Published online 10 December 2014 *Nucleic Acids Research, 2015, Vol. 43, No. 1* ***95–103***
doi: 10.1093/nar/gku1288

Leveraging cross-link modification events in CLIP-seq for motif discovery

Emad Bahrami-Samani[1], Luiz O.F. Penalva[2], Andrew D. Smith[1] and Philip J. Uren[1,*]

[1]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA and [2]Children's Cancer Research Institute and Department of Cellular and Structural Biology, University of Texas Health Science Center, San Antonio, TX 78229, USA

Received May 08, 2014; Revised November 04, 2014; Accepted November 25, 2014

ABSTRACT

High-throughput protein–RNA interaction data generated by CLIP-seq has provided an unprecedented depth of access to the activities of RNA-binding proteins (RBPs), the key players in co- and post-transcriptional regulation of gene expression. Motif discovery forms part of the necessary follow-up data analysis for CLIP-seq, both to refine the exact locations of RBP binding sites, and to characterize them. The specific properties of RBP binding sites, and the CLIP-seq methods, provide additional information not usually present in the classic motif discovery problem: the binding site structure, and cross-linking induced events in reads. We show that CLIP-seq data contains clear secondary structure signals, as well as technology- and RBP-specific cross-link signals. We introduce Zagros, a motif discovery algorithm specifically designed to leverage this information and explore its impact on the quality of recovered motifs. Our results indicate that using both secondary structure and cross-link modifications can greatly improve motif discovery on CLIP-seq data. Further, the motifs we recover provide insight into the balance between sequence- and structure-specificity struck by RBP binding.

INTRODUCTION

During the past decade, an increasing amount of attention has been drawn toward the complex mechanisms of post-transcriptional regulation of gene expression, heavily driven by interactions between RNA and RNA-binding proteins (RBPs) (1–3). It is now clear that post-transcriptional regulation is a major modulator of phenotype, with important roles in practically all biological processes, and implications for many human diseases (4–6). Approaches such as RNAcompete have uncovered complex relationships between RBPs and their binding specificities (7), underscoring the need for a greater understanding of their activities. Despite substantial progress, much remains to be learned about post-transcriptional regulation. One of the prime movers of recent and continuing genome-wide insights is CLIP-seq (cross-linking with immunoprecipitation followed by high-throughput sequencing), which allows a high-resolution investigation of the binding sites for a given RBP. Three variants of CLIP-seq have been developed: high-throughput sequencing CLIP, termed HITS-CLIP (3); Photoactivatable-ribonucleoside-enhanced CLIP, termed PAR-CLIP (8); and individual nucleotide resolution CLIP-seq, termed iCLIP (9). Each of these CLIP-seq variants allows target mRNAs to be identified, but importantly the resolution is high enough to localize binding to within a small window. Beyond this, statistical models are needed to exactly localize the binding site. Even with iCLIP, which inherently gives single-nucleotide resolution for the cross-link location, background noise and sequencing artifacts will mean that not every locus identified by the assay is a *bona fide* binding site, nor will the localization always be perfect, as cross-linking biases are well known (10,11).

In addressing the problem of simultaneous binding site characterization and localization, much of the analysis methodology has been borrowed from the field of transcription factors, for example from the analysis of sequences identified in ChIP-seq experiments. Binding sites for RBPs though are different from transcription factor binding sites: they tend to be shorter, and have characteristic secondary structures (12). Most existing motif-discovery methods concentrate exclusively on primary sequence, but the shorter length of RBP binding sites, coupled with the abundance of highly similar non-binding sites and the proclivity toward low sequence specificity in some RBPs present difficulties for such approaches. Despite mounting evidence that RNA secondary structure plays a role in RBP binding site selection, few motif discovery tools consider it (13,14).

The CLIP-seq experimental procedure causes sequencing reads to exhibit characteristic substitutions and deletions relative to the reference genome at the cross-link location. We call these substitutions and deletions diagnostic events (DEs). (8,10,15,16). Although the cross-link location is not

*To whom correspondence should be addressed. Tel: +1 213 740 2416; Fax: +1 213 740 8631; Email: uren@usc.edu

necessarily within the binding site, it is generally in close proximity. These properties of CLIP-seq reads have been employed to localize the cross-linking location (17,18), but their distribution relative to the binding site has not previously been exploited for motif-discovery in a probabilistic model, and tools using them have been restricted to individual variants of CLIP-seq.

Here, we describe the Zagros algorithm for simultaneous motif characterization and binding site localization from CLIP-seq data that uses a model specifically designed for CLIP-seq derived RNA binding sites. Our method models sequence, secondary structure and technology-specific cross-linking events using a formalization where we treat the locations of motif occurrences as missing data. Zagros is the first motif-discovery method that is able to exploit all of the additional information inherent in CLIP-seq data of any variety, and we demonstrate that this extra information has utility in successfully recovering RBP motifs in CLIP-seq data. Zagros source code licensed under the GNU General Public License (version 3) is freely available for download from http://smithlabresearch.org.

Using our method, we show that motifs enriched in CLIP-seq datasets represent a trade-off between sequence and structure specificity, suggesting that RBPs with highly specific sequence motifs require less structural constraints on their binding sites to achieve their specificity.

MATERIALS AND METHODS

Data sets

To investigate the properties of CLIP-seq, the sequence and secondary structure preference of a range of RBPs, and to evaluate our proposed method we collected a set of public data derived from 20 studies, covering iCLIP (9,10,19–21), HITS-CLIP (22–32) and PAR-CLIP (8,33–35). This collection constitutes data profiling 40 RBPs (36 human and 4 mouse). A complete description is provided in supplementary materials. Sequence data was mapped to hg19 and mm9 using Novoalign (Novocraft, http://www.novocraft.com). Definitions of genes, exons and UTRs were taken from RefSeq (36).

