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

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
[2]Children's Cancer Research Institute, University of Texas Health Science Center

Contents

1 Supplementary materials

1.1 Public CLIP-Seq data used

The full list of data used is given in Table S1.

RBP	Technology	Cell	Citation
Ago	HITS-CLIP	HeLa	[4]
Ago{1..4}	PAR-CLIP	HEK293	[9]
IGF2BP{1..3}	PAR-CLIP	HEK293	[9]
PUM2	PAR-CLIP	HEK293	[9]
QKI	PAR-CLIP	HEK293	[9]
TNRC6{A..C}	PAR-CLIP	HEK293	[9]
hnRNP H1	HITS-CLIP	HEK293	[11]
hnRNP a2b1	HITS-CLIP	HEK293	[10]
hnRNP {a1,F,M,U}	HITS-CLIP	HEK293	[10]
Lin28	HITS-CLIP	HEK293, hESC	[18]
MOV10	PAR-CLIP	HEK293	[26]
Ago2, HuR	HITS-CLIP, PAR-CLIP	HEK293	[12]
hnRNP C	iCLIP	HeLa	[13]
HuR	PAR-CLIP	HeLa	[16]
HuR	PAR-CLIP	HEK293	[21]
HuR	iCLIP	HeLa	[30]
TIA1, TIAL1	iCLIP	HeLa	[32]
PTB	HITS-CLIP	HeLa	[33]
TDP43	iCLIP	SH-SY5Y	[28]
Ago2	HITS-CLIP	HEK293	[29]
hTra2	RIP-Seq	HeLa	[29]
Ago2	HITS-CLIP	mESC	[17]
TDP43	HITS-CLIP	Mouse brain	[23]
Nova	HITS-CLIP	Mouse brain	[35]
Nova	iCLIP	Mouse brain	[27]
Mbnl1	HITS-CLIP	Mouse brain, C2C12, heart, muscle	[31]

Table S1: List of CLIP-Seq data sets used

1.2 Previously described RBP binding preferences

To evaluate the performance of Zagros on existing CLIP-Seq datasets we compared the motif recovered with a consensus sequence built from previously described binding preferences of each RBP examined. These consensus sequences and their origins are given in Table S2.

RBP	Consensus
HuR	UUUUU [24]
PTB	CUCUCU [22]
QKI2	ACUAA [7]
Nova	YCAY [2, 5]
TIA1	UUUUA [32, 6]
TIAL1	UUUUA [32]
PUM2	UGUAUAUA [8]
TDP43	UGUGU [14]
hnRNP C	UUUUU [34]
hnRNP H	GGGA [3]
IGF2BP1,2,3	CATH [9, 25]

Table S2: Previously described consensus sequences used in evaluating performance

2 Supplementary methods

Here we present the complete formalization of our method. The data available for the solution of the motif-finding problem is the primary sequences, the secondary structure of those sequences, and the location of the cross-link induced read artifacts, which we call diagnostic events. Given that the primary sequence will always be used, the optional use of the other two pieces of information gives us four different ways our algorithm can be run. The first of these uses only sequence information; the second uses sequence information and pre-computed base-pair probabilities informing us about the structure of the sequences; the third uses sequence information and the locations of the diagnostic events; and finally, the fourth way of running the algorithm uses all of sequence, structure and diagnostic events information. Since we present results for each of these four approaches, we will fully describe here both the model and how the algorithm works in each case. Throughout, we will assume the length of the RBP binding motif is fixed at w nucleotides. Further, we assume that any given sequence may either contain an occurrence of the motif, or it might not – this is the so-called zero-or-one-occurrence-per-sequence assumption, or ZOOPS. Much literature exists on the problem of motif discovery, which we will attempt not to rehash here; for further background and details please refer to [15, 19, 1].

2.1 Observed data

Let $S = \{S_1, S_2, \ldots, S_n\}$ be a set of unaligned RNA sequences over the alphabet $\Sigma = \{A, C, G, U\}$, obtained from a CLIP-seq experiment. Without loss of generality, to make the notations simpler we assume a fixed length, m, for all the sequences. Let T represent the secondary structure of the sequences in S, such that $T_{ij} = 1$ if nucleotide j in sequence i is paired, and $T_{ij} = 0$ otherwise. The structure of an RNA molecule is completely determined by its primary sequence; in describing our model and algorithms then, we will assume that T is known whenever S is also known (in practice, we estimate T using McCaskill's algorithm – see Section 2.6.3).

2.2 Latent data

Under the zero-or-one-occurrence-per-sequence (ZOOPS) model, each sequence contains either zero or one occurrences of the RBP binding motif; we treat the presence and location of the motif as latent data. Let $X_{ij} = 1$ if the motif occurrence for sequence i is at position j, and $X_{ij} = 0$ otherwise. For notational convenience, we also define $O_i = 1$ if there is a motif occurrence in sequence i and $O_i = 0$ otherwise. Notice that O is completely specified by X, since $O_i = \sum_{j=1}^{m-w+1} X_{ij}$.

2.3 General model description

In all cases, our model will contain a representation of the sequence of the RBP binding site. This will be augmented by additional information about cross-linking and structure when such information is used. To avoid repetition, we start by describing the sequence component of the model here, and defer description of the cross-link and structure components until Sections 2.5 and 2.6 respectively. Further, the ZOOPS assumption requires augmenting the model with an additional parameter, which we also describe here as it is common to all four variants of Zagros.

We model the sequence of motif occurrences using a position weight matrix (product multinomial distribution), M and background distribution f. More specifically, $M = (M_k)_{k=1}^{w}$, where for any $b \in \Sigma$, we have $M_k(b) = \Pr\big(b \text{ appears at position } k \text{ of the motif}\big)$, and $f(b) = \Pr\big(b \text{ appears in background}\big)$. To account for the ZOOPS assumption, we introduce the model parameter γ, which is the probability that a sequence contains the motif. Throughout, we will refer to the model as Θ. Hence, our basic model, which we will extend later for structure and diagnostic events, is $\Theta = \{M, f, \gamma\}$.

2.4 Sequence-only motif discovery

Considering only the sequence, the model parameter set is defined as $\Theta = \{M, f, \gamma\}$. The complete-data likelihood for this model is

$$\begin{aligned}
\Pr(S, X|\Theta) &= \prod_{i=1}^{n} \Pr(S_i, X_i|\Theta) \\
&= \prod_{i=1}^{n} \Pr(S_i, X_i|\Theta, O_i) \Pr(O_i|\Theta) \\
&= \Pr(O|\Theta) \prod_{i=1}^{n} \Pr(S_i, X_i|O_i, \Theta) \\
&= \Pr(O|\Theta) \prod_{i=1}^{n} \Pr(S_i|O_i, X_i, \Theta) \Pr(X_i|O_i, \Theta), \\
&= \Pr(O|\Theta) \Pr(X|O, \Theta) \prod_{i=1}^{n} \Pr(S_i|O_i = 0, \Theta)^{(1-O_i)} \prod_{j=1}^{m-w+1} \Pr(S_i|X_{ij} = 1, \Theta)^{X_{ij}}. \qquad (1)
\end{aligned}$$

Here, we have two priors, $\Pr(O|\Theta)$ and $\Pr(X|O, \Theta)$. The first is the prior probability of a motif occurrence, given that the model is known, while the second is the prior probability of the occurrence indicators, given knowledge of the model and which sequences contain the motif occurrences. In the context of the sequence-only model, these two priors may be calculated as

$$\Pr(O|\Theta) = \Pr(O|\gamma) = \gamma^q (1-\gamma)^{n-q},$$

where we define q to be the number of sequences which contain an occurrence of the motif, i.e. $q = \sum_{i=1}^{n} O_i$, and

$$\Pr(X|O, \Theta) = \prod_{i=1}^{n} \prod_{j=1}^{m-w+1} \left(\frac{1}{m-w+1}\right)^{X_{ij}}.$$

The complete-data likelihood also contains expressions for the probability of observing a sequence, given the location of the motif in the sequence and the model are known, i.e. $\Pr(S_i|X_{ij} = 1, \Theta)$, as well as the special case where no motif occurrence exists in the sequence, i.e. $\Pr(S_i|O_i = 0, \Theta)$. Again, in the context of the sequence-only model, we may calculate these as

$$\Pr(S_i|X_{ij} = 1, \Theta) = \prod_{l=1}^{m} \prod_{b=A}^{U} (\Psi_{blj})^{\nu_{bli}},$$

and

$$\Pr(S_i|O_i = 0, \Theta) = \prod_{l=1}^{m} \prod_{b=A}^{U} f(b)^{\nu_{bli}}.$$

Here we introduce the indicator variable ν, which is defined such that $\nu_{bli} = 1$ if base b appears at position l of sequence i and, $\nu_{bli} = 0$ otherwise. Finally, we have $\Psi_{b,l,j}$, which is the foreground PWM (i.e. M) when l falls within the motif occurrence starting at position j, or the background model f if it does not. More formally,

$$\Psi_{blj} = \begin{cases} M_{l-j+1}(b) & \text{if } j \le l \le j+w-1, \\ f(b) & \text{otherwise.} \end{cases}$$

We employ the expectation maximization (EM) algorithm to estimate parameters for our model, Θ. The algorithm iterates between an expectation step, in which the expected value of the log-likelihood is calculated using current estimates of the parameters and a maximization step, where the model parameters are optimized to find the maximum of this function.

