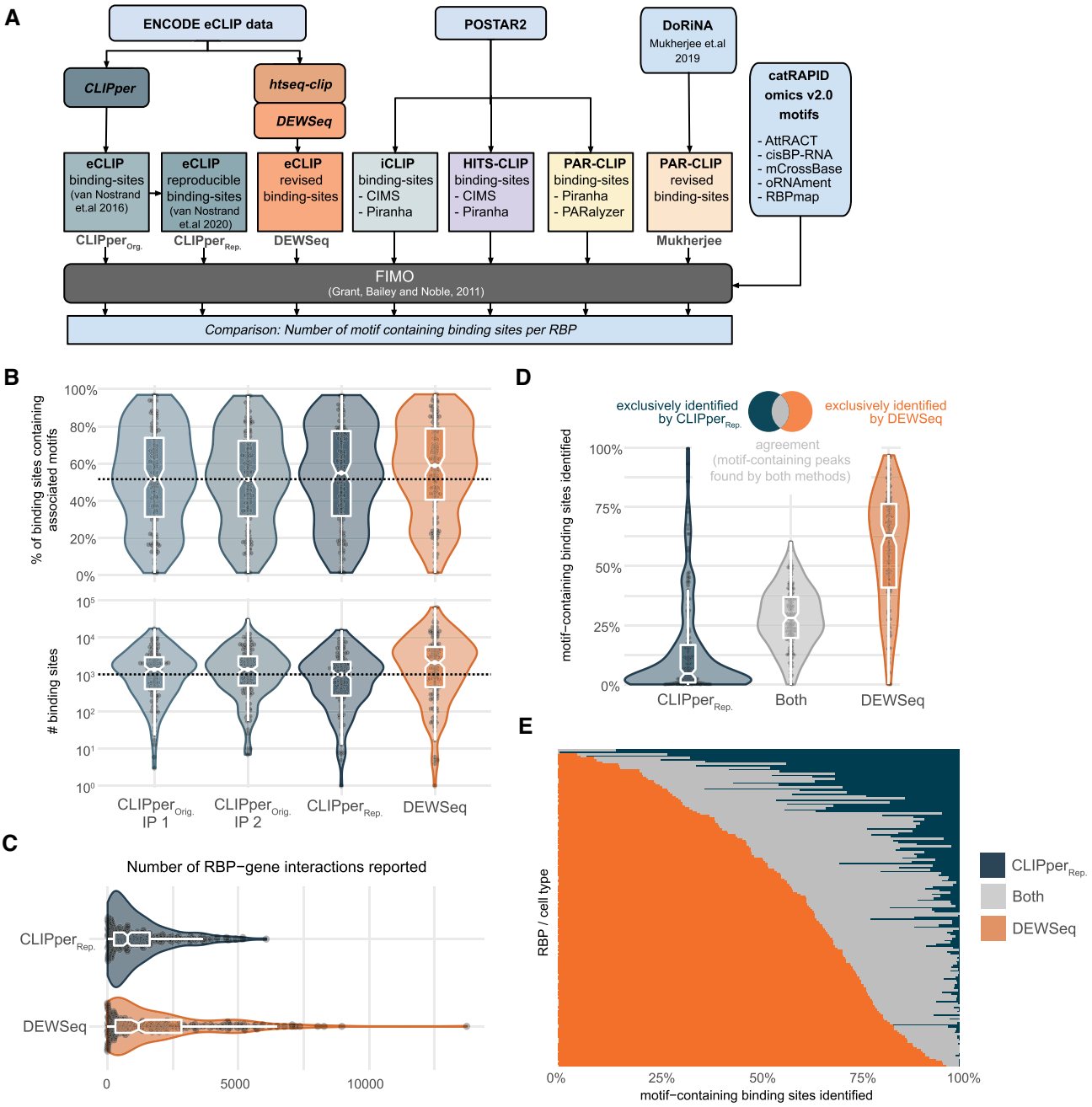


Figure 3. Overview and results for benchmarking workflow on ENCODE eCLIP datasets for proteins with known RNA sequence motifs. (**A**) ENCODE eCLIP datasets were reanalysed with *DEWSeq* and compared to '*CLIPper original*' (*CLIPper*$_{Orig.}$) (10,32) and '*CLIPper reproducible*' (*CLIPper*$_{Rep.}$) dataset (15) analyses. Other CLIP binding sites from iCLIP, HITS-CLIP and PAR-CLIP were extracted from POSTAR2 (51). Additional PAR-CLIP datasets from DoRiNA (46) were included in the analyses. Binding sites from all datasets were analysed with FIMO (34) using known RNA sequence motifs from catRAPID omics v2.0 (32). (**B**) Top panel shows violin and boxplots of the number of motif-containing binding sites in datasets detected with FIMO and catRAPID omics v2.0 motifs. Bottom panel violin and boxplots show the percentage of motif-containing binding sites to the total number of binding sites for each method. (**C**) Number of reported RBP-gene interactions. (**D**) Exclusiveness of motif-containing binding sites for datasets with known motifs. Left shows binding sites exclusive for *CLIPper*$_{Rep.}$ dataset, the middle the binding sites which were found in both, and right for sites found exclusively in *DEWSeq*. (**E**) Heatmap of exclusiveness and overlap for motif-containing *CLIPper*$_{Rep.}$ and *DEWSeq* binding sites.