Defining sets of target 3'UTRs

To investigate the structure preferences of hexamers in RBP targets, for each CLIP-seq dataset we defined a set of target 3′ UTRs. To do this, we binned CLIP-seq reads in 1nt bins (iCLIP) or 20nt bins (PAR-CLIP, HITS-CLIP), and retained only those bins that could be uniquely assigned to a single transcript. For each 3′ UTR we found the bin with the largest number of reads. We then ranked 3′ UTRs by the count of reads in the bin with the most reads, and selected the top 1000 3′ UTRs as our target set for each RBP. The non-target set is simply any 3′ UTR region (as defined by refseq) that is not contained in the target set.

Calculating secondary structure

In our model we represent secondary structure using base pairing probabilities (for more detail, see supplementary section 2.6.3). For an RNA sequence, we calculate base pairing probabilities using McCaskill's algorithm (37). These probabilities are then input either with just the sequences, or with both the sequences and the cross-link modification events, and are considered part of the data. Our model includes parameters for the base-pair-probability of each position within the motif; these parameters are learned from the data using expectation maximization.

Simulated data

Each simulated dataset was produced by randomly selecting 500 segments, each of length 50 bp, from human 3′ UTRs (RefSeq), and planting within each a motif occurrence from a randomly generated position weight matrix with an information content of approximately 0.5 bits per column. Structure was imposed by taking a short segment upstream of the placed motif and planting its reverse complement downstream of the motif, forming a hairpin loop. The offset of diagnostic events from the start of the planted motif was simulated (once per dataset) as a uniform random variable on the range −8 to +8. The number of diagnostic events placed for a motif occurrence, where diagnostic events were present, was sampled from the empirical distribution observed in the same 50 bp window of the same 3′ UTR in CLIP-seq datasets. A more detailed description of the simulation procedure is given in supplementary methods. In total we simulated 2200 datasets. Of these, 1100 have no particular structure imposed on the motif occurrence, but have a DE-fraction ranging from 0 (no sequences with diagnostic events) to 1 (all sequences having diagnostic events), with a step of 0.1 and 100 simulations for each level. The remaining 1100 have no diagnostic events, but a structure-fraction ranging from 0 (no sequences have motif occurrences with any particular structure imposed on them) to 1 (the motif occurrence in every sequence is forced to adopt a ssRNA conformation), with a step size of 0.1 and 100 simulations for each level.

Recovered PWMs were compared to the simulated PWM by calculating the Kullback–Leibler divergence of the recovered PWM from the simulated one. We estimated a false discovery threshold for KLD by randomly shuffling the recovered and simulated motifs, and considered a recovered motif to match the simulated motif if $P < 0.05$.

CLIP-seq derived data

To evaluate the performance of Zagros on recovering previously reported RBP binding motifs from CLIP-seq data, we defined a set of CLIP-seq target sequences for each dataset. As with the simulated data, these sequences were each 50bp in length, and we selected 500 sequences per datasets. To select these sequences, we binned CLIP-seq reads for each dataset into 50 bp bins, and selected the top 500 bins with the greatest number of CLIP reads.

Sequence vs. structure specificity

We ran zagros using both structure and diagnostic events. We used a motif length of 6 for all datasets to avoid any biases in counting oligomer occurrences that would be introduced by varying lengths. We retained only those datasets

where the recovered motif contained a match to the previously reported consensus. Base-pair probabilities for each recovered motif are the average across all positions in the motif.

RESULTS AND DISCUSSION

CLIP-seq data encodes RBP-specific sequence and structure signals

Prior to developing our model, we sought to investigate the interaction of sequence and secondary structure in RBP targets identified from CLIP-seq data. Since structure is inherently a property of sequence, our purpose here was to determine to what extent the local sequence of the binding site directly informs its structure. If structure can trivially be inferred from the sequence of the motif, then explicitly modeling it is unlikely to be beneficial in motif-discovery. Conversely, if the interaction between sequence and structure in CLIP-seq targets is identifiably different from non-targets, then leveraging this may improve motif-discovery.

Earlier work has shown that matches to a known consensus sequences for an RBP in target genes derived from RIP-Chip show different RNA structure to those in non-target genes (13). We extended this by determining the structural bias of all hexamers in CLIP-seq target sets, while remaining agnostic of what the RBP binding motif was. To do this, we selected a set of target 3′ UTRs for each RBP, and counted the number of times each hexamer occurred in the target set in single- or double-stranded conformation, and also for non-targets. We computed an odds-ratio for all hexamers: the odds of the hexamer being single-stranded in target 3′ UTRs (target structure) against the odds of it being single-stranded in non-target 3′ UTRs (background structure). In each dataset, for each hexamer/ratio, we calculated a *P*-value using Fisher's exact test, where our null hypothesis is that the odds-ratio does not significantly deviate from 1. Figure 1A shows the results for three example RBPs, HuR, IGF2BP1, and TIAL1, with hexamers that show the most strongly significant deviation from their background structure being close matches to previously reported sequence motifs for these proteins (8,20,38–40). To determine whether this trend was present in a range of CLIP-seq datasets, we calculated *P*-values in this fashion for all hexamers across all CLIP-seq datasets. From these, we determined the frequency of *P*-values (Figure 1B). As a control, we also did this using randomly selected 3' UTRs. We noted an enrichment of significant *P*-values when using CLIP-derived target 3′ UTRs which was not present with randomly selected 3′ UTRs, suggesting that targets identified by CLIP-seq are enriched for structural motifs, as well as sequence motifs.