2.4.1 Expectation-maximization – E-step

The log of the complete-data likelihood is easily derived from Equation 1 as

$$\begin{aligned}\log \ell(\Theta|S,X) &= \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} X_{ij}\log\left(\frac{1}{m-w+1}\right)\\ &\quad + q\log\gamma(n-q)\log(1-\gamma)\\ &\quad + \sum_{i=1}^{n}(1-\textstyle\sum_{j=1}^{m-w+1}X_{ij})\displaystyle\sum_{l=1}^{m}\sum_{b=A}^{U}\nu_{bli}\log f(b)\\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} X_{ij}\sum_{l=1}^{m}\sum_{b=A}^{U}\nu_{bli}\log\Psi_{bl'}. \end{aligned} \tag{2}$$

Let us define $Q(\Theta,\Theta^{(t)})$ as the expected value of the log-likelihood function with respect to X, given our current model estimate, which we will represent as $\Theta^{(t)}$, and our observed data S. Hence,

$$\begin{aligned} Q(\Theta,\Theta^{(t)}) &= E_{X|\Theta^{(t)},S}(\log\ell(\Theta|S,X))\\ &= \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij})\log\left(\frac{1}{m-w+1}\right)\\ &\quad + E_{X|\Theta^{(t)},S}(q)\log\gamma(n-E_{X|\Theta^{(t)},S}(q))\log(1-\gamma)\\ &\quad + \sum_{i=1}^{n}(1-\textstyle\sum_{j=1}^{m-w+1}E_{X|\Theta^{(t)},S}(X_{ij}))\displaystyle\sum_{l=1}^{m}\sum_{b=A}^{U}\nu_{bli}\log f(b)\\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij})\sum_{l=1}^{m}\sum_{b=A}^{U}\nu_{bli}\log\Psi_{bl'}. \end{aligned} \tag{3}$$

Note here that we can replace the expected value of q with the explicit summation, i.e.

$$E_{X|\Theta^{(t)},S}(q) = \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij}).$$

Notice also that, since X_{ij} are Bernoulli random variables, $E_{X|\Theta^{(t)},S}(X_{ij}) = \Pr(X_{ij}=1|S_i,\Theta^{(t)})$, which we may compute as

$$\Pr(X_{ij}=1|S_i,\Theta^{(t)}) = \frac{\Pr(S_i|X_{ij}=1,\Theta^{(t)})\Pr(X_{ij}=1|\Theta^{(t)})}{\Pr(S_i|\Theta^{(t)},O_i=0)\Pr(O_i=0|\Theta^{(t)}) + \sum_{j'=1}^{m-w+1}\Pr(S_i|X_{ij'}=1,\Theta^{(t)})\Pr(X_{ij'}=1|\Theta^{(t)})}.$$

Here, the prior probability on occurrence at any given position can easily be calculated by noting that

$$\Pr(X_{ij}=1|\Theta^{(t)}) = \Pr(X_{ij}=1|O_i=1,\Theta^{(t)})\Pr(O_i=1|\Theta^{(t)}) = \frac{\gamma^{(t)}}{m-w+1}.$$

The prior on non-occurrence of the motif is also straight-forward: $\Pr(O_i=0|\Theta^{(t)}) = 1-\gamma^{(t)}$. The remaining two probabilities on observing the sequence, given either where the motif occurs, or the special case of non-occurrence, have the same form as in the complete-data likelihood, i.e.

$$\Pr(S_i|X_{ij}=1,\Theta^{(t)}) = \prod_{l=1}^{m}\prod_{b=A}^{U}(\Psi_{blj}^{(t)})^{\nu_{bli}},$$

and

$$\Pr(S_i|O_i = 0, \Theta^{(t)}) = \prod_{l=1}^{m} \prod_{b=A}^{U} f^{(t)}(b)^{\nu_{bli}}.$$

The same can be said for the definition of $\Psi_{blj}^{(t)}$, which is analogously defined as

$$\Psi_{blj}^{(t)} = \begin{cases} M_{l-j+1}^{(t)}(b) & \text{if } j \leq l \leq j+w-1, \\ f^{(t)}(b) & \text{otherwise.} \end{cases}$$

2.4.2 Expectation-maximization – M-step

Maximizing for $\Theta^{(t+1)}$ with respect to $Q(\Theta, \Theta^{(t)})$ is strightforward. For M, the MLE is

$$\hat{M}_k(b) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m-w+1} (I_{\{s_{i,j+k-1}=b\}}) \Pr(X_{ij} = 1|\Theta^{(t)}), \tag{4}$$

where $I_{\{s_{i,j}=b\}}$ is the indicator function, equal to 1 when sequence s_i contains base b at position j, and 0 otherwise. The MLE estimate for f is defined analogously as

$$\hat{f}(b) = \frac{1}{n(m-w)} \sum_{i=1}^{n} \sum_{l=1}^{m} \sum_{j=1}^{m-w+1} (I_{\{s_{i,l}=b\}})(I_{\{l \notin [j,j+w-1]\}}) \Pr(X_{ij} = 1|\Theta^{(t)}). \tag{5}$$

Finally, the MLE for the ZOOPS model parameter, γ, is

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m-w+1} \Pr(X_{ij} = 1|\Theta^{(t)})}{n}. \tag{6}$$

2.5 Sequence and diagnostic events motif discovery

CLIP-seq uses UV light to induce cross-links between RBPs and RNA molecules at the point of interaction. A high density of reads at a given genomic locus indicates that locus is likely to be a binding site. Individual reads can contain artifacts that localize the cross-link location to a single nucleotide (we call these artifacts "diagnostic events"; see the main manuscript for details of how we extract these from reads). In bringing the locations of diagnostic events into our algorithm to assist in locating RBP binding sites, we make some simplifying assumptions:

- each sequence comes from a distinct RNA transcript, and every one of these transcripts was cross-linked to the RBP of interest.
- if an RNA transcript is cross-linked, it will be cross-linked at exactly one location.
- all sequences have equal affinity for the RBP, given that they contain an occurrence of the motif (put another way, we do not assume some sequences with motif occurrence are more informative than others).

Let C_i be the (unknown) location of the cross-link for the i^{th} sequence. We assume the distances of the motif occurrences from the cross-link locations (i.e. $|C_i - j + g_2|$, given that $X_{ij} = 1$) follow a geometric distribution parameterised by g_1. Here, g_2 is an offset from the start of the motif occurrence to account for our observation, from CLIP-seq data, that the most likely cross-link location is not always at the motif start location. It is worth pointing out that we do not treat C as latent data. We include the number of diagnostic events in each sequence i at each position j as a fixed parameter of our model, D_{ij}. Hence, the expanded model is defined as follows: $\Theta = \{M, f, \gamma, g_1, g_2, D\}$. We bring the influence of cross-linking information into our algorithm via the prior on motif occurrence locations, $\Pr(X|O, \Theta)$. As such, the form of the complete-data likelihood for the model remains

unchanged from Equation 1. The only difference is in computing the prior on X, which can be decomposed as

$$\begin{aligned}\Pr(X|O,\Theta) &= \prod_{i=1}^{n}\Pr(X_i|O_i,\Theta) \\ &= \prod_{i=1}^{n}\Pr(X_i|O_i=1,\Theta)^{O_i}\Pr(X_i|O_i=0,\Theta)^{1-O_i} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{m-w+1}\Pr(X_{ij}=1|O_i=1,\Theta)^{X_{ij}}. \end{aligned} \quad (7)$$

Notice here that $\Pr(X_i|O_i=0,\Theta)=1$. As noted above, we assume that the distance of motif occurrences from the cross-link location follows a geometric distribution; we integrate over all possible cross-link locations, hence

$$\Pr(X_{ij}=1|O_i=1,\Theta) = \sum_{l=1}^{m}\Pr(C_i=l|O_i=1,\Theta)\left[g_1(1-g_1)^{|l-(j+g_2)|}\right]^K.$$

Here, K is 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. In practice, we set this heuristically using an exhaustive search of possible values and picking the one that maximized the number of motifs recovered from our collection of CLIP-seq data. To make sure that we are not over-fitting this parameter, we performed 1000 bootstrap samples of our data collection and a value of 1.1 proved to be the optimal value. However, this parameter can be adjusted by users if they wish to do so. Using our selection of current publicly available datasets, the default value seems to be optimal, however as CLIP experiment improves the users might feel the need to increase this value for their data sets. The probability of the cross-link location for sequence i being at position l is estimated from the diagnostic events, so

$$\Pr(C_i=l|O_i=1,\Theta)=\frac{D_{ij}+\epsilon}{\sum_{j'=1}^{m}(D_{ij'}+\epsilon)}, \quad (8)$$

or put less formally, the fraction of diagnostic events observed at that location. Here ϵ is a pseudo-count to avoid division by 0 in any sequences with no diagnostic events – in practice we set $\epsilon=1$.

