A novel method for the identification of conserved structural patterns in RNA: From small scale to high-throughput applications

Marco Pietrosanto[1], Eugenio Mattei[1], Manuela Helmer-Citterich[1,*] and Fabrizio Ferrè[2]

[1]Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy and [2]Department of Pharmacy and Biotechnology (FaBiT), University of Bologna Alma Mater, Via Belmeloro 8/2, 40126 Bologna, Italy

ABSTRACT

Functional RNA regions are often related to recurrent secondary structure patterns (or motifs), which can exert their role in several different ways, particularly in dictating the interaction with RNA-binding proteins, and acting in the regulation of a large number of cellular processes. Among the available motif-finding tools, the majority focuses on sequence patterns, sometimes including secondary structure as additional constraints to improve their performance. Nonetheless, secondary structures motifs may be concurrent to their sequence counterparts or even encode a stronger functional signal. Current methods for searching structural motifs generally require long pipelines and/or high computational efforts or previously aligned sequences. Here, we present BEAM (BEAr Motif finder), a novel method for structural motif discovery from a set of unaligned RNAs, taking advantage of a recently developed encoding for RNA secondary structure named BEAR (Brand nEw Alphabet for RNAs) and of evolutionary substitution rates of secondary structure elements. Tested in a varied set of scenarios, from small- to large-scale, BEAM is successful in retrieving structural motifs even in highly noisy data sets, such as those that can arise in CLIP-Seq or other high-throughput experiments.

INTRODUCTION

The notion of motifs (or patterns) in biological molecules, defined as local recurring elements in functionally related entities, either due to evolutionary relationships or through convergence, has been exploited successfully in the past by computational methods aimed at functional characterization. Motifs can be detected (with relative ease) at the primary sequence level, but they almost always have a structural meaning, being clusters of spatially close residues working in concert to achieve a given function. The bioinformatics field of motif finding in proteins and DNA is well developed, providing several tools, approaches and databases (1–3), while fewer resources are available for structural motif finding in RNAs. Such tools can be particularly useful in helping the functional characterization of non-coding RNAs (ncRNAs), for which information about the involved specific sequences and structures is still scarce. ncRNAs are involved in a wide range of biological functions through diverse molecular mechanisms often involving the interaction with one or more RNA binding protein (RBP) partners, with other RNAs or with the genomic DNA (4,5). Experimental and computational techniques are becoming available to depict, in high-throughput settings and at high resolution, protein–RNA interactions, chromatin–RNA interactions and RNA secondary structures, allowing the identification of binding partners, binding sites and function determinants. Protein–RNA interactions are central to many cellular processes (6–9). The complexity of the protein–RNA interaction network is starting to be fully appreciated thanks to several technological advances (10). Generally, sequence-level binding preferences are often found, allowing the definition of sequence motifs and the usage of sequence-only based tools such as MEME (1) or cERMIT (11). Still, these sequence determinants frequently must be carried by a specific structural context (12–14), while in other cases it is the RNA secondary structure that dictates the interaction specificity: for example, some proteins tend to recognize complex secondary structure elements such as stem-loops and bulges (15). The RBP–RNA binding is therefore heterogeneous in nature and different RBP domains are governed by different rules. The influence of the RNA structural context upon protein binding, and

the impact on motif-finding methods has been recently reviewed (16).

Given the importance of the structural context of functional motifs in RNA molecules, a number of methods for approaching the RNA motif-finding problem that include the secondary structure are available (for two recent reviews see (17,18)). FOLDALIGN and its variants (19,20), comRNA (21), RNAProfile (22,23), RSmatch (24), RNAmine (25), MEMERIS (26), CMfinder (27), Seed (28), GeRNAMo (29), RNApromo (30), SCARNA_LM (31), GraphProt (32) are all tools that take advantage of secondary structure information for tackling the motif-finding problem, employing different approaches and to different extents. Some other methods were developed specifically for the identification of protein-binding motifs, e.g. RNAcontext (33), the algorithm by Li *et al.* (34), mCarts (35), RBPmotif (36) and Zagros (37). The underlying algorithms can vary: expectation maximization (MEMERIS), covariance models (CMfinder), stochastic context-free grammars (RNApromo), graph matching (comRNA, RNAmine), graph kernels (GraphProt), fold-and-align methods (FOLDALIGN), conditional random fields (SCARNA_LM), hidden Markov models (mCarts), genetic programming (GeRNAMo) and others. The nature of the secondary structure information needed by these methods can also vary: some need pre-computed structures, or perform a minimum free energy prediction on-the-fly, others employ base-pairing probabilities, while others try to build the secondary structure simultaneously with the motif finding procedure. Some methods seek for purely structural motifs, while other can consider sequence information as well. Finally, many algorithms are limited in searching motifs having a specific nature, e.g. only in single-stranded regions (MEMERIS), or in regions containing a limited and/or fixed number of hairpins (CMfinder, FOLDALIGN, RNAProfile), or starting from and expanding well-conserved stem structures (RNApromo, RNAmine).

When the algorithm requires the RNA secondary structure, it is often converted into formats that are more informative than the standard dot-bracket notation. Nevertheless, going beyond the dot-bracket notation generally increases algorithm complexities and computational times. Graph representations provide very accurate results, but are usually computationally expensive as well as limited to topological assertions that hardly detect structural similarities that find their reasons in biological relations, and models of RNA structure evolution are not implemented when comparing RNA secondary structures. To solve this issue, we recently proposed Brand nEw Alphabet for RNAs (BEAR), a representation of the RNA secondary structure by an alphabet of characters describing secondary structure elements and their size, and computed substitution matrix-like rates of variation of these structural elements in functionally related RNAs (38). Having an informative string-based representation of the secondary structure and a substitution matrix, it becomes possible to apply standard algorithms for sequence alignment to the problem of RNA structural comparison (38,39).

Here we present BEAM (BEAr Motif finder), a method that explores sets of unaligned RNAs sharing a biological property (e.g. the ability to bind a specific RNA-binding protein) looking for the most represented local secondary structure motifs, and evaluating their significance with respect to a common background. BEAM employs the BEAR secondary structure notation and its associated similarity matrix of secondary structure elements, in order to capture motifs by structural similarities that derive from evolutionary related ncRNAs in a way that covers topological comparison, yet expands it by considering the evolutionary history behind the abstraction of structure representation. BEAM is able to identify structurally similar sites shared by hundreds or thousands of RNAs, and the extension of the motifs is not subject to limitations (other than those imposed by the user). Hence, it is a tool suitable for low-, medium- and high-throughput settings such as those in CLIP-seq analysis (40).