FASTKD2

FASTKD2 (FAST kinase domain-containing protein 2) is a mitochondrial RBP that has been shown to interact with a defined set of mitochondrial transcripts (54,55). In this analysis, we compared the number of FASTKD2-bound regions of *DEWSeq* against *CLIPper*$_{Rep.}$ results. In HepG2 cells, *CLIPper*$_{Rep.}$ reports 7 out of 451 bound genes as mitochondrial, compared to *DEWSeq* with 16 out of 268 bound genes being mitochondrial (Figure 4B). A similar pattern emerges in the K562 cell line, where the numbers for *CLIPper* and *DEWSeq* were 19 out of 364 and 29 out of 426, respectively. Fisher's exact test confirmed a significant enrichment in the number of FASTKD2-bound mitochondrial genes reported by *DEWSeq* compared to *CLIPper*$_{Rep.}$ results in both cell lines (Figure 4B, Supplementary Table S5, Sheet 2). Example genomic tracks with FASTKD2 targets with FASTKD2 HepG2

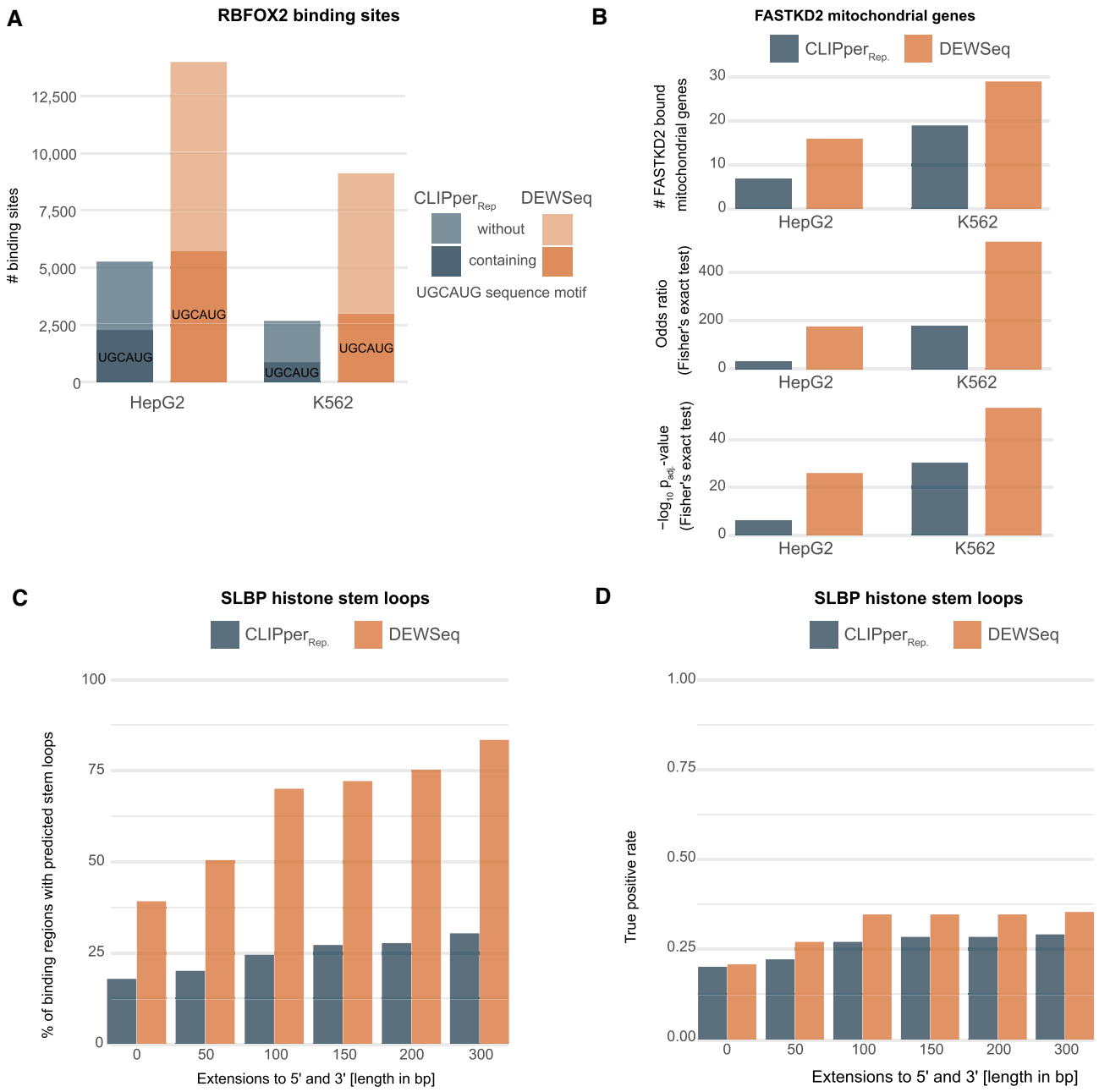


Figure 4. Binding site comparisons. (**A**) Comparison of RBFOX2 binding regions and RBFOX2 binding regions containing a UGCAUG motif in *'CLIPper reproducible'* (*CLIPper*$_{Rep.}$) and *DEWSeq*. (**B**) FASTKD2 binding site enrichment for mitochondrial genes compared to other chromosomal locations for *CLIPper*$_{Rep.}$ and *DEWSeq* binding sites. Supplementary Table S5, Sheet 2 contains the complete enrichment analysis results across all chromosomes for all FASTKD2 samples. (**C**) SLBP stem–loops found with 3′ and 5′ extensions of *DEWSeq* and *CLIPper*$_{Rep.}$ binding sites. Left panel shows the percentage of binding regions containing the predicted stem–loops. (**D**) True positive rate (sensitivity) with respect to reference histone mRNA 3′ UTR stem–loop regions retrieved from the Rfam database (for Rfam ID: RF00032).

crosslink sites, *DEWSeq* and *CLIPper*$_{Rep.}$ enriched regions are visualised in Supplementary Figure S4d, e.