2.5.1 Expectation Maximization – E-step

The log-likelihood of the expanded model is easily arrived at by replacing the prior on X in Equation 2 with the one given by Equation 7, hence

$$\begin{aligned}\log\ell(\Theta|S,X) &= \sum_{i=1}^{n}\sum_{j=1}^{m-w+1}X_{ij}\log\left(\sum_{l=1}^{m}\Pr(C_i=l|O_i=1,\Theta^{(t)})\left[g_1(1-g_1)^{|l-(j+g_2)|}\right]^K\right) \\ &+ q\log\gamma+(n-q)\log(1-\gamma) \\ &+ \sum_{i=1}^{n}(1-\textstyle\sum_{j=1}^{m-w+1}X_{ij})\displaystyle\sum_{l=1}^{m}\sum_{b=A}^{U}\nu_{bli}\log f(b) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{m-w+1}X_{ij}\sum_{l=1}^{m}\sum_{b=A}^{U}\nu_{bli}\log\Psi_{bl'}, \end{aligned} \quad (9)$$

and so

$$
\begin{aligned}
Q(\Theta, \Theta^{(t)}) &= E_{X|\Theta^{(t)},S}(\log \ell(\Theta|S, X)) \quad (10)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} \Pr(X_{ij}=1|\Theta^{(t)}, S_i) \log\left(\sum_{l=1}^{m} \Pr(C_i = l|O_i = 1, \Theta^{(t)}) \left[g_1(1-g_1)^{|l-(j+g_2)|}\right]^K\right)\\
&\quad + E_{X|\Theta^{(t)},S}(q)\log\gamma + (n - E_{X|\Theta^{(t)},S}(q))\log(1-\gamma)\\
&\quad + \sum_{i=1}^{n}(1 - \textstyle\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij}))\sum_{l=1}^{m}\sum_{b=A}^{U}\nu_{bli}\log f(b)\\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij})\sum_{l=1}^{m}\sum_{b=A}^{U}\nu_{bli}\log\Psi_{bl'}, \quad (11)
\end{aligned}
$$

where

$$
\Pr(X_{ij}=1|\Theta^{(t)}, S_i) = \frac{\Pr(S_i|X_{ij}=1,\Theta^{(t)})\Pr(X_{ij}=1|\Theta^{(t)})}{\Pr(S_i|O_i^{(t)}=0,\Theta^{(t)})\Pr(O_i^{(t)}=0|\Theta^{(t)}) + \sum_{j'=1}^{m-w+1}\Pr(S_i|X_{ij'}=1,\Theta^{(t)})\Pr(X_{ij'}=1|\Theta^{(t)})}.
$$

Here $\Pr(S_i|X_{ij}=1,\Theta^{(t)})$, $\Pr(S_i|O_i=0,\Theta^{(t)})$ and $\Pr(O_i^{(t)}=0|\Theta^{(t)})$ are computed exactly as in the sequence-only method. The only change is that

$$
\Pr(X_{ij}=1|\Theta^{(t)}) = \gamma\sum_{l=1}^{m}\Pr(C_{i=l})\left[g_1^{(t)}(1-g_1^{(t)})^{|l-(j+g_2^{(t)})|}\right]^K. \quad (12)
$$

2.5.2 Expectation Maximization – M-step

Maximization of the sequence component of the model, i.e. M, follows the same procedure described in section 2.4.2. With respect to diagnostic events, there are two free parameters of the model that need to be maximized: g_1 and g_2. There is no closed form maximum likelihood estimator for either of these parameters. To maximize g_2, we perform an exhaustive search of all possible values in the range $-8 \leq g_2 \leq 8$. We maximize g_1 by using the Newton-Raphson algorithm to find the root of the first derivative of Q with respect to g_1. This is a numerical approach that iteratively refines an initial estimate of g_1, $g_1^{(0)}$ as follows:

$$
g_1^{(t)} = g_1^{(t-1)} - \frac{Q'(\Phi^{(t-1)}, \Phi^{(t-2)})}{Q''(\Phi^{(t-1)}, \Phi^{(t-2)})}. \quad (13)
$$

Where $\Phi^{(t)} = \{\hat{M}, \hat{f}, \hat{\gamma}, g_1^{(t)}, \hat{g_2}, D\}$. This process continues until $Q(\Phi^{(t)}, \Phi^{(t-1)}) - Q(\Theta^{(t-1)}, \Theta^{(t-2)}) < \nu$, where ν is some fixed precision threshold. Here

$$
\begin{aligned}
Q'(\Phi, \Phi^{(t)}) &= \frac{\partial Q(\Phi, \Phi^{(t)})}{\partial g_1}\\
&= \sum_{i}^{n}\sum_{j}^{m-w+1}\Bigg[\Pr(X_{ij}=1|\Phi^{(t)}, S_i)\\
&\quad \frac{\sum_{l=1}^{m}\Pr(C_i=l|O_i=1,\Phi^{(t)})[Kg_1^{K-1}(1-g_1)^{K|l-(j+g_2)|} - g_1^K(K|l-(j+g_2)|)(1-g_1)^{K|l-(j+g_2)|-1}]}{\sum_{l=1}^{m}\Pr(C_i=l|O_i=1,\Phi^{(t)})g_1^K(1-g_1)^{K|l-(j+g_2)|}}\Bigg]\\
&= K\sum_{i}^{n}\sum_{j}^{m-w+1}\Bigg[\Pr(X_{ij}=1|\Phi^{(t)}, S_i)\\
&\quad \left(\frac{1}{g_1} - \frac{\sum_{l=1}^{m}\Pr(C_i=l|O_i=1,\Phi^{(t)})(|l-(j+g_2)|)(1-g_1)^{K|l-(j+g_2)|-1}}{\sum_{l=1}^{m}\Pr(C_i=l|O_i=1,\Phi^{(t)})(1-g_1)^{K|l-(j+g_2)|}}\right)\Bigg], \quad (14)
\end{aligned}
$$

and

$$
\begin{aligned}
Q''(\Phi, \Phi^{(t)}) &= \frac{\partial^2 Q(\Phi, \Phi^{(t)})}{\partial (g_1)^2} \\
&= K \sum_{i}^{n} \sum_{j}^{m-w+1} \Bigg[\Pr(X_{ij} = 1 | \Phi^{(t)}, S_i) \\
&\quad \left(- \frac{1}{(g_1)^2} - \frac{D \frac{\partial N}{\partial g_1} - N \frac{\partial D}{\partial g_1}}{(\sum_{l=1}^{m} \Pr(C_i = l | O_i = 1, \Phi^{(t)})(1 - g_1)^{K|l-(j+g_2)|})^2} \right) \Bigg],
\end{aligned} \tag{15}
$$

where

$$
N = \sum_{l=1}^{m} \Pr(C_i = l | O_i = 1, \Phi^{(t)})(|l - (j + g_2)|)(1 - g_1)^{K|l-(j+g_2)|-1} \tag{16}
$$

$$
D = \sum_{l=1}^{m} \Pr(C_i = l | O_i = 1, \Phi^{(t)})(1 - g_1)^{K|l-(j+g_2)|} \tag{17}
$$

$$
\frac{\partial N}{\partial g_1} = -K \sum_{l=1}^{m} \Pr(C_i = l | O_i = 1, \Phi^{(t)})(|l - (j + g_2)|)(|l - (j + g_2)| - 1)(1 - g_1)^{K|l-(j+g_2)|-2} \tag{18}
$$

$$
\frac{\partial D}{\partial g_1} = -K \sum_{l=1}^{m} \Pr(C_i = l | O_i = 1, \Phi^{(t)})(|l - (j + g_2)|)(1 - g_1)^{K|l-(j+g_2)|-1}. \tag{19}
$$

In practise, we found that Zagros was relatively insensitive to the choice of g_1, but optimizing this parameter increases the asymptotic runtime. As a trade-off, we computed the optimal value of g_1 as described above for all of the CLIP-seq datasets in our collection and found the mode of this distribution. By default, Zagros uses this value for g_1 and does not optimize the parameter, however we still provide an option for the user to maximize g_1 via Newton-Raphson if they wish.

