BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 14. NO 1. 74–89 doi:10.1093/bfgp/elu047

High-throughput characterization of protein–RNA interactions

Kate B. Cook, Timothy R. Hughes and Quaid D. Morris

Advance Access publication date 13 December 2014

Abstract

RNA-binding proteins (RBPs) are important regulators of eukaryotic gene expression. Genomes typically encode dozens to hundreds of proteins containing RNA-binding domains, which collectively recognize diverse RNA sequences and structures. Recent advances in high-throughput methods for assaying the targets of RBPs *in vitro* and *in vivo* allow large-scale derivation of RNA-binding motifs as well as determination of RNA–protein interactions in living cells. In parallel, many computational methods have been developed to analyze and interpret these data. The interplay between RNA secondary structure and RBP binding has also been a growing theme. Integrating RNA–protein interaction data with observations of post-transcriptional regulation will enhance our understanding of the roles of these important proteins.

Keywords: *RNA-binding proteins; RBP target identification; high-throughput sequencing; RNA secondary structure*

INTRODUCTION: RNA-BINDING PROTEINS, RNA-BINDING DOMAINS AND RNA SECONDARY STRUCTURE

RNA-binding proteins (RBPs) have diverse roles in post-transcriptional gene expression, including regulation of alternative splicing, RNA export and localization, RNA stability and translation [1]. Many RBPs regulate multiple cellular processes [2–5], and are implicated in human diseases including cancer and neurological disorders [6]. RBP functionality in gene regulation is naturally dependent on their ability to selectively recognize and bind target RNAs within the cell; consequently, elucidation of RBP specificity is an area of active research. Recent technological developments have allowed characterization of RNA–protein interactions at an unprecedented scale.

Here we introduce the major classes of sequence- and structure-specific RNPs, and discuss aspects of RNA secondary structure that impact RBP binding; knowledge of how proteins interact with RNA is important for the interpretation of high-throughput data. We then review current methods for high-throughput experimental determination of the RNA targets of RBPs *in vitro* and *in vivo*, as well as methods to determine the proteins bound to an RNA molecule. Finally, we discuss computational methods for analyzing high-throughput data and predicting RBP binding.

RNA recognition by RNA-binding domains

Different classes of RNA-binding domains (RBDs) use different strategies for binding to RNA. Nonetheless, there are some general features of RBP–RNA interactions: RNA recognition is commonly a combination of recognition of the RNA by the overall protein fold (involving hydrogen bonds with backbone atoms) as well as specific amino acid side chain–nucleotide interactions. Target specificity is often accomplished by way of hydrogen bonding and electrostatic interactions; the latter also contribute to the protein's affinity for RNA along with

Corresponding author. Quaid D. Morris, Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, 160 College Street, Room 616, Toronto, ON M5S 3E1, Canada. Tel.: +1-416-978-8568; Fax: +1-416-978-8287; E-mail: quaid.morris@utoronto.ca

Kate Cook is a PhD student in the Department of Molecular Genetics at the University of Toronto. She studies targeting and specificity of RNA-binding proteins.

Timothy R. Hughes is a Professor in the Donnelly Centre and Molecular Genetics at the University of Toronto, and is a fellow of the Canadian Institute for Advanced Research Genetic Networks program.

Quaid Morris is an Associate Professor in the Donnelly Centre, Molecular Genetics, Computer Science, and Electrical and Computer Engineering at the University of Toronto, and a visiting professor at the Centre for Genomic Regulation in Barcelona.

© The Author 2014. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

stacking interactions [7]. Individual RNA-binding domains typically only contact a few nucleotides, and combinations of RBDs within the same protein are frequently observed, presumably to increase affinity and specificity. Structures of the most common RNA-binding domains are shown in Figure 1 and described below. Not all sequence-specific RBPs contain canonical RBDs [8, 9]. The full extent of the RNA-binding proteome is an area of current research, discussed in a further section.

RNA recognition motif

The RNA recognition motif (RRM) is found in all domains of life; in metazoans, it is the most common RNA-binding domain. A single RRM spans ~90 amino acids and contains two conserved motifs, RNP-1 and RNP-2, consisting respectively of 8 and 6 mostly positively charged or aromatic amino acids [10, 11].