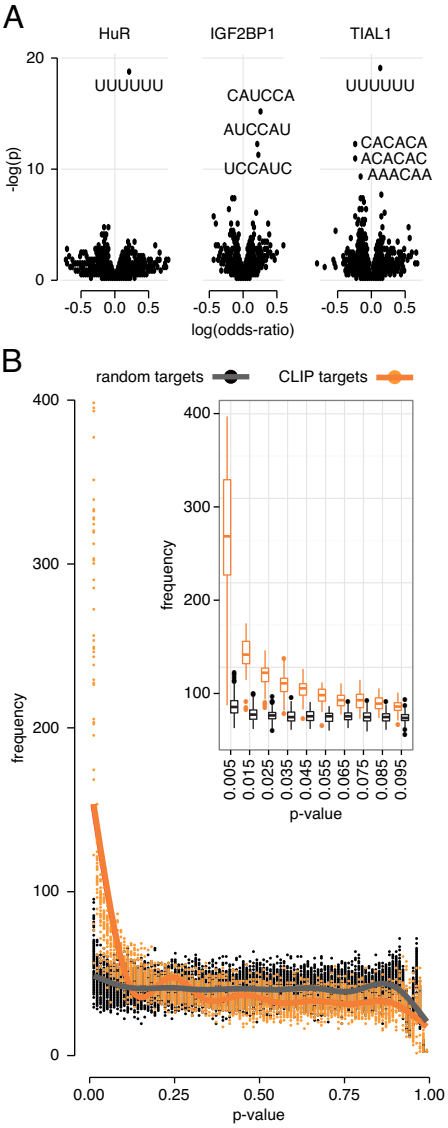

Figure 1. Sequence elements in CLIP-derived target sets show structural preferences not trivially determined by sequence. (**A**) Log odds-ratios are the odds of an occurrence of a particular hexamer being single-stranded in the 3′ UTR of target genes for the indicated RBP versus non-target genes. Significance (*P*-values, *y*-axis) is determined using Fisher's exact test. The most significant hexamers are highlighted for each RBP, and show a close match to reported consensus binding preferences. Data for HuR, IGF2BP1, and TIAL1 from (16), (8) and (20) respectively. (**B**) Each *P*-value is derived by testing the odds-ratio of a particular hexamer in a particular CLIP dataset, as in panel A (for all datasets, and all hexamers). For each CLIP-seq dataset, we calculated the frequency of *P*-values in discrete bins of width 0.01, and plotted these. Red points correspond to 3′ UTRs of genes determined to be targets in each CLIP-seq dataset, blue points were computed in the same way using randomly selected 3′ UTR sequences, and do not show an enrichment for significant associations between sequence and structure. Lines were fit using LOESS regression.

CLIP-seq diagnostic events follow RBP- and technology-specific patterns

CLIP-seq has several stages, from UV irradiation to high-throughput sequencing of the short isolated RNA segments. Of the three CLIP-seq variants, only iCLIP was designed to have single-nucleotide resolution, achieved through read truncation at the cross-link location (9). Still, single-nucleotide resolution has been achieved in PAR-CLIP and HITS-CLIP by exploiting characteristic deletions and mutations caused by cross-linking (17,18). The reverse transcriptase used in HITS-CLIP sometimes skips the crosslinked RNA nucleotide bound to the protein residue, resulting in a nucleotide deletion, while in PAR-CLIP, 4sU

labeled transcripts sometimes induce T → C conversions in the sequencing step, identifiable as mismatches to the reference genome when mapping. We refer to these deletions, mutations and truncations collectively as 'diagnostic events' (DEs). The relative prevalence of each type of event varies by technology (see Figure 2A).

HITS-CLIP and iCLIP have similar mapping rates, however only 8–20% of the mapped reads in HITS-CLIP have a deletion at the cross-link site, whereas for iCLIP around 99% of mapped reads appear to have been truncated at the cross-link site. More specifically, in 99% of the reads, we did not find any deletions, which can either mean that reverse-transcriptase has read through the cross-link location with no deletion or it has been halted at the cross-link location. Only in the latter case the read in fact contains the diagnostic event. Sugimoto *et al.* (10) estimated that 82% of the reads are truncated at the cross-link location, but for simplicity, Zagros considers all of the reads without deletions to have a diagnostic event at the truncation site. For more information on diagnostic events on iCLIP data refer to supplementary information, Section 2.9. PAR-CLIP has a higher percentage of reads containing diagnostic events than HITS-CLIP, but suffers lower mapping rates than HITS-CLIP and iCLIP. Since DEs for iCLIP are identified as the 5′ mapping location, in the absence of a deletion, each read can have at most one DE. For HITS-CLIP and PAR-CLIP, it is possible that sequencing errors or even multiple actual cross-links could lead to multiple potential DEs per sequence. Fortunately for our analyses, usually only one DE can be found (Figure 2B). To investigate the distribution of diagnostic events around binding sites, we searched for occurrences of simple consensus sequences, defined from literature (see supplementary materials), around CLIP sites identified for each RBP in our data. We observed that diagnostic events are often enriched at a particular distance from the consensus match, and follow a characteristic distribution. We noted that distance from the start of the consensus match to the most likely position of a diagnostic event is specific to the RBP (Figure 2C). Efforts to model CLIP-seq diagnostic events must account for the variability imparted by RBP and technology.

A probabilistic model of sequence, structure and diagnostic events

Our model of sequence and structure builds upon the widely used classic mixture model introduced by Lawrence and Reilley (41) with many notable extensions (42). Let S be a set of n sequences such that each member $S_i \in S$ has the same length m. Define the set $X = \{X_1, \ldots, X_n\}$ of motif occurrence indicators in correspondence with S so that $X_{ij} = 1$ exactly when a motif occurrence starts at position j in S_i. We use the "zero-or-one occurrence per sequence" assumption (ZOOPS), so $\|X_i\|_1 \in \{0, 1\}$, as used in the well-known MEME program (43). We augment this data with secondary structure indicators $T = \{T_1, \ldots, T_n\}$, with T_{ij} indicating that position j of S_i has a paired structural state, which implicitly assumes a single underlying secondary structure for each sequence. In explaining our model, we assume a fixed motif width w and use the notation $S_{i\{X_i\}}$ to denote the w consecutive positions of S_i beginning at the unique position j such that $X_{ij} = 1$; if $\|X_i\|_1 = 0$, then $S_{i\{X_i\}}$ is empty. We also define $S_{i\{\overline{X_i}\}}$ as the concatenated positions of S_i not in $S_{i\{X_i\}}$.