2.6 Sequence and structure motif discovery

From a theoretical standpoint, we consider secondary structure to be an inherent property of the sequences, so accounting for structure does not introduce any new observed data. We will describe how our model works assuming structure is fully determined by the sequence; by this we mean that, given a sequence S_i, the base-pairing state (paired or unpaired) of each nucleotide within that sequence can be exactly determined. At the end of this subsection we will discuss the practicalities of determining the structure of the sequences. In order to account for secondary structure, we augment M and f as follows:

$$
\begin{aligned}
M_k(b, \tau) &= \Pr\big(b \text{ appears at position } k \text{ in the motif with pairing state } \tau\big), \\
f(b, \tau) &= \Pr\big(b \text{ appears in the background with pairing state } \tau\big),
\end{aligned}
$$

where $\tau \in \{0, 1\}$, either paired or unpaired. The form of the complete-data likelihood is unchanged from Equation 1; all that is required is changes in the equations for computing $\Pr(S_i | X_{ij} = 1, \Theta)$ and $\Pr(S_i | O_i = 0, \Theta)$, i.e.

$$
\begin{aligned}
\Pr(S_i | X_{ij} = 1, \Theta) &= \prod_{l=1}^{m} \prod_{b=A}^{U} \prod_{\tau=0}^{1} (\Psi_{blj\tau})^{\nu_{bli\tau}} \\
\Pr(S_i | O_i = 0, \Theta) &= \prod_{l=1}^{m} \prod_{b=A}^{U} \prod_{\tau=0}^{1} f(b, \tau)^{\nu_{bli\tau}},
\end{aligned} \tag{20}
$$

where

$$\Psi_{bl j\tau} = \begin{cases} M_{l-j+1}(b,\tau) & \text{if } j \le l \le j+w-1, \\ f(b,\tau) & \text{otherwise.} \end{cases}$$

and $\nu_{bli\tau} = 1$ if base l in sequence i is b and has pairing state τ, and $\nu_{bli\tau} = 0$ otherwise.

2.6.1 Expectation-maximization – E-step

The log-likelihood is largely unchanged from Equation 2; i.e.

$$\begin{aligned} \log \ell(\Theta|S,X) &= \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} X_{ij} \log\left(\frac{1}{m-w+1}\right) \\ &\quad + q\log\gamma + (n-q)\log(1-\gamma) \\ &\quad + \sum_{i=1}^{n}(1-\textstyle\sum_{j=1}^{m-w+1} X_{ij}) \displaystyle\sum_{l=1}^{m}\sum_{b=A}^{U}\sum_{\tau=0}^{1} \nu_{bli\tau}\log f(b,\tau) \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} X_{ij} \sum_{l=1}^{m}\sum_{b=A}^{U}\sum_{\tau=0}^{1} \nu_{bli\tau}\log\Psi_{bl j\tau}. \end{aligned} \tag{21}$$

As such, $Q(\Theta,\Theta^{(t)})$ is also similar, and is defined as

$$Q(\Theta,\Theta^{(t)}) = E_{X|\Theta^{(t)},S}(\log\ell(\Theta|S,X)) \tag{22}$$

$$\begin{aligned} &= \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij}) \log\left(\frac{1}{m-w+1}\right) \\ &\quad + E_{X|\Theta^{(t)},S}(q)\log\gamma^{(t)} + (n - E_{X|\Theta^{(t)},S}(q))\log(1-\gamma^{(t)}) + \\ &\quad + \sum_{i=1}^{n}(1-\textstyle\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij})) \displaystyle\sum_{l=1}^{m}\sum_{b=A}^{U}\sum_{\tau=0}^{1} \nu_{bli\tau}\log f^{(t)}(b,\tau) \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij}) \sum_{l=1}^{m}\sum_{b=A}^{U}\sum_{\tau=0}^{1} \nu_{bli\tau}\log\Psi^{(t)}_{bl j\tau}, \end{aligned} \tag{23}$$

where, just as previously, $E_{X|\Theta^{(t)},S}(q) = \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij})$, and $E_{X|\Theta^{(t)},S}(X_{ij}) = \Pr(X_{ij} = 1|S_i,\Theta^{(t)})$. Further, the same form applies for the calculation of the occurrence probabilities, so

$$\Pr(X_{ij}=1|S_i,\Theta^{(t)}) = \frac{\Pr(S_i|X_{ij}=1,\Theta^{(t)})\Pr(X_{ij}=1|\Theta^{(t)})}{\Pr(S_i|\Theta^{(t)},O_i=0)\Pr(O_i=0|\Theta^{(t)}) + \sum_{j'=1}^{m-w+1}\Pr(S_i|X_{ij'}=1,\Theta^{(t)})\Pr(X_{ij'}=1|\Theta^{(t)})}.$$

Here we amend the sequence components of the likelihood to account for structure by introducing τ, hence

$$\begin{aligned} \Pr(S_i|X_{ij}=1,\Theta^{(t)}) &= \prod_{l=1}^{m}\prod_{b=A}^{U}\prod_{\tau=0}^{1}(\Psi^{(t)}_{bl j\tau})^{\nu_{bli\tau}}, \\ \Pr(S_i|O_i=0,\Theta^{(t)}) &= \prod_{l=1}^{m}\prod_{b=A}^{U}\prod_{\tau=0}^{1} f^{(t)}(b)^{\nu_{bli\tau}}, \end{aligned} \tag{24}$$

where

$$\Psi^{(t)}_{blj\tau} = \begin{cases} M^{(t)}_{l-j+1}(b,\tau) & \text{if } j \le l \le j+w-1, \\ f^{(t)}(b,\tau) & \text{otherwise.} \end{cases} \tag{25}$$

The prior probabilities on occurrence of the motif, i.e. $\Pr(X_{ij} = 1|\Theta^{(t)})$ and $\Pr(O_i = 0|\Theta^{(t)})$ are unchanged from section 2.4.1.

2.6.2 Expectation-maximization – M-step

Maximizing the ZOOPS model parameter is unchanged from section 2.4.2. Only minor changes are needed to the maximization step for M and f:

$$\hat{M}_k(b,\tau) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m-w+1} \nu_{bji\tau} \Pr(X_{ij} = 1|\Theta^{(t)}), \tag{26}$$

and

$$\hat{f}(b,\tau) = \frac{1}{n(m-w)}\sum_{i=1}^{n}\sum_{l=1}^{m}\sum_{j=1}^{m-w+1} \nu_{bli\tau}(I_{\{l\notin[j,j+w-1]\}}) \Pr(X_{ij} = 1|\Theta^{(t)}). \tag{27}$$

2.6.3 Determining the secondary structure of an RNA sequence

The above description assumes an exact pairing state (paired or unpaired) can be assigned to each nucleotide in each sequence. One way to do this is to compute the minimum free energy structure for each sequence. However, the minimum free energy structure can be misleading. In reality, there is an ensemble of folds that a given transcript may adopt, and each one has a particular probability of occurring. Rather than using point estimates of the pairing state of nucleotides, we elected to instead use base-pair probabilities to represent structure. The changes to the above description to facilitate this are trivial: we re-designate the indicator variable $\nu_{bli\tau}$ as follows:

$$\nu_{bli\tau} = \Pr(\text{base } l \text{ in sequence } i \text{ has pairing state } \tau)\nu_{bli}, \tag{28}$$

where ν_{bli} retains the interpretation of Section 2.4. We calculate these base-pair probabilities using McCaskill's algorithm. We used the implementation from the RNA Vienna package [20], modifying its interface to allow its efficient use and embedding in our package. There are three main reasons (among many others) why we modified RNAFold, rather than use the original: (1) to allow input either in BED format (with sequences extracted from chromosome fasta files), or in fasta format – RNAFold cannot directly accommodate this; (2) to make the output conform to the expected input format that Zagros uses; and (3) to efficiently use McCaskill's algorithm for calculating base pair probabilities without the overhead of ancillary computations performed by RNAfold that are not required for Zagros. This is a pre-processing step. Zagros does not do any computational folding. The user is free to compute base-pair probabilities in any way they wish, though we provide the code we wrote to do the above, and the relevant code from RNAFold, as a convenience to the user.

2.7 Sequence, structure, and diagnostic events motif discovery

The modifications described in sections 2.5 and 2.6 which respectively include information about crosslinking using the prior on motif occurrence indicators and include information about secondary structure by expanding the sequence component of the model are completely orthogonal. As such, it is straight-forward to combine them. The combined model will be $\Theta = \{M, f, \gamma, g_1, g_2, D\}$, as in Section 2.5, but using the definitions for M and f from Section 2.6.