Structurally, the RRM consists of a four-strand antiparallel β-sheet backed by two α-helices. RRMs are highly versatile in their mode of RNA recognition: in canonical RRM–RNA interactions, the bound RNA lies across the β-sheet and contacts one or more key residues in the conserved RNP-1 and RNP-2 motifs; however, the amount of the β-sheet surface directly contacting the RNA varies considerably [12]. Noncanonical RRM–RNA interactions can involve interactions with loop regions or amino acids N- or C-terminal to the RRM domain. In some cases, the β-sheet surface is not involved in RNA binding at all [13].

The β-sheet surface of a single RRM can contact up to four nucleotides, while engaging the loop regions external to the β-sheet can allow binding of up to six nucleotides [12]. Approximately 40% of RRM-containing proteins contain multiple RRM domains [12, 14]. Two RRMs can be separated by a flexible linker, can be arranged as a continuous RNA-binding platform either oriented in the same direction (Figure 1A, [15]) or forming an RNA-binding cleft (Figure 1B, [16]) or can interact back to back, forcing the RNA to loop around the protein (Figure 1C, [17]). Presumably, other binding modes are possible; as in Nucleolin, U1A/SNRPA [18] or other proteins that bind RNA in the context of secondary structures (Figure 1G–H).

K homology domain

K homology (KH) domains, named after the founding member of the family, hnRNP K [19], are also widely present in all three domains of life, although the KH type most commonly found in prokaryotes adopts a different fold than the majority eukaryotic type [20]. The typical metazoan genome encodes several dozen proteins containing KH domains [14]. KH domains are about 70 amino acids in size, and bind RNA inside a cleft composed of two α-helices, a variable loop sequence containing a conserved GXXG motif, and a β-strand (Figure 1D). This binding cleft can accommodate four RNA bases, and KH domains are often combined in multiples to enhance affinity and specificity of binding [21].

Double-stranded RNA binding domain

Double-stranded RNA binding domains (dsRBDs) are involved in diverse aspects of post-transcriptional regulation, including RNA editing [22], miRNA biogenesis [23] and RNA localization [24, 25]. The human genome encodes 18 proteins containing dsRBDs. The domain is 65–70 amino acids in size and consists of two α-helices packed against a three-strand antiparallel β-sheet. Portions of both α-helices and a loop region between two of the β-strands are involved in RNA binding, and mutation of the amino acids in these regions can affect binding [24, 26]. Structural details of the dsRBD–RNA interaction were recently and extensively reviewed in [27].

Because of the nature of the A-form RNA double helix, in which the major groove is narrow and deep[28], it is generally assumed that dsRBDs recognize only the double-stranded RNA (dsRNA) shape and are not sequence specific. Nonetheless, dsRBD containing proteins do recognize specific target RNAs, which may be the result of recognizing mismatches or bulges in RNA duplexes: a recent study of Staufen targets observed that dsRNA stems with specific numbers of base pairs and few unbalanced unpaired bases were enriched in the bound transcripts [25]. Intriguingly, a structure of ADAR2 bound to RNA (Figure 1I) displays sequence-specific contacts between the protein and the minor groove [29].

Pumilio homology domain

In contrast to the other major RNA binding domains, for which the elucidation of a recognition code has proven elusive, RNA recognition by Pumilio Homology Domains (PUM-HD) is understood to the extent that custom proteins can be designed to bind new sequences. The PUM-HD typically consists of 8 PUF (Pumilio and FBF) repeats

