DEWSeq addresses the pervasive disparity of crosslink site distributions in the discovery of RNA-binding protein binding sites in eCLIP data

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

* contributed equally

[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
[+] Correspondence

Keywords

RNA biology, RNA-binding proteins, eCLIP, ENCODE, iCLIP, PAR-CLIP, CLIPper, DEWSeq, htseq-clip, Sequencing, RNA motifs

Abstract

Enhanced cross-linking and immunoprecipitation (eCLIP) sequencing is a powerful approach to transcriptome-wide detect binding sites of RNA-binding proteins (RBPs). High background noise of CLIP protocols and current peak calling strategies, which compare each sequenced library to a single control, result in low replicability and high false-positives rates. Moreover, identified crosslink sites can profoundly deviate from experimentally established functional elements of well-studied RBPs. Here, we present the DEWSeq method to detect significantly enriched crosslink regions using a sliding-window approach that makes full use of replicate information and input control data. We benchmarked DEWSeq on 107 RBPs for which ENCODE produced eCLIP data and associated sequence motifs from catRAPID omics v2.0 were available. Using DEWSeq, the number of motif-containing binding sites detected increased 1.68 fold relative to standard eCLIP processing, accompanied by a slight improvement in the fraction of binding sites containing known RNA sequence motifs from 56.1% to 59.1%. In particular, DEWSeq results show strong

enrichment in known histone stem loop structures for the histone stem-loop binding protein (SLBP) and in mitochondrial targets for the mitochondrial RBP FASTKD2 compared to the standard method. Comparison with HITS-CLIP, PAR-CLIP and iCLIP data also show a consistent improvement of overlap with these orthogonal methods when using DEWSeq on eCLIP data compared to the standard data analysis approach.
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 such as known RNA sequence motifs and secondary structures. We envision that the binding site datasets produced as part of this study will be a valuable resource for the community.

Introduction

RNA-binding proteins play major roles in biological processes such as splicing (Ule and Blencowe 2019), polyadenylation, nuclear export, subcellular localisation, stabilisation/degradation, and translation (Glisovic et al. 2008). In recent years, thousands of proteins have been identified to bind to RNA (Gebauer et al. 2021; Hentze et al. 2018), many of which with unknown RNA targets. To identify RNA sites bound by an RBP of interest, several related cross-linking and immunoprecipitation (CLIP) high-throughput sequencing methods have been developed (Hafner et al. 2021). These methods exploit the phenomenon that UV light induces covalent RNA-protein crosslinks between RNA nucleotides and protein amino acids in immediate contact with each other (Ule et al. 2003). Over the years, variants of CLIP-based sequencing methods have been developed: HITS-CLIP (Licatalosi et al. 2008) sequences the cross-linked RNA fragment, PAR-CLIP detects mutations induced at the cross-link site (Hafner et al. 2010), and individual-nucleotide methods iCLIP (König et al. 2012), eCLIP (Van Nostrand et al. 2016) and further derivatives such as irCLIP (Zarnegar et al. 2016), seCLIP (Van Nostrand et al. 2017) and easyCLIP (Porter et al. 2021) optimise the protocols for reverse transcription truncations for identification of RNA-protein crosslink sites.

Enhanced CLIP (eCLIP) introduced changes to library generation and a size-matched input (SMI) control to address background noise and false positives in CLIP data (Van Nostrand et al. 2016). The ENCODE Consortium has generated the largest public set of eCLIP data,

covering 150 classical RNA-binding proteins in two cell types, processed with the computational peak-calling analysis pipeline CLIPper (Van Nostrand et al. 2016). Detected binding sites were referenced against SMI controls individually for each replicate and extended by 50 nucleotides to the 5' end for functional analyses (Van Nostrand, Freese, et al. 2020). Re-analyses of these data revealed the variable reproducibility of reported binding sites between replicates (Fig. 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) showed almost perfect reproducibility 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*$_{Rep.}$) data set was introduced, featuring only binding sites that were detected in both replicates (Van Nostrand, Freese, et al. 2020). This requirement 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 CLIP data results is the identification of crosslink peaks that are offset from the RBP binding motif. 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 (Porter et al. 2021; Konig et al. 2011; Van Nostrand et al. 2016). RBFOX2's crosslink site pattern around discovered binding sites show binding affinity towards RNA sequence motifs (Fig. 1b), similarly, hnRNPC contacts RNA in a motif- and position-dependent context, displaying bell shaped cross-link distribution around the expected binding site (Zarnack et al. 2013). 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 (Supplementary Fig. 1a,b): For example, the stem-loop binding protein (SLBP) protein, a protein which binds conserved 3'UTR stem-loop structures in histone genes (Dominski et al. 1999), shows systematic crosslink site enrichment upstream of the actual stem loop (Fig. 1c). CSTF2, known to interact with the AAUAAA polyadenylation signal (Nourse, Spada, and Danckwardt 2020), conversely displays crosslink enrichment downstream of its binding motif (Fig. 1d), while U2AF2 binds either on

top or downstream of its uridine/cytidine-rich motifs (Mackereth et al. 2011) (Fig. 1e). Other RBPs such as HNRNPL (Smith et al. 2013) (Fig. 1f), CPEB4 (Schelhorn et al. 2014) (Supplementary Fig. 1a, b) or the non-classical RBP ENO1 show enrichment of crosslink sites rather than peak shape behaviour cross link patterns (Van Nostrand et al. 2016; Huppertz et al. 2020). HNRNPL crosslink sites are even depleted over its binding motif (Fig. 1f and Supplementary Fig. 1b). Generally, the crosslink sites enrich over or in proximity of the binding motif (Supplementary Fig. 1a,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 in the study (Hauer et al. 2015, 2016). Without prior knowledge of the RBPs, such shifts in positioning and varying crosslink site accumulation behaviour can lead to misinterpretation of the binding sites.

Given the varying cross-linking behaviour of each protein and the low reproducibility of binding site detections in CLIP-sequencing experiments, we developed a method, *DEWSeq*, that allows accurate and robust identification of RBP interactions in CLIP data by detecting areas enriched in crosslink sites compared to the control. *DEWSeq* is a sliding-window-based approach that uses information across multiple replicates and control experiments for significance testing. The same approach has proven to be advantageous for transcription factor binding site analysis compared to individual replicate peak-calling (Lun and Smyth 2016). To test *DEWSeq* and to facilitate a comprehensive analysis of RBP-RNA interactions, we benchmarked it on the eCLIP data for classical RBPs provided by ENCODE. Notably, 107 out of 150 RBPs in the dataset have experimentally determined RNA sequence motifs and one, SLBP, recognises a known secondary structure motif. We used these validated RNA motifs in enriched eCLIP sequences as a proxy for the biological relevance of the sequence. This compilation represents, to the best of our knowledge, the biggest eCLIP benchmark based on known sequence motifs. We show that RNA binding regions identified by *DEWSeq* show a consistent improvement in assignments considering HITS-CLIP, iCLIP and PAR-CLIP experiments.

Results

We developed *DEWSeq* as a new R/Bioconductor statistical analysis package for the robust detection of RBP binding regions from i/eCLIP data sets. The *DEWSeq* workflow starts from the output of an accompanying Python package for post-processing *CLIP alignment files, *htseq-clip (Sahadevan et al. 2022)*, 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 significant testing using *DESeq2 (Love, Huber, and Anders 2014)*, result summarization and binding site visualisation (Fig. 1g). Similar to the *csaw* package (Lun and Smyth 2016) for ChIP-seq data, *DEWSeq* incorporates biological variation with significance testing, which reduces the false discovery rate (Lun and Smyth 2016), with the difference that *DEWSeq* is tailored to single-nucleotide position data. *DEWSeq* uses SMI controls as input control parameters compared to the previously used ranked-based metric for the specificity of binding (Wheeler, Van Nostrand, and Yeo 2018).

223 ENCODE eCLIP datasets covering 150 RBPs were processed and analysed with *htseq-clip/DEWSeq.* We compared these analyses to the published '*CLIPper original' (CLIPper$_{Orig}$)*, which features peak-calling on individual replicates and a 50 nt extension of the peaks in 5' direction introduced for functional analysis (Van Nostrand, Freese, et al. 2020), and *CLIPper$_{Rep.}$* which is a refinement of *CLIPper$_{Orig}$* across replicates, including the 50 nt extension results (Van Nostrand et al. 2016; Van Nostrand, Freese, et al. 2020). As end points for biological relevance, we utilised experimentally determined RNA sequence motifs and the known secondary structure for SLBP.

Evaluation strategy

To identify positions of known motifs within binding regions identified by *CLIPper* and *DEWSeq* from ENCODE datasets, we used FIMO (Grant, Bailey, and Noble 2011) and a curated set of motifs for 107 of the published ENCODE RBPs from *cat*RAPID omics v2.0 (Armaos et al. 2021) (Supplementary File 1). To compare these results to other orthogonal datasets, we retrieved iCLIP, HITS-CLIP, PAR-CLIP binding sites from POSTAR2 (Zhu et al. 2019), and other curated PAR-CLIP binding sites (Mukherjee et al. 2019) and repeated the motif scanning step. Fig. 2a presents an overview of the RNA sequence motif-based benchmarking strategy we adopted in this manuscript. Accuracy of binding site detection is

estimated as the proportion of motif-containing binding sites to the total reported binding sites.

Performance comparison DEWSeq/CLIPper

Around half (52.2%) of the *CLIPper*$_{Orig}$ binding sites include a known binding motif of the RBPs under investigation (Fig. 2b bottom). *CLIPper*$_{Rep.}$ slightly increases the motif-containing binding site fraction to 54.7%, but at the expense of reducing the total number of motif-containing regions from 2,326 to 1,618 median binding sites per RBP (Fig. 2b top). While DEWSeq binding sites show a motif-containing rate (58.6%) similar to *CLIPper*$_{Rep.}$, it recognizes a total number of binding sites (3,023 median) that is profoundly higher than the less stringent *CLIPper*$_{Orig.}$ set (complete results from this analysis are provided in Supplementary Table 1). Thus, *DEWSeq* achieves an increase in the number of detected binding sites without reducing the proportion of motif-containing sites, as well as without any apparent systemic bias towards gene regions or gene types (Supplementary Fig. 2). We thus posit that the detection quality is at least as good as that for the *CLIPper*$_{Rep.}$ approach, while detection rate is 1.68 fold higher.

Motif exclusiveness analysis

Following the motif-containing binding site analysis, we assessed what proportion of binding-sites that *DEWSeq* and *CLIPper* have in common and also the proportion of binding sites that are exclusive to each of the respective methods. We compared *CLIPper*$_{Rep.}$ and *DEWSeq* as the best-performing representatives for the two analysis strategies (data for all runs are provided in Supplementary Table 2, Sheet 1), and focussed on the binding sites with known motifs. In this analysis, all detected motifs 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 (Fig. 2c,d). A substantial fraction of motif-containing binding regions was detected exclusively by *DEWSeq* (median across RBP–cell type experiments: 62.1%). This greatly exceeds the fraction agreed on by both methods (28.6%) and those found exclusively by *CLIPper*$_{Rep.}$ (3.5%) (Supplementary Table 2, Sheet 1). Some RBPs, such as hnRNPC, show very good agreement of binding sites between both methods (HepG2 cell line, Fig. 2d). The fraction of binding sites containing known RNA sequence motifs is also slightly improved from a

median of 56.1% to 59.1%. Importantly, *DEWSeq* increases the number of detected motif-containing binding sites by 1.68 fold and consequently increases the discovery of RBP-RNA gene interactions from a median of 1050 to 1764 per RBP and cell type (Fig. 2e).

RBFOX2

RBFOX2 (RNA Binding Fox-1 Homolog 2) is an alternative splicing regulator that binds to UGCAUG motifs (Kuroyanagi 2009). It was used for benchmarking in the original eCLIP manuscript (Van Nostrand et al. 2016). Here, we compared the number of RBFOX2 binding regions containing the UGCAUG motif reported by both, *DEWSeq* and $CLIPper_{Rep.}$ results. Fig. 3a 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 approximately the same for both *DEWSeq* and $CLIPper_{Rep.}$ results: 56.0% and 59.1% in HepG2 cell lines and 45.6% and 45.8% 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*. In HepG2 cell line, DEWSeq reports 8,947 UGCAUG motif-containing regions compared to 4,463 motif-containing regions for *CLIPper,* and in K562 cells DEWSeq reports 4,560 motif containing regions in comparison to 1,723 from *CLIPper* (Supplementary Table 3, Sheet 1).

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 (Popow et al. 2015; Jourdain et al. 2015). 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 449 bound genes as mitochondrial, compared to *DEWSeq* with 16 out of 268 bound genes scored as mitochondrial (Fig. 3b). A similar pattern emerges in the K562 cell line, where the numbers for *CLIPper* and *DEWSeq* were 19 out of 354 and 29 out of 426, respectively. Fisher's Exact test confirmed a significant enrichment in the number of FASTKD2-bound mitochondrial genes in *DEWSeq* compared to *CLIPper* results in both cell lines (Fig. 3b, Supplementary Table 3, Sheet 2).

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 (Wang et al.