SLBP

SLBP (Stem-Loop Binding Protein) is an RBP that binds to a conserved stem–loop structure motif at the 3′ end of mRNAs that encode replication-dependent histones (2,56,57). To the best of our knowledge, it represents the only RBP included in the ENCODE project that recognises a secondary structure motif. We scanned SLBP binding regions and surroundings (binding site were extended with 50, 100, 150, 200 and 300 nt in both 5′ and 3′ direction) from both *CLIPper*$_{Rep.}$ and *DEWSeq* results for the SLBP stem–loop structure binding site using the Infernal suite and histone 3′ UTR stem–loop covariance model (Rfam ID: RF00032).

A higher proportion of *DEWSeq* binding sites (39.2%) contain predicted SLBP binding structures, as compared to *CLIPper*$_{Rep.}$ binding regions (17.9%) (Figure 4C). This trend becomes more pronounced with the extension of binding regions in both 5′ and 3′ directions (Figure 4C and Supplemen-

tary Figure S4), as *DEWSeq* discovers increasingly more stem–loops: 83.5% of detected binding sites are in proximity to histone stem–loops, whereas only 30.4% of $CLIPper_{Rep.}$ binding sites are in the vicinity of known targets, suggesting a significant decrease of false positives for *DEWSeq*.

Further, we calculated the true positive rate (sensitivity) of these predicted stem–loop structures using histone mRNA 3′ UTR stem–loop annotations from the Rfam database as a reference set (Figure 4D). *DEWSeq* without extension shows a marginal increase in true positive rate compared to $CLIPper_{Rep.}$ (from 0.201 to 0.208), with slight improvement in extensions (Supplementary Table S6).

In addition to the increased presence of expected stem–loop structures, we also noted that $CLIPper_{Rep.}$ identified binding of SLBP to mRNAs deriving from a total of 44 histone genes (66.7% of its target genes being histones), while *DEWSeq* identified binding of SLBP to a total of 53 histone mRNAs (71.6% of its target genes being histones). For reference, the HGNC histone gene set contains a total of 118 genes. Supplementary Figure S4f, g shows genomic tracks for two histone genes with SLBP K562 crosslink sites, *DEWSeq* and $CLIPper_{Rep.}$ enriched regions.

Evaluation of eCLIP compared to iCLIP, HITS-CLIP and PAR-CLIP

To validate the newly discovered binding sites, we used the motif exclusivity benchmark to compare eCLIP binding sites assigned by *DEWSeq* and $CLIPper_{Rep.}$, respectively, to sites from iCLIP, HITS-CLIP and PAR-CLIP protocols retrieved from the POSTAR2 (51) and DoRiNA (46) databases. To address differences in detection methods for these different CLIP protocols, we employed multiple established analysis methods: Piranha (43) and CIMS (44) for iCLIP and HITS-CLIP data, and Piranha (43), PARalyzer (45) as well as Mukherjee's method (46) for PAR-CLIP data.

Figure 5A–C left panels show the comparison of $CLIPper_{Rep.}$ results to iCLIP, HITS-CLIP and PAR-CLIP, and right panels show the comparison of *DEWSeq* results to the same. eCLIP yields more motif-containing binding sites overall, with a substantially higher number identified by *DEWSeq* compared to $CLIPper_{Rep.}$ (Figure 5A–C, Supplementary Figure S5a–c and Supplementary Table S4, Sheet 2). Interestingly, the overlaps of binding sites found in either eCLIP and iCLIP, HITS-CLIP, or PAR-CLIP are modest compared to binding sites detected in one of the protocols alone. *DEWSeq* recovers a median of 60 (2.1% of total) motif-containing binding sites found by other methods, compared to a median of 18 (1.0% of total) for $CLIPper_{Rep.}$ (Figure 5D and Supplementary Figure S5d).

Comparison of eCLIP TARDBP regions and motifs to HyperTRIBE and STAMP datasets

We compared TARDBP (TDP-43) binding regions and motif coordinates from *DEWSeq*, $CLIPper_{Orig.}$ and $CLIPper_{Rep.}$ results to the single nucleotide edit sites from the HyperTRIBE and STAMP protocols (47). These protocols identified 12,339 and 11,419 unique RNA edit sites, respectively (Supplementary Table S7, Sheet 1). Out of these, 831 (6.73%) HyperTRIBE edit sites were found to be within *DEWSeq* binding regions, whereas this number ranged between 179 and 221 (1.79% and 1.45%) in $CLIPper_{Orig.}$ results to 182 (1.47%) in $CLIPper_{Rep.}$ results. A similar trend can be observed for STAMP edit sites, where 586 (5.13%) sites were found to be within *DEWSeq* binding regions whereas the number of edit sites in $CLIPper_{Orig.}$ and $CLIPper_{Rep.}$ binding regions were 208, 168 and 191 (1.82%, 1.47% and 1.67%) respectively. Out of the 831 HyperTRIBE edit sites in *DEWSeq* regions, 519 (4.21%) were found to have a motif within ±50 bp neighbourhood compared to 165 (1.34%) and 122 edit sites (0.99%) in $CLIPper_{Orig.}$ binding regions and 133 (1.08%) edit sites in $CLIPper_{Rep.}$ results. Here again STAMP edit sites showed a similar tendency: 416 (3.64%) edit sites out of 586 edit sites in *DEWSeq* binding regions were in the neighbourhood of a motif, compared to 162 (1.42%) and 130 (1.14%) edit sites in $CLIPper_{Rep.}$ binding regions and 141 (1.23%) edit sites in $CLIPper_{Rep.}$ binding regions (Supplementary Table S7, Sheet 1). Additionally, this result also shows that there are 312 (2.53%) HyperTRIBE edit sites within *DEWSeq* binding regions that are not within the neighbourhood of a motif. For $CLIPper_{Rep.}$ and $CLIPper_{Orig.}$ results, these values are 49 (0.40%), 56 (0.45%) and 57 (0.46%) respectively. This difference can also be seen for STAMP data, where 170 (1.49%) of edit sites fall within *DEWSeq* binding regions, but outside motif neighbourhood, compared to 50 (0.44%) edit sites within $CLIPper_{Rep.}$ binding regions and 46 (0.40%) and 38 (0.33%) within $CLIPper_{Orig.}$ binding regions (Supplementary Table S7, Sheet 1).