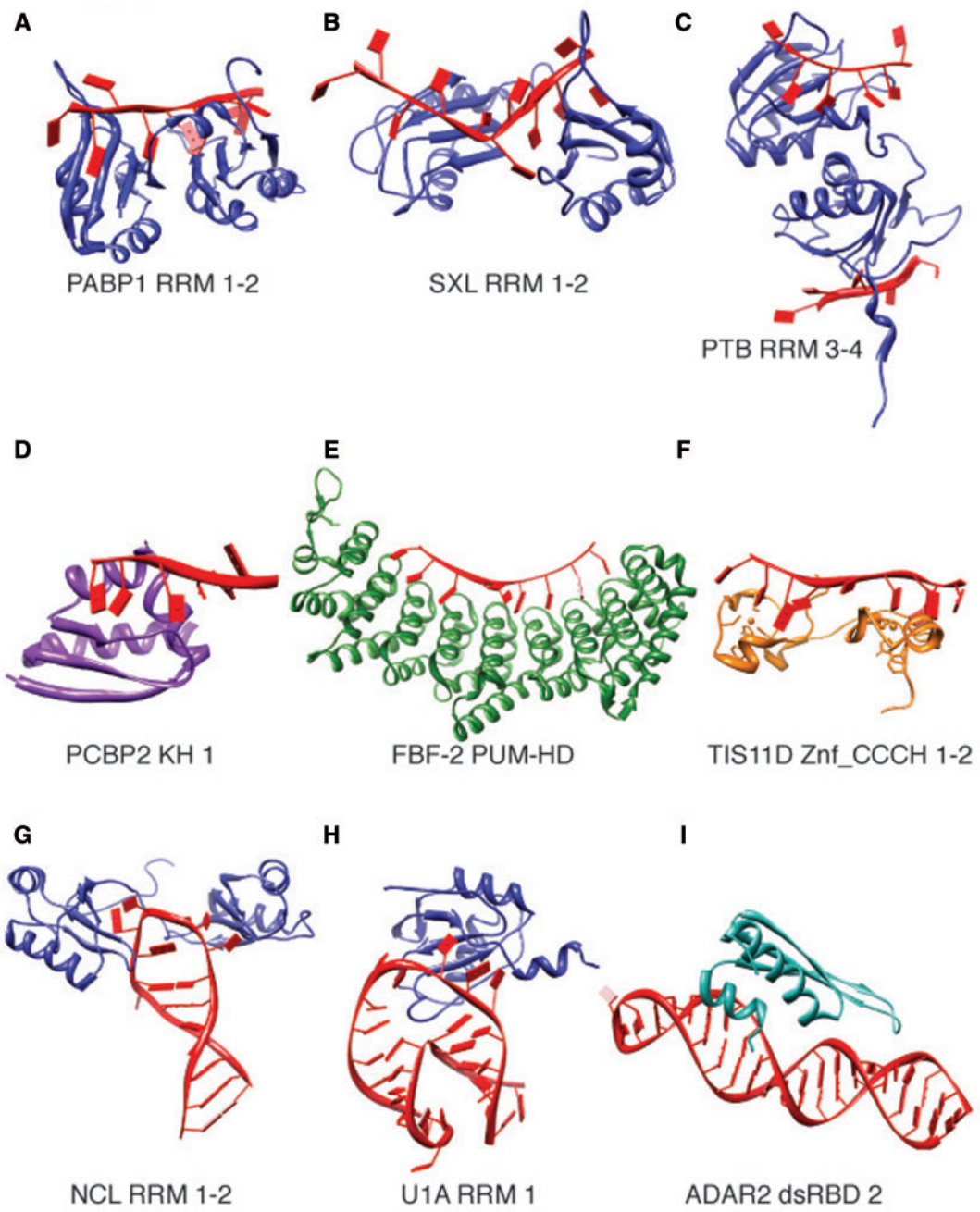

Figure 1: RNA-binding domains use a variety of strategies for binding RNA. (**A–C**), Different arrangements of two RRM domains. (A) RRMs 1–2 of PABPI are arranged to form a flat RNA-binding surface (PDB ID: 1CVJ). (B) RRMs 1–2 of SXL form an RNA-binding cleft (1B7F). (C) RRMs 3–4 of PTB are arranged back to back (2ADC). (**D–F**) Examples of other RNA-binding domains. (D) KH domain 1 of PCBP2 forms an RNA-binding cleft (2PY9). (E) The Puf repeats of the FBF-2 PUM-HD form a concave RNA-binding surface (3K62). (F) The two CCCH zinc fingers of TIS11D/ZFP36L2 (1RGO). (**G–I**) RBPs binding to structured RNA. (G) Hairpin loop recognition by RRMs 1–2 of Nucleolin (1RKJ). (H) Bulge loop recognition by RRM 1 of U1A/SNRPA (1AUD). (I) dsRNA binding by the dsRBD domain of ADAR2 (2L2K).

of a 36 amino acid motif, and the entire domain forms a curved structure that binds RNA in the concave side of the domain, while the convex side mediates protein–protein interactions [30]. Each PUF repeat contacts two RNA nucleotides, and recognizes its target nucleotides using only a few well-conserved amino acids. Because of its modular design, a recognition code for the PUF

repeat has been developed, and custom PUM-HD domains have been designed to bind new motifs [31]. PUM-HD domains have also been engineered to bind to cytosine (which is not observed in any of the natural PUF binding specificities) and to bind targets longer than 8 bases by increasing the number of repeats [32, 33].