1996; Glisovic et al. 2008; Zanier et al. 2002). 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 Infernal suite and Histone 3' UTR stem-loop covariance model (Rfam: RF00032).

A higher proportion of *DEWSeq* binding sites (39.2%) contain predicted SLBP binding structures, as compared to *CLIPper*$_{Rep.}$ binding regions (17.9%) (Fig. 3c). This trend becomes more pronounced with the extension of binding regions in both 5' and 3' directions (Fig. 3c and Supplementary Fig. 3), 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 3' UTR stem-loop annotations from the Rfam database as a reference set (Fig. 3d). *DEWSeq* without extension shows a marginal increase in true positive rate compared to *CLIPper*$_{Rep.}$ (0.201 vs 0.208), with slight improvement in extensions (Supplementary Table 4).

Comparison of eCLIP to iCLIP, HITS-CLIP and PAR-CLIP

To validate 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 (Zhu et al. 2019) and DoRiNA (Mukherjee et al. 2019) databases. To address differences in detection methods for these different CLIP protocols, we employed multiple established analysis methods: Piranha (Uren et al. 2012) and CIMS (Moore et al. 2014) for iCLIP and HITS-CLIP data, and Piranha (Uren et al. 2012), PARalyzer(Corcoran et al. 2011) as well as Mukherjee (Mukherjee et al. 2019) for PAR-CLIP data.

Fig. 4a-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*$_{Rep.}$ (Fig. 4a-c, Supplementary Table 2, 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 binding-sites found by other methods when compared to *CLIPper*$_{Rep.}$ (Fig. 4d).

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 accumulation behaviour and location relative to known RNA-binding sites (Supplementary Fig. 1). Notably, RBPs display accumulation of crosslink sites centred (e.g. RBFOX2, hnRNPC), one-sided (e.g. CSTF2), upstream (e.g. HNRNPA1, TROVE2), downstream (e.g. U2AF2) or even enrichment around but depletion of the crosslink sites around the associated RNA sequence motif (e.g. HNRNPL, CPEB4) (Fig. 1d-f and Supplementary Fig. 1a,b). In the case of SLBP, which binds to 3' UTR histone stem loops (Wang et al. 1996), crosslink sites accumulate upstream of the binding motif locations (Fig. 1c). For these reasons, the majority of CLIP protocols were benchmarked on selected RBPs like RBFOX2 or hnRNPC, which do show peak-like behaviour on top of the known target sites, opting for peak-callers for analysis.

To improve the reliable identification of binding sites, we developed a computational method that detects enriched regions of crosslink sites in single-nucleotide resolution CLIP, *DEWSeq*. Similar to *csaw* for ChIP-seq (Lun and Smyth 2016), *DEWSeq* takes into account biological variability between replicates for significance testing the IP samples against SMI input controls. We reanalyzed 223 ENCODE eCLIP data sets from 150 RBPs, of which 107 have experimentally determined RNA sequence motifs and one, SLBP, targets a known RNA secondary structure (Dominski et al. 1999). We have created, to the best of our knowledge, the biggest eCLIP benchmark data set using known sequence motif targets.

We show that *DEWSeq*, even when operating on the minimal working requirement of two IP samples and one input control, outperforms the single replicate peak calling strategy of *CLIPper*, *CLIPper*$_{Orig}$ (Van Nostrand et al. 2016) and *CLIPper*$_{Rep.}$ (Van Nostrand, Freese, et al. 2020), with regard to both the number of binding sites and the percentage of motif-containing sites (which approximates true positive rate) for the majority of RBPs in the ENCODE data set. *DEWSeq* discovers numerous binding sites not found by *CLIPper*$_{Rep.}$,

whereas CLIPper outperforms *DEWSeq* only in a handful of cases. Overall, *DEWSeq* increases the number of reported RBP-gene interactions.

We used the RFAM covariance model for SLBP's known histone 3' UTR stem-loop structure targets (Wang et al. 1996) to estimate accuracy and sensitivity of the binding site assignments. 39.2% of *DEWSeq*-identified binding regions contain the histone stem loop, compared to 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 loops, whereas *CLIPper* levels off at ~30%. *CLIPper* shows only minimal improvement even with 300 nt extensions to both sides. This suggests the superior potential of DEWSeq in identifying the target mRNA genes using secondary structure signals as a reliable detection approach orthogonal to the sequencing information. Our benchmark of *DEWSeq* parameters highlights that bigger window sizes are beneficial. For SLBP, true positive rates are comparable between the methods and that the length of binding sites is not the primary driving factor for detecting binding sites.

Although *DEWSeq* does have minimal binding site length due to its selection of window size, for motif calling, *CLIPper* extends its binding sites by 50 base pairs upstream (Dominguez et al. 2018; Van Nostrand, Freese, et al. 2020). 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 structure show differences in relative positions, up- or downstream of the target site, which justifies a broader searching frame.

DEWSeq does consistently improve the overlap with binding sites of other CLIP protocols compared to $CLIPper_{Rep.}$, although the overlap across protocols is alarmingly low overall. This study is a meta analysis of binding sites, generated by different protocols, methods and analysis tools. The general trend indicates that compared to iCLIP, HITS-CLIP and PAR-CLIP datasets, ENCODE eCLIP dataset contains a higher number of motif-containing binding sites. A further large-scale investigation is needed to study the differences between CLIP type protocols. Additionally, the binding sites discovered by *DEWSeq* are proportionally found in other CLIP protocols, providing independent validation. Noticeably, our meta analysis is only focused on sequence motifs containing binding-sites.

When, like in ChIP-seq, bell-shape-like patterns are displayed in eCLIP or iCLIP raw data, peak-callers should perform well. However, given the evidence provided by our

analysis, eCLIP data can also show a general enrichment of crosslink sites around sequence motifs. Given the evidence for crosslink site patterns for well-known RBPs, and 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. Without knowing the underlying biology of the RNA-binding for the protein under investigation, the increased true-positive rate of *DEWSeq* will help with functional analysis downstream. We highly recommend for any CLIP type experiments to have at least 3 replicates for sample and controls as this will drastically improve binding site detection with *DEWSeq*. *DEWSeq* is extremely scalable, easy-to-use, open-source, well documented, and is designed to work with limitations of the individual-nucleotide resolution CLIP protocols.

Finally, we would strongly advise that CLIP type protocols and analysis methods would be benchmarked on proteins with different crosslinking behaviour, taking into account biological variation.

Materials and Methods

eCLIP data

In order to compare the results from our newly developed *htseq-clip/DEWSeq* packages, we chose the eCLIP data (Van Nostrand et al. 2016) published by the ENCODE Project (ENCODE Project Consortium 2012). 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 (Van Nostrand, Pratt, et al. 2020). For data reanalysis using *DEWSeq*, we downloaded alignment files (.bam) with reads mapped to the GRCh38 genome annotations from ENCODE project data portal (Sloan et al. 2016). Bam file accessions and additional details are given in Supplementary Table 5.

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 of called on individual IP samples with respect to SMI controls (Van Nostrand et al. 2016) (referred to as *'CLIPper original' (CLIPper$_{Orig}$)* in this manuscript) and the refined set of binding sites based on stringent

thresholds, IDR analysis and blacklist region removal (Van Nostrand, Freese, et al. 2020) (referred to as '*CLIPper reproducible*' (*CLIPper*$_{Rep.}$) in this manuscript). 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(Sahadevan et al. 2022)* 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 and 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* to analyse the crosslink site count matrix from *htseq-clip* and to test for differential enrichment in IP samples in comparison with SMI samples. For statistical testing, *DEWSeq* utilises *DESeq2 (Love, Huber, and Anders 2014)*, a well known R/Biocondutor package primarily used for the analysis of differentially expressed genes in RNA-seq data. After *DESeq2* initial pruning, normalisation and dispersion estimation, *DEWSeq* uses a custom one-tailed test for detecting significant crosslink regions enriched in IP samples over SMI, followed by two multiple hypothesis correction steps. In the first 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) (Ignatiadis et al. 2016). Finally, all enriched windows passing user specified enrichment thresholds are merged into regions. *DEWSeq* is available as an open-source R/Bioconductor package (https://bioconductor.org/packages/release/bioc/vignettes/DEWSeq/inst/doc/DEWSeq.html). (Sahadevan, Sekaran, and Schwarzl 2022) describes a sample pipeline for the analysis of eCLIP/iCLIP data using *htseq-clip* and *DEWSeq*.

In this study, 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 (Love, Huber, and Anders 2014) 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: either correct for dependencies between overlapping windows using Bonferroni correction or using 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 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 (Supplementary Fig. 4) and parameters 100 nucleotide 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"*.

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, whose authors collected and curated motifs from a comprehensive set of sources including ATtRACT, cisBP-RNA, mCrossBase, oRNAment and RBPmap (Armaos et al. 2021). Supplementary Table. 5, 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 (Tremblay 2020). 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 (Grant, Bailey, and Noble 2011) from the MEME Suite (Bailey et al. 2009) 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 described by Nostrand *et al.* (Dominguez et al. 2018; Van Nostrand, Freese, et al. 2020). 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.

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 cmsearch tool from Infernal suite (Nawrocki, Kolbe, and Eddy 2009) and covariance models from Rfam database (Kalvari et al. 2021) to scan for secondary structures in enriched regions from both *CLIPper*$_{Rep.}$ and *DEWSeq* results. In the first step, we used *mergeBed* from *bedtools* suite (Quinlan and Hall 2010) to merge overlapping regions or regions separated by a maximum of 10 nucleotides in *CLIPper*$_{Rep.}$ SLBP results. In case of *DEWSeq* SLBP regions, no such merging step was necessary. In the next step, we sequentially extended the length of both DEWSeq and ENCODE regions by up to 300 nt (50, 100, 150, 200, 300) using *slopBed* tool from the bedtool suite. With this extension step, we aim 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 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 bash *awk* command to obtain the genomic location of the model hits and extracted unique hits based on the genomic coordinates. The reference set of Histone 3' UTR stem-loop regions were retrieved from the Rfam database (Rfam id: RF00032).

Comparison to orthogonal datasets

We retrieved iCLIP (Konig et al. 2011), HITS-CLIP (Licatalosi et al. 2008), and PAR-CLIP (Hafner et al. 2010) interaction datasets from the POSTAR2 database (previously known as CLIPdb) (Hu et al. 2017). For each CLIP experiment type, POSTAR2 provided peak calling results from *Piranha* (Uren et al. 2012), as well as the more specialised *CIMS* (cross-link-induced mutation sites) (Moore et al. 2014) and PARalyzer data analysis pipelines (Corcoran et al. 2011). 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 (Mukherjee et al. 2019), 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 prioritised Piranha in our figures since it had been applied to all three experimental methods in POSTAR2, providing more directly comparable results. 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 *DEWSeq* and *CLIPper*$_{Rep.}$ results.

Comparing gene region and gene type enrichment using OLOGRAM

To determine whether *DEWSeq* or *CLIPper* results were biassed towards gene regions (5' UTR, exon, 3' UTR) or gene types (e.g. protein coding RNAs, non coding RNA, mtRNA, ...) we performed gene region and gene type enrichment analysis in both results using OLOGRAM (Ferré et al. 2019). 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 HepG2 (ENCODE accession: ENCFF533XPJ, ENCFF321JIT) and K562 (ENCODE accession: ENCFF286GLL, ENCFF986DBN) cell lines available from 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.

Availability

DEWSeq is available as an R/Bioconductor package. Reported *DEWSeq* binding sites are available as Supplementary File 2.

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 and advice.

Figure Legends

Figure 1a. Reproducibility of binding sites between ENCODE eCLIP data sets 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 data set ENCSR756CKJ (K562) and ENCSR987FTF (HepG2) relative to known RBFOX2 UGCAUG motif. **c.** Crosslink site distribution of SLBP eCLIP data set relative to known 3' UTR histone stem-loop (ENCODE eCLIP data set ENCSR483NOP, K562 cell line). **d.** CSTF2 crosslink site distribution for ENCODE eCLIP data set ENCSR384MWO (HepG2 cell line) relative to known AAUAAA polyadenylation signal. **e.** U2AF2 crosslink site distribution for ENCODE eCLIP data set ENCSR202BFN (HepG2 cell line) relative to uridine/cytidine-rich motifs. **f.** HNRNPL crosslink site distribution for ENCODE eCLIP data set ENCSR724RDN (HepG2 cell line) relative to CA repeat motifs. **g.** Overview of DEWSeq workflow

Figure 2. Overview and results for benchmarking workflow on ENCODE eCLIP data sets for proteins with known RNA sequence motifs. a. ENCODE eCLIP data sets were reanalysed with *DEWSeq* and was compared to '*CLIPper original*' ($CLIPper_{Orig.}$) (Armaos et al. 2021; Van Nostrand et al. 2016) and '*CLIPper reproducible*' ($CLIPper_{Rep.}$) data set (Van Nostrand, Freese, et al. 2020) analysis. Other CLIP binding sites from iCLIP, HITS-CLIP and PAR-CLIP were extracted from POSTAR2. Additional PAR-CLIP data sets from DoRiNA (Mukherjee et al. 2019) were included in the analysis. Binding sites from all data-sets were analysed with FIMO (Grant, Bailey, and Noble 2011) using known RNA sequence motifs from catRAPID omics v2.0 (Armaos et al. 2021). **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 total number of binding sites for each method. **c.** Exclusiveness

of motif-containing binding sites for data sets with known motifs. Left shows binding sites exclusive for *CLIPper*$_{Rep.}$ data set, the middle the binding-sites which were found in both, and right for sites found exclusively in *DEWSeq*. **d.** Heatmap of exclusiveness and overlap for motif containing *CLIPper*$_{Rep.}$ and *DEWSeq* binding sites. **e.** Number of reported RBP-gene interactions.