Since the HyperTRIBE and STAMP datasets were generated using poly(A) mRNAs, the edit sites found within *DEWSeq* and $CLIPper_{Rep.}$ enriched regions were further intersected with gene feature annotations such as exons, 5′ UTR, CDS and 3′ UTR to compute the fraction of edit sites within these features. Out of all the edit sites within *DEWSeq* and $CLIPper_{Rep.}$ enriched regions, greater than 98% of sites from both HyperTRIBE and STAMP experiment lie within exons. Additionally, a considerable proportion of these (68–90% in HyperTRIBE data and greater than 93% in STAMP data) are within 3′ UTR regions, reflecting the trend described in the original publication (47) (Supplementary Table S7, Sheet 2).

Comparison of eCLIP to RNA interference (RNAi) data

As an additional orthogonal evaluation, we investigated whether the sets of target genes identified by eCLIP agreed with those identified in RNA interference experiments using small hairpin RNA (shRNA) knockdowns performed by the Graveley laboratory as part of the ENCODE Project (15). We found a relatively low agreement with a median Jaccard index across RBPs of 0.029 for $CLIPper_{Rep.}$, which was significantly increased at a median of 0.035 for *DEWSeq* ($P = 0.043$, Mann–Whitney *U* test) (Figure 5E and Supplementary Table S8). These results also show that, on median, *DEWSeq* results show a better overlap with RNAi results, on median, recovering up to 26% of RNAi target compared to $CLIPper_{Rep.}$ results which recover only 13% of the RNAi target genes. $CLIPper_{Orig.}$ results on median recover 17.5% and 20% of the RNAi targets respectively (Supplementary Figure S6a). Conversely, we also observed that the proportion of common eCLIP and RNAi targets to the total number of genes with eCLIP binding regions is relatively low. For *DEWSeq* results, RNAi targets covered only 5% of the total number of genes with a binding region, whereas for $CLIPper_{Rep.}$