Zinc fingers

Zinc fingers are a large and diverse class of domains with the common property of coordinating zinc. The different types of zinc fingers have varying three-dimensional structure and likely have independent evolutionary origins. Nonetheless, several types act as DNA-, RNA-, and protein-binding domains. The extent to which individual zinc finger proteins bind and recognize each class of biopolymers is unknown, however there are some trends: C2H2 zinc fingers are usually DNA binding, while CCCH zinc fingers are primarily single-stranded RNA binding. CCHC zinc knuckles in viral and metazoan proteins also bind RNA; however, RNA binding by metazoan CCHC zinc knuckles is understood only in the context of proteins that also contain another RBD [14, 34]. In support of these trends, a recent mass spectrometry-based study (see below) observed significant enrichment for CCHC, CCCH and several smaller zinc finger families, but not C2H2 zinc fingers, in the mRNA-bound proteome [35].

The human genome encodes ~60 proteins with CCCH zinc fingers (a larger number than contain KH domains), of which 11 have evidence of single-stranded RNA (ssRNA) binding [14]. RNA recognition by CCCH proteins is accomplished through stacking interactions and hydrogen bonds, particularly between the RNA and backbone atoms, so the overall fold of the protein is likely to be important for RNA recognition [36].

The role of RNA secondary structure in RBP binding

RNA structure is a critical aspect of describing, measuring and predicting RBP binding: many RBPs recognize specific structures, while those that bind single-stranded RNA presumably compete with RNA structures. In addition, RBP binding likely has some impact on RNA structure, and *in vivo* RNA structure is impacted both by the presence of ATP-dependent RNA helicases, as well as a number of proteins that could bind co-transcriptionally. Despite the development of high-throughput methods for measurement of RNA structure, there is ongoing controversy regarding the degree of RNA structure present in cells and the accuracy of both computational algorithms and these experimental techniques to predict RNA folding.

Experimental determination of RNA secondary structure

The secondary structure of an RNA molecule can be determined by footprinting techniques: cutting the RNA using RNases specific to ssRNA or dsRNA, or small molecule reagents that cleave or modify RNA at positions in a manner proportional to their accessibility [37]. The cleaved or modified sequences are traditionally separated on a sequencing gel to determine the positions of more or less accessible nucleotides. The first genome-wide application of this strategy was FragSeq, which consists of fragmenting RNA using nuclease S1 (preferring ss or accessible RNA), ligating adaptors to the 5′ phosphate produced and high-throughput sequencing to identify cleavage locations [38]. A similar method, PARS, uses fragmentation with two complementary enzymes: RNase V1, which preferentially cleaves dsRNA, and nuclease S1. The PARS score is the log of the ratio of V1/S1 reads at each position, and reflects the tendency for that base to be double-stranded [39]. Small molecule reagents have also been used in a genome-wide fashion. Dimethyl sulphate (DMS) has been used to profile RNA secondary structure *in vivo* in *Arabidopsis* [40] and yeast and human cells [41].

The overall trends identified by these studies vary. RNA accessibility around the start codon was associated with translational efficiency (as measured by ribosome profiling) in yeast using PARS [39] and in *Arabidopsis* using DMS [40]; however, overall mRNA structural accessibility did not correlate with translation efficiency in yeast using DMS [41]. Using PARS, Kertesz et al. observed a higher level of base pairing in yeast coding sequences as compared with untranslated regions [39]. This contrasts with the results obtained for human PARS data [42] as well as data from *Arabidopsis* using DMS [40] all of which observe coding regions as being more single stranded. Computational predictions in both yeast (K.B.C., unpublished observation) and mammals [43] support a relatively less structured coding sequence on average. One possible application of

these methods is to determine the impact of protein binding on RNA secondary structure, as binding by both ssRNA-preferring RBPs and RBPs that recognize structured RNA is likely to have an impact on RNA structure in an 'induced fit' fashion.