Figure 3. Binding site comparisons. a. Comparison of RBFOX2 binding regions and RBFOX2 binding regions containing TGCAGT 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 3, 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 3' UTR stem-loop regions retrieved from Rfam database (for Rfam id: RF00032).

Figure 4. 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 limited only to the common RBPs (in ENCODE dataset and these methods) with motifs in catRAPID omics v2.0 motifs. Stacked bar plots (plots: **a-c**) show comparisons of iCLIP, HITS-CLIP and PAR-CLIP motif containing regions (in percentage) against *CLIPper*$_{Rep.}$ and *DEWSeq* 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. Left panels show comparisons to *CLIPper*$_{Rep.}$ results and right panels show comparisons to *DEWSeq* results. **b.** comparison of POSTAR2 HITS-CLIP binding regions from Piranha and CIMS pipelines with *CLIPper*$_{Rep.}$ results and *DEWSeq* results. Left panels show comparisons to *CLIPper*$_{Rep.}$ results and right panels show comparisons to *DEWSeq* results. **c.** comparison of PAR-CLIP binding regions from Piranha, PARalyzer and (Mukherjee et al. 2019) pipelines with eCLIP *CLIPper*$_{Rep.}$

results and *DEWSeq* results. Left panels show comparisons to *CLIPper*$_{Rep.}$ results and right panels show comparisons to *DEWSeq* results. **d.** Violin plots showing the absolute number of motif-containing regions in common either between iCLIP, HITS-CLIP, PAR-CLIP datasets and *CLIPper*$_{Rep.}$ results or between iCLIP, HITS-CLIP, PAR-CLIP datasets and DEWSeq results. The dotted horizontal line indicates the median value for comparisons between iCLIP, HITS-CLIP, PAR-CLIP datasets and *CLIPper*$_{Rep.}$ results.

Supplementary Material

Supplementary Table 1. Results of RNA sequence motif benchmark on RNA-binding protein binding site detections methods

Supplementary Table 2. Results of motif exclusiveness analysis, comparison to iCLIP, HITS-CLIP and PAR-CLIP datasets (Sheet 1), Overlap of iCLIP, HITS-CLIP, PAR-CLIP datasets from POSTAR with ENCODE eCLIP (Sheet 2)

Supplementary Table 3. RBFOX2 UGCAUG motif counts (Sheet 1), Enrichment analysis results across all chromosomes for all FASTKD2 samples (Sheet 2)

Supplementary Table 4. Analysis of histone stem-loop found in SLBP eCLIP binding sites and extensions to 5' and 3' end.

Supplementary Table 5. Details of ENCODE eCLIP data used (Sheet 1). Total number of RBPs per motif length in the data source and the number of RBPs in common with the ENCODE dataset (Sheet 2).

Supplementary File 1: RBP motifs in .meme file format and as motif logos.

Supplementary File 2: DEWSeq enriched binding regions in narrowPeak format

Supplementary Figure 1. Heatmaps of crosslink sites around experimentally derived RNA sequence motifs. Each row displays a motif for an RNA-binding protein. **a.** clustered with WARD2 to group similar binding patterns and **b.** clustered with seriation to reveal the broader structure of crosslinking sites around motifs.

Supplementary Figure 2. Functional analysis of **a.** gene feature or **b.** gene type enrichment of 'CLIPper reproducible' (*CLIPper*$_{Rep.}$) and **b.** DEWSeq binding sites with OLOGRAMS.

Supplementary Figure 3. SLBP eCLIP binding-sites detected by **a.** 'CLIPper reproducible' and **b.** DEWSeq positions respective to the known 3' histone mRNA stem loop targets

Supplementary Figure 4. Test of different DEWSeq parameters **a.** in combination **b.** for robustness

References

Armaos, Alexandros, Alessio Colantoni, Gabriele Proietti, Jakob Rupert, and Gian Gaetano Tartaglia. 2021. "catRAPID Omics v2.0: Going Deeper and Wider in the Prediction of protein–RNA Interactions." *Nucleic Acids Research* 49 (W1): W72–79.

Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. "MEME SUITE: Tools for Motif Discovery and Searching." *Nucleic Acids Research* 37 (Web Server issue): W202–8.

Corcoran, David L., Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L. Skalsky, Jack D. Keene, and Uwe Ohler. 2011. "PARalyzer: Definition of RNA Binding Sites from PAR-CLIP Short-Read Sequence Data." *Genome Biology* 12 (8): R79.

Dominguez, Daniel, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, et al. 2018. "Sequence, Structure, and Context Preferences of Human RNA Binding Proteins." *Molecular Cell* 70 (5): 854–67.e9.

Dominski, Z., L. X. Zheng, R. Sanchez, and W. F. Marzluff. 1999. "Stem-Loop Binding Protein Facilitates 3'-End Formation by Stabilizing U7 snRNP Binding to Histone Pre-mRNA." *Molecular and Cellular Biology* 19 (5): 3561–70.

ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74.

Ferré, Q., G. Charbonnier, N. Sadouni, F. Lopez, Y. Kermezli, S. Spicuglia, C. Capponi, B. Ghattas, and D. Puthier. 2019. "OLOGRAM: Determining Significance of Total Overlap Length between Genomic Regions Sets." *Bioinformatics* , November. https://doi.org/10.1093/bioinformatics/btz810.

Gebauer, Fátima, Thomas Schwarzl, Juan Valcárcel, and Matthias W. Hentze. 2021. "RNA-Binding Proteins in Human Genetic Disease." *Nature Reviews. Genetics* 22 (3): 185–98.

Glisovic, Tina, Jennifer L. Bachorik, Jeongsik Yong, and Gideon Dreyfuss. 2008. "RNA-Binding Proteins and Post-Transcriptional Gene Regulation." *FEBS Letters*. https://doi.org/10.1016/j.febslet.2008.03.004.

Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. 2011. "FIMO: Scanning for Occurrences of a given Motif." *Bioinformatics* 27 (7): 1017–18.

Hafner, Markus, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. 2021. "CLIP and Complementary Methods." *Nature Reviews Methods Primers* 1 (1): 1–23.

Hafner, Markus, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, et al. 2010. "Transcriptome-Wide Identification of RNA-Binding Protein and microRNA Target Sites by PAR-CLIP." *Cell* 141 (1): 129–41.

Hauer, Christian, Tomaz Curk, Simon Anders, Thomas Schwarzl, Anne-Marie Alleaume, Jana Sieber, Ina Hollerer, et al. 2015. "Improved Binding Site Assignment by High-Resolution Mapping of RNA-Protein Interactions Using iCLIP." *Nature Communications* 6 (August): 7921.

Hauer, Christian, Jana Sieber, Thomas Schwarzl, Ina Hollerer, Tomaz Curk, Anne-Marie Alleaume, Matthias W. Hentze, and Andreas E. Kulozik. 2016. "Exon Junction Complexes Show a Distributional Bias toward Alternatively Spliced mRNAs and against mRNAs Coding for Ribosomal Proteins." *Cell Reports* 16 (6): 1588–1603.

Hentze, Matthias W., Alfredo Castello, Thomas Schwarzl, and Thomas Preiss. 2018. "A Brave New World of RNA-Binding Proteins." *Nature Reviews. Molecular Cell Biology* 19 (5): 327–41.

Hu, Boqin, Yu-Cheng T. Yang, Yiming Huang, Yumin Zhu, and Zhi John Lu. 2017. "POSTAR: A Platform for Exploring Post-Transcriptional Regulation Coordinated by RNA-Binding Proteins." *Nucleic Acids Research* 45 (D1): D104–14.

Huppertz, Ina, Joel I. Perez-Perri, Panagiotis Mantas, Thileepan Sekaran, Thomas Schwarzl, Lyudmila Dimitrova-Paternoga, Janosch Hennig, Pierre A. Neveu, and Matthias W. Hentze. 2020. "RNA Regulates Glycolysis and Embryonic Stem Cell Differentiation via Enolase 1." *Cold Spring Harbor Laboratory*.

https://doi.org/10.1101/2020.10.14.337444.
Ignatiadis, Nikolaos, Bernd Klaus, Judith B. Zaugg, and Wolfgang Huber. 2016. “Data-Driven Hypothesis Weighting Increases Detection Power in Genome-Scale Multiple Testing.” *Nature Methods* 13 (7): 577–80.
Jourdain, Alexis A., Mirko Koppen, Christopher D. Rodley, Kinsey Maundrell, Naïg Gueguen, Pascal Reynier, Adela M. Guaras, et al. 2015. “A Mitochondria-Specific Isoform of FASTK Is Present in Mitochondrial RNA Granules and Regulates Gene Expression and Function.” *Cell Reports* 10 (7): 1110–21.
Kalvari, Ioanna, Eric P. Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, et al. 2021. “Rfam 14: Expanded Coverage of Metagenomic, Viral and microRNA Families.” *Nucleic Acids Research* 49 (D1): D192–200.
König, Julian, Kathi Zarnack, Nicholas M. Luscombe, and Jernej Ule. 2012. “Protein–RNA Interactions: New Genomic Technologies and Perspectives.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3141.
Konig, Julian, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J. Turner, Nicholas M. Luscombe, and Jernej Ule. 2011. “iCLIP - Transcriptome-Wide Mapping of Protein-RNA Interactions with Individual Nucleotide Resolution.” *Journal of Visualized Experiments*. https://doi.org/10.3791/2638.
Kuroyanagi, Hidehito. 2009. “Fox-1 Family of RNA-Binding Proteins.” *Cellular and Molecular Life Sciences: CMLS* 66 (24): 3895–3907.
Licatalosi, Donny D., Aldo Mele, John J. Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A. Clark, et al. 2008. “HITS-CLIP Yields Genome-Wide Insights into Brain Alternative RNA Processing.” *Nature* 456 (7221): 464–69.
Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.
Lun, Aaron T. L., and Gordon K. Smyth. 2016. “Csaw: A Bioconductor Package for Differential Binding Analysis of ChIP-Seq Data Using Sliding Windows.” *Nucleic Acids Research* 44 (5): e45.
Mackereth, Cameron D., Tobias Madl, Sophie Bonnal, Bernd Simon, Katia Zanier, Alexander Gasch, Vladimir Rybin, Juan Valcárcel, and Michael Sattler. 2011. “Multi-Domain Conformational Selection Underlies Pre-mRNA Splicing Regulation by U2AF.” *Nature* 475 (7356): 408–11.
Moore, Michael J., Chaolin Zhang, Emily Conn Gantman, Aldo Mele, Jennifer C. Darnell, and Robert B. Darnell. 2014. “Mapping Argonaute and Conventional RNA-Binding Protein Interactions with RNA at Single-Nucleotide Resolution Using HITS-CLIP and CIMS Analysis.” *Nature Protocols*. https://doi.org/10.1038/nprot.2014.012.
Mukherjee, Neelanjan, Hans-Hermann Wessels, Svetlana Lebedeva, Marcin Sajek, Mahsa Ghanbari, Aitor Garzia, Alina Munteanu, et al. 2019. “Deciphering Human Ribonucleoprotein Regulatory Networks.” *Nucleic Acids Research* 47 (2): 570–81.
Nawrocki, Eric P., Diana L. Kolbe, and Sean R. Eddy. 2009. “Infernal 1.0: Inference of RNA Alignments.” *Bioinformatics* 25 (10): 1335–37.
Nourse, Jamie, Stefano Spada, and Sven Danckwardt. 2020. “Emerging Roles of RNA 3′-End Cleavage and Polyadenylation in Pathogenesis, Diagnosis and Therapy of Human Disorders.” *Biomolecules*. https://doi.org/10.3390/biom10060915.
Popow, Johannes, Anne-Marie Alleaume, Tomaz Curk, Thomas Schwarzl, Sven Sauer, and Matthias W. Hentze. 2015. “FASTKD2 Is an RNA-Binding Protein Required for Mitochondrial RNA Processing and Translation.” *RNA* 21 (11): 1873–84.
Porter, Douglas F., Weili Miao, Xue Yang, Grant A. Goda, Andrew L. Ji, Laura K. H. Donohue, Maria M. Aleman, Daniel Dominguez, and Paul A. Khavari. 2021. “easyCLIP Analysis of RNA-Protein Interactions Incorporating Absolute Quantification.” *Nature Communications* 12 (1): 1569.
Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42.
Sahadevan, Sudeep, Thileepan Sekaran, Nadia Ashaf, Marko Fritz, Matthias W. Hentze, Wolfgang Huber, and Thomas Schwarzl. 2022. “Htseq-Clip: A Toolset for the Preprocessing of eCLIP/iCLIP Datasets.” *Bioinformatics* , November. https://doi.org/10.1093/bioinformatics/btac747.
Sahadevan, Sudeep, Thileepan Sekaran, and Thomas Schwarzl. 2022. “A Pipeline for Analyzing eCLIP and iCLIP Data with Htseq-Clip and DEWSeq.” *Methods in Molecular Biology* 2404: 189–205.
Schelhorn, Constanze, James M. B. Gordon, Lidia Ruiz, Javier Alguacil, Enrique Pedroso, and Maria J. Macias. 2014. “RNA Recognition and Self-Association of CPEB4 Is Mediated by Its Tandem RRM Domains.” *Nucleic Acids Research* 42 (15): 10185–95.
Sloan, Cricket A., Esther T. Chan, Jean M. Davidson, Venkat S. Malladi, J. Seth Strattan, Benjamin C. Hitz, Idan Gabdank, et al. 2016. “ENCODE Data at the ENCODE Portal.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1160.
Smith, Sarah A., Debashish Ray, Kate B. Cook, Michael J. Mallory, Timothy R. Hughes, and Kristen W. Lynch. 2013. “Paralogs hnRNP L and hnRNP LL Exhibit Overlapping but Distinct RNA Binding Constraints.” *PloS*