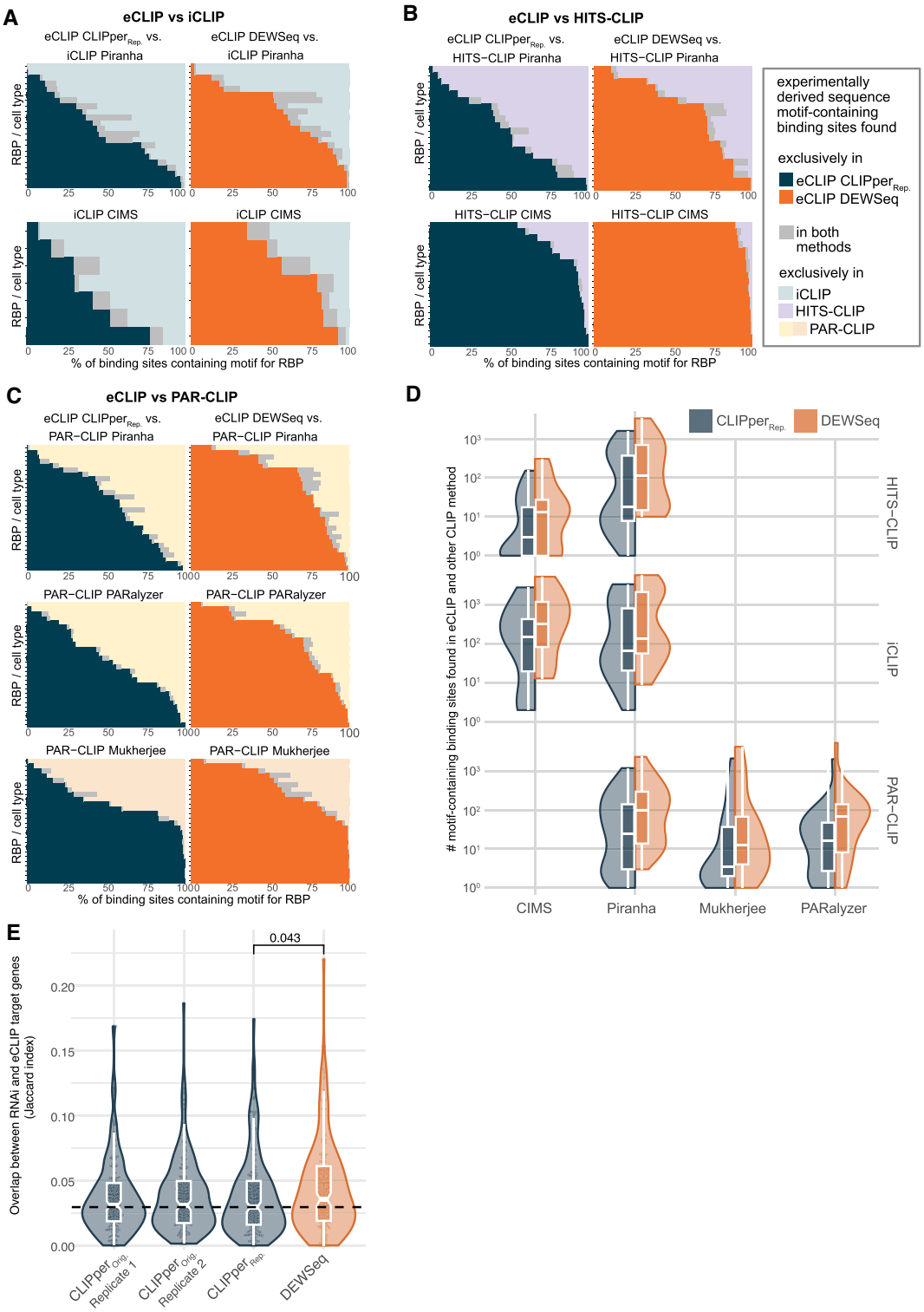


Figure 5. Binding site exclusiveness. Comparison of motif-containing binding sites from various CLIP datasets against eCLIP *'CLIPper reproducible'* (*CLIPper*$_{Rep.}$) and *DEWSeq* results. The comparisons are for common RBPs (in ENCODE dataset and these methods) with motifs in catRAPID omics v2.0. Stacked bar plots (plots: A–C) show comparisons of iCLIP, HITS-CLIP and PAR-CLIP motif-containing regions (in percentage) against *CLIPper*$_{Rep.}$ (left panels) and *DEWSeq* (right panels) results. Blue bars depict percentage motifs containing regions exclusive to *CLIPper*$_{Rep.}$ results, orange bars depict percentage motifs containing regions exclusive to *DEWSeq* results and gray bars depict percentage of common motifs. (**A**) Comparison of POSTAR2 iCLIP binding regions from Piranha and CIMS pipelines with eCLIP *CLIPper*$_{Rep.}$ results and *DEWSeq* results. (**B**) Comparison of POSTAR2 HITS-CLIP binding regions from Piranha and CIMS pipelines with *CLIPper*$_{Rep.}$ results and *DEWSeq* results. (**C**) Comparison of PAR-CLIP binding regions from Piranha, PARalyzer and Mukherjee (46) pipelines with eCLIP *CLIPper*$_{Rep.}$ results and *DEWSeq* results. (**D**) Violin plots showing the absolute number of motif-containing regions in common either between iCLIP, HITS-CLIP, PAR-CLIP datasets analysed with CIMS, Piranha, PARalyzer or Mukherjee and ENCODE eCLIP data analysed with *CLIPper*$_{Rep.}$ and *DEWSeq*, respectively. (**E**) Violin plots showing Jaccard index values for overlap between RNAi targets and genes with binding regions for *CLIPper*$_{Orig.}$, *CLIPper*$_{Rep.}$ and *DEWSeq* respectively. Vertical dotted black line shows the median Jaccard index for *CLIPper*$_{Rep.}$ results. The *p*-value on top (between *CLIPper*$_{Rep.}$ and *DEWSeq* results) is derived from Mann–Whitney *U* test.

results and *CLIPper*$_{Orig.}$ results the overlaps were slightly smaller, ranging between 3.9% and 4.2% (Supplementary Figure S6b).

Discussion

CLIP-mapped crosslink sites for RBPs frequently fall outside of their biological binding motifs. The crosslink sites of individual-nucleotide resolution CLIP methods such as eCLIP show significant variability of site distributions and locations relative to known sequence motifs bound by RNA-binding proteins (Figure 2 and Supplementary Figure S1). Notably, RBPs display accumulation of crosslink sites either centred on the motif (e.g. RBFOX2, hnRNPC), displaced to one side (e.g. CSTF2), immediately upstream (e.g. HNRNPA1, TROVE2), or immediately downstream (e.g. U2AF2) of the motif. Some also show crosslink site enrichment surrounding the motif, but depletion directly at the associated RNA sequence motif (e.g. HNRNPL, CPEB4) (Figure 1D–F, Figure 2 and Supplementary Figure S1a, b). In the case of SLBP, which binds to histone mRNA 3′ UTR stem–loops (56), crosslink sites accumulate upstream of the binding motif locations (Figure 1C). However, the majority of CLIP protocols were primarily benchmarked on selected RBPs like RBFOX2 or hnRNPC which display peak-like behaviour on top of the known target sites, justifying the choice of peak-callers for data analysis.

To increase robustness to the observed crosslinking patterns and therefore to improve the reliability of binding site identification, we developed a computational method called *DEWSeq*, that detects enriched regions of crosslink sites in single-nucleotide resolution CLIP. Similar to *csaw* for ChIP-seq (49), *DEWSeq* takes into account biological variation between replicates for significance testing of the IP samples against size-matched input (SMI) controls. We reanalysed 223 ENCODE eCLIP datasets covering 150 RBPs in either one or both of the two cell lines (K562 and HepG2), of which 107 RBPs had known experimentally determined RNA sequence motifs and one, SLBP, is known to target a specific RNA stem–loop secondary structure (18). Using these sequence and structural motifs, we have performed, to the best of our knowledge, the most comprehensive eCLIP benchmarking study to date.

We showed that *DEWSeq*, even when operating on the minimal working requirement of two IP samples and one control sample, outperforms the single-replicate peak calling strategy of *CLIPper*$_{Orig.}$ (10) and *CLIPper*$_{Rep.}$ (15) (Figure 3B). This is the case both for the number of motif-containing binding sites detected (a median 1.8-fold or 2.3-fold improvement, respectively) and for the percentage of sites that contain a motif, which approximates the true positive rate, for the majority of RBPs in the ENCODE dataset (a median 7.1% or 3.9% improvement, respectively). *DEWSeq* discovers numerous motif-containing binding sites not found by *CLIPper*$_{Rep.}$, whereas CLIPper$_{Rep.}$ outperforms *DEWSeq* only in a handful of cases (Figure 3D, E). Overall, *DEWSeq* results increased the number of reported RBP-gene interactions 1.55-fold (median across RBPs and cell types) (Figure 3C).