We tested BEAM on a number of artificial and real cases, verified its robustness to noisy data sets and the impact of imprecise secondary structure predictions on the results. Clearly, the requirement of a known or predicted secondary structure might limit BEAM applicability, but we believe that this would not be a major hindrance thanks to recent technology advances that are quickly leading toward an era when high-quality RNA secondary structure information will be available for entire transcriptomes (41). BEAM source code is freely available at https://github.com/noise42/beam.

MATERIALS AND METHODS

Data set preparation

For the studies carried out in this work, we used Rfam seed 11.0 (2) and the data available in the DoRiNA database (42). For Rfam RNAs we followed the same procedure used in (38) to fold RNAs, by combining RNAfold (43) ability to use constraints in secondary structure predictions and specific information retrieved from the secondary structure consensus of the Rfam seed. For each of the 2208 Rfam families (from now on RF) we required the RF alignments to be annotated with a *consensus* structure and with the *per-row* nucleotide conservation, and filtered out those RFs that bore no constraints, remaining with 1694 families. We selected as constrained only those nucleotides that were annotated as highly conserved in the seed alignment. For every RNA we then removed the gaps created by the alignment, and folded the sequence. In this way every single RNA sequence was folded independently and with different sets of constraints in order to capture variations between RNAs belonging to the same family.

Each RNA secondary structure was then converted into the 85-character alphabet defined in the BEAR format (38). The employed BEAR Encoder (Available at http://bioinformatica.uniroma2.it/BEAR) produces output files in the FASTB format, which is similar to a FASTA with an additional third row for the dot bracket notation and a fourth row for the BEAR notation (39).

Secondary structure motif model

Let $S^i = \{s^i_1, s^i_2, s^i_3, \ldots, s^i_W\}$ be a substructure of width W of the BEAR-encoded structure of an RNA i in a collection
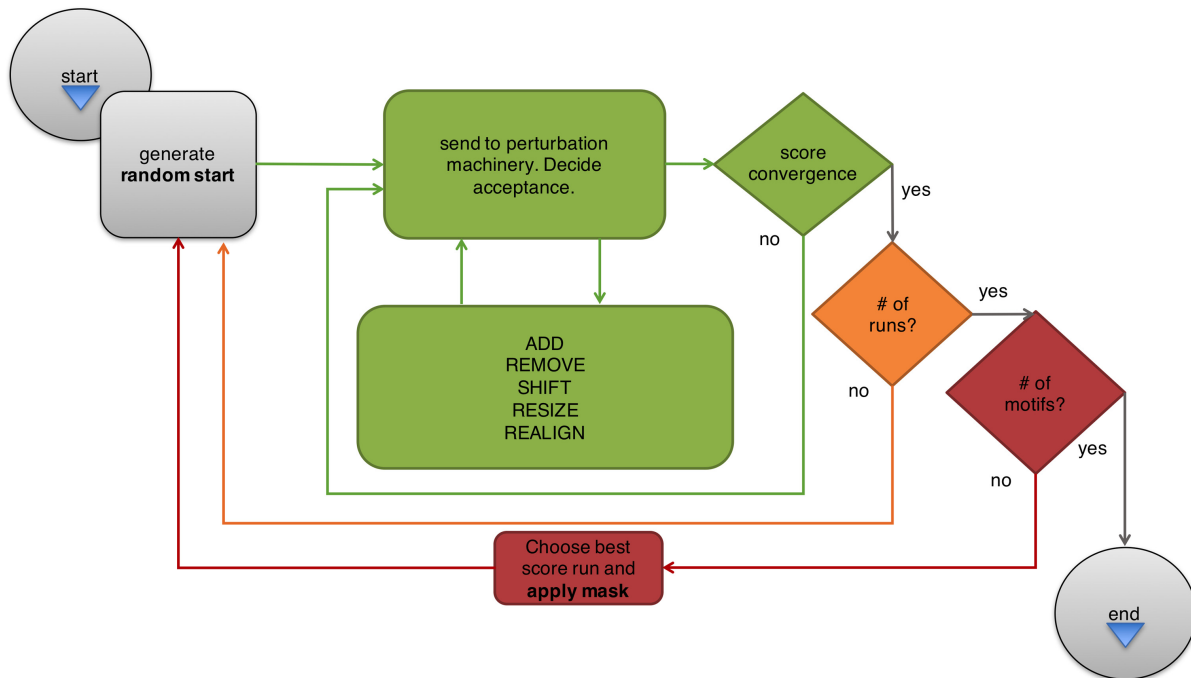
Figure 1. The flowchart illustrates the main steps of the BEAM algorithm. Starting from a random subset of sequences, the composed motif is altered through small perturbations of the type {add, remove, shift, resize, realign}. Every perturbation modifies the set of currently selected RNAs and their substructures, and the step is accepted as per metropolis rule. This corresponds to dynamically look for the best motif in terms of BEAM score. A second layer of computation (in orange) starts different runs from different starting points in order to better sample the search space. The third and final layer (in red) uses the best motif from all the previous steps and hides every instance of it in the input from a subsequent series of runs.

of N RNAs, which are supposed to be functionally related. A motif, or system state, is then composed of M BEAR-encoded structures $A = \{S^1,\ldots,S^M\}$, out of the N RNAs in the collection, each of width W. This motif definition corresponds to a ZOOPS model (Zero or One Occurrence Per Sequence). Also note that the motifs thusly defined are gapless. A $85xW$ Position Frequency Matrix (from now on PFM) is derived from A, with each entry p_{kj} being the relative frequency of BEAR character $k \in \{BEAR\}$ in position $j \in [1,W]$ of A. We then assign to the alignment A the Beam Score $BS(A)$: $BS(A) = \sum_{i=1}^{M} \sum_{j=1}^{W} \sum_{k=1}^{85} MBR(S_j^i, BEAR_k) p_{kj}$.

Here MBR (substitution Matrix of BEAR-encoded RNA secondary structures) is an 85×85 symmetric substitution matrix reporting substitution rates between each pair of BEAR characters, computed from Rfam seed alignments (38).