Computational prediction of RNA secondary structure

Computational prediction of mRNA secondary structure generally conforms to one of the two approaches. The first relies on the assumption that thermodynamically stable structures are more likely to exist than unstable structures, exemplified by the Zuker MFOLD algorithm [44] and extended using approaches that consider all possible structures using partition function approaches [45–47]. Many of these algorithms have been implemented in various packages including the Vienna RNA package [48] and the RNAstructure Web servers [49].

As an alternative to the free-energy based algorithms, covariation-based approaches take advantage of the fact that functional RNA secondary structures are more likely to be conserved through evolution. Covariation algorithms use a number of simplifying heuristics (reviewed in [50]), as simultaneous folding and alignment of RNA sequences is computationally costly [51]. While covariation algorithms have been successfully applied to define many noncoding RNA families [52], large numbers of related sequences are required for input. As well, care must be taken in interpreting the results, as the results from covariation methods may be affected by the choice of alignment method if it is not selected to minimize spurious alignments [53], and covariation methods may over-predict structure because their statistical scoring procedure is biased toward predicting base pairing [54, 55].

Benchmarking the accuracy of mRNA secondary structure estimates

Given the inconsistencies among the experimental methods for assessing mRNA secondary structure and uncertainty about the accuracy of computational predictions, it is important to evaluate the accuracy of both these types of estimates. However, doing so has been troublesome because of the lack of gold standards for mRNA secondary structures. Classic RNA secondary structure benchmarks are likely inappropriate because they are composed of highly structured ncRNAs like ribosomal RNAs and ribozymes. In addition, mRNAs are longer than most well-characterized ncRNAs, such that windowed approaches (e.g. the RNAplfold algorithm) are often preferred both for their speed and potentially increased accuracy [56].

Lange and colleagues performed an analysis to determine the accuracy of predicted secondary structures using yeast PARS data [39] and a curated set of structured cis-regulatory elements, and found that more accurate secondary structures were predicted using local (i.e. windowed) folding with window sizes of 100-150 nt, and that the edges of windows were predicted with less accuracy [57]. Estimates for the optimal window size based on siRNA efficacy (which depends on the accessibility of the target RNA) range from 80 nt to 800 nt [58, 59]. Li, Kazan and colleagues applied the observation that accessible sites are more likely to be bound by RBPs to compare the ability of PARS and the RNAplfold algorithm to score accessible sequences and separate bound and unbound transcripts from RIP-chip data [60]. RNAplfold performed significantly better except in the case when only nucleotides with robust PARS data were considered, at which point the difference was not statistically significant. A method that combined PARS data and computational predictions [61] performed better than RNAplfold on some RBPs but results on the entire benchmark were not reported. As such, because of the shallowness of current high-throughput experimental techniques for RNA secondary structure determination and the indirect nature of the data produced, it is likely that computational predictions will continue to be important. However, the optimal parameters (i.e., window size) for making those predictions remain to be determined.

EXPERIMENTAL CHARACTERIZATION OF RBP-RNA INTERACTIONS

High-throughput characterization of RBP–RNA interactions can be broken down into *in vitro* approaches (Figure 2), which determine the specificity of RBPs free from interacting proteins and other cellular factors, and *in vivo* approaches (Figure 3), which measure a snapshot of RBP binding to expressed RNAs. Here we also discuss some aspects of analysis of *in vivo* RBP-RNA data, as identifying bona fide protein binding sites can be challenging. Proteome-wide methods of identifying RBP–RNA interactions are described, and computational aspects of RNA motif finding are discussed.