We used the Rfam covariance model for SLBP's known histone 3′ UTR stem–loop structure targets (56) to estimate accuracy and sensitivity of the binding site assignments. 39.2% of *DEWSeq*-identified binding regions contain the histone stem–loop, compared to only 17.9% for *CLIPper*$_{Rep.}$. Interestingly, when searching the surrounding areas (up to 300 nt) of the crosslink sites, *DEWSeq* was able to detect >80% of all known stem–loops, whereas *CLIPper*$_{Rep.}$ levels off at ∼30%. *CLIPper*$_{Rep.}$ shows only minimal improvement even with 300 nt extensions to both sides. The far better exclusion of false positives and an improvement in detecting true positives suggests the superior potential of *DEWSeq* in identifying the target mRNA genes using secondary structure signals (Figure 4C, D). Our benchmark of *DEWSeq* parameters highlights that bigger window sizes are beneficial (Supplementary Figure S6a, b), however bigger window sizes are not the driving factor for identifying true positives in the case of SLBP (Figure 4C, D). For SLBP, we observe that the true positive rates are comparable between the methods and that extending the window around stem–loop structures does not lead to the detection of more binding sites for *CLIPper*$_{Rep.}$.

Although *DEWSeq* does have a minimal binding site length due to its window size parameter, for motif calling, *CLIPper* extends its binding sites by 50 base pairs upstream (15,36). The reasoning is that the 5′ end of the *CLIPper* peak represents UV crosslink sites between protein and RNA, implying that the actual binding motif can be upstream. Crosslink site distributions around known RNA sequence motifs and secondary structures show differences in relative positions, up- or downstream of the target site, which justifies a broader searching frame. Based on our findings presented here, we conclude that although the strategy of extending binding sites only to upstream works in general, it could also potentially lead to false interpretation of the data especially in cases where the functional motif/structure is downstream of the crosslink sites, or are unknown.

DEWSeq does consistently improve the overlap with binding sites from other CLIP protocols compared to *CLIPper*$_{Rep.}$, although the overlap across protocols is very low overall (Figure 5A–C). Our study includes a meta-analysis of binding sites generated using different protocols, methods and analysis tools. The general trend indicates that compared to iCLIP, HITS-CLIP and PAR-CLIP datasets, ENCODE eCLIP dataset analysed with *DEWSeq* contains a higher number of motif-containing binding sites. Also, binding sites discovered exclusively by *DEWSeq* can be found in other CLIP protocols, providing independent validation. However, a further large-scale investigation is needed to study the differences between CLIP-type protocols. Comparison of TARDBP (TDP-43) binding regions and motif positions to edit sites from HyperTRIBE and STAMP protocols also shows a similar trend, although the agreement between the orthogonal approaches is minimal (Supplementary Table S7). We further support this argument by computing the overlap between genes with eCLIP binding regions and RNAi targets. These results also show that compared to *CLIPper*$_{Rep.}$ and *CLIPper*$_{Orig.}$ results, *DEWSeq* results significantly improved the overlap to RNAi targets despite the fact that the overlap between eCLIP and RNAi targets was low (Figure 5E, Supplementary Table S8).

CLIP peak caller methods should perform well on datasets where the data shows a bell shaped curve, similar to ChIP-seq data. However, given the evidence provided by our analysis (Figure 1B–F, Figure 2 and Supplementary Figure S1), eCLIP data can also show a general enrichment of crosslink sites adjacent to sequence motifs. Based on this evidence, we concluded that contrary to ChIP-seq, testing for enrichments of crosslink sites (IP over SMI control) with broader sliding windows is more appropriate for the analysis of eCLIP

data. Though the underlying biology of the RNA-binding behaviour of the protein under investigation is key for understanding CLIP data, the reduction in the number of false positives and the considerable increase in the number of motif-containing binding sites provided by *DEWSeq* should help to improve functional analyses downstream. To fully capitalise on the potential of *DEWSeq* and for result reproducibility, we further highly recommend that any CLIP-type experiments should be performed with at least 3 replicates both for samples and controls, as this will drastically improve the statistical power for reliable binding site detection with *DEWSeq* in a way that standard eCLIP data processing using *CLIPper* cannot provide. *DEWSeq* is highly scalable, easy-to-use, open-source, fully documented and is designed to circumvent the limitations of the individual-nucleotide resolution CLIP protocols outlined above. Finally, based on the results shown, we strongly advise that CLIP-type protocols and analysis methods should be evaluated on RNA-binding proteins with a variety of crosslinking and binding behaviours, thereby taking into account structural and functional biological differences.

Data availability

DEWSeq along with extensive documentation is available as an R/Bioconductor package. Reported *DEWSeq* binding sites are available in the Supplementary Data. Additional files for performing analysis with *DEWSeq* are available in Zenodo at https://doi.org/10.5281/zenodo.8416672.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank Alessio Colantoni and Alexandros Armaos for providing the curated RBP motif dataset from Armaos *et al.* (2021). We thank Thileepan Sekaran for bioinformatics help. We thank Michael Uhl for creating Galaxy wrappers for *htseq-clip* and *DEWSeq*.

Funding

Thomas Schwarzl was supported by a fellowship from the EMBL Interdisciplinary Postdoc (EIPOD) programme under Marie Sklodowska-Curie Actions COFUND programme [291772]; Benjamin Lang has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 793135 [DeepRNA]; Matthias W. Hentze was generously supported by the Manfred Lautenschläger Foundation as well as by funding from the DFG [SFB1550]; Gian G. Tartaglia was supported by the ERC [ASTRA_855923]; PNRR grant from National Center for Gene Therapy and Drugs based on RNA Technology [CN00000041, EPNR-RCN3]; IVBM4PAP_101098989. Funding for open access charge: ERC [ASTRA_855923].

Conflict of interest statement

None declared.