2.7.1 Expectation-maximization – E-step

The log-likelihood takes the prior on motif occurrences from Equation 9, and the sequence/structure terms from Equation 21; i.e.

$$\begin{aligned}\log \ell(\Theta|S,X) &= \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} X_{ij} \log\left(\sum_{l=1}^{m} \Pr(C_i = l|O_i = 1,\Theta^{(t)}) \left[g_1(1-g_1)^{|l-(j+g_2)|}\right]^K\right) \\ &+ q\log\gamma + (n-q)\log(1-\gamma) \\ &+ \sum_{i=1}^{n}(1-\textstyle\sum_{j=1}^{m-w+1} X_{ij}) \displaystyle\sum_{l=1}^{m}\sum_{b=A}^{U}\sum_{\tau=0}^{1} \nu_{b,l,i,\tau}\log f(b,\tau) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} X_{ij} \sum_{l=1}^{m}\sum_{b=A}^{U}\sum_{\tau=0}^{1} \nu_{b,l,i,\tau}\log\Psi_{b,l,j,\tau}. \end{aligned} \tag{29}$$

Hence $Q(\Theta,\Theta^{(t)})$ is defined as

$$\begin{aligned} Q(\Theta,\Theta^{(t)}) &= E_{X|\Theta^{(t)},S}(\log\ell(\Theta|S,X)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} \Pr(X_{ij}=1|\Theta^{(t)},S_i)\log\left(\sum_{l=1}^{m}\Pr(C_i=l|O_i=1,\Theta^{(t)})\left[g_1(1-g_1)^{|l-(j+g_2)|}\right]^K\right) \\ &+ E_{X|\Theta^{(t)},S}(q)\log\gamma^{(t)} + (n - E_{X|\Theta^{(t)},S}(q))\log(1-\gamma^{(t)}) \\ &+ \sum_{i=1}^{n}(1-\textstyle\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij})) \displaystyle\sum_{l=1}^{m}\sum_{b=A}^{U}\sum_{\tau=0}^{1}\nu_{b,l,i,\tau}\log f^{(t)}(b,\tau) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{m-w+1} E_{X|\Theta^{(t)},S}(X_{ij}) \sum_{l=1}^{m}\sum_{b=A}^{U}\sum_{\tau=0}^{1}\nu_{b,l,i,\tau}\log\Psi^{(t)}_{b,l,j,\tau}, \end{aligned} \tag{30}$$

2.7.2 Expectation-maximization – M-step

Maximization of the sequence/structure component of the model and the component associated with cross-linking is orthogonal, hence M and f are maximized as in section 2.6.2; g_1 and g_2 are optimized as in section 2.5.2.

2.8 Calculating hexamer structural preference in CLIP-seq

First we define target and non-target sets for each CLIP experiment on a particular RBP. For each CLIP-seq dataset we define a set of target 3'UTRs by binning 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 that RBP of interest that the CLIP experiment was carried out on. The non-target set is simply any 3' UTR regions (as defined by refseq) that are not contained in the target set. Secondly, we define paired and unpaired states for hexamer occurrences, meaning an occurrence of a hexamer is called single stranded if and only if the average base pairing probability over k nucleotides is less than 0.5, and double stranded otherwise. Now, for each hexamer, we obtained the set of occurrences in the whole region of all the 3'UTRs. This comprises our reference sets of occurrences for each hexamer. Then for each hexamer, we calculated the number of times it appears as double or single stranded in these reference sets. Let's call these numbers R^+ and R^- respectively. Then we calculate the same numbers, this time not for all the 3'UTRs but only for target set of a particular RBP in a CLIP experiment. For each hexamer and each CLIP experiment, let the counts of double and single stranded occurrences in target 3'UTRs be E^+ and E^-. For each hexamer and each RBP, the counts of

double and single stranded occurrences of that hexamer in non-target 3'UTRs are defined as $C^+ = R^+ - E^+$ and $C^- = R^- - E^-$.

Now for each hexamer we can produce a contingency table, with the number of times a hexamer in 3'UTRs tends to be in paired or unpaired states (C^+ and C^-), and the number of times this hexamer in relation to binding to protein is in paired or unpaired states (E^+ and E^-). The odds-ratio from this contingency table represents the tendency of the hexamer to appear as either single- or double-stranded in target 3' UTRs, as opposed to non-target 3' UTRs. Significance is determined via Fisher's exact test, which gives us a *p*-value on the odds-ratio for each hexamer. To establish what the p-value distribution is under the null hypothesis, we do the same analysis, but select the target set randomly, rather than using CLIP-seq data. We plotted the distribution of *p*-values in both cases (CLIP derived targets, and random targets), and noted a strong enrichment for highly significant p-values was present with CLIP-derived targets, but not random targets.

2.9 Diagnostic events in iCLIP data

In 1% of iCLIP reads, we detected a deletion, indicating that reverse-transcriptase read through the cross-link location. Hence the read is not truncated at the cross-link location and does not have a diagnostic event. In the remaining 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 does the read in fact contain the diagnostic event. Sugimoto *et al.* [27], using previously published Nova HITS-CLIP and mRNA-seq, estimated that 82% of the reads are truncated at the cross-link location. More formally, they estimated f, the proportion of read-through cDNAs in the total Nova iCLIP library, to be 18% according to the following formula

$$f = \frac{p(\text{iCLIP}) - p(\text{BG})}{p(\text{RT}) - p(\text{BG})}, \tag{31}$$

where $p(\text{iCLIP})$ is the proportion of cDNAs with deletions in the first 25 nucleotides for Nova iCLIP data, $p(\text{RT})$ is the proportion of cDNAs with deletions in the first 25 nucleotides for read-through cDNAs from Nova CLIP data and $p(\text{BG})$ is the proportion of cDNA with deletions in the first 25 nucleotides of mRNA-Seq cDNAs, which was used to estimate the background occurrence of deletions. Thus, they estimated that 82% of cDNAs were lost in CLIP cDNA cloning protocol due to truncations. However, there is no way to distinguish between the truncated reads and the ones that have read-through without any deletion, using a single iCLIP data set. For simplicity, we considered all of the reads without deletions to have a diagnostic event at the truncation site.

2.10 Generation of simulated data

2.10.1 Position weight matrix / motif

For each simulated dataset, we generated a random position weight matrix from which motif occurrences were drawn. For all simulations, the PWMs had length six, and an average information content of 0.5 bits per column (within a tolerance of 0.005 bits). The target average information content was achieved using a rejection sampling approach.

2.10.2 Sequence data

We first took the set of all human 3' UTRs from refseq and collapsed these into non-overlapping genomic regions. We retained all regions of length greater than 50bp. For each simulation, we selected 500 of these regions randomly, and for each region randomly selected a 50bp sub-sequence. We also randomly generate a PWM. For each of the 500 sequences, we generated a motif occurrence from the PWM and randomly place this into the sequence.

When imposing structure on a motif occurrence, we select a subsequence either upstream or downstream of the motif occurrence (with equal probability if the occurrence is sufficiently far from the edge of the sequence, or specifically one or the other if it is close) such that (1) the selected subsequence has length 10, and (2) the subsequence does not overlap the motif, but is no more than 15 bases from the motif. We then place the reverse

complement of this subsequence on the opposite side of the motif such that it also does not overlap the motif, but is no more than 15 bases from the selected subsequence. In this way, there is high probability that the local sequence will give rise to a computational fold with a stem loop encompassing the motif occurrence.

2.10.3 Diagnostic events

We consider diagnostic events to have a distance from motif occurrences that follows a geometric distribution. We also consider that this distance distribution may be centered on a position that is offset from the start of the occurrence. We call this offset delta. This is also the generative model we use to simulate diagnostic events. For each simulated dataset, we randomly generate a value for delta (random uniform distribution on -8 to $+8$ relative to the motif occurrence start), and a value for the probability parameter of the geometric distribution (also random uniform in range 0 to 1). For each diagnostic event we place into the dataset, we simulate its location relative to the selected motif location by drawing a distance from the geometric distribution specified by p, and adding delta to this. In addition, we consider that some fraction of diagnostic events will be noise. For all simulations performed in preparing this manuscript, we fixed this fraction at 20%. For noise events, we place the diagnostic event at a position selected uniformly at random from all loci in the sequence. The number of diagnostic events that were placed into each sequence was drawn from the empirical distribution of diagnostic event counts in real CLIP-Seq dataset as follows: we randomly selected a CLIP-Seq dataset and counted the number of diagnostic events that were present in this dataset within the 50bp region that the sequence was drawn from (see section 2.10.2).

2.11 Evaluating motifs recovered from simulated data

To determine how close to the planted motif a recovered motif was, we calculated the KL-divergence of the recovered motif from the planted motif. Rather than reporting raw KL-divergence values though, which are difficult to interpret, we instead report fractions of the number of simulated datasets on which the motif was successfully recovered. To determine whether a motif was recovered or not, we established a threshold KL-D value below which we would consider the motif to be recovered, and above which we would consider that the motif was not recovered. To find this threshold, we calculated the KL-divergence between the simulated position weight matrix and the set of all position weight matrices recovered by Zagros from simulated datasets – this allowed us to estimate the background distribution of KL-divergence values. Using this background distribution, we calculated a p-value for the motif recovered for each simulated datasets. If the p-value was below 0.05 we considered that the motif had been recovered, otherwise we considered it not-recovered.