We augment the usual motif model $M = (M_k)_{k=1}^{w}$ and background model f to account for secondary structure: f and the M_k are multinomial distributions over {A, C, G, U} × {paired, unpaired}, as explained in detail in supplementary methods. Then

$$\begin{aligned}\Pr(S, T, X|M, f) &= \textstyle\prod_{i=1}^{n} \Pr(S_i, T_i, X_i|M, f)\\ &= \Pr(X|M, f, O)\Pr(O|M, f) \times\\ &\quad \textstyle\prod_{i=1}^{n} \Pr(S_{i\{X_i\}}, T_{i\{X_i\}}|M, X, O_i) \times\\ &\quad \Pr(S_{i\{\overline{X_i}\}}, T_{i\{\overline{X_i}\}}|f, X, O_i),\end{aligned}$$

where $O_i = 1$ if sequence i contains a motif occurrence and 0 otherwise. When dealing only with sequence and structure, the prior $\Pr(X|M, f, O)$ is uniform, and can be disregarded. For any sequence s having structural indicators t,

$$\Pr(s, t|M) = \textstyle\prod_{k=1}^{w} M_k(s_k, t_k)$$

and

$$\Pr(s, t|f) = \textstyle\prod_{k=1}^{|s|} f(s_k, t_k).$$

Next we make an additional augmentation to the model to account for cross-linking. We assume each sequence contains one cross-link location. We estimate the probabilities of where the cross linking is located using diagnostic events as follows:

$$\Pr(C_i = l|O_i = 1, D) = \frac{D_{ij} + \epsilon}{\sum_{j'=1}^{m} D_{ij'} + \epsilon},$$

where D_{ij} is the count of diagnostic events at location j in sequence i (we treat D as a fixed model parameter), C_i is the location of the cross-link in sequence i, and ε is a pseudocount to avoid zero-counts.

We bring information about diagnostic events in via the prior on motif occurrences, $\Pr(X|M, f, O, D, g)$, where $g = \{g_1, g_2\}$ models the distance between cross-link site and motif occurrences, as

$$\Pr(X|O, D, g) = \prod_{i=1}^{n} \prod_{j=1}^{m-w+1} \Pr(X_{ij} = 1|O_i = 1, D, g)^{X_{ij}},$$

where

$$\begin{aligned}\Pr(X_{ij} = 1|O_i = 1, D, g) &= \sum_{l=1}^{m} \Pr(C_i = l|O_i = 1, D) \times\\ &\quad \left[g_1(1 - g_1)^{|l-(j+g_2)|}\right]^K.\end{aligned}$$

Note that $\Pr(X|M, f, O, D, g) = \Pr(X|O, D, g)$, since the prior has no dependence on the sequence/structure parameters of the model. We decided to model the relationship between diagnostic events and binding site positions using a geometric distribution, where $g = \{g_1, g_2\}$ has probability g_1 and location parameter g_2 to indicate the typical offset between the cross-link site and the start of the binding site. The reason for using a geometric distribution is primarily based on observing what seems to be an exponential decay in the

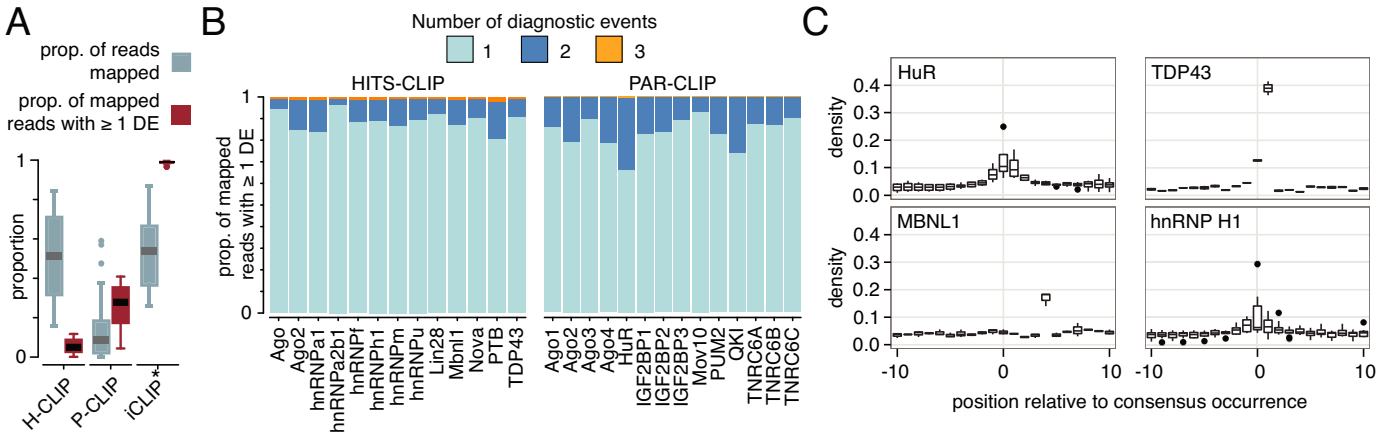

Figure 2. Properties of diagnostic events in CLIP-seq data. (**A**) Percent of reads mapped, and mapped reads with diagnostic events for HITS-CLIP, PAR-CLIP and iCLIP (*In iCLIP all the reads without deletion, are assumed to be truncated at cross- link location.). (**B**) Number of diagnostic events within reads that contain at least one event. Most of the reads have exactly one diagnostic events. (**C**) Density of diagnostic events relative to occurrences of an expected consensus sequence in the top 1000 CLIP-identified target sequences for HuR, hnRNP H1, MBNL1 and TDP43; data from (21,23,25,34).

density of diagnostic events moving away from what we believe to be true binding sites in many of the higher quality data sets. This also makes sense conceptually in iCLIP, since the rate parameter can be viewed as the probability that the truncation will mistakenly occur one base too soon or too late.

Our exploration of CLIP-seq data highlighted the fact that cross-linking is often quite noisy. To balance the impact of CLIP-seq cross-link events against sequence and structure information, we include the parameter K – a tuning parameter that modulates the impact of diagnostic events on the algorithm. Low values (near 0) reduce the impact of diagnostic events, while higher values increase it. For all results reported here, we fixed this at 1.1 (see supplementary for details on selection of this), which is the default in our implementation. This is a somewhat conservative value, but users can increase this if they feel their data is of higher quality and as the CLIP-seq assay continues to improve.