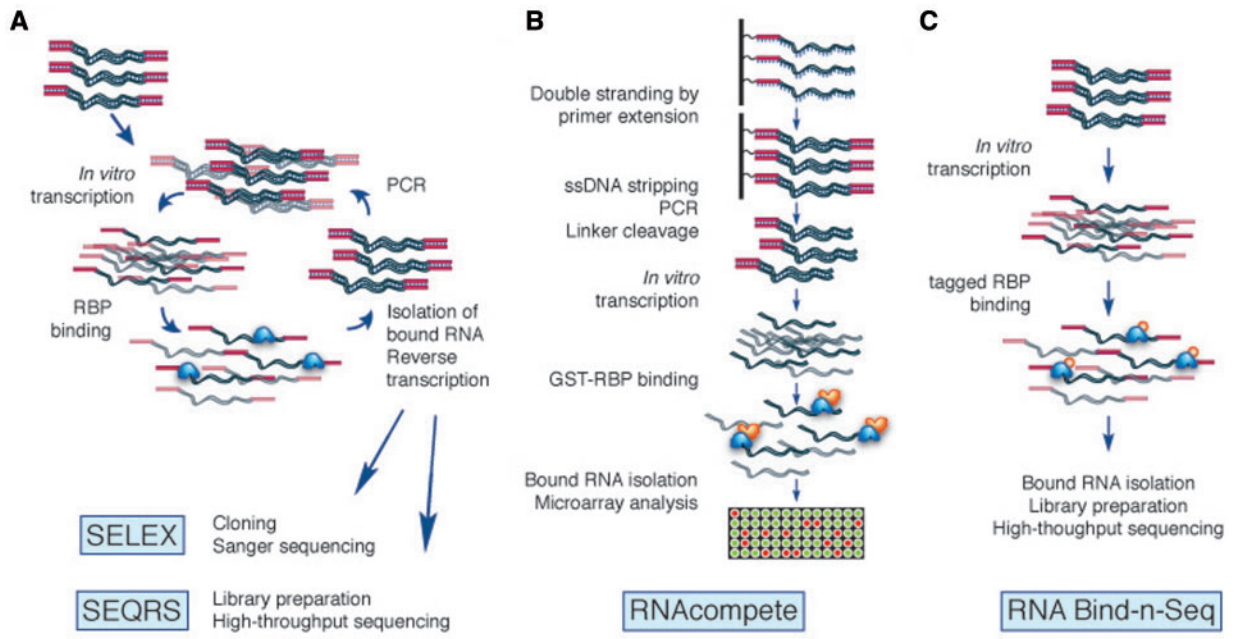

Figure 2: In *vitro* methods for determining RBP targets. (**A**) SELEX consists of several rounds of binding and amplification of RNA molecules. SEQRS modifies traditional SELEX by sequencing the bound pool of RNA at each round. (**B**) RNAcompete queries a designed RNA pool under competitive conditions and assays the bound RNAs using a microarray. (**C**) RNA Bind-n-Seq assays RNA binding by incubating RNA and various amounts of protein and sequencing the bound RNAs.