The BEAM algorithm

BEAM follows a heuristically tweaked simulated annealing algorithm to look for the system state A having the best score $BS(A)$. Given N BEAR-encoded RNA secondary structures, a random initial state is drawn. From there, and for a fixed maximum number of steps (or until score convergence), a random perturbation is drawn from a uniform distribution (a flowchart is shown in Figure 1). Let us now call A^t a system state at time t. The possible transitions $A^t \rightarrow A^{t+1}$ are drawn with equal probability from the following pool of perturbations:

- *Add*: add a window to an RNA, randomly sampled from the N in the data set, that did not already have one. The window is placed in the interval which, aligned with the current PFM and scored with MBR, gives the highest score.
- *Remove*: randomly remove a window from an RNA.
- *Shift*: shift all the current windows left or right by a Poisson-extracted discrete magnitude ($\lambda = 1$).
- *Resize*: shrink or enlarge all the current windows by a Poisson-extracted discrete magnitude ($\lambda = 1$).
- *Re-align*: Randomly selects an existing window and individually sets all the others to the RNA interval on the sequence that scores better (in MBR terms) with the chosen window. This step is equivalent to making bigger jumps in the search space (as opposed to the classical formulation of simulated annealing which requires detailed balance – and this is why we call it heuristic). The reason for this step is the need to speed up the process of exploring local *maxima*. Since the trade-off between exploration depth and run-time is one of the problems we have to address, this choice seemed to fit in the design. The possible flaws generated by not satisfying detailed balance are patched with the starting from different random initial system states and retaining of the best end-point of each parallel run.

Each perturbation generates a new system state, having a difference ΔBS in score with respect to the previous one. The new state is accepted or rejected using the following relation: if $\Delta BS > 0$ accept otherwise accept the step with probability $p_{accept} = exp(-|\Delta BS| / kT)$. The final state is then
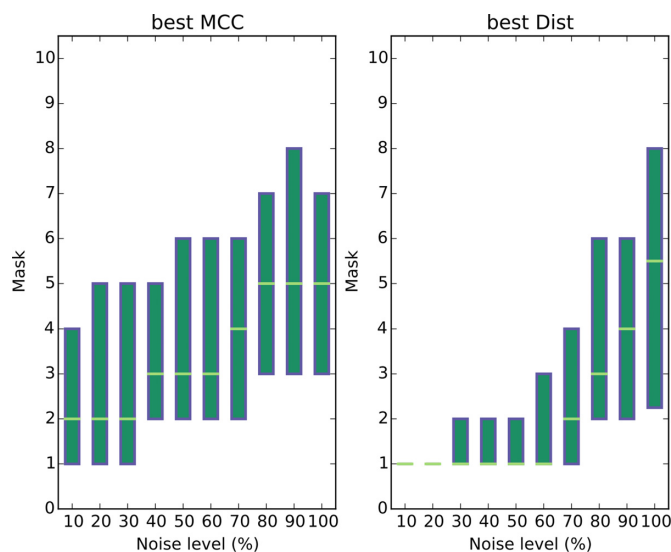
Figure 2. Benchmark for the robustness test on Rfam, with whisker-less boxplots. Left boxplot indicates the rank of the motif with the best MCC against the amount of noise. The method can recover the motif with the best MCC within the 4th or 5th in the output up to 70% of noise. Right boxplot indicates the rank of the motif most similar to the RF motif. Even here we can give a range of confidence up until about 70% of noise. All error bars showed are sample standard deviations.

Figure 3. The Iron Responsive Elements (IRE) motif. Top panel: two foldings for IRE RNAs. Bottom panel: Logo in qBEAR notation showing the structural motif that is composed of both of the foldings combined in the same RNA subset.

recorded, as well as the motif coordinates in each RNA included in the model. If more motifs are requested for the same data set, these coordinates are masked (i.e. excluded) in the ensuing runs.

Every motif is reported as its PFM and its consensus, using for clarity a simplified version of the BEAR alphabet denoted as qBEAR (quick BEAR), and its significance is evaluated on a background distribution (see the Supplementary Materials for details).

Assessing BEAM robustness at different noise levels

To assess the BEAM ability of finding motifs when not all the RNAs in the input set actually contain a motif instance, we created a series of artificial data sets by injecting a fixed proportion of noise, i.e. RNAs that do not contain the motif. For every RF we evaluated two motifs from the BEAM run: the one with the highest discriminative power between true and false positives computed as the Matthews Correlation Coefficient (MCC) and the one most similar to the RF motif, using as background RNAs with a similar length and fraction of paired nucleotides (these groups are called L-S bins, a detailed description can be found in the Supplementary Materials and in Figure S1). The original RNAs of the foreground were checked to appear only once.

The noise sets were prepared by injecting progressive noise (RNAs not from the same RF but from the same L-S bin) in the data set (10%, 20%, ..., 100%). The total number of structures was maintained as to avoid the possible confounding effect of different sized data sets. To achieve this, we removed an equal amount of the original RNAs at random. While the previously removed original sequences were kept out of the following level of noise, the injected noise was extracted *de novo* every time. To avoid composi-
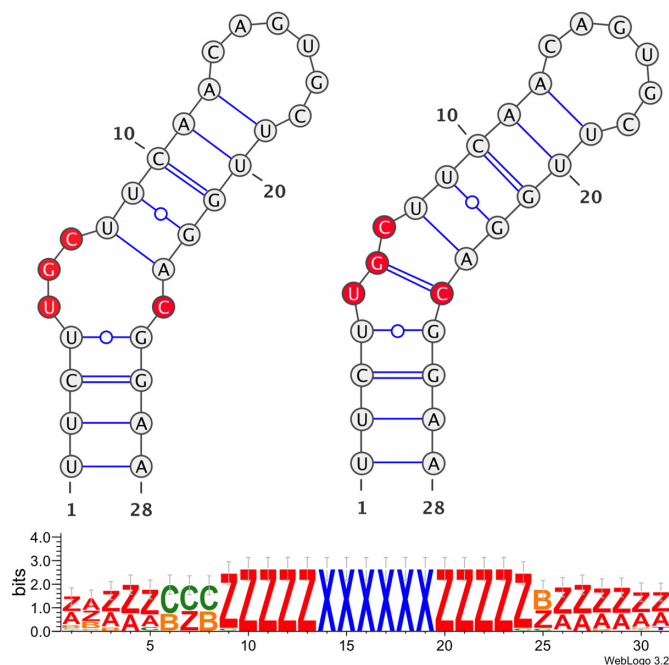
tion bias due to the removed RNAs, for each level of noise we created 10 different *replicas*. Consequently, for each RF we built 100 data sets (10 *replicas* for each of the 10 levels of noise). Note that the 100% noise data sets do not retain anything of the original RF, hence being representatives of full background data sets. All the RFs with a number of seed members between 10 and 250 were used in this test.