Our goal in characterizing binding specificity from the data is to find estimates for the model parameters M and f. The traditional formulation, without secondary structure indicators or diagnostic event locations, treats the sites X as missing data, and uses a method like EM or Gibbs Sampling. Our model has two important properties that facilitate using an EM algorithm to estimate the model parameters:

- We assume the values of the structure indicators T are not dependent on the model parameters. We also have a method to compute expected values for T: we can use the partition function algorithm due to McCaskill (37), as explained in supplementary methods. Hence, in using EM to estimate M and f, we need only recompute expectations for the values of X at each iteration, while T remains static. We also remark here that in theory we could re-estimate values for T using a restricted partition function algorithm (44), but doing so would be computationally prohibitive.
- The contribution of diagnostic events, given the motif locations is independent of the sequences S and secondary structure indicators T, as well as the model parameters M and f. This makes it easy to decompose the estimation procedure into steps involving S and T and other steps involving D. This has intuitive appeal: if every observed diagnostic event corresponded to a functional binding event, and if every functional binding event resulted in a diagnostic event, then we should be able to estimate the binding site indicators X using data from D alone, only involving the parameters g.

Use of RNA secondary structure and diagnostic events improves motif-discovery in simulated data

To examine the extent to which our expanded model improves motif recovery, we produced a set of simulated datasets. For each simulation we generated a random position weight matrix of length six, and used this to plant occurrences into 500 sequences of length 50nt, that were randomly selected from human 3′ UTRs. For equal proportions of the datasets, we fixed the secondary structure of the occurrence to be single-stranded in 100% of the occurrences, 90% of the occurrences and so on down to 0% of the occurrences (we call this the structure-fraction of the dataset). We also simulated diagnostic events with distance geometrically distributed around some offset from the motif occurrence. This offset is fixed for each dataset, but varies uniformly at random in all dataset on the range ± 8nt. Similar to the structure, for equal proportions of the datasets, we planted these diagnostic events for 100% of the motif occurrences, 90% of the occurrences and so on down to 0% of the occurrences (we call this the DE-fraction of the dataset). Further details of the simulation process are given in 'Materials and Methods' section.

For each simulated dataset we ran Zagros in four different ways: with just the simulated sequences; with the sequences and base-pair probabilities (RNA secondary structure); with the sequences and diagnostic event locations; and with the sequences, structure, and diagnostic events. Although Zagros can be run in these four different modes, when we refer simply to Zagros, we mean the version us-

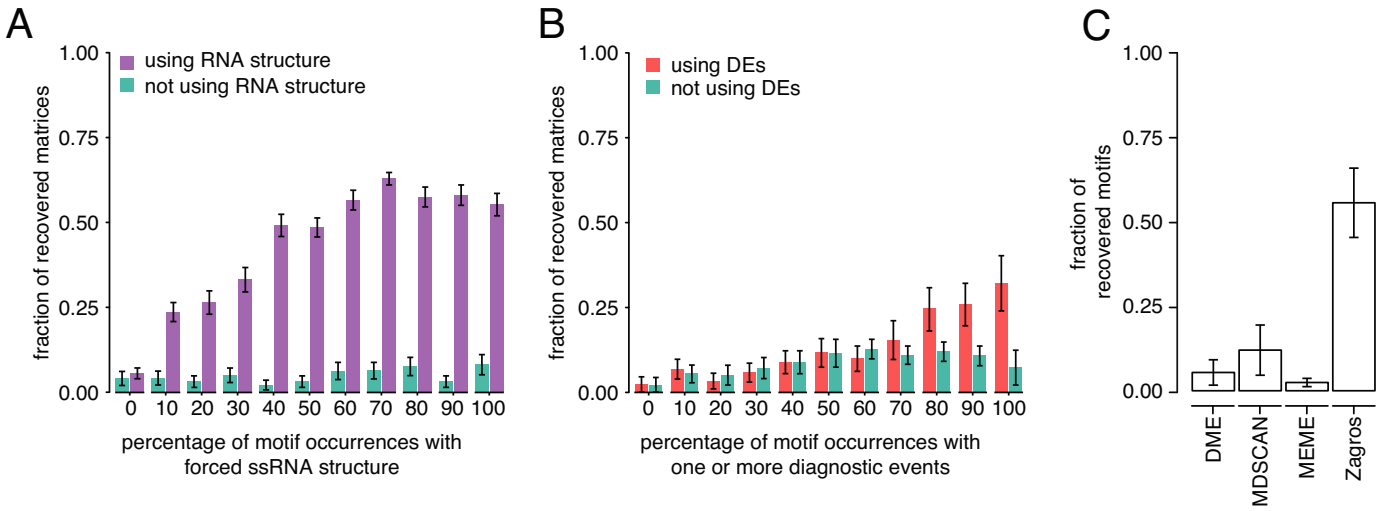

Figure 3. Use of RNA structure and diagnostic events improves motif-discovery performance on simulated data and outperforms other methods. (**A**) Proportion of recovered motifs as a function of the fraction of planted motif occurrences that are forced to adopt a specific RNA secondary structure (in this case, ssRNA). Zagros was run with either just sequence data, or both the sequences and the structure (base-pair probabilities). For each fraction of occurrences with planted motifs, we simulated 500 random datasets. Error-bars are 95% confidence interval from 1000 bootstrap samples. (**B**) As in panel (A), but as a function of the fraction of planted motif occurrences that have one or more diagnostic events. Zagros was run either with just sequence data, or sequence and diagnostic events data. (**C**) Comparison with other methods.

ing all three of sequence, structure and DEs. In each case, we determined whether the motif recovered by Zagros was a match to the planted one (see 'Materials and Methods' section for details), and calculated the fraction of datasets for which each method was able to recover the planted motif. As the fraction of motif occurrences with fixed structure increases, the ability of Zagros to recover the motif if structural information is provided also increases, while it remains stationary if only sequence information is provided (see Figure 3A). Similar, although less dramatic, improvements in performance are observed as the fraction of motif occurrences with diagnostic events increases (Figure 3B). It is possible to increase the impact of diagnostic events by increasing the value of the parameter *K*. Although this results in stronger performance on simulated data, we elected not to do this as it has an adverse impact on the algorithm's performance with CLIP-seq derived data, which has higher levels of more heterogeneous cross-link noise.