One 8 (11): e80701.
Tremblay, B. J. M. 2020. “Universalmotif: Import.” *Modify, and Export Motifs with R*.
Ule, Jernej, and Benjamin J. Blencowe. 2019. “Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution.” *Molecular Cell* 76 (2): 329–45.
Ule, Jernej, Kirk B. Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B. Darnell. 2003. “CLIP Identifies Nova-Regulated RNA Networks in the Brain.” *Science* 302 (5648): 1212–15.
Uren, Philip J., Emad Bahrami-Samani, Suzanne C. Burns, Mei Qiao, Fedor V. Karginov, Emily Hodges, Gregory J. Hannon, Jeremy R. Sanford, Luiz O. F. Penalva, and Andrew D. Smith. 2012. “Site Identification in High-Throughput RNA–protein Interaction Data.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts569.
Van Nostrand, Eric L., Peter Freese, Gabriel A. Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M. Blue, et al. 2020. “A Large-Scale Binding and Functional Map of Human RNA-Binding Proteins.” *Nature* 583 (7818): 711–19.
Van Nostrand, Eric L., Thai B. Nguyen, Chelsea Gelboin-Burkhart, Ruth Wang, Steven M. Blue, Gabriel A. Pratt, Ashley L. Louie, and Gene W. Yeo. 2017. “Robust, Cost-Effective Profiling of RNA Binding Protein Targets with Single-End Enhanced Crosslinking and Immunoprecipitation (seCLIP).” *Methods in Molecular Biology* 1648: 177–200.
Van Nostrand, Eric L., Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Y. Fang, Balaji Sundararaman, Steven M. Blue, et al. 2016. “Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (eCLIP).” *Nature Methods* 13 (6): 508–14.
Van Nostrand, Eric L., Gabriel A. Pratt, Brian A. Yee, Emily C. Wheeler, Steven M. Blue, Jasmine Mueller, Samuel S. Park, et al. 2020. “Principles of RNA Processing from Analysis of Enhanced CLIP Maps for 150 RNA Binding Proteins.” *Genome Biology* 21 (1): 90.
Wang, Z. F., M. L. Whitfield, T. C. Ingledue 3rd, Z. Dominski, and W. F. Marzluff. 1996. “The Protein That Binds the 3’ End of Histone mRNA: A Novel RNA-Binding Protein Required for Histone Pre-mRNA Processing.” *Genes & Development* 10 (23): 3028–40.
Wheeler, Emily C., Eric L. Van Nostrand, and Gene W. Yeo. 2018. “Advances and Challenges in the Detection of Transcriptome-Wide Protein-RNA Interactions.” *Wiley Interdisciplinary Reviews. RNA* 9 (1). https://doi.org/10.1002/wrna.1436.
Zanier, Katia, Ingrid Luyten, Catriona Crombie, Berndt Muller, Daniel Schümperli, Jens P. Linge, Michael Nilges, and Michael Sattler. 2002. “Structure of the Histone mRNA Hairpin Required for Cell Cycle Regulation of Histone Gene Expression.” *RNA* 8 (1): 29–46.
Zarnack, Kathi, Julian König, Mojca Tajnik, Iñigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, Nicholas M. Luscombe, and Jernej Ule. 2013. “Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements.” *Cell* 152 (3): 453–66.
Zarnegar, Brian J., Ryan A. Flynn, Ying Shen, Brian T. Do, Howard Y. Chang, and Paul A. Khavari. 2016. “irCLIP Platform for Efficient Characterization of protein–RNA Interactions.” *Nature Methods* 13 (6): 489–92.
Zhu, Yumin, Gang Xu, Yucheng T. Yang, Zhiyu Xu, Xinduo Chen, Binbin Shi, Daoxin Xie, Zhi John Lu, and Pengyuan Wang. 2019. “POSTAR2: Deciphering the Post-Transcriptional Regulatory Logics.” *Nucleic Acids Research* 47 (D1): D203–11.

Abbreviations

adj	adjusted
Orig.	original
IP	immunoprecipitated
CIMS	cross-link-induced mutation sites
CLIP	cross-linking immunoprecipitation
nt	nucleotide(s)

PAR	photoactivatable ribonucleoside enhanced
Rep.	reproducible
SMI	size-matched input
TPM	tags per million
UTR	untranslated region

Figure 1

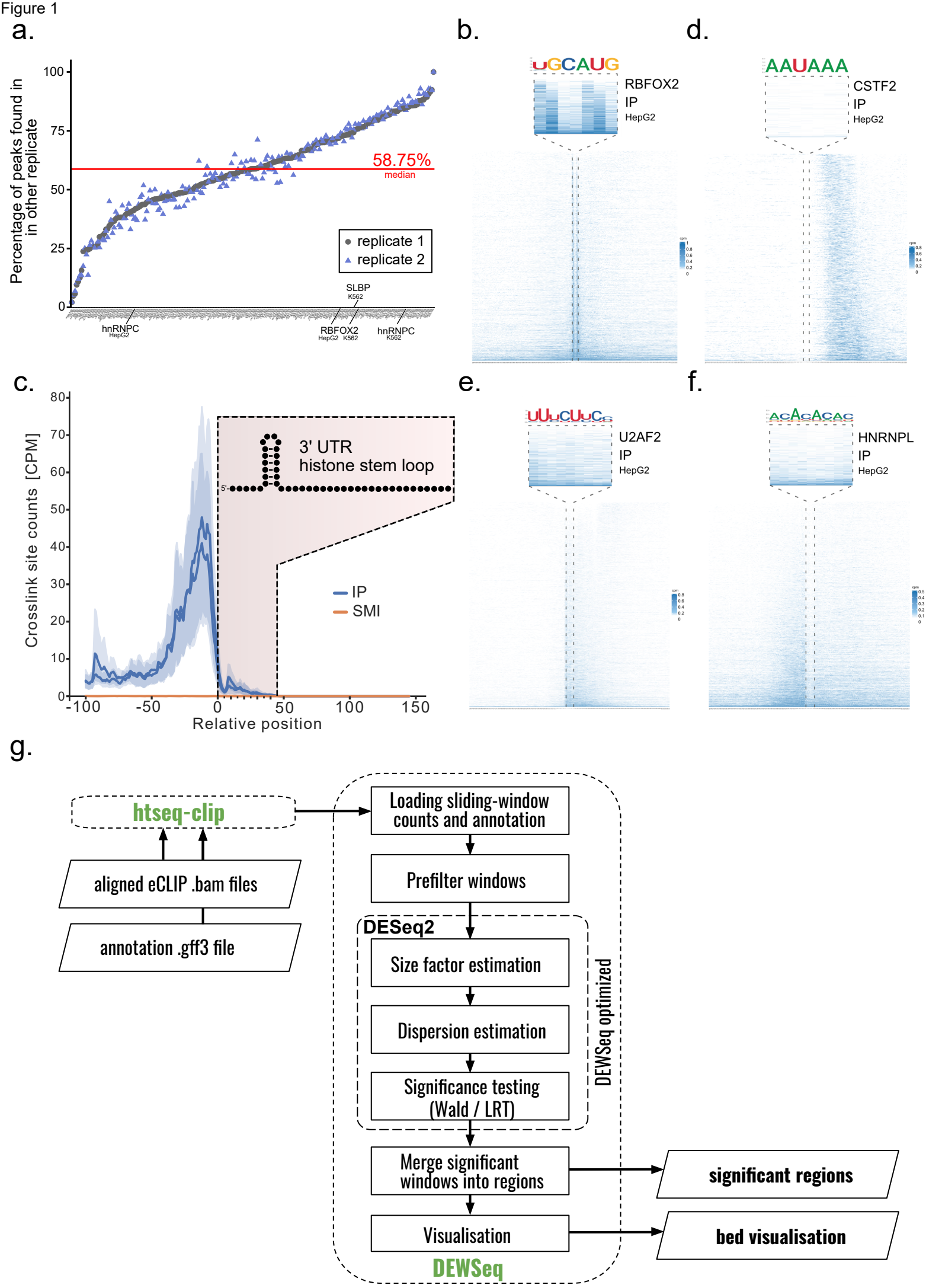
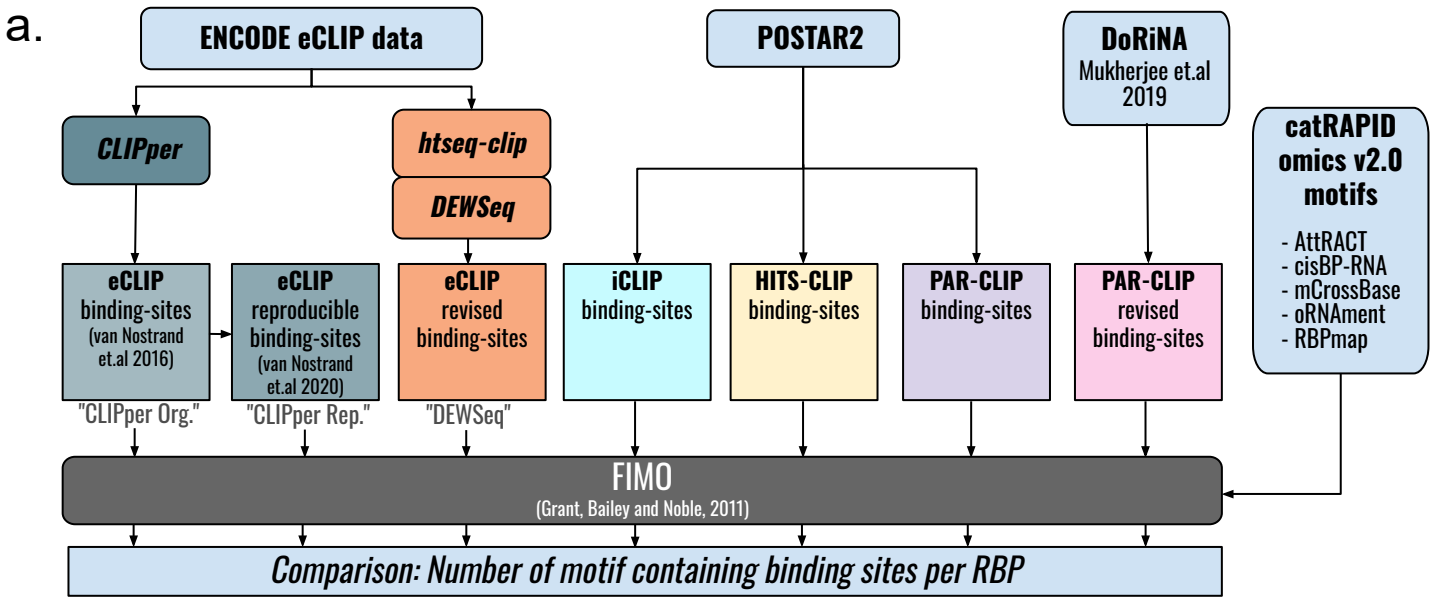
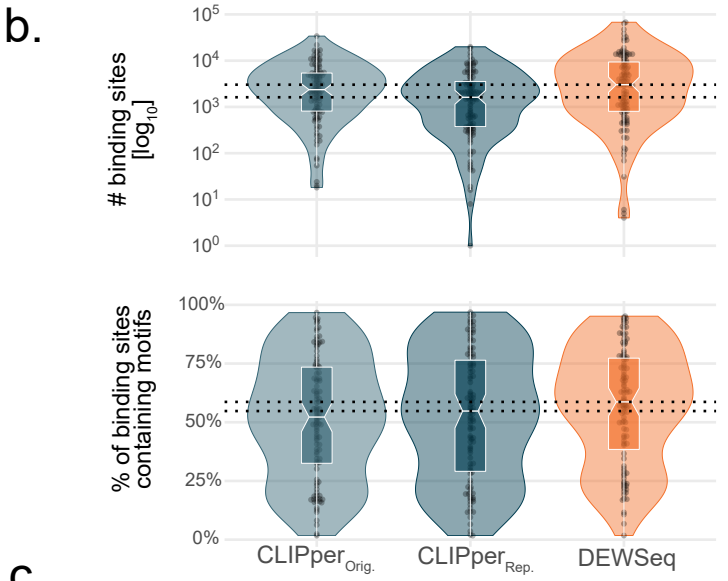
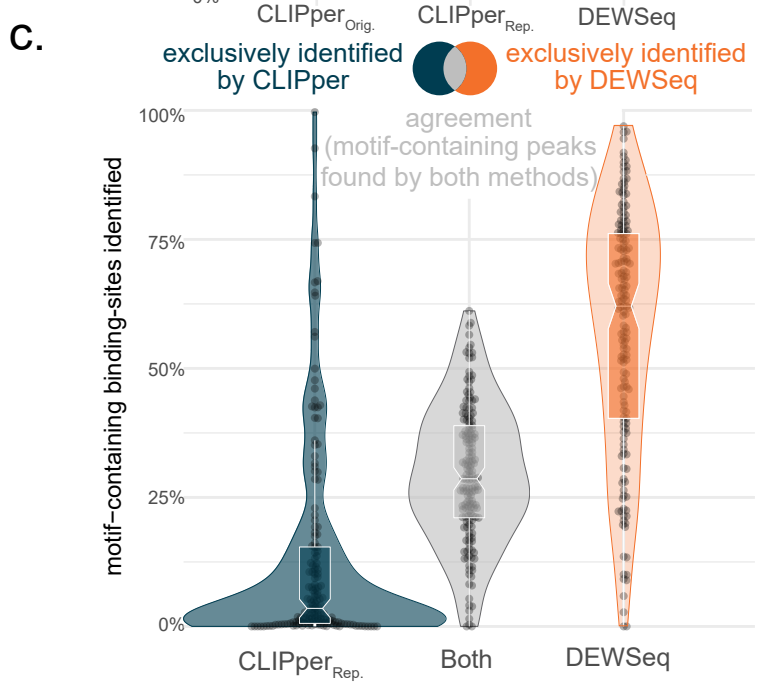
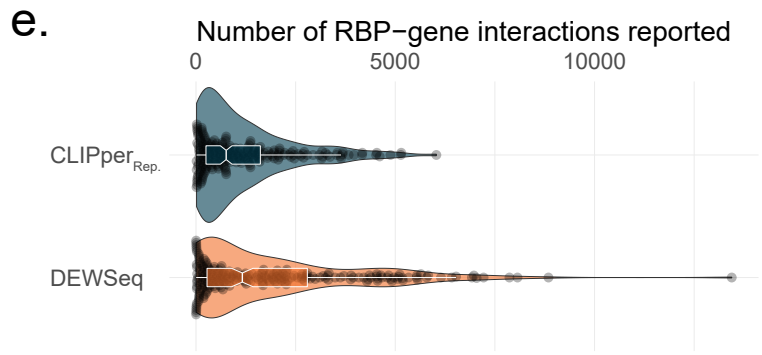