Generation of artificial large data sets

To demonstrate the capability of BEAM of working with larger data sets (thousands of RNAs after filtering, reduction of redundancy, etc.) we took a series of mouse RNAs contained in the data set GSE37114 in the GEO database (44). This is a large data set (about 180k RNAs) of LIN28A interactors (45). Of all the RNAs, a subset of 10k was chosen until we had a 'ground' set without any significant motif (MCC $\approx 10^{-2}$). We then proceeded with the insertion of a known motif (from now on denoted as the gold structure) with this pipeline: (i) Fold the ground set of 10k sequences with RNAfold with default parameters; (ii) For every RNA, spot the less affecting zone of the fold where to insert the gold structure, defined as the area of the fold that was more distant from every non-branching hairpin; (iii) Insert the nucleotide sequence of the gold structure (i.e. the gold sequence) into the less affecting zone of a randomly selected 10% of the ground set RNAs, then refold the resulting sequences; (iv) Repeat step (iii) for different density levels, from 10% to 100% in steps of 10, creating 10 data sets. The procedure is repeated for 100 different hairpin gold structures sampled from Rfam, generating a total of 1000 data sets (10 for each gold structure). Details on the procedure can be found in the Supplementary Materials. A back-

ground set was also generated by sampling another subset of 10k RNA from LIN28A and shuffling its sequences with the software uShuffle (46) before folding with RNAfold.

We then ran BEAM on each data set, evaluating its ability of retrieving the gold structure by means of the MCC and the structural distance between the gold and the predicted motif. Prediction accuracy was also related to the folding accuracy, i.e. how similar is the folded inserted gold sequence to the gold structure. The precise definition of these measures and a more detailed description of the results are provided in the Supplementary Materials and in Figure S2.

High-Throughput protein–RNA interaction data sets

We selected human and mouse protein–RNA interaction data detected by PAR-CLIP or HITS-CLIP from the Do-RiNA database, taking every available data set but filtering out miRNA-related ones. Given the interaction data detected for a given RNA binding protein, we analyzed separately those mapping on CDSs and those mapping into UTRs, as reported in the GENCODE annotation file (gencode.v19.hg19 and gencode.vM1.mm9). This is done under the assumption that a specific RBP function should be carried out on the same area of an RNA.

RESULTS AND DISCUSSION

Method overview

We developed a new algorithm (BEAM) to identify structural motifs in a set of RNAs. With this tool we want to address the problem of finding local structural recurrences in a set of (supposedly) functionally related RNAs. The model we present is entirely based on secondary structures. This choice is due to the observation that there are cases in which structural similarity arises even in absence of significant sequence identity (2).

The algorithm is based on a simulated annealing approach, exploiting the recently developed BEAR encoding (38) and its associated RNA structural elements substitution matrix (MBR) to derive a scoring function. We first verified the BEAM ability in finding known motifs, then assessed its robustness to background noise by artificially injecting in RNA sets structures not containing the motif. Finally, we applied BEAM to the identification of structural motifs in large and noisy data sets, both artificial and composed of RNAs bound by a given protein, as detected in CLIP-Seq experiments.

Using BEAM on small data sets

General performances on Rfam. Firstly, we tested the ability of BEAM in detecting structural similarities in relatively small sets, such as known RNA families, as reported in Rfam. Families were selected by considering noise insertion limitations (see Materials and Methods), resulting in ∼800 RFs. The boxplot on the left in Figure 2 represents the motif (here called mask, meaning that each identified motif was masked in the ensuing runs) at which the highest MCC is found, considering the noise percentage in the input set. Note that 100% means that the input set is fully composed of random RNAs, therefore not having a single RNA of the

original considered family. The median is centered on the 2nd motif in the output up to 30% of noise, but the method can recover the motif with the best MCC within the 4th or 5th in the output up to 70% of noise. The boxplot on the right is representing the mask (motif) that is most similar (in terms of PFM distance, defined in the Supplementary Materials) to the one we retrieve at 0% noise (gold standard), for each RNA family. This plot is assessing the robustness of the method because it shows how, even in a (likely) situation containing many unrelated RNAs and only a small core (up to 20%) of RNAs encoding the motif, our tool is able to retrieve a correct result in the 4th–5th position of the output list.

Iron responsive element (IRE - RF00037). Iron Responsive Elements (IRE) are ∼30 nt long sequences with a small conserved motif in primary structure (typically CAGUGN) (47). RNAs containing IREs are involved in iron metabolism in animal cells (48). The current known structural context of IREs, as reported by the Rfam family RF00037, is a 6-nt loop with a 3-nt internal loop at the 5′ and an opposing bulge with an alternative form where the central nucleotide in the internal loop is bound to the bulge. According to the MBR substitution scores, both variants have similar scores and BEAM successfully retrieved both these foldings in the same motif (Figure 3).

IRES – BEAM as structural classifier. The previous examples showed the ability of BEAM in identifying known structural motifs in data sets composed by up to few hundreds of RNAs, and its robustness when RNAs not containing motif instances contaminate the input set. Here, we tested BEAM for its ability to distinguish between distinct structural motifs (belonging to different RNAs) in the same input set. The purpose of this kind of analysis can be the classification of RNAs based on structural characteristics. We tested two Internal Ribosome Entry Site (IRES) families, described in Rfam in the RFs RF00223 and RF00224 (respectively bip IRES and FGF-2 IRES), which we combined in the same input set. This set is composed of 9 (RF00223) and 6 RNAs (RF00224) seed sequences. BEAM was asked to retrieve 10 motifs: every motif was then scored with the percentage of RNAs included in the final PFM that belonged to the most represented family. This so-called 'purity' of the PFM represents the ability of BEAM to distinguish between two families that differ in structure, but have a very deceptive sequence similarity. In fact we ran the same test with the popular sequence-based motif identification algorithm MEME (1), and the results showed that MEME cannot reliably distinguish between these two families, providing, as a proof of concept, a case where a motif-finding method purely based on secondary structure patterns can be better suited. In Table 1, we report the purity value of the ranked motifs, where a value of 1 means full homogeneity of a single RF members while lower values come from mixed situations where the motif finder could not be able to distinguish between the two RNA families.