Figure 3C shows the accuracy of Zagros (using sequence, structure and diagnostic events) compared to three well-known motif discovery tools (MEME, DME and MDSCAN). We ran all of these methods on the same simulated datasets described above. As shown in Figure 3C, using the extra information of structure and diagnostic events allows Zagros to clearly achieve the best performance of the methods tested. The stark contrast is due to the low information content of the motif, making discovery by sequence alone highly challenging. The bars show the fraction of motifs recovered by each program; error bars indicate the standard deviation.

Zagros recovers previously validated motifs from CLIP-seq data

We compared the performance of Zagros to DME, MDSCAN and MEME on a set of datasets derived from a range of CLIP-seq experiments comprised of multiple replicates. We selected a subset of 19 RBPs from this data for which a sequence preference had previously been reported (excluding miRNA-associated RBPs, such as Ago2; list of selected datasets and details of sequence selection provided in supplementary materials). We report results for those replicates where at least one of the tested methods was able to recover the expected motif. When one of the algorithms designed for transcription factor binding sites finds the reverse compliment of a motif, we count that as a success, since they could easily be modified to be strand specific. Figure 4 shows twelve example datasets and the top-scoring motif recovered for each dataset; the previously described binding site for each is also shown. Note that in all the examples, the match to the previously described site is obtained by Zagros using all three of sequence, structure and diagnostic events.

Although the other tested programs sometimes recover the expected motif, Zagros achieves the most consistent recovery. In some cases, it is clear that just the addition of structure is sufficient, such as with HuR. In other cases, such as for the IGF2BP proteins and hnRNPC, Zagros achieves a close match to the expected motif only through the use of both structure and diagnostic events. In general, the combination of both extra pieces of information gives the best result. Logos for all tested datasets are provided in Supplementary Figures S1–S5.

Zagros uncovers a potential link between sequence- and structure-specificity in RBP binding sites

We applied Zagros to investigate the relationship between structure and sequence specificity of RBP binding. We determined the sequence specificity of the motif reported by Zagros as the frequency with which the consensus sequence from the motif occurred in either human or mouse exons, depending on the organism for which the CLIP experiment

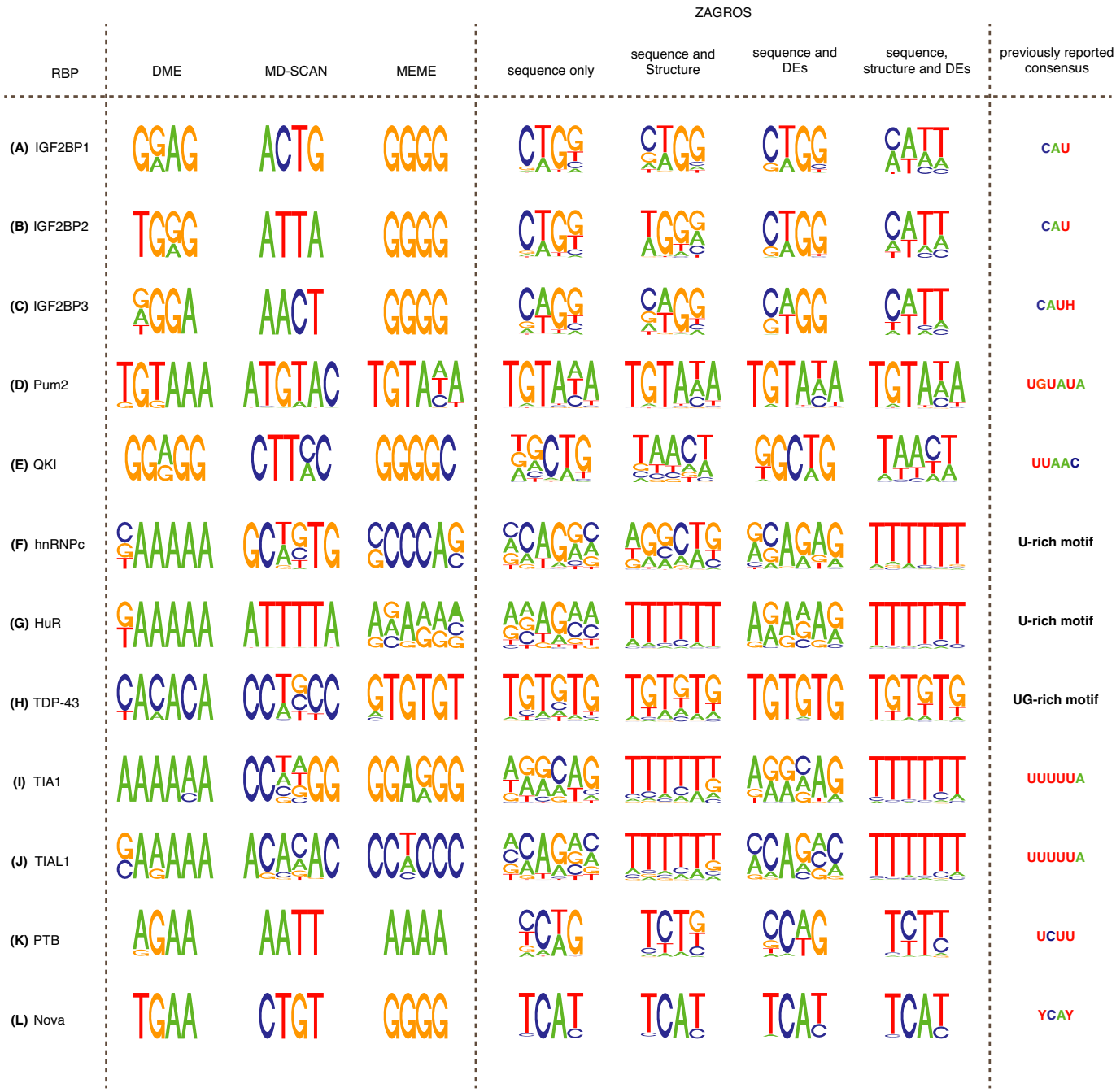

Figure 4. Use of RNA structure and diagnostic events improves recovery of expected motifs from CLIP-seq data Top-scoring motifs recovered by Zagros on twelve example CLIP-seq datasets. For each dataset we show the motif recovered by DME, MDSCAN and MEME in addition to each version of Zagros. The data shown in this figure is obtained from: A, B, C, D, E (8) – F (9) – G (35) – H (21) – I, J (20) – K (29) – L (22).