2.12 Evaluating motifs recovered from CLIP-seq data

To evaluate motifs that were recovered from CLIP-Seq data we compiled a set of previously reported consensus sequences for each RBP (see section 1.2). For each recovered position weight matrix, we converted this into a consensus sequence by taking the most likely nucleotide at each position. We then determined the best alignment of this consensus sequence to the previously reported consensus, where best is defined as the alignment with the maximum matching bases. We allowed gaps only at the start/end of the sequences, not within. All recovered motifs were hexamers.

2.13 Calculation of sequence-motif specificity

In order to calculate the sequence motif specificity, for each of the motifs found by Zagros, we obtained a consensus sequence. Then we looked for occurrences of these consensus sequences in the exons defined by refseq. The trend of sequence specificity is obtained from the fact that more occurrences correspond to low sequence specificity and vice versa.

3 Supplementary Results

3.1 Results of CLIP-seq data

Figures S1-S5 show the result of running DME, MD-SCAN and MEME along with Zagros in all four different modes on all the CLIP data sets we have. We present results for all replicates, though we eliminate those for which no method was able to recover the expected motif. As shown in these figures, the superiority of Zagros using sequence, structure and diagnostic events is obvious.

It is worth noting that there are differences in the results recovered by MEME and Zagros using sequence-only information. This is to be expected. Despite the fact that they use the same model and both employ expectation maximization, in practice there will be differences in implementation details, not to mention the accumulation of heuristics and optimizations in MEME as a result of its long history.

Furthermore, there is noticeable variation in the results obtained by all of the programs on datasets for the same RBP arising from different labs. This is again expected. CLIP-seq is a very challenging assay to perform, and the skill of the technician is instrumental in achieving a good outcome. It is not surprising then that variability in data quality also results from this. Some datasets are so clean as to allow identification of the motif almost by visual inspection, while others are extremely challenging and even with advanced statistical methods it might not be possible to extract the expected motif above all others. A range of intermediate possibilities also exist. Add to this minor differences in the laboratory procedures used, and even different cell types in some cases, and it is not at all surprising that results are variable.

Figure S1: Top-scoring motifs recovered by Zagros on IGF2BP{1..3} CLIP-seq datasets [9]. For each dataset we show the motif recovered by DME, MD-SCAN and MEME in addition to each version of Zagros.

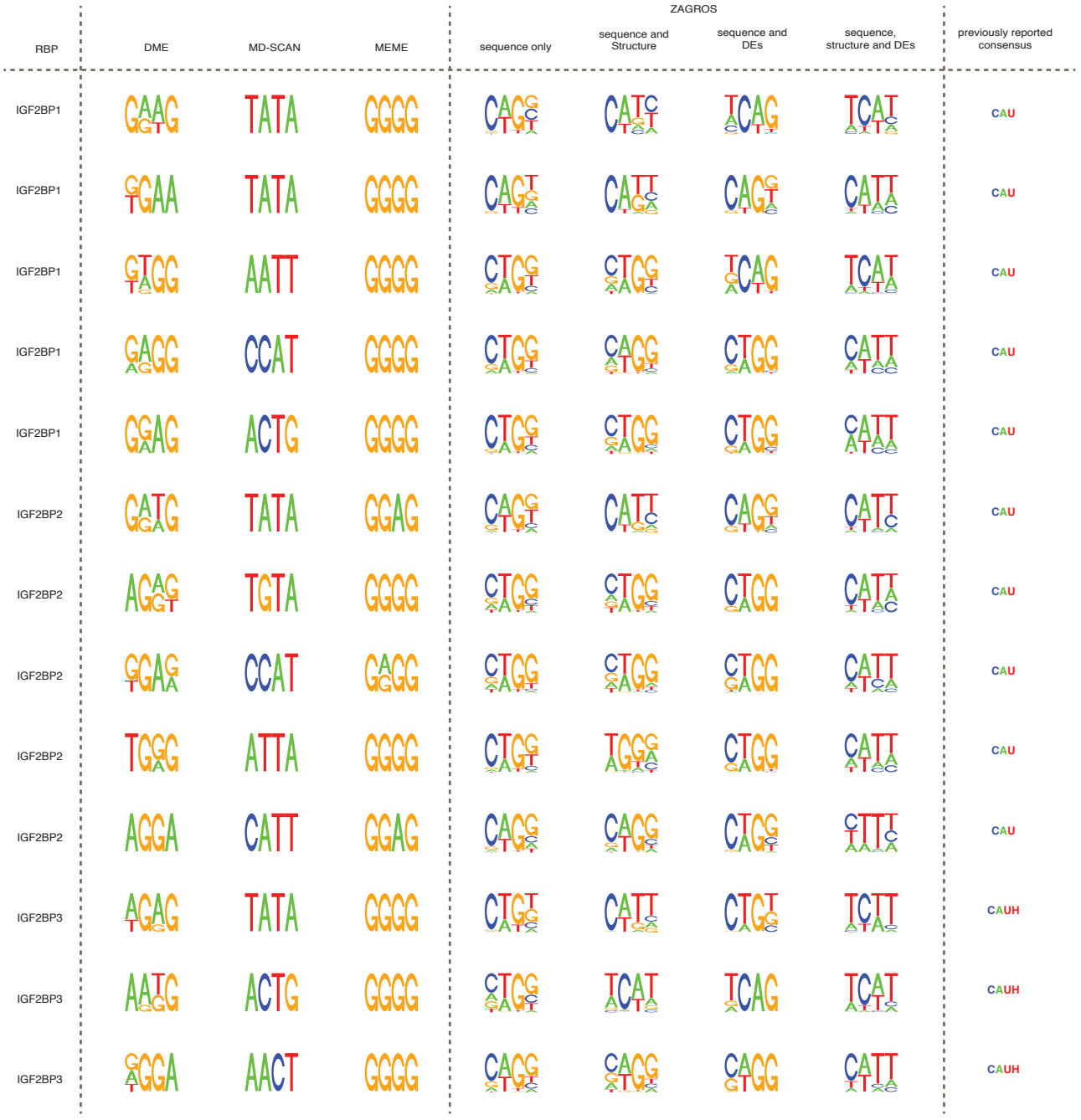


Figure S2: Top-scoring motifs recovered by Zagros on PUM2, QKI [9], hnRNPc [13] and HuR CLIP-seq datasets (The first HuR is obtained from [21] and the rest is from [16]). For each dataset we show the motif recovered by DME, MD-SCAN and MEME in addition to each version of Zagros.