After the first five motifs, BEAM started to retrieve less significant motifs contained in both families. This is expected since after having masked every highly family-specific motif, only trivial structures remain. With a
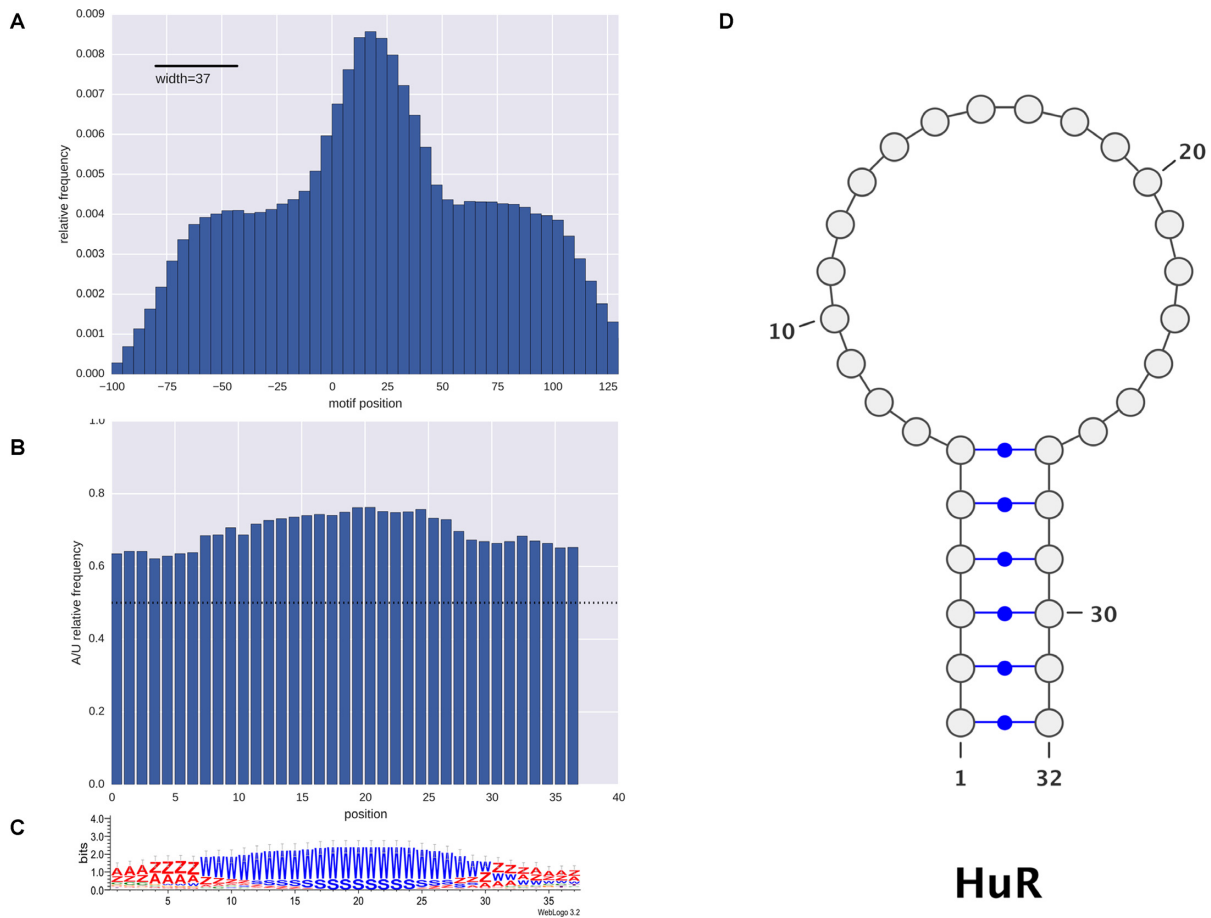
Figure 4. The HuR binding motif. (**A**) Histogram showing the motif position for the HuR data set. The zero of the x-axis corresponds to the start of the binding site reported in the corresponding BED file for every RNA; (**B**) A/U percentage of primary sequence motif corresponding to the structural motif found. The horizontal line at 50% is the value expected by chance; (**C**) Logo in qBEAR notation showing the structural motif; (**D**) Structural motif representation, showing the long loop that composes the structural signal.

Table 1. The table shows the fraction of RNAs belonging to the major RF found with the motif over the total RNAs retrieved (purity)

Motif rank	BEAM (noiseless)	MEME (noiseless)	CMfinder (noiseless)	BEAM (60% added noise)
1	1	0.60	0.6	1
2	1	1	1	1
3	1	0.60	1	1
4	1	0.69	0.60	1
5	1	1	1	1
6	0.50	1	0.64	0.33
7	0.57	0.60	1	0.61
8	0.78	1	0.64	0.44
9	0.80	1	1	0.33
10	0.67	1	1	0.67
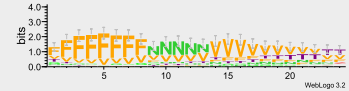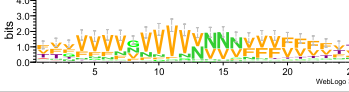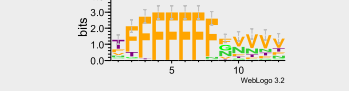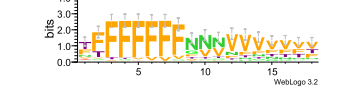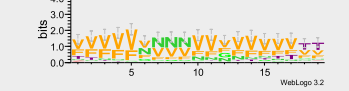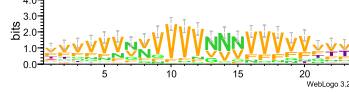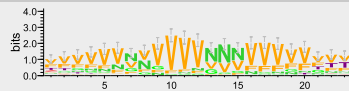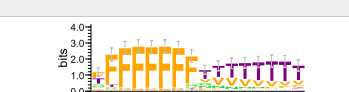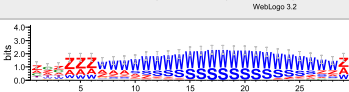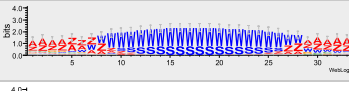
In this case BEAM manages to find characteristic motifs for RF00223 or RF00224 in the first five ranked motifs, while structural motifs found using the covariance model-based algorithm CMfinder, or nucleotide motifs (found with MEME) were not able to discriminate consistently between the two families. The same results apply with a 60% of added noise.