was conducted. Consensus sequences that occur more frequently are considered to have low-specificity, while those that occurred relatively rarely we considered to have high-specificity. We then ordered RBPs by their mean specificity and plotted the mean base-pair probabilities recovered by Zagros for each (Figure 5). There was a negative correlation (Spearman's correlation coefficient: −0.36) between sequence specificity and mean base-pair probability. In their seminal work, Schneider *et al.* (45) found that binding sites tend to contain enough information for them to be recognized, but not more than is required. RBP binding sites tend to be short and degenerate, leading to generally low sequence-based information. It is reasonable to assume then that other features, such as secondary structure, contribute the necessary information to overcome the deficit and allow the sites to be recognized. Importantly though, the work of Schneider *et al.* indicates that (owing to random drift), the additional information contributed will not be more than is required for recognition. Our results are supportive of this

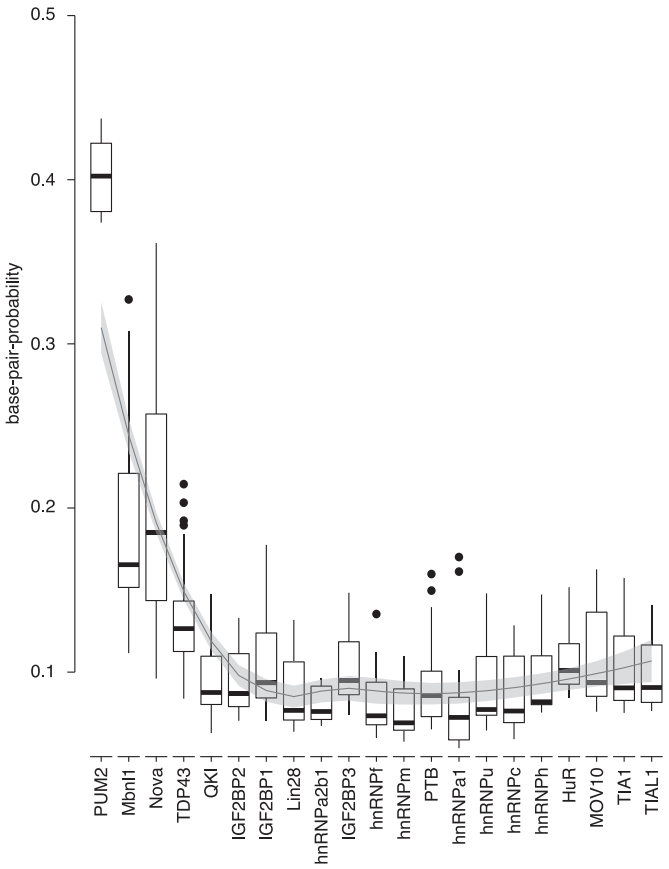

Figure 5. Use of RNA structure and diagnostic events reveals a potential link between sequence- and structure-specificity in RBP binding sites The average base-pair probability of the recovered motifs for RRM-containing RBPs. The *x*-axis is sorted by the average specificity of the sequence component of the recovered motifs; those motifs with high sequence specificity (i.e. for which matches to the sequence component of their recovered motif are rare) on the left, and those with low specificity (i.e. for which matches to their sequence are common) are on the right.

position, as we observe that motifs with more informative sequence are less structured, and vice versa.

CONCLUSION

Motif discovery in CLIP-seq data is a challenging problem due to the relatively short length of RBP binding motifs, potentially low levels of sequence specificity, and biases in the CLIP-seq protocol. However, there are opportunities to improve performance by leveraging attributes specific to RBP binding and CLIP-seq data.

We have demonstrated here that 3′ UTRs identified as targets by CLIP-seq contain distinct structural signals that are not trivially dependent upon sequence, providing additional high-throughput evidence that structure plays a role in defining RBP binding sites. We also showed that cross-linking induced diagnostic events follow RBP-specific patterns and are enriched around motif occurrences.

Our proposed model brings together the sequence and structure of RBP binding sites, and augments this with technical information from the CLIP-seq assay (cross-linking induced events, or diagnostic events). Our model, when fit using the expectation maximization algorithm, showed much improved performance in simulated datasets with structural specificity and informative diagnostic events. Moreover, we demonstrated that this approach recovers meaningful motifs from CLIP-seq datasets.

Finally, our finding of a correlation between the sequence-specificity of motifs recovered for RRM-containing RBPs (as measured by the frequency that the consensus occurs in exonic sequences) and the strength of mean base-pair probability for the motif support the hypothesis that sequence and structure function in concert to achieve RBP binding specificity.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health [R01HG006015].
Conflict of interest statement. None declared.