References

1. Ule,J. and Blencowe,B.J. (2019) Alternative splicing regulatory networks: functions, mechanisms, and evolution. *Mol. Cell*, **76**, 329–345.
2. Glisovic,T., Bachorik,J.L., Yong,J. and Dreyfuss,G. (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, **582**, 1977–1986.
3. Gebauer,F., Schwarzl,T., Valcárcel,J. and Hentze,M.W. (2021) RNA-binding proteins in human genetic disease. *Nat. Rev. Genet.*, **22**, 185–198.
4. Hentze,M.W., Castello,A., Schwarzl,T. and Preiss,T. (2018) A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.*, **19**, 327–341.
5. Hafner,M., Katsantoni,M., Köster,T., Marks,J., Mukherjee,J., Staiger,D., Ule,J. and Zavolan,M. (2021) CLIP and complementary methods. *Nat. Rev. Methods Primers*, **1**, 20.
6. Ule,J., Jensen,K.B., Ruggiu,M., Mele,A., Ule,A. and Darnell,R.B. (2003) CLIP identifies nova-regulated RNA networks in the brain. *Science*, **302**, 1212–1215.
7. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X., *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
8. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.-C., Munschauer,M., *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
9. König,J., Zarnack,K., Luscombe,N.M. and Ule,J. (2012) Protein–RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.*, **13**, 77–83.
10. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C., Elkins,K., *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.
11. Zarnegar,B.J., Flynn,R.A., Shen,Y., Do,B.T., Chang,H.Y. and Khavari,P.A. (2016) irCLIP platform for efficient characterization of protein–RNA interactions. *Nat. Methods*, **13**, 489–492.
12. Van Nostrand,E.L., Nguyen,T.B., Gelboin-Burkhart,C., Wang,R., Blue,S.M., Pratt,G.A., Louie,A.L. and Yeo,G.W. (2017) Robust, cost-effective profiling of RNA binding protein targets with single-end enhanced crosslinking and immunoprecipitation (seCLIP). *Methods Mol. Biol.*, **1648**, 177–200.
13. Porter,D.F., Miao,W., Yang,X., Goda,G.A., Ji,A.L., Donohue,L.K.H., Aleman,M.M., Dominguez,D. and Khavari,P.A. (2021) easyCLIP analysis of RNA-protein interactions incorporating absolute quantification. *Nat. Commun.*, **12**, 1569.
14. Buchbender,A., Mutter,H., Sutandy,F.X.R., Körtel,N., Hänel,H., Busch,A., Ebersberger,S. and König,J. (2020) Improved library preparation with the new iCLIP2 protocol. *Methods*, **178**, 33–48.
15. Van Nostrand,E.L., Freese,P., Pratt,G.A., Wang,X., Wei,X., Xiao,R., Blue,S.M., Chen,J.-Y., Cody,N.A.L., Dominguez,D., *et al.* (2020) A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **583**, 711–719.
16. Konig,J., Zarnack,K., Rot,G., Curk,T., Kayikci,M., Zupan,B., Turner,D.J., Luscombe,N.M. and Ule,J. (2011) iCLIP - transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J. Vis. Exp.*, **50**, e2638.
17. Zarnack,K., König,J., Tajnik,M., Martincorena,I., Eustermann,S., Stévant,I., Reyes,A., Anders,S., Luscombe,N.M. and Ule,J. (2013) Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, **152**, 453–466.
18. Dominski,Z., Zheng,L.X., Sanchez,R. and Marzluff,W.F. (1999) Stem-loop binding protein facilitates 3'-end formation by stabilizing U7 snRNP binding to histone pre-mRNA. *Mol. Cell. Biol.*, **19**, 3561–3570.