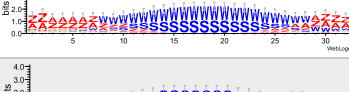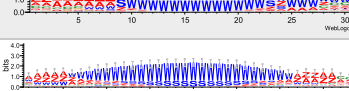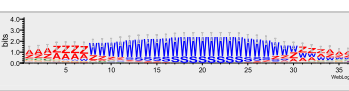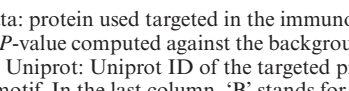sequence-based approach instead, MEME does not provide a reliable output since characteristic motifs (with purity equal to 1) are interspersed with motifs common to both families, providing another example where structural analysis is able to capture features that are elusive at sequence level. Moreover, we tested if this discrepancy remained when using another structural approach, CMfinder (27). This covariance model-based algorithm does not perform as well as BEAM in this case.

The same test was done by adding 9 random sequences to the total of 15, (about a 60%, arbitrary) and the same conclusions apply (note that this is different from the whole performance test, since here the noise is added to the total and is not replacing any of the original RNAs; this choice is due to the small size of this set).

Table 2. Performance on CLIP-Seq data sets

data	map	motif	*P*-value	coverage	input	Uniprot	pos
HNRNP hg19	utr		6,62E-03	0,46	2397	P14866	3'
SFRS1 hg19	cds		5,24E-03	0,48	12 175	Q07955	3'
EZH2 mm9	cds		7,88E-03	0,55	1244	Q61188	3'
LIN28A hg19	cds		1,01E-02	0,55	5662	Q9H9Z2	3'
Mbnl1 mm9	utr		1,49E-02	0,63	275	Q9JKP5	3'
AGO2 (Haecker) hg19	cds		2,76E-03	0,52	5955	Q9UKV8	3'
AGO2 (Karginov) hg19	utr		5,21E-03	0,49	14 021	Q9UKV8	3'
HuR hg19	utr		2,20E-02	0,62	9257	Q15717	5'/3'
CAPRIN1 hg19	cds		1,04E-02	0,68	4639	Q921F2	5'/3'
HNRNPC hg19	utr		2,17E-02	0,36	1597	P07910	B
TIA1 hg19	utr		1,20E-02	0,32	9949	P31483	B
FUS hg19	utr		3,17E-02	0,40	10 468	P35637	B
TIAL1 hg19	utr		4,03E-02	0,45	17 379	Q01085	B
ELAVL1 hg19	utr		1,14E-02	0,34	9929	Q15717	B
HuR hg19	utr		1,37E-02	0,33	9257	Q15717	B

The table columns are as follows: data: protein used targeted in the immunoprecipitation, and species (hg19: human, mm9: mouse); map: region in which the motif was found; *P*-value: motif *P*-value computed against the background distribution; coverage: fraction of input RNAs in which the motif is found; input: total number of input RNAs; Uniprot: Uniprot ID of the targeted protein; pos: positional preference of the found structural motif with respect to the experimentally defined binding motif. In the last column, 'B' stands for a positional preference of the motif on the binding site reported by DoRiNA, '3'' stands for a positional preference downstream of the binding site, '5'' indicates a positional preference upstream of the binding site.
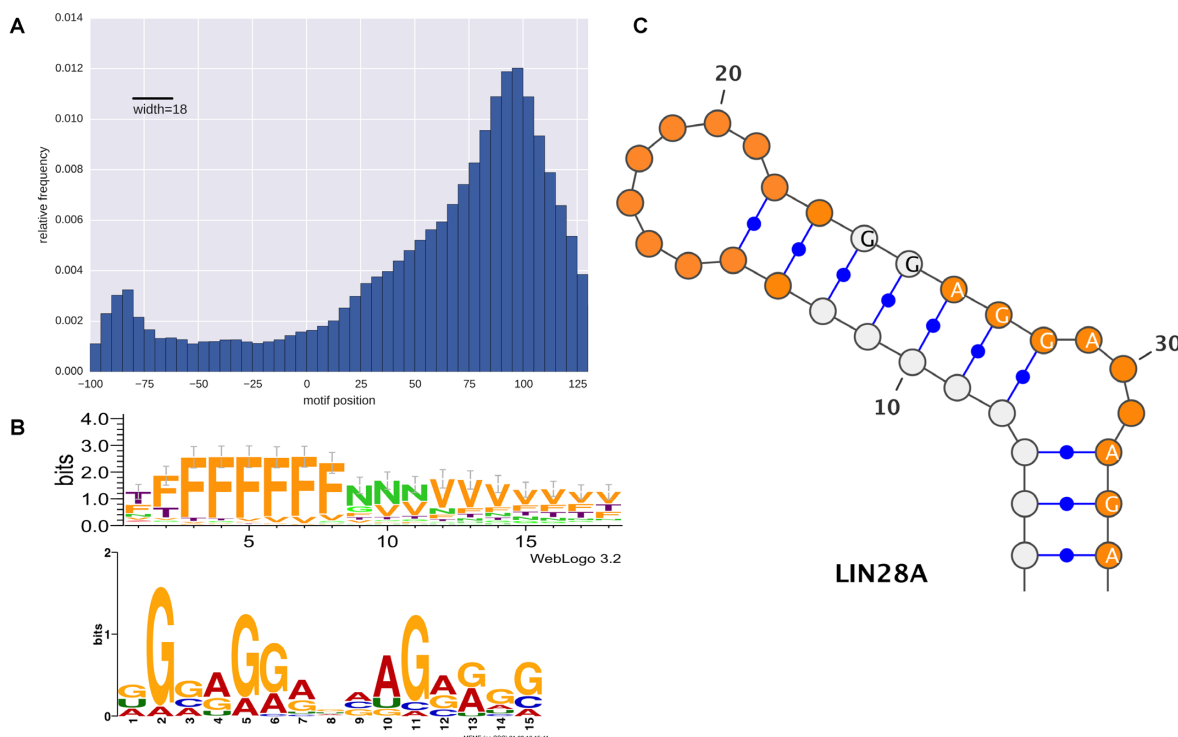
Figure 5. The Lin28A binding motif. (**A**) Histogram showing the motif position for the LIN28A data set. The zero of the x-axis corresponds to the start of the binding site reported in the corresponding BED file for every RNA; (**B**) (top) Logo in qBEAR notation (bottom) primary sequence motif corresponding to the structural motif found; (**C**) Structural motif representation, showing the two independent structural motifs separated by a variable stretch of nucleotides, along with the primary sequence motif mapped around the internal loop.