REFERENCES

1. Tenenbaum,S.A., Carson,C.C., Lager,P.J. and Keene,J.D. (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl Acad. Sci. U.S.A.*, **97**, 14085–14090.
2. Ule,J., Jensen,K., Mele,A. and Darnell,R.B. (2005) CLIP: A method for identifying protein–RNA interaction sites in living cells. *Methods*, **37**, 376–386.
3. 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.
4. Lukong,K.E., wei Chang,K., Khandjian,E.W. and Richard,S. (2008) RNA-binding proteins in human genetic disease. *Trends Genet.*, **24**, 416–425.
5. Lunde,B.M., Moore,C. and Varani,G. (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.
6. Moore,M.J. (2005) From birth to death: the complex lives of eukaryotic mRNAs. *Science*, **309**, 1514–8.
7. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., Li,X., Gueroussov,S., Albu,M., Zheng,H., Yang,A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.
8. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Jr,M.A., Jungkamp,A.-C., Munschauer,M., Ulrich,A. *et al.* (2010) Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, **141**, 129–141.
9. Konig,J., Zarnack,K., Rot,G., Curk,T., Kayikci,M., Zupan,B., Turner,D., Luscombe,N. and Ule,J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct. Mol. Biol.*, **17**, 909–915.
10. Sugimoto,Y., Konig,J., Hussain,S., Zupan,B., Curk,T., Frye,M. and Ule,J. (2012) Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.*, **13**, R67.
11. Friedersdorf,M. and Keene,J. (2014) Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol.*, **15**, R2.
12. Zhang,C., Lee,K.-Y., Swanson,M.S. and Darnell,R.B. (2013) Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Res.*, **41**, 6793–6807.
13. Li,X., Quon,G., Lipshitz,H.D. and Morris,Q. (2010) Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, **16**, 1096–1107.

14. Kazan,H., Ray,D., Chan,E.T., Hughes,T.R. and Morris,Q. (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, **6**, e1000832.
15. Granneman,S., Kudla,G., Petfalski,E. and Tollervey,D. (2009) Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9613–9618.
16. Kishore,S., Jaskiewicz,L., Burger,L., Hausser,J., Khorshid,M. and Zavolan,M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, **8**, 559–564.
17. Zhang,C. and Darnell,R.B. (2011) Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotech.*, **29**, 607–614.
18. Corcoran,D., Georgiev,S., Mukherjee,N., Gottwein,E., Skalsky,R., Keene,J. and Ohler,U. (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, **12**, R79.
19. Uren,P.J., Burns,S.C., Ruan,J., Singh,K.K., Smith,A.D. and Penalva,L.O.F. (2011) Genomic analyses of the RNA binding protein Hu Antigen R (HuR) identify a complex network of target genes and novel characteristics of its binding sites. *J. Biol. Chem.,* **286**, 37063–37066.
20. Wang,Z., Kayikci,M., Briese,M., Zarnack,K., Luscombe,N.M., Rot,G., Zupan,B., Curk,T. and Ule,J. (2010) iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol.*, **8**, e1000530.
21. Tollervey,J.R., Curk,T., Rogelj,B., Briese,M., Cereda,M., Kayikci,M., Konig,J., Hortobagyi,T., Nishimura,A.L., Zupunski,V. *et al.* (2011) Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.*, **14**, 452–458.
22. Zhang,C., Frias,M.A., Mele,A., Ruggiu,M., Eom,T., Marney,C.B., Wang,H., Licatalosi,D.D., Fak,J.J. and Darnell,R.B. (2010) Integrative modeling defines the nova splicing-regulatory network and its combinatorial controls. *Science*, **329**, 439–443.
23. Wang,E.T., Cody,N.A., Jog,S., Biancolella,M., Wang,T.T., Treacy,D.J., Luo,S., Schroth,G.P., Housman,D.E., Reddy,S. *et al.* (2012) Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell*, **150**, 710–724.
24. Yeo,G.W., Coufal,N.G., Liang,T.Y., Peng,G.E., Fu,X.-D. and Gage,F.H. (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat. Struct. Mol. Biol.*, **16**, 130–137.
25. Katz,Y., Wang,E.T., Airoldi,E.M. and Burge,C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.
26. Huelga,S.C., Vu,A.Q., Arnold,J.D., Liang,T.Y., Liu,P.P., Yan,B.Y., Donohue,J.P., Shiue,L., Hoon,S., Brenner,S. *et al.* (2012) Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep.*, **1**, 167–178.
27. Li,F., Zheng,Q., Vandivier,L.E., Willmann,M.R., Chen,Y. and Gregory,B.D. (2012) Regulatory impact of RNA secondary structure across the arabidopsis transcriptome. *Plant Cell Online*, **24**, 4346–4359.
28. Chi,S.W., Zang,J.B., Mele,A. and Darnell,R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.
29. Xue,Y., Zhou,Y., Wu,T., Zhu,T., Ji,X., Kwon,Y.-S., Zhang,C., Yeo,G., Black,D.L., Sun,H. *et al.* (2009) Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell*, **36**, 996–1006.
30. 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.
31. Leung,A. K.L., Young,A.G., Bhutkar,A., Zheng,G.X., Bosson,A.D., Nielsen,C.B. and Sharp,P.A. (2011) Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 237–244.
32. Polymenidou,M., Lagier-Tourenne,C., Hutt,K.R., Huelga,S.C., Moran,J., Liang,T.Y., Ling,S.-C., Sun,E., Wancewbicz,E., Mazur,C. *et al.* (2011) Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat. Neurosci.*, **14**, 459–468.
33. Sievers,C., Schlumpf,T., Sawarkar,R., Comoglio,F. and Paro,R. (2012) Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res.*, **40**, e160–e160.
34. Lebedeva,S., Jens,M., Theil,K., Schwanhäusser,B., Selbach,M., Landthaler,M. and Rajewsky,N. (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell*, **43**, 340–352.
35. Mukherjee,N., Corcoran,D.L., Nusbaum,J.D., Reid,D.W., Georgiev,S., Hafner,M., Jr,M.A., Tuschl,T., Ohler,U. and Keene,J.D. (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell*, **43**, 327–339.
36. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
37. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
38. Ray,D., Kazan,H., Chan,E.T., Castillo,L.P., Chaudhry,S., Talukder,S., Blencowe,B.J., Morris,Q. and Hughes,T.R. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.*, **27**, 667–670.
39. Scheibe,M., Butter,F., Hafner,M., Tuschl,T. and Mann,M. (2012) Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions. *Nucleic Acids Res.*, **40**, 9897–9902.
40. Kim,H., Headey,S., Yoga,Y., Scanlon,M., Gorospe,M., Wilce,M. and JA,W. (2013) Distinct binding properties of TIAR RRMs and linker region. *RNA Biol.*, **10**, 579–589.
41. Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins-Structu. Funct. Bioinform.*, **7**, 41–51.
42. Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.
43. Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. In: Machine Learning, pp. 51–80.
44. Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
45. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.