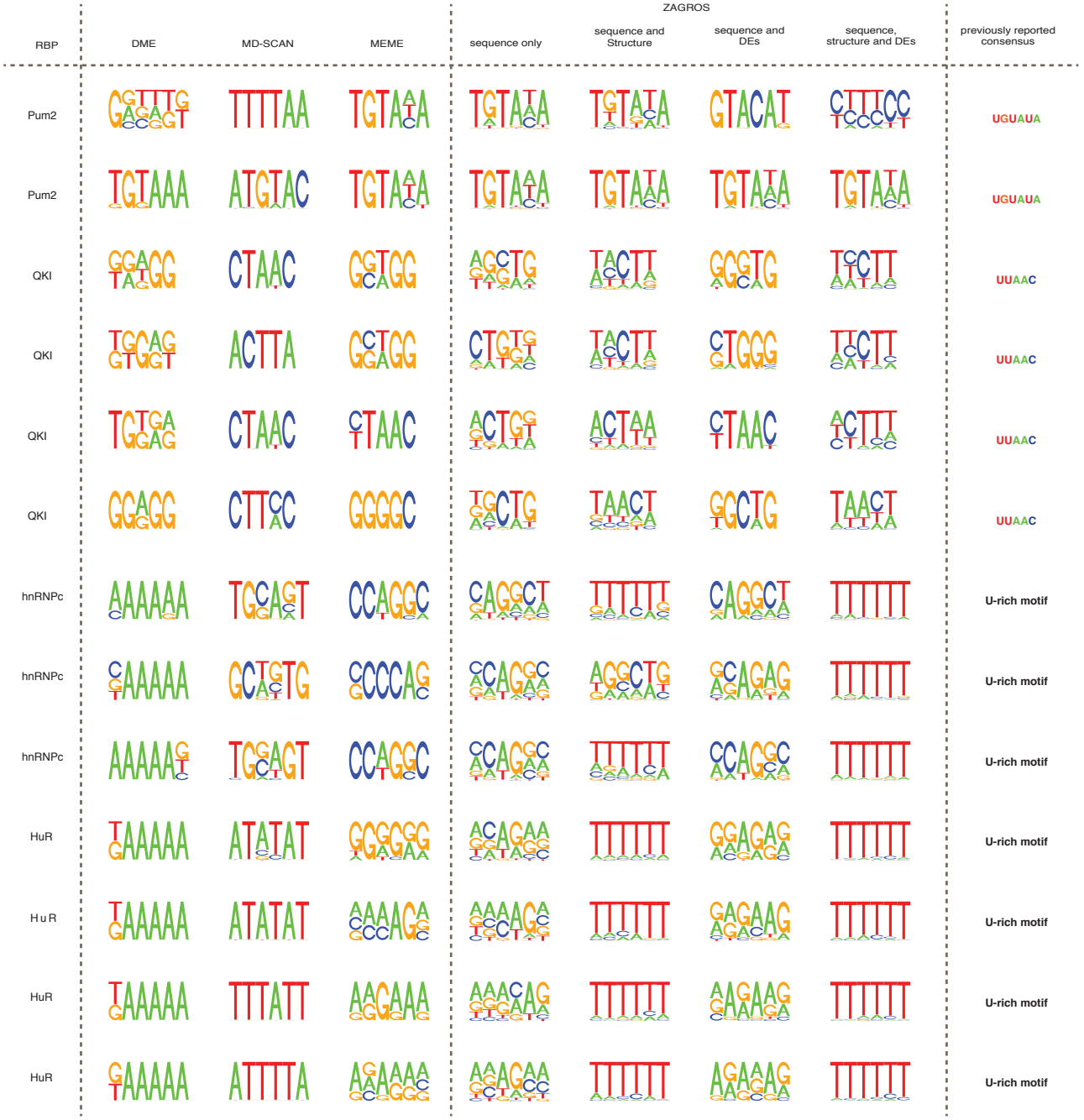


Figure S3: Top-scoring motifs recovered by Zagros on TDP-43 CLIP-seq datasets [28]. For each dataset we show the motif recovered by DME, MD-SCAN and MEME in addition to each version of Zagros.

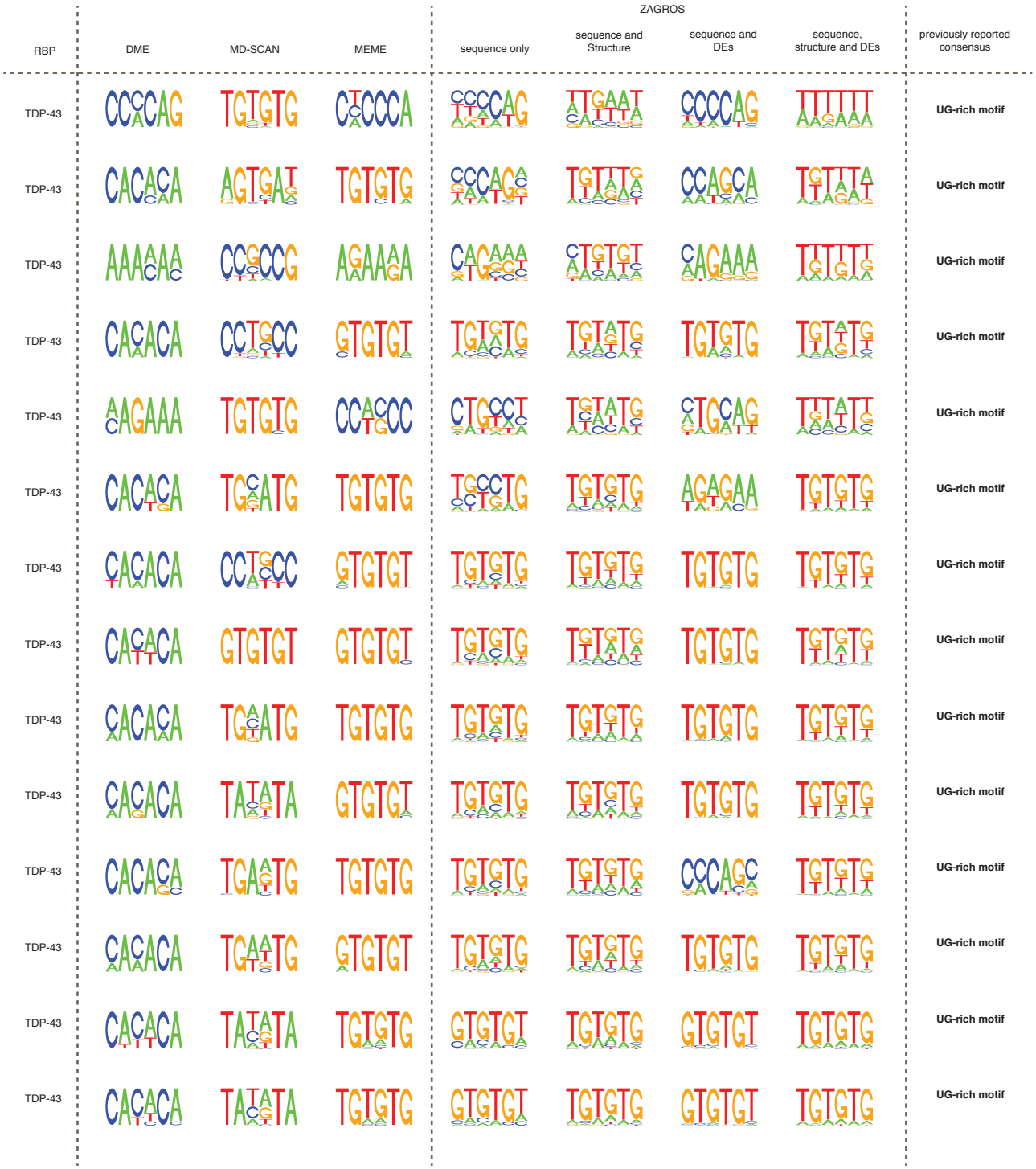
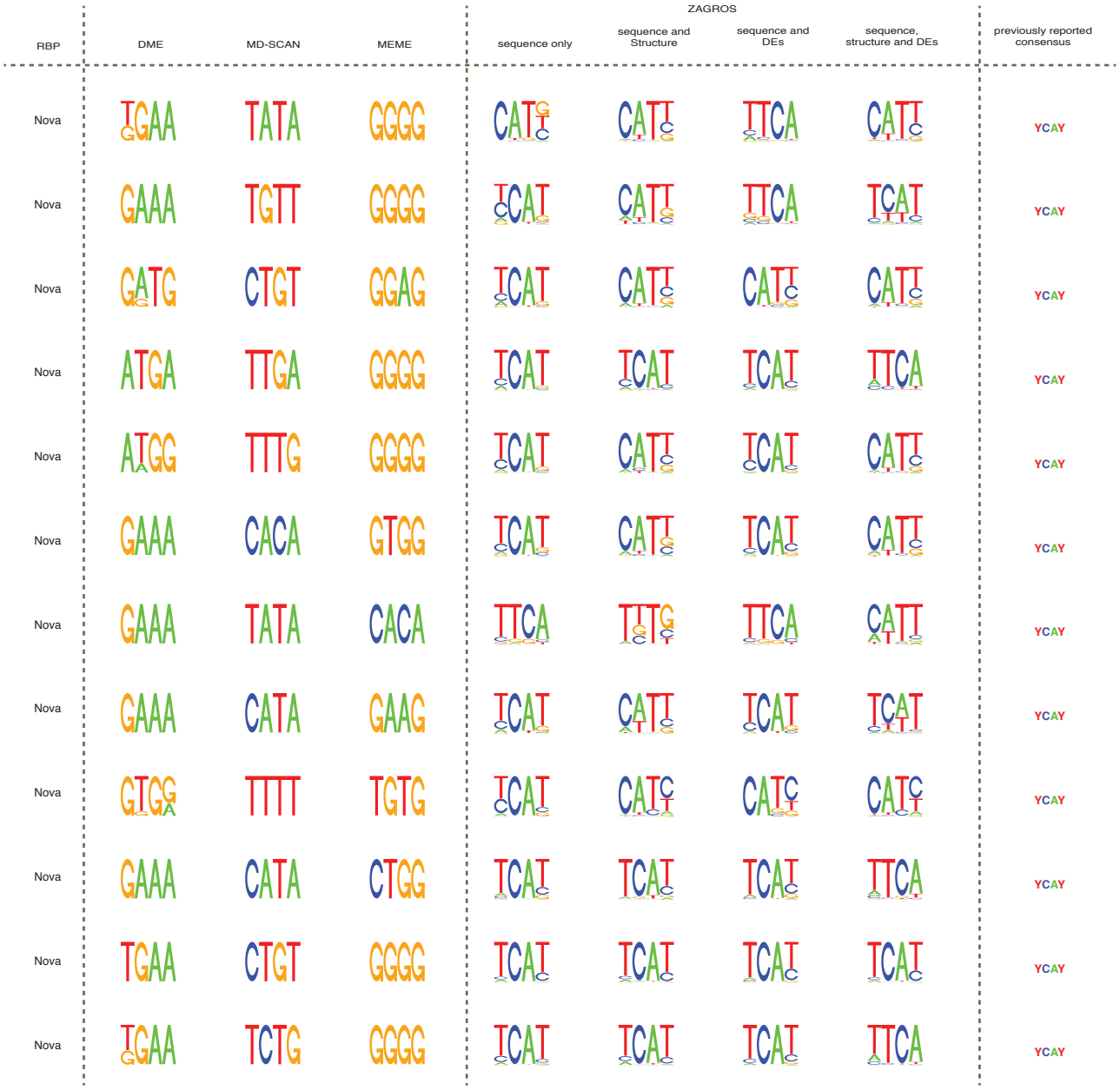


Figure S4: Top-scoring motifs recovered by Zagros on TIA1, TIAL1 [32] and PTB [33] CLIP-seq datasets. For each dataset we show the motif recovered by DME, MD-SCAN and MEME in addition to each version of Zagros.