Performance of BEAM on HT data sets

We then verified the ability of BEAM in identifying structural motifs in large artificial data sets where the amount of noise could be high, thus simulating a realistic CLIP-Seq scenario. We did this by planting sequences corresponding to structural motifs into a sampling of 10k RNAs, at different density levels (i.e. varying the number of RNAs containing the motif), to assess the BEAM robustness at different noise levels. BEAM was able to retrieve the motif in data sets in which down to only 20% of the RNAs were containing the planted motif, and this was also true when the planted motif was not folded correctly by RNAfold, up to 40% of structural alteration with respect to the true structure. Moreover, we used the same data sets to compare BEAM with CMfinder, obtaining positive results (the latter is actually a test on small inputs since CMfinder does not scale very well over some hundreds of sequences, see Supplementary Materials). We were able to show that BEAM is as much accurate, if not better, than CMfinder (Supplementary Figure S4), with similar running times on smaller data sets, and remarkably better running times on larger ones (Supplementary Figure S5). We then launched BEAM on several of the PAR-CLIP/HITS-CLIP data sets present in the database DoRiNA (42) and reported for each set those motifs that had a 1-tailed *P*-value < 0.05, coverage > 20% (as per results of our benchmark on large data sets, see Supplementary Materials) and with fallout (or False Positive Rate) < coverage (data not shown) for a total of ∼40 motifs identified for 20 RBPs. We also verified the structural

motif position with respect to the binding site detected by the analysis of PAR-CLIP/HITS-CLIP data (Supplementary Figure S3). A number of the identified motifs shows clear positional presence, either upstream, downstream or overlapping the binding site. On a total of 12 RBPs with positional preference (Table 2), 5 showed the structural motif overlapping the binding site reported by the corresponding experiment, 6 downstream and outside of the binding site and 2 both on the 5′ and 3′ extensions outside of the binding site (HuR data sets returned two significant motifs, one on the binding site and one on the 5′ and 3′ ends). One possible interpretation is that these RBPs recognize structural motifs as signals and use them to bind respectively downstream, upstream or right at them.

The RBP resulting in signals that fall both in 5′ and 3′ ends may belong to the class of homo-dimers, with repeated domains allowing multiple contacting points. It is the case of CAPRIN1, known mRNA transport regulator (49), which may form homo-multimers. We further investigated structural motifs identified for HuR and LIN28A, since sequence preferences for the binding to these two proteins are known and reported in the literature. HuR was reported to bind selectively AU-rich elements composed by 17–20 nt of ssRNA in UTRs (50,51) in human, and BEAM identified a loop motif composed of 21 nt (present in 33% of the total data set), having a positional preference coinciding with the beginning of CLIP-Seq binding site. Moreover, the structural motif has a higher percentage of AU than expected by chance (Figure 4). About half of the RNAs containing this motif have the structural motif positioned

at 5'/3' ends (1784 over ~3100). LIN28A in human contains a zinc-finger domain that is known to bind 5'-GGAG-3' (or more generally 5'-NGNNG-3' (52)). The reported structural motif (a 3 nt internal loop on a branching stem) has an underlying sequence motif (computed with MEME) that matches (Figure 5). Moreover, this structure is located downstream of the binding site, as Zinc Fingers (ZF) are known to contact these motifs to guide the binding of the CSD domain in an upstream loop, which is located at variable distance from the ZF contact (53,54). The loop has been identified by our method but is not significant in terms of score, hence it is not reported. Our results indicate that only 4.3% (134 over ~3100) of sequences containing the structural motif have the primary sequence motif (that is the most significant according to MEME), revealing a possible structure signal that cannot be captured by sequence alone.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Bailey,T.L., Boden,M., Buske,F.a, Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
2. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
3. Sonnhammer,E.L.L., Eddy,S.R. and Durbin,R. (1997) Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins Struct. Funct. Genet.*, **28**, 405–420.
4. Fallmann,J., Sedlyarov,V., Tanzer,A., Kovarik,P. and Hofacker,I.L. (2015) AREsite2: An enhanced database for the comprehensive investigation of AU/GU/U-rich elements. *Nucleic Acids Res.*, **44**, D90–D95.
5. Leibovich,L. and Yakhini,Z. (2012) Efficient motif search in ranked lists and applications to variable gap motifs. *Nucleic Acids Res.*, **40**, 5832–5847.
6. Lukong,K.E., Chang,K.W., Khandjian,E.W. and Richard,S. (2008) RNA-binding proteins in human genetic disease. *Cell*, **24**, 416–425.
7. Kishore,S., Luber,S. and Zavolan,M. (2010) Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Brief. Funct. Genomics*, **9**, 391–404.
8. Singh,S. (2002) RNA-protein interactions that regulate pre-mRNA splicing. *Gene Expr.*, **10**, 79–92.
9. Licatalosi,D.D. and Darnell,R.B. (2010) RNA processing and its regulation: Global insights into biological networks. *Nat. Rev. Genet.*, **11**, 75–87.
10. Ferrè,F., Colantoni,A. and Helmer-Citterich,M. (2016) Revealing protein–lncRNA interaction. *Brief. Bioinform.*, **17**, 106–116.
11. Georgiev,S., Boyle,A.P., Jayasurya,K., Ding,X., Mukherjee,S. and Ohler,U. (2010) Evidence-ranked motif identification. *Genome Biol.*, **11**, 1–17.
12. Buckanovich,R.J. and Darnell,R.B. (1997) The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Mol. Cell. Biol.*, **17**, 3194–3201.
13. Hiller,M., Zhang,Z., Backofen,R. and Stamm,S. (2007) Pre-mRNA secondary structures influence exon recognition. *PLoS Genet.*, **3**, e204.
14. Meisner,N., Auer,M., Jaritz,M. and Stadler,P.F. The effect of RNA secondary structures on RNA-Ligand binding and the modifier RNA mechanism: A quantitative model. *Bioinformatics*, **345**, 3–12.
15. Cusack,S. (1999) RNA–protein complexes. *Curr. Opin. Struct. Biol.*, **9**, 66–73.
16. Li,X., Kazan,H., Lipshitz,H.D. and Morris,Q.D. (2014) Finding the target sites of RNA-binding proteins. *Wiley Interdiscip. Rev. RNA*, **5**, 111–130.
17. Badr,G., Al-Turaiki,I. and Mathkour,H. (2013) Classification and assessment tools for structural motif discovery algorithms. *BMC Bioinformatics*, **14**(Suppl. 9), S4.
18. Achar,A. and Sætrom,P. (2015) RNA motif discovery: A computational overview. *Biol. Direct.*, **10**, 61–86.
19. Gorodkin,J., Heyer,L.J. and Stormo,G.D. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.*, **25**, 3724–3732.
20. Gorodkin,J., Stricklin,S.L. and Stormo,G.D. (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **29**, 2135–2144.
21. Ji,Y., Xu,X. and Stormo,G.D. (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, **20**, 1591–1602.
22. Pavesi,G., Mauri,G., Stefani,M. and Pesole,G. (2004) RNAProfile: An algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **32**, 3258–3269.
23. Zambelli,F. and Pavesi,G. (2015) In: Picardi,E (ed). *RNA Bioinformatics*. Springer, NY, pp. 49–62.
24. Liu,Y., Zhao,Q., Zhang,H., Xu,R., Li,Y. and Wei,L. (2015) A new method to predict RNA secondary structure based on RNA folding simulation. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, doi:10.1109/TCBB.2015.2496347.
25. Hamada,M., Tsuda,K., Kudo,T., Kin,T. and Asai,K. (2006) Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*, **22**, 2480–2487.
26. Hiller,M., Pudimat,R., Busch,A. and Backofen,R. (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, **34**, e117.
27. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
28. Anwar,M., Nguyen,T. and Turcotte,M. (2006) Identification of consensus RNA secondary structures using suffix arrays. *BMC Bioinformatics*, **7**, 1–15.
29. Michal,S., Ivry,T., Schalit-Cohen,O., Sipper,M. and Barash,D. (2007) Finding a common motif of RNA sequences using genetic programming: The GeRNAMo system. *IEEE/ACM Trans Comput Biol Bioinform.*, **4**, 596–610.
30. Rabani,M., Kertesz,M. and Segal,E. (2008) Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl. Acad. Sci.*, **105**, 467–479.
31. Tabei,Y. and Asai,K. (2009) A local multiple alignment method for detection of non-coding RNA sequences. *Bioinformatics*, **25**, 1498–1505.
32. Maticzka,D., Lange,S.J., Costa,F. and Backofen,R. (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.*, **15**, R17.
33. 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.
34. 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.
35. 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.