a.
Percentage of peaks found in in other replicate
100
75
50
25
0
58.75%
median
replicate 1
replicate 2
SLBP
K562
hnRNPC
HepG2
RBFOX2
HepG2 K562
hnRNPC
K562
b.
UGCAUG
RBFOX2
IP
HepG2
d.
AAUAAA
CSTF2
IP
HepG2
c.
Crosslink site counts [CPM]
80
70
60
50
40
30
20
10
0
-100
-50
0
50
100
150
Relative position
3' UTR
histone stem loop
5'
IP
SMI
e.
UUuCUuCc
U2AF2
IP
HepG2
f.
AcAcAcAc
HNRNPL
IP
HepG2
g.
htseq-clip
aligned eCLIP .bam files
annotation .gff3 file
Loading sliding-window counts and annotation
Prefilter windows
DESeq2
Size factor estimation
Dispersion estimation
Significance testing (Wald / LRT)
DEWSeq optimized
Merge significant windows into regions
Visualisation
DEWSeq
significant regions
bed visualisation

Figure 2


a.
ENCODE eCLIP data
POSTAR2
DoRiNA
Mukherjee et.al 2019
CLIPper
htseq-clip
DEWSeq
catRAPID omics v2.0 motifs
- AttRACT
- cisBP-RNA
- mCrossBase
- oRNAment
- RBPmap
eCLIP binding-sites (van Nostrand et.al 2016)
eCLIP reproducible binding-sites (van Nostrand et.al 2020)
eCLIP revised binding-sites
iCLIP binding-sites
HITS-CLIP binding-sites
PAR-CLIP binding-sites
PAR-CLIP revised binding-sites
"CLIPper Org."
"CLIPper Rep."
"DEWSeq"
FIMO
(Grant, Bailey and Noble, 2011)
Comparison: Number of motif containing binding sites per RBP


b.
binding sites [log10]
% of binding sites containing motifs
CLIPperOrig.
CLIPperRep.
DEWSeq


c.
exclusively identified by CLIPper
exclusively identified by DEWSeq
agreement (motif-containing peaks found by both methods)
motif−containing binding-sites identified
CLIPperRep.
Both
DEWSeq


e.
Number of RBP−gene interactions reported
0
5000
10000
CLIPperRep.
DEWSeq

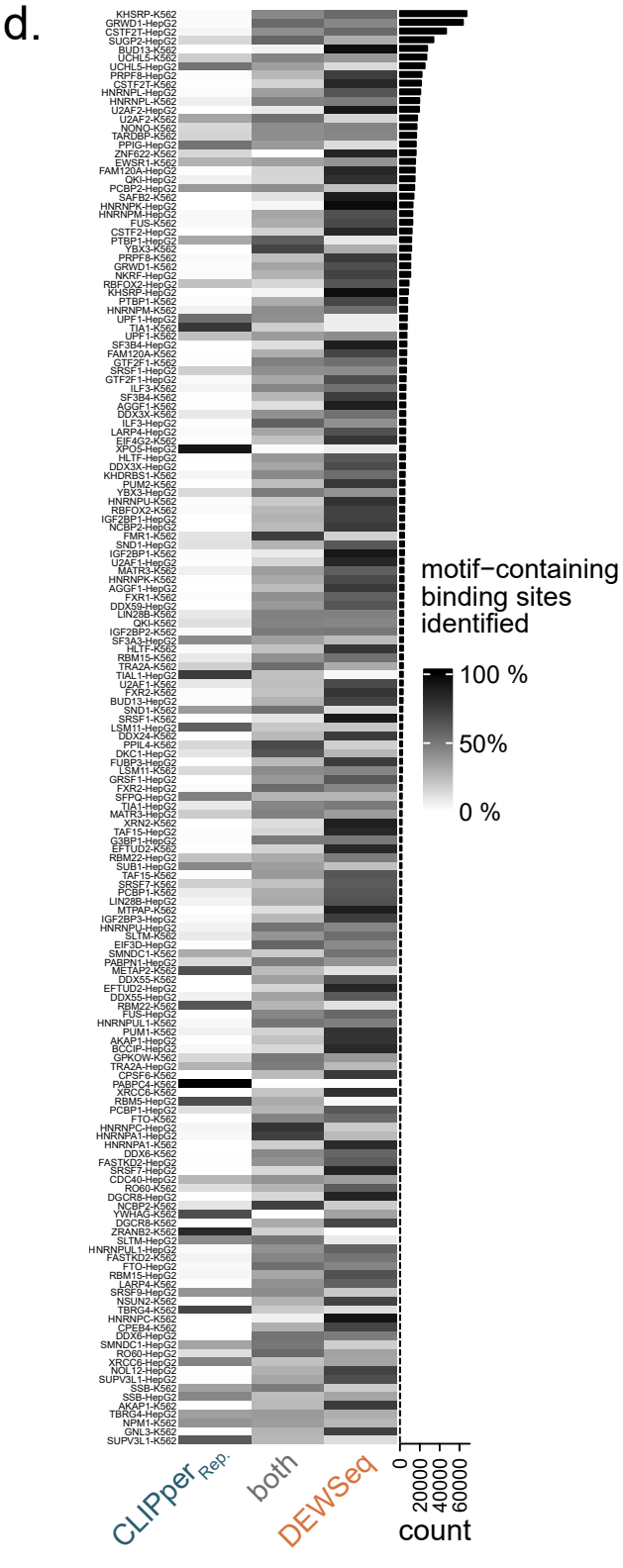

d.
motif−containing binding sites identified
100 %
50%
0 %
CLIPperRep.
both
DEWSeq
count