Figure S5: Top-scoring motifs recovered by Zagros on Nova [35] CLIP-seq datasets. For each dataset we show the motif recovered by DME, MD-SCAN and MEME in addition to each version of Zagros.



References

[1] TL Bailey and C Elkan. The value of prior knowledge in discovering motifs with meme. In *Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology*, volume 3, page 21, 1995.

[2] R J Buckanovich and R B Darnell. The neuronal rna binding protein nova-1 recognizes specific rna targets in vitro and in vivo. *Molecular and Cellular Biology*, 17(6):3194–201, 1997.

[3] Massimo Caputi and Alan M. Zahler. Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. *Journal of Biological Chemistry*, 276(47):43850–43859, 2001.

[4] Sung Wook Chi, Julie B. Zang, Aldo Mele, and Robert B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, 07 2009.

[5] B. Kate Dredge and Robert B. Darnell. Nova regulates gabaa receptor $\gamma 2$ alternative splicing via a distal downstream ucau-rich intronic splicing enhancer. *Molecular and Cellular Biology*, 23(13):4687–4700, 2003.

[6] Patrik Förch, Oscar Puig, Nancy Kedersha, Concepción Martínez, Sander Granneman, Bertrand Séraphin, Paul Anderson, and Juan Valcárcel. The apoptosis-promoting factor tia-1 is a regulator of alternative pre-mrna splicing. *Molecular Cell*, Volume 6(Issue 5):1089–1098, 2000.

[7] Andre Galarneau and Stephane Richard. Target RNA motif and target mrnas of the quaking star protein. *Nat Struct Mol Biol*, 12(8):691–698, 08 2005.

[8] Alessia Galgano, Michael Forrer, Lukasz Jaskiewicz, Alexander Kanitz, Mihaela Zavolan, and André P. Gerber. Comparative analysis of mrna targets for human puf-family proteins suggests extensive interaction with the mirna regulatory system. *PLoS ONE*, 3(9):e3164, 09 2008.

[9] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano Jr., Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129 – 141, 2010.

[10] Stephanie C Huelga, Anthony Q Vu, Justin D Arnold, Tiffany Y Liang, Patrick P Liu, Bernice Y Yan, John Paul Donohue, Lily Shiue, Shawn Hoon, Sydney Brenner, et al. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell reports*, 1(2):167–178, 2012.

[11] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Meth*, 7(12):1009–1015, 12 2010.

[12] Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid, and Mihaela Zavolan. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Meth*, 8(7):559–564, 07 2011.

[13] J Konig, K Zarnack, G Rot, T Curk, M Kayikci, B Zupan, DJ Turner, NM Luscombe, and J Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17:909–915, 2010.

[14] Pan-Hsien Kuo, Chien-Hao Chiang, Yi-Ting Wang, Lyudmila G. Doudeva, and Hanna S. Yuan. The crystal structure of tdp-43 rrm1-dna complex reveals the specific recognition for ug- and tg-rich nucleic acids. *Nucleic Acids Research*, 2014.

[15] Charles E. Lawrence and Andrew A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins-structure Function and Bioinformatics*, 7:41–51, 1990.

[16] Svetlana Lebedeva, Marvin Jens, Kathrin Theil, Björn Schwanhäusser, Matthias Selbach, Markus Landthaler, and Nikolaus Rajewsky. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Molecular Cell*, 43(3):340 – 352, 2011.

[17] Anthony K L Leung, Amanda G Young, Arjun Bhutkar, Grace X Zheng, Andrew D Bosson, Cydney B Nielsen, and Phillip A Sharp. Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat Struct Mol Biol*, 18(2):237–244, 02 2011.

[18] Fan Li, Qi Zheng, Lee E Vandivier, Matthew R Willmann, Ying Chen, and Brian D Gregory. Regulatory impact of RNA secondary structure across the arabidopsis transcriptome. *The Plant Cell Online*, 24(11):4346–4359, 2012.

[19] J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. *Journal of the American Statistical Association*, 90(432):1156–1170, 1995.

[20] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[21] Neelanjan Mukherjee, David L. Corcoran, Jeffrey D. Nusbaum, David W. Reid, Stoyan Georgiev, Markus Hafner, Manuel Ascano Jr., Thomas Tuschl, Uwe Ohler, and Jack D. Keene. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Molecular Cell*, 43(3):327 – 339, 2011.

[22] Florian C. Oberstrass, Sigrid D. Auweter, Michèle Erat, Yann Hargous, Anke Henning, Philipp Wenter, Luc Reymond, Batoul Amir-Ahmady, Stefan Pitsch, Douglas L. Black, and Frédéric H.-T. Allain. Structure of ptb bound to rna: Specific binding and implications for splicing regulation. *Science*, 309(5743):2054–2057, 2005.

[23] Magdalini Polymenidou, Clotilde Lagier-Tourenne, Kasey R Hutt, Stephanie C Huelga, Jacqueline Moran, Tiffany Y Liang, Shuo-Chien Ling, Eveline Sun, Edward Wancewicz, Curt Mazur, Holly Kordasiewicz, Yalda Sedaghat, John Paul Donohue, Lily Shiue, C Frank Bennett, Gene W Yeo, and Don W Cleveland. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat Neurosci*, 14(4):459–468, 04 2011.

[24] Debashish Ray, Hilal Kazan, Esther T Chan, Lourdes Pena Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J Blencowe, Quaid Morris, and Timothy R Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotech*, 27(7):667–670, 07 2009.

[25] Marion Scheibe, Falk Butter, Markus Hafner, Thomas Tuschl, and Matthias Mann. Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions. *Nucleic Acids Research*, 40(19):9897–9902, 2012.

[26] Cem Sievers, Tommy Schlumpf, Ritwick Sawarkar, Federico Comoglio, and Renato Paro. Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Research*, 40(20):e160–e160, 2012.

[27] Yoichiro Sugimoto, Julian Konig, Shobbir Hussain, Blaz Zupan, Tomaz Curk, Michaela Frye, and Jernej Ule. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biology*, 13(8):R67, 2012.

[28] James R Tollervey, Tomaz Curk, Boris Rogelj, Michael Briese, Matteo Cereda, Melis Kayikci, Julian Konig, Tibor Hortobagyi, Agnes L Nishimura, Vera Zupunski, Rickie Patani, Siddharthan Chandran, Gregor Rot, Blaz Zupan, Christopher E Shaw, and Jernej Ule. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat Neurosci*, 14(4):452–458, 04 2011.

[29] Philip J. Uren, 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. Site identification in high-throughput RNA–protein interaction data. *Bioinformatics*, 28(23):3013–3020, 2012.

[30] Philip J. Uren, Suzanne C. Burns, Jianhua Ruan, Kusum K. Singh, Andrew D. Smith, and Luiz O. F. Penalva. 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. *Journal of Biological Chemistry*, 2011.

[31] Eric T Wang, Neal AL Cody, Sonali Jog, Michela Biancolella, Thomas T Wang, Daniel J Treacy, Shujun Luo, Gary P Schroth, David E Housman, Sita Reddy, et al. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell*, 150(4):710–724, 2012.

[32] Zhen Wang, Melis Kayikci, Michael Briese, Kathi Zarnack, Nicholas M. Luscombe, Gregor Rot, Blaž Zupan, Tomaž Curk, and Jernej Ule. iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol*, 8(10):e1000530, 10 2010.

[33] Yuanchao Xue, Yu Zhou, Tongbin Wu, Tuo Zhu, Xiong Ji, Young-Soo Kwon, Chao Zhang, Gene Yeo, Douglas L. Black, Hui Sun, Xiang-Dong Fu, and Yi Zhang. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Molecular Cell*, 36(6):996 – 1006, 2009.

[34] Kathi Zarnack, Julian König, Mojca Tajnik, Iñigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, Nicholas M. Luscombe, and Jernej Ule. Direct competition between hnrnp c and u2af65 protects the transcriptome from the exonization of alu elements. *Cell*, 152(3):453–466, 2013.

[35] Chaolin Zhang, Maria A. Frias, Aldo Mele, Matteo Ruggiu, Taesun Eom, Christina B. Marney, Huidong Wang, Donny D. Licatalosi, John J. Fak, and Robert B. Darnell. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science*, 329(5990):439–443, 2010.