36. Kazan,H. and Morris,Q. (2013) RBPmotif: a web server for the discovery of sequence and structure preferences of RNA-binding proteins. *Nucleic Acids Res.*, **41**, W180–W186.

37. Bahrami-Samani,E., Penalva,L.O.F., Smith,A.D. and Uren,P.J. (2015) Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Res.*, **43**, 95–103.

38. Mattei,E., Ausiello,G., Ferrè,F. and Helmer-Citterich,M. (2014) A novel approach to represent and compare RNA secondary structures. *Nucleic Acids Res.*, **42**, 6146–6157.

39. Mattei,E., Pietrosanto,M., Ferrè,F. and Helmer-Citterich,M. (2015) Web-Beagle: a web server for the alignment of RNA secondary structures: Figure 1. *Nucleic Acids Res.*, **43**, W493–W497.

40. Änkö,M.L. and Neugebauer,K.M. (2012) RNA-protein interactions in vivo: Global gets specific. *Trends Biochem. Sci.*, **37**, 255–262.

41. Bai,Y., Dougherty,L. and Xu,K. (2014) Towards an improved apple reference transcriptome using RNA-seq. *Mol. Genet. Genomics*, **289**, 427–438.

42. Blin,K., Dieterich,C., Wurmus,R., Rajewsky,N., Landthaler,M. and Akalin,A. (2014) DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.*, **43**, D160–D167.

43. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neuböck,R. and Hofacker,I.L. (2008) The vienna RNA websuite. *Nucleic Acids Res.*, **36**, W70–W74.

44. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

45. Cho,J., Chang,H., Kwon,S.C., Kim,B., Kim,Y., Choe,J., Ha,M., Kim,Y.K. and Kim,V.N. (2012) LIN28A is a suppressor of ER-associated translation in embryonic stem cells. *Cell*, **151**, 765–777.

46. Jiang,M., Anderson,J., Gillespie,J. and Mayne,M. (2008) uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, **9**, 1–11.

47. Walden,W.E., Selezneva,A.I., Dupuy,J., Volbeda,A., Fontecilla-Camps,J.C., Theil,E.C. and Volz,K. (2006) Structure of dual function iron regulatory protein 1 complexed with ferritin IRE-RNA. *Science*, **314**, 1903–1908.

48. Piccinelli,P. and Samuelsson,T. (2007) Evolution of the iron-responsive element. *RNA*, **13**, 952–966.

49. Ellis,J.A. and Luzio,J.P. (1995) Identification and characterization of a novel protein (p137) which transcytoses bidirectionally in Caco-2 cells. *J. Biol. Chem.*, **270**, 20717–20723.

50. de Silanes,I.L., Zhan,M., Lal,A., Yang,X. and Gorospe,M. (2004) Identification of a target RNA motif for RNA-binding protein HuR. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 2987–2992.

51. 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.

52. Heo,I., Joo,C., Kim,Y.-K., Ha,M., Yoon,M.-J., Cho,J., Yeom,K.-H., Han,J. and Kim,V.N. (2009) TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell*, **138**, 696–708.

53. Nam,Y., Chen,C., Gregory,R.I., Chou,J.J. and Sliz,P. (2011) Molecular basis for interaction of let-7 microRNAs with Lin28. *Cell*, **147**, 1080–1091.

54. Zeng,Y., Yao,B., Shin,J., Lin,L., Kim,N., Song,Q., Liu,S., Su,Y., Guo,J.U., Huang,L. *et al.* (2016) Lin28A binds active promoters and recruits Tet1 to regulate gene expression. *Mol. Cell*, **61**, 153–160.