Figure 3

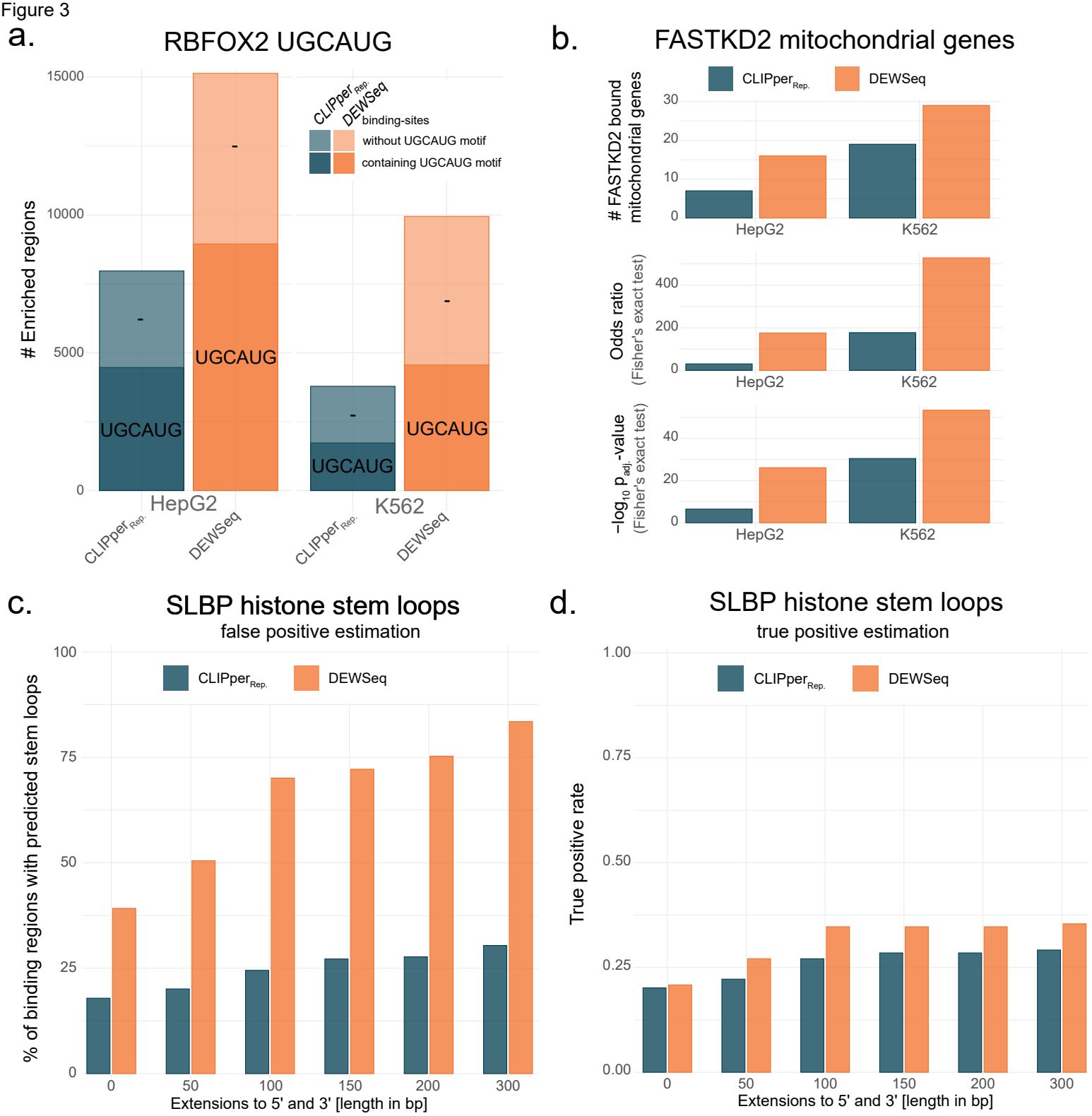
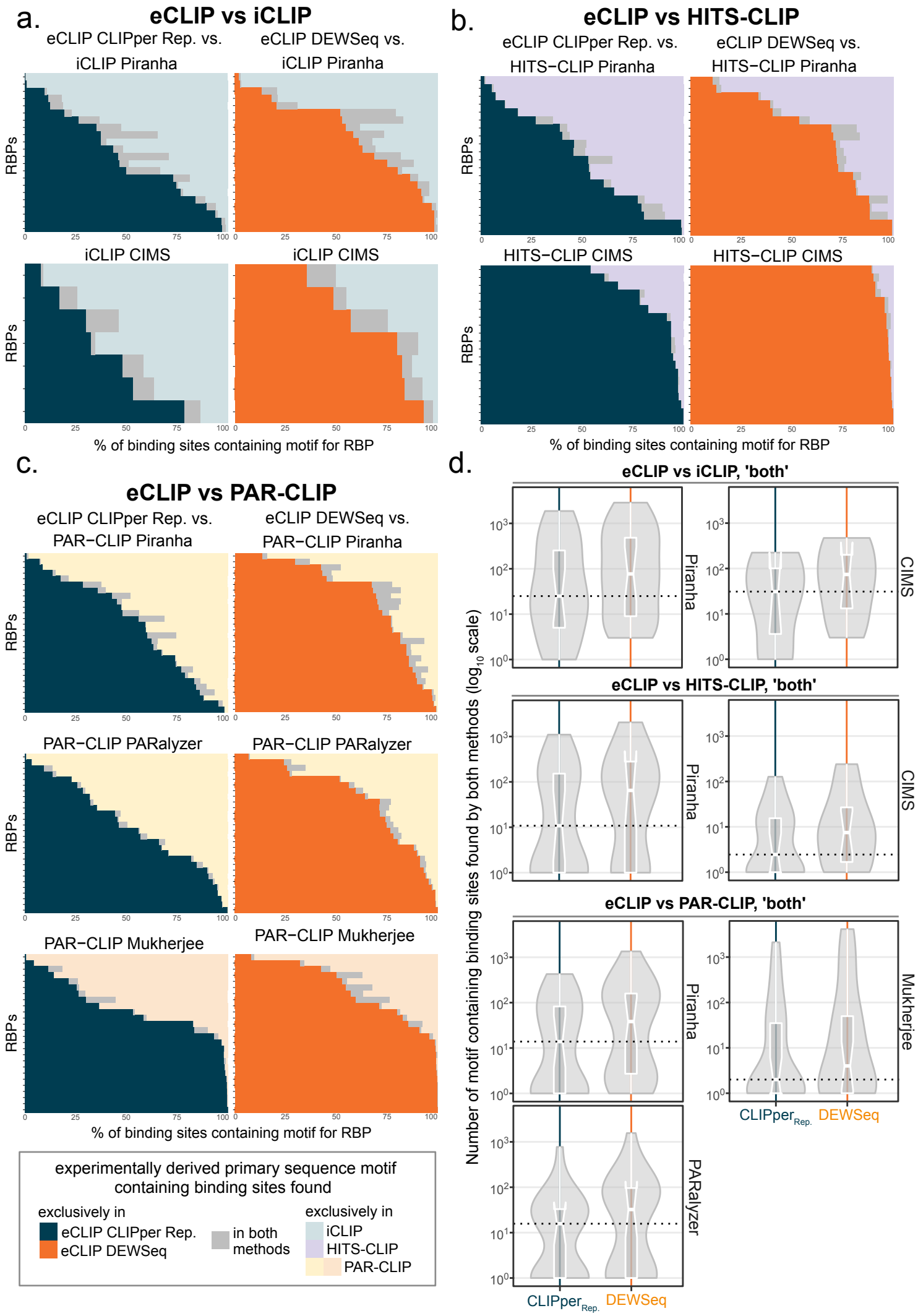
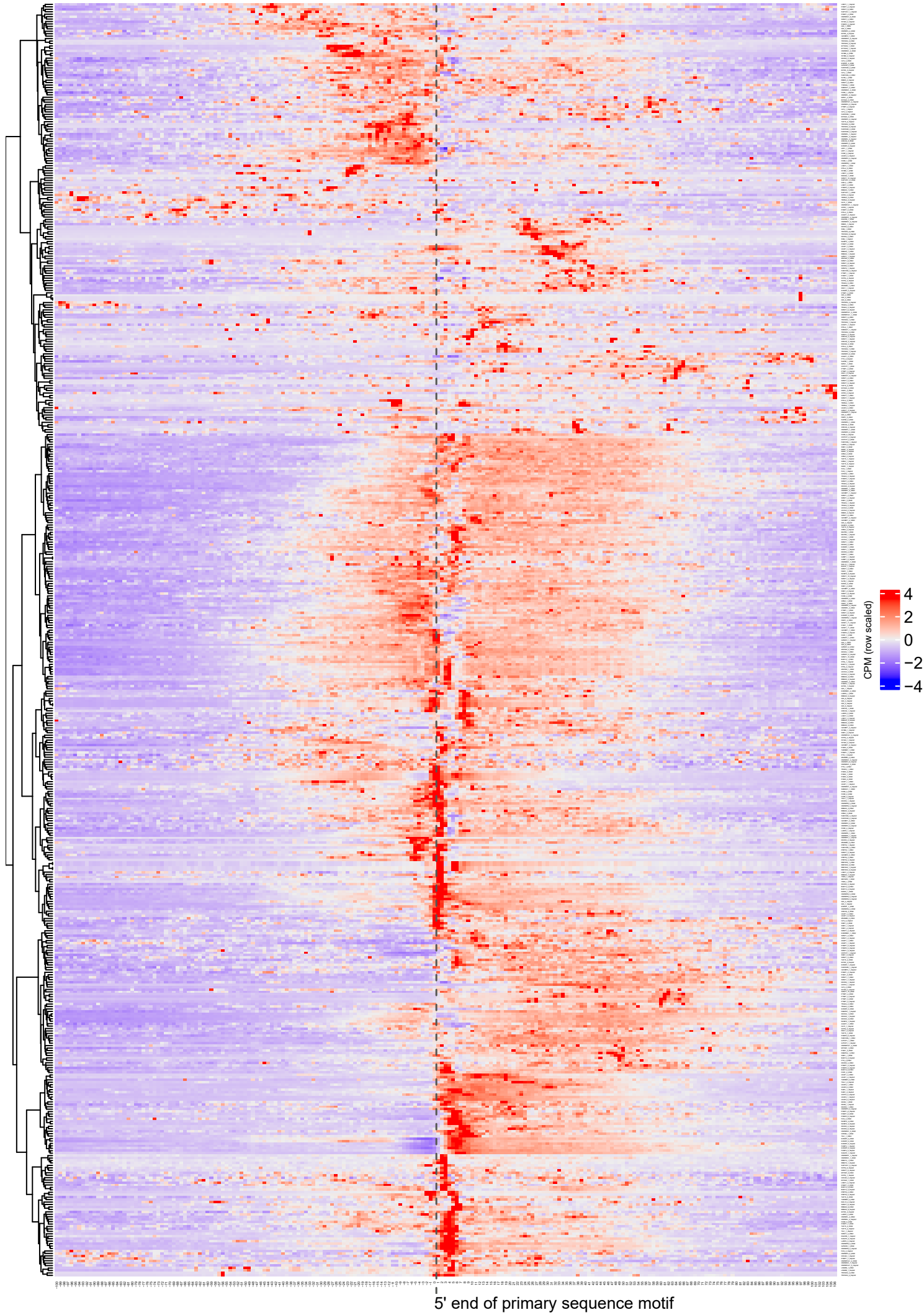

a.
RBFOX2 UGCAUG
Enriched regions
15000
10000
5000
0
CLIPper Rep.
DEWSeq
binding-sites
without UGCAUG motif
containing UGCAUG motif
UGCAUG
HepG2
K562
b.
FASTKD2 mitochondrial genes
FASTKD2 bound mitochondrial genes
Odds ratio (Fisher's exact test)
−log10 padj.-value (Fisher's exact test)
CLIPperRep.
DEWSeq
HepG2
K562
c.
SLBP histone stem loops
false positive estimation
% of binding regions with predicted stem loops
Extensions to 5' and 3' [length in bp]
d.
SLBP histone stem loops
true positive estimation
True positive rate
Extensions to 5' and 3' [length in bp]

Figure 4

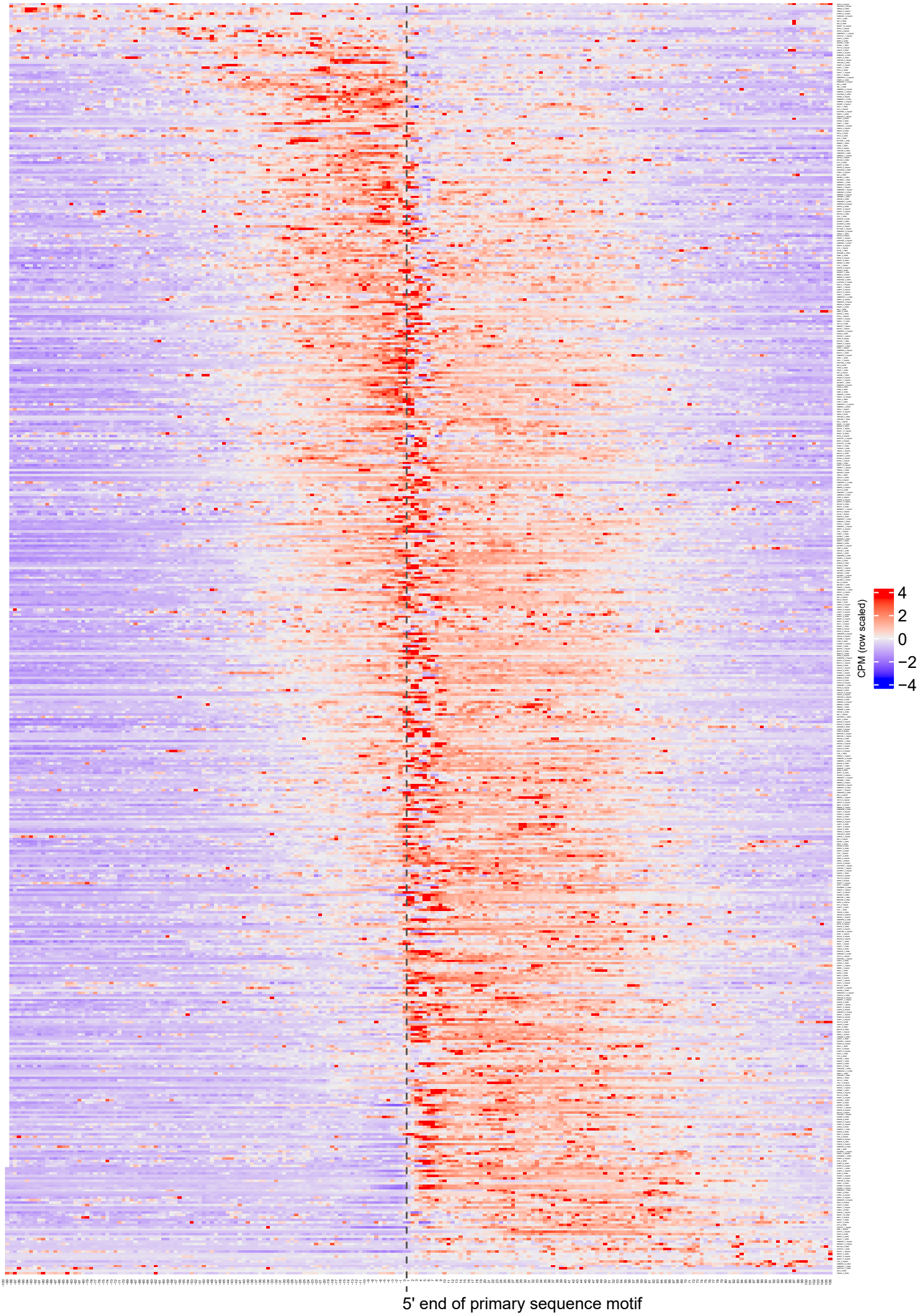

a.
eCLIP vs iCLIP
eCLIP CLIPper Rep. vs. iCLIP Piranha
eCLIP DEWSeq vs. iCLIP Piranha
iCLIP CIMS
iCLIP CIMS
RBPs
% of binding sites containing motif for RBP
b.
eCLIP vs HITS-CLIP
eCLIP CLIPper Rep. vs. HITS−CLIP Piranha
eCLIP DEWSeq vs. HITS−CLIP Piranha
HITS−CLIP CIMS
HITS−CLIP CIMS
RBPs
% of binding sites containing motif for RBP
c.
eCLIP vs PAR-CLIP
eCLIP CLIPper Rep. vs. PAR−CLIP Piranha
eCLIP DEWSeq vs. PAR−CLIP Piranha
PAR−CLIP PARalyzer
PAR−CLIP PARalyzer
PAR−CLIP Mukherjee
PAR−CLIP Mukherjee
RBPs
% of binding sites containing motif for RBP
experimentally derived primary sequence motif containing binding sites found
exclusively in
eCLIP CLIPper Rep.
eCLIP DEWSeq
in both methods
exclusively in
iCLIP
HITS-CLIP
PAR-CLIP
d.
eCLIP vs iCLIP, 'both'
eCLIP vs HITS-CLIP, 'both'
eCLIP vs PAR-CLIP, 'both'
Number of motif containing binding sites found by both methods (log10 scale)
Piranha
CIMS
Mukherjee
PARalyzer
CLIPperRep.
DEWSeq

Supplementary Figure 1

a.