19. Nourse,J., Spada,S. and Danckwardt,S. (2020) Emerging roles of RNA 3′-end cleavage and polyadenylation in pathogenesis, diagnosis and therapy of Human disorders. *Biomolecules*, **10**, 915.
20. Mackereth,C.D., Madl,T., Bonnal,S., Simon,B., Zanier,K., Gasch,A., Rybin,V., Valcárcel,J. and Sattler,M. (2011) Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature*, **475**, 408–411.
21. Smith,S.A., Ray,D., Cook,K.B., Mallory,M.J., Hughes,T.R. and Lynch,K.W. (2013) Paralogs hnRNP L and hnRNP LL exhibit overlapping but distinct RNA binding constraints. *PLoS One*, **8**, e80701.
22. Schelhorn,C., Gordon,J.M.B., Ruiz,L., Alguacil,J., Pedroso,E. and Macias,M.J. (2014) RNA recognition and self-association of CPEB4 is mediated by its tandem RRM domains. *Nucleic Acids Res.*, **42**, 10185–10195.
23. Huppertz,I., Perez-Perri,J.I., Mantas,P., Sekaran,T., Schwarzl,T., Russo,F., Ferring-Appel,D., Koskova,Z., Dimitrova-Paternoga,L., Kafkia,E., *et al.* (2022) Riboregulation of Enolase 1 activity controls glycolysis and embryonic stem cell differentiation. *Mol. Cell*, **82**, 2666–2680.
24. Hauer,C., Curk,T., Anders,S., Schwarzl,T., Alleaume,A.-M., Sieber,J., Hollerer,I., Bhuvanagiri,M., Huber,W., Hentze,M.W., *et al.* (2015) Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP. *Nat. Commun.*, **6**, 7921.
25. Hauer,C., Sieber,J., Schwarzl,T., Hollerer,I., Curk,T., Alleaume,A.-M., Hentze,M.W. and Kulozik,A.E. (2016) Exon junction complexes show a distributional bias toward alternatively spliced mRNAs and against mRNAs coding for ribosomal proteins. *Cell Rep.*, **16**, 1588–1603.
26. Van Nostrand,E.L., Pratt,G.A., Yee,B.A., Wheeler,E.C., Blue,S.M., Mueller,J., Park,S.S., Garcia,K.E., Gelboin-Burkhart,C., Nguyen,T.B., *et al.* (2020) Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol.*, **21**, 90.
27. Sloan,C.A., Chan,E.T., Davidson,J.M., Malladi,V.S., Seth Strattan,J., Hitz,B.C., Gabdank,I., Narayanan,A.K., Ho,M., Lee,B.T., *et al.* (2016) ENCODE data at the ENCODE portal. *Nucleic Acids Res.*, **44**, D726–D732.
28. Sahadevan,S., Sekaran,T., Ashaf,N., Fritz,M., Hentze,M.W., Huber,W. and Schwarzl,T. (2022) Htseq-clip: a toolset for the preprocessing of eCLIP/iCLIP datasets. *Bioinformatics*, **39**, btac747.
29. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
30. Ignatiadis,N., Klaus,B., Zaugg,J.B. and Huber,W. (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods*, **13**, 577–580.
31. Sahadevan,S., Sekaran,T. and Schwarzl,T. (2022) A pipeline for analyzing eCLIP and iCLIP data with Htseq-clip and DEWSeq. *Methods Mol. Biol.*, **2404**, 189–205.
32. Armaos,A., Colantoni,A., Proietti,G., Rupert,J. and Tartaglia,G.G. (2021) catRAPID omics v2.0: going deeper and wider in the prediction of protein–RNA interactions. *Nucleic Acids Res.*, **49**, W72–W79.
33. Tremblay,B.J. (2020) universalmotif: Import, Modify, and Export Motifs with R. https://doi.org/10.18129/B9.bioc.universalmotif.
34. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
35. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W8.
36. Dominguez,D., Freese,P., Alexis,M.S., Su,A., Hochman,M., Palden,T., Bazile,C., Lambert,N.J., Van Nostrand,E.L., Pratt,G.A., *et al.* (2018) Sequence, structure, and context preferences of Human RNA binding proteins. *Mol. Cell*, **70**, 854–867.
37. Bailey,T.L. (2021) STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, **37**, 2834–2840.
38. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
39. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
40. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Lamkiewicz,K., Marz,M., Griffiths-Jones,S., Toffano-Nioche,C., Gautheret,D., Weinberg,Z., *et al.* (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, **49**, D192–D200.
41. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
42. Hu,B., Yang,Y.-C.T., Huang,Y., Zhu,Y. and Lu,Z.J. (2017) POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.*, **45**, D104–D114.
43. Uren,P.J., Bahrami-Samani,E., Burns,S.C., Qiao,M., Karginov,F.V., Hodges,E., Hannon,G.J., Sanford,J.R., Penalva,L.O.F. and Smith,A.D. (2012) Site identification in high-throughput RNA–protein interaction data. *Bioinformatics*, **28**, 3013–3020.
44. Moore,M.J., Zhang,C., Gantman,E.C., Mele,A., Darnell,J.C. and Darnell,R.B. (2014) Mapping argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat. Protoc.*, **9**, 263–293.
45. Corcoran,D.L., Georgiev,S., Mukherjee,N., Gottwein,E., Skalsky,R.L., Keene,J.D. and Ohler,U. (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, **12**, R79.
46. Mukherjee,N., Wessels,H.-H., Lebedeva,S., Sajek,M., Ghanbari,M., Garzia,A., Munteanu,A., Yusuf,D., Farazi,T., Hoell,J.I., *et al.* (2019) Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Res.*, **47**, 570–581.
47. Abruzzi,K., Ratner,C. and Rosbash,M. (2023) Comparison of TRIBE and STAMP for identifying targets of RNA binding proteins in human and drosophila cells. *RNA*, **29**, 1230–1242.
48. Ferré,Q., Charbonnier,G., Sadouni,N., Lopez,F., Kermezli,Y., Spicuglia,S., Capponi,C., Ghattas,B. and Puthier,D. (2019) OLOGRAM: determining significance of total overlap length between genomic regions sets. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btz810.
49. Lun,A.T.L. and Smyth,G.K. (2016) csaw: a bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res.*, **44**, e45.
50. Wheeler,E.C., Van Nostrand,E.L. and Yeo,G.W. (2018) Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. *WIREs RNA*, **9**, e1436.
51. Zhu,Y., Xu,G., Yang,Y.T., Xu,Z., Chen,X., Shi,B., Xie,D., Lu,Z.J. and Wang,P. (2019) POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.*, **47**, D203–D211.
52. Kuroyanagi,H. (2009) Fox-1 family of RNA-binding proteins. *Cell. Mol. Life Sci.*, **66**, 3895–3907.
53. Diesh,C., Stevens,G.J., Xie,P., De Jesus Martinez,T., Hershberg,E.A., Leung,A., Guo,E., Dider,S., Zhang,J., Bridge,C., *et al.* (2023) JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol.*, **24**, 74.
54. Popow,J., Alleaume,A.-M., Curk,T., Schwarzl,T., Sauer,S. and Hentze,M.W. (2015) FASTKD2 is an RNA-binding protein required for mitochondrial RNA processing and translation. *RNA*, **21**, 1873–1884.
55. Jourdain,A.A., Koppen,M., Rodley,C.D., Maundrell,K., Gueguen,N., Reynier,P., Guaras,A.M., Enriquez,J.A., Anderson,P., Simarro,M., *et al.* (2015) A mitochondria-specific isoform of FASTK is present in mitochondrial RNA granules and regulates gene expression and function. *Cell Rep.*, **10**, 1110–1121.

56. Wang,Z.F., Whitfield,M.L., Ingledue,T.C. 3rd, Dominski,Z. and Marzluff,W.F. (1996) The protein that binds the 3' end of histone mRNA: a novel RNA-binding protein required for histone pre-mRNA processing. *Genes Dev.*, **10**, 3028–3040.

57. Zanier,K., Luyten,I., Crombie,C., Muller,B., Schümperli,D., Linge,J.P., Nilges,M. and Sattler,M. (2002) Structure of the histone mRNA hairpin required for cell cycle regulation of histone gene expression. *RNA*, **8**, 29–46.

Received: February 17, 2023. **Revised:** September 4, 2023. **Editorial Decision:** October 5, 2023. **Accepted:** October 18, 2023