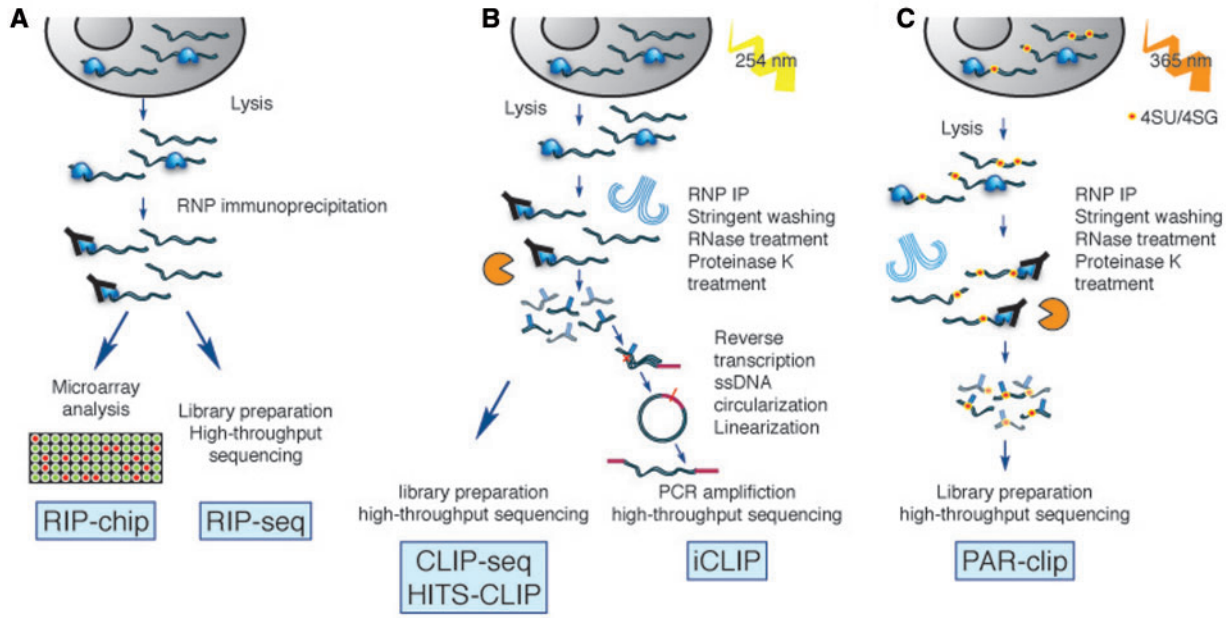

Figure 3: *In vivo* methods for determining RBP targets. (**A**) RIP-chip and RIP-seq determine bound RNAs by analyzing immunoprecipitated RNPs by microarrays or high-throughput sequencing. (**B**) UV cross-linking and immunoprecipitation allows more stringent washing and RNase treatment of bound RNAs. iCLIP identifies binding sites more precisely by taking advantage of the fact that the amino acid tag left by proteinase K treatment terminates reverse transcription. (**C**) PAR-CLIP is another modification of CLIP-seq that first treats the cell with a modified nucleoside (4SU or 6SG), which is incorporated into transcribed RNA. The modified nucleotide can be cross-linked using longer wavelength UV radiation.

In vitro approaches

Systematic evolution of ligands by exponential enrichment (SELEX), also known as *in vitro* selection, is a common method for determining the consensus binding motif for an RBP [62, 63]. In a SELEX experiment, a pool of randomized RNA oligos is incubated with the RBP of interest, bound RNA is reverse transcribed, amplified by PCR and transcribed *in vitro*, and the process is repeated three or more times, with each selection increasing the proportion of high-affinity binding sites in the pool. Selected sequences are traditionally cloned and sequenced by Sanger sequencing. Although not strictly a high-throughput technique, SELEX has been used to ascertain high-affinity motifs for >70 metazoan proteins by various laboratories [14]. In addition, SELEX has been performed in parallel on a number of yeast RBPs [64]. While SELEX is a powerful technique for determining the optimal motif, the highest-affinity motif may not reflect the entirety of biologically functional binding sites [62, 63], and SELEX does not give quantitative information about the protein's affinity for sub-optimal motifs.

SEQRS is a method that applies high-throughput sequencing to SELEX to monitor the enrichment of optimal and alternate binding motifs by sequencing after each round of selection [65]. Similar methods have been applied to measure the DNA-binding specificities of transcription factors in large numbers [66]. Using SEQRS, Campbell and colleagues determined the binding specificity of the *Caenorhabditis elegans* PUM domain RBP FBF-2 to a 20 nt randomized pool alone and in the presence of a non–RNA-binding peptide fragment of the CPEB protein CPB-1, and found that FBF-2's binding specificity was altered in the presence of the CPB-1 fragment. As well, they reported an alternate binding mode for FBF-2 with a slightly different motif.

RNAcompete [67] is a method for determining the binding specificity of RBPs by incubating a purified GST-tagged RBP of interest with a pool of ~40 nt RNAs designed to compactly and robustly represent all short sequences up to 9 bases. The binding reaction is performed with a vast excess of RNA so that RNA molecules compete for binding to the protein, and thus relative abundance can be used to assess relative affinity. After a single-step selection, the bound RNAs are washed, eluted and hybridized to a microarray for detection. RNAcompete was applied to a panel of 200 eukaryotic RBPs to determine their RNA sequence specificities [68]. Furthermore, protein sequence homology-based rules for predicting motifs of closely related RBPs were developed, allowing motifs for 30% of metazoan multi-RBD proteins to be inferred.

Neither SEQRS nor RNAcompete, in their current versions, are optimal for determining the structural specificity of RBPs. SEQRS uses a 20 nt pool, which is not long enough to encompass all but the simplest of secondary structures, while RNAcompete, owing to its microarray-based strategy, is able to robustly represent only primary RNA sequence in the ~244 k RNA sequences in the pool, and so the current version of the RNAcompete pool was designed to avoid highly structured