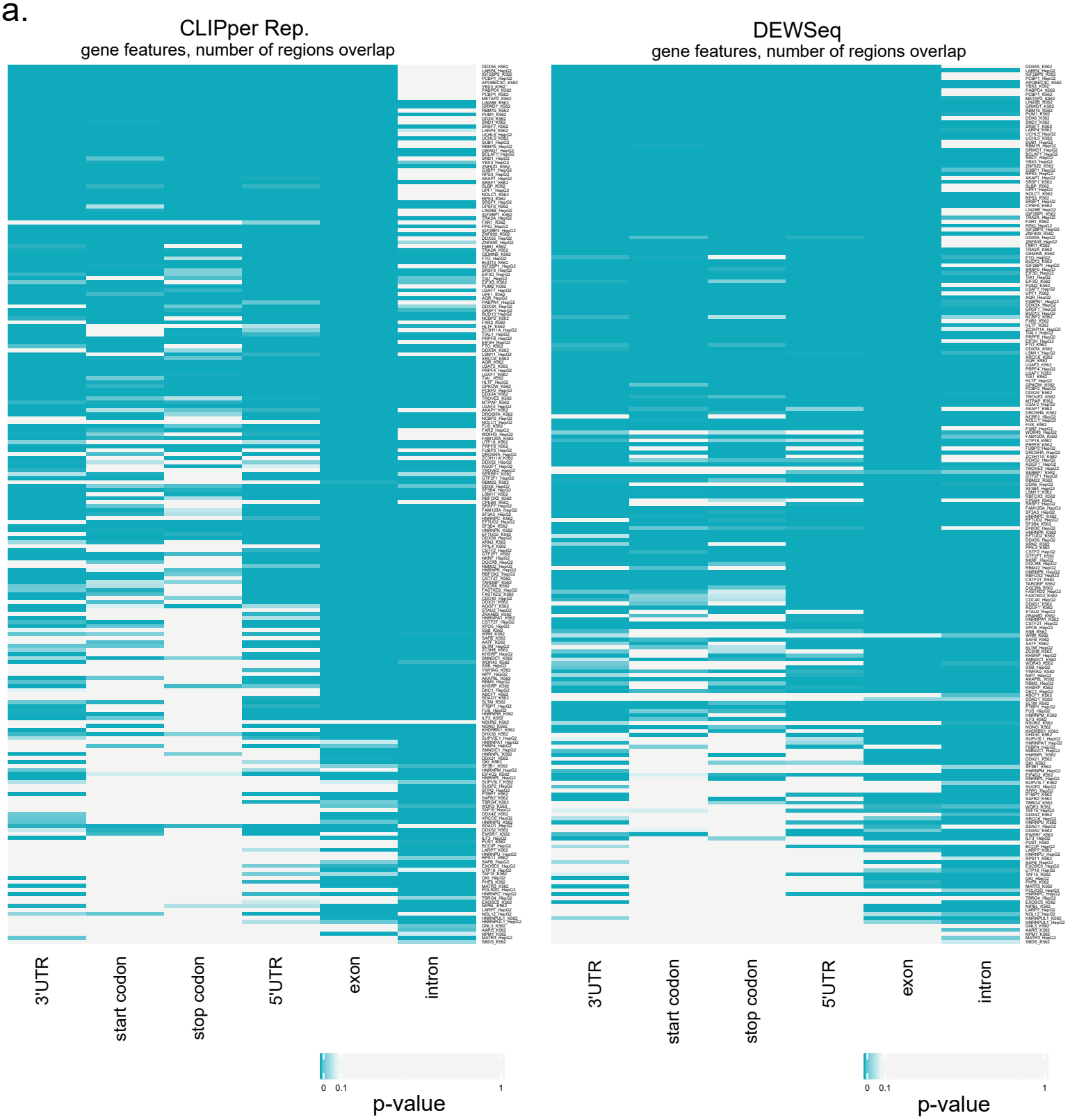


b. eCLIP crosslink (truncation) sites around RBP motifs, seriation clustered



Supplementary Figure 2

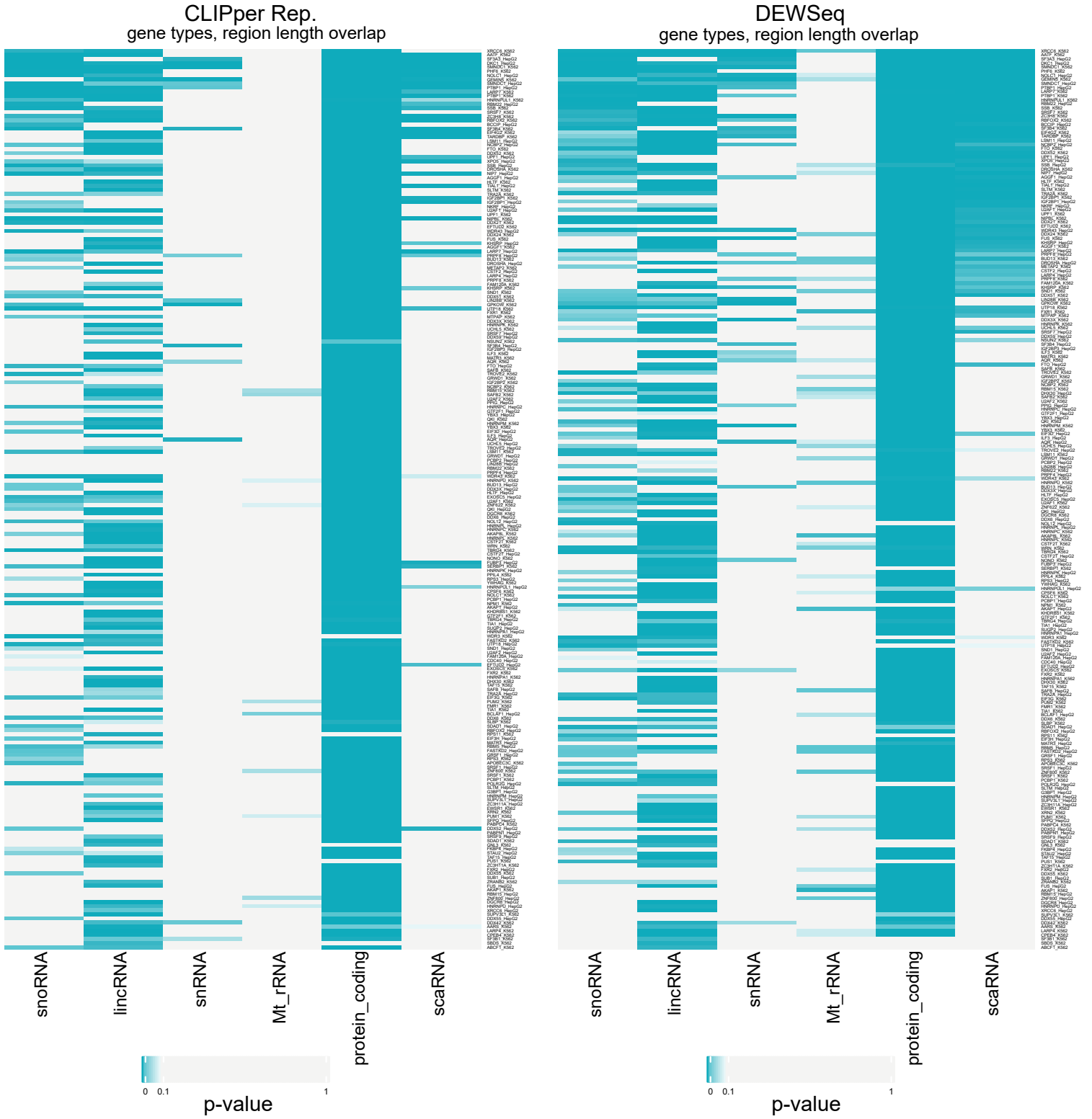
a.

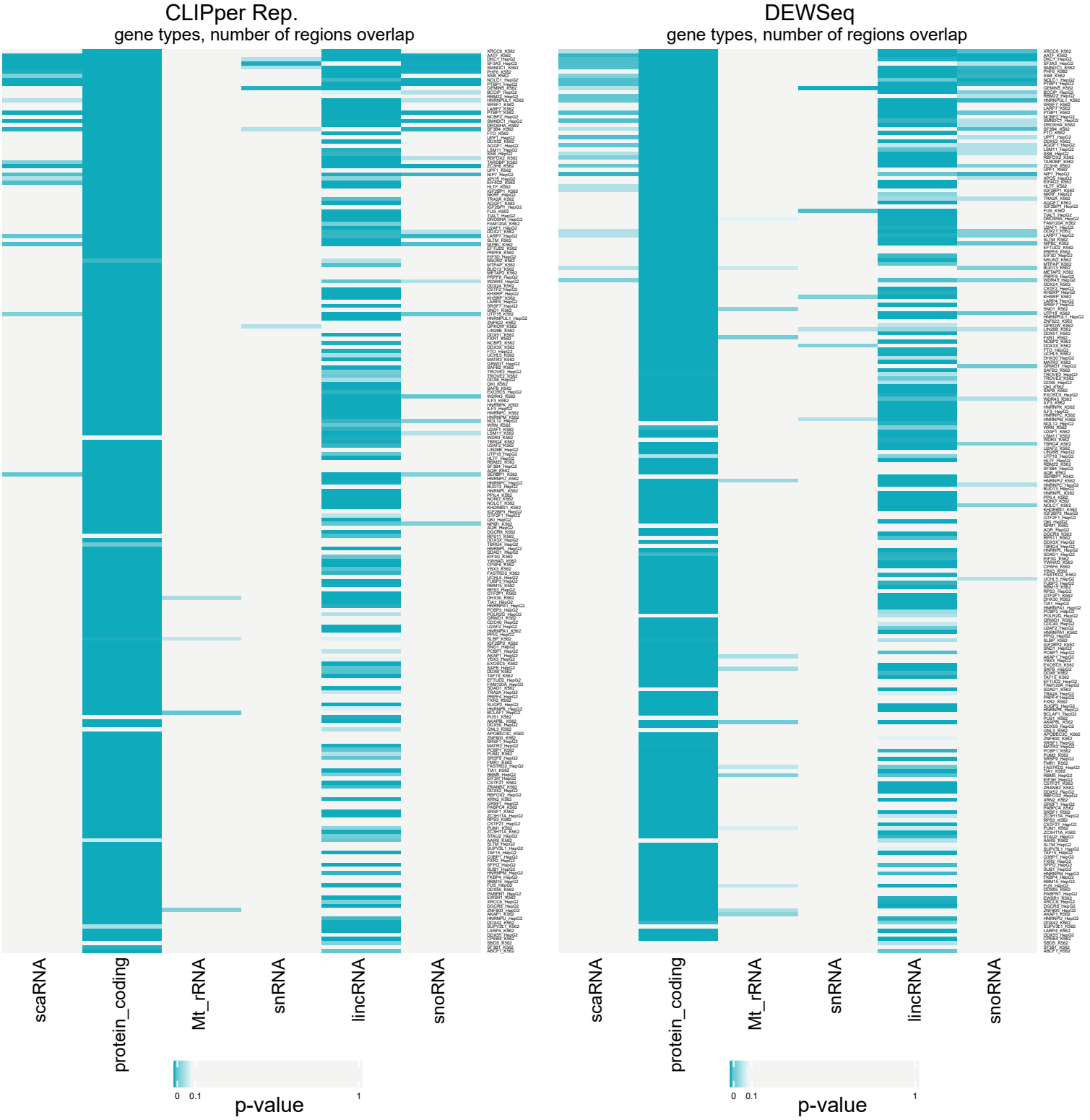


CLIPper Rep.
gene features, region length overlap
3'UTR
start codon
stop codon
5'UTR
exon
intron
0
0.1
1
p-value

DEWSeq
gene features, region length overlap
3'UTR
start codon
stop codon
5'UTR
exon
intron
0
0.1
1
p-value

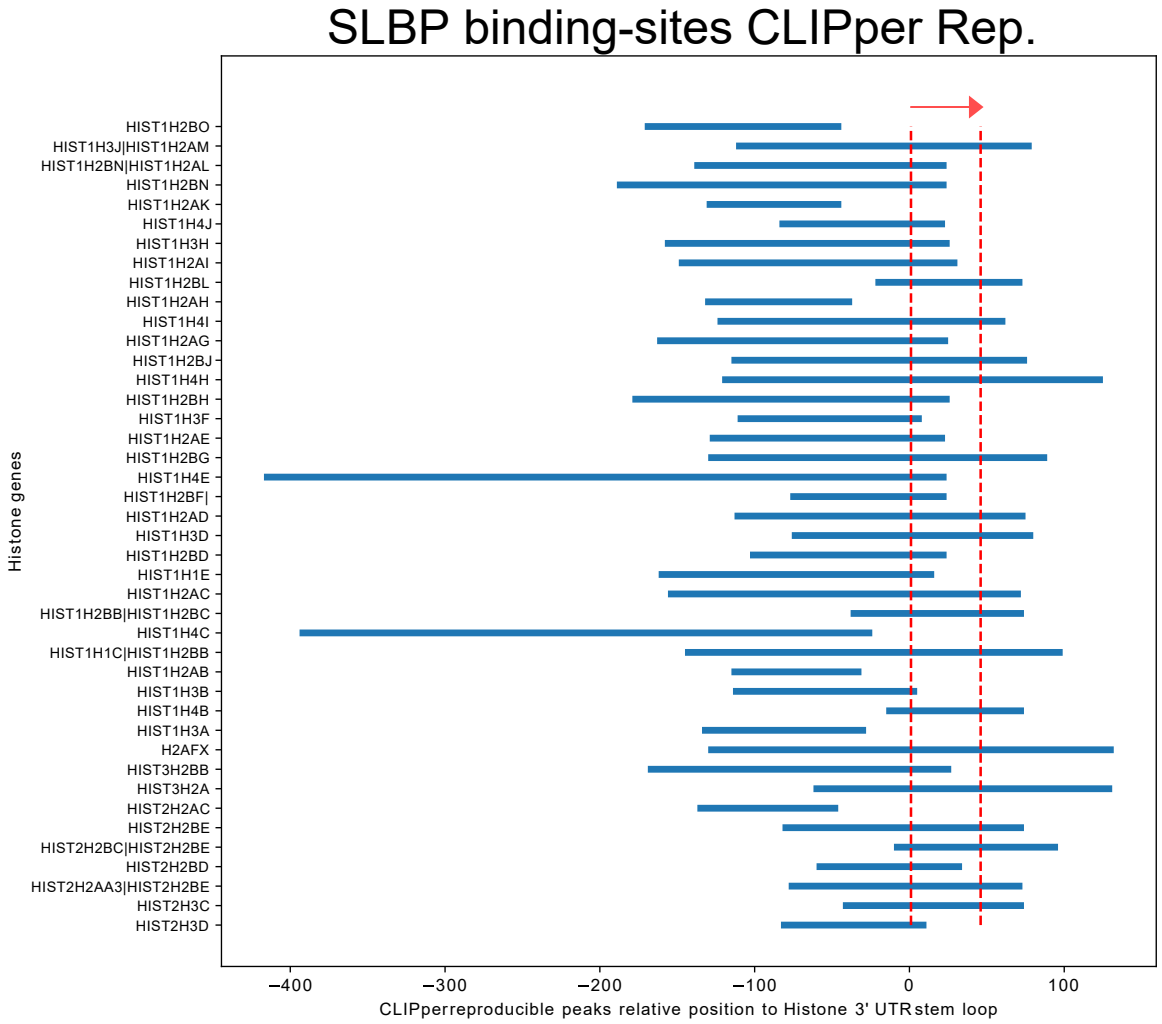
b.


CLIPper Rep.
gene types, region length overlap
DEWSeq
gene types, region length overlap
snoRNA
lincRNA
snRNA
Mt_rRNA
protein_coding
scaRNA
0
0.1
1
p-value

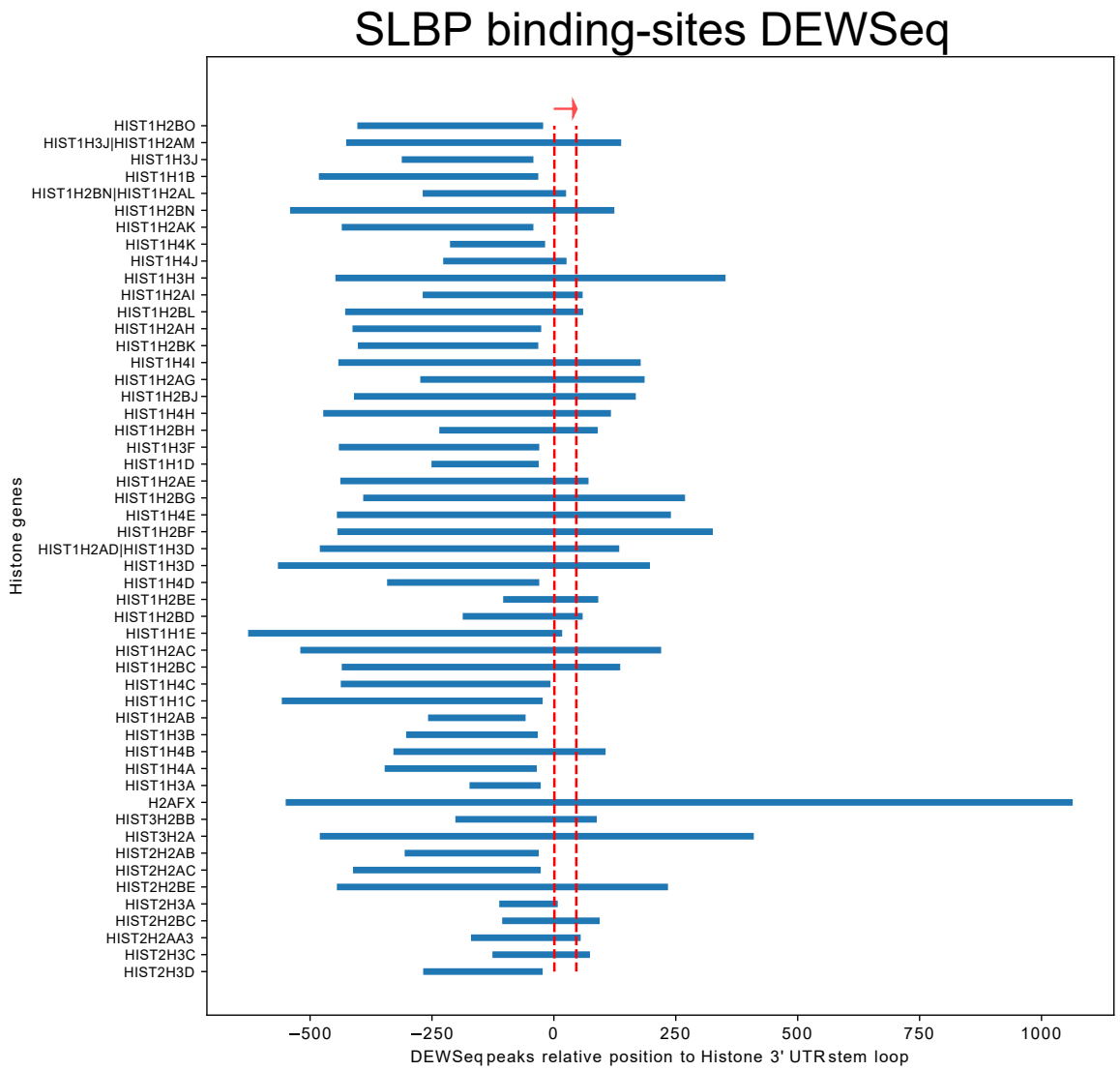

CLIPper Rep.
gene types, number of regions overlap
DEWSeq
gene types, number of regions overlap
scaRNA
protein_coding
Mt_rRNA
snRNA
lincRNA
snoRNA
0
0.1
1
p-value

Supplementary Figure 3

a.

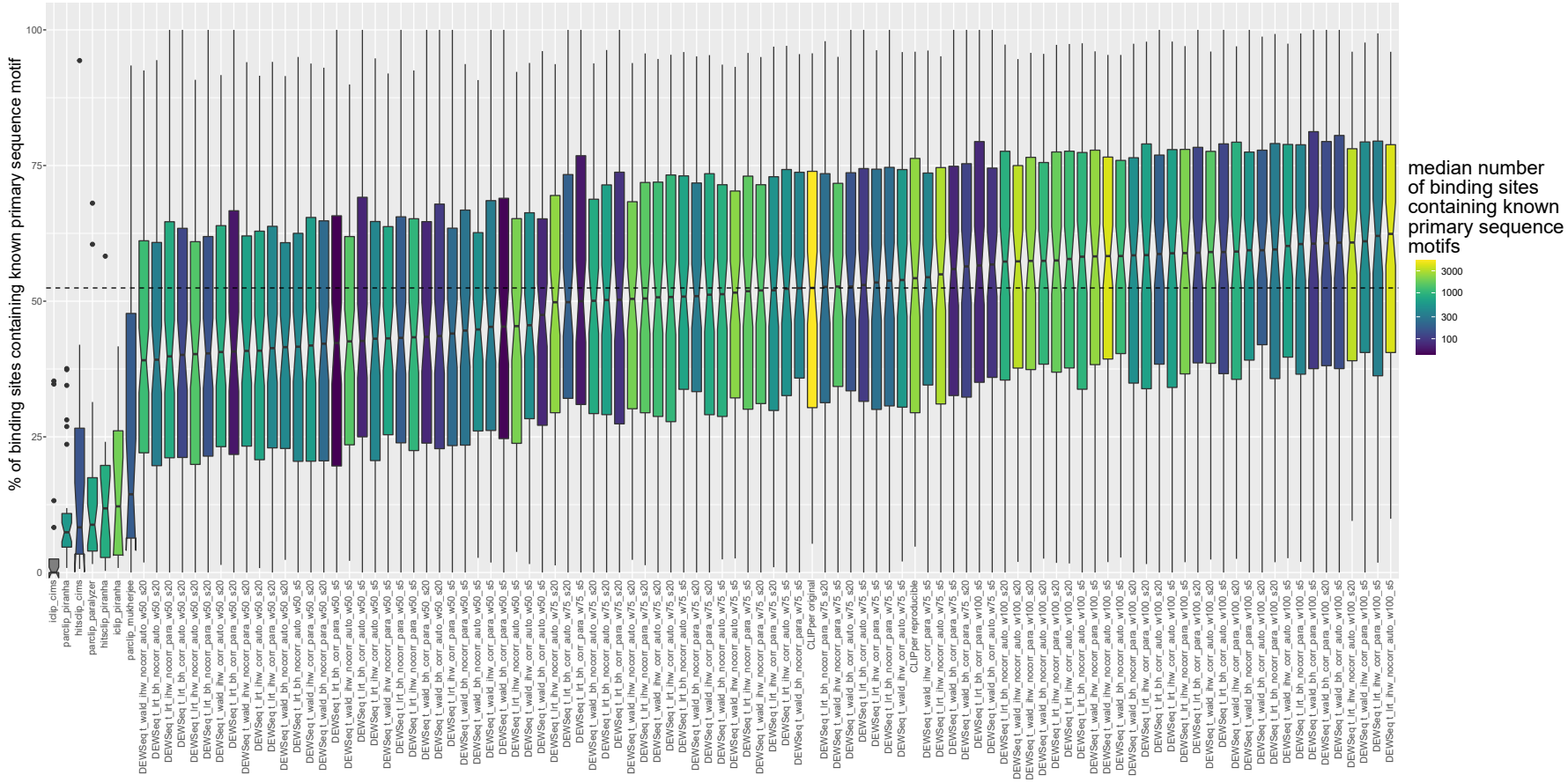
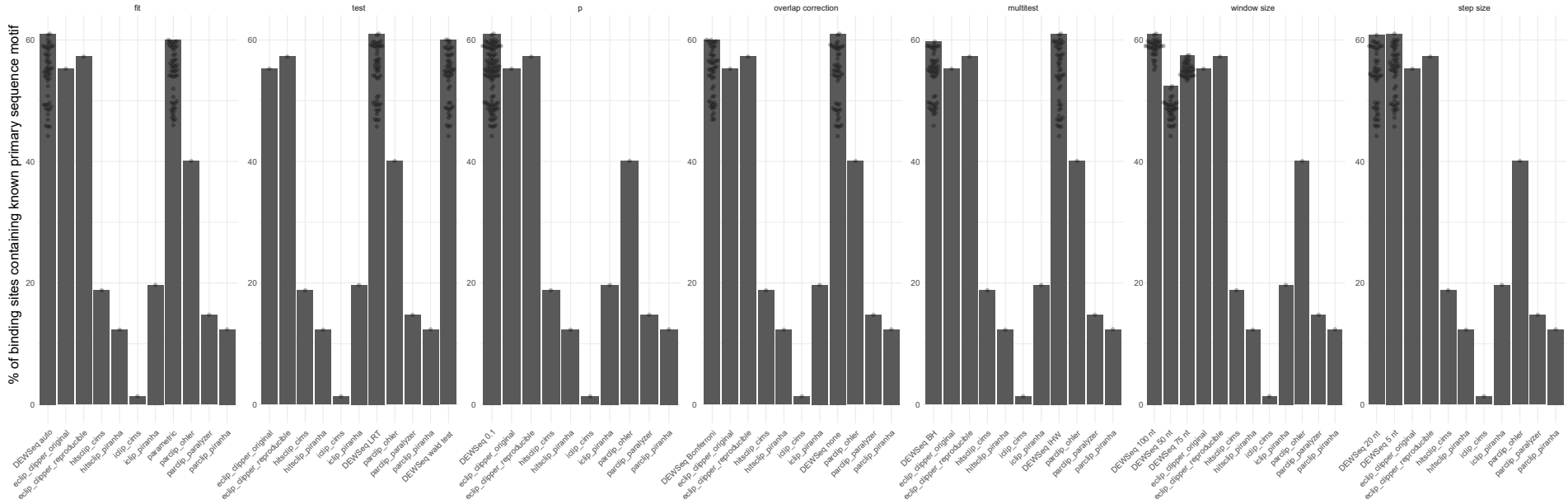


b.



Supplementary Figure 4

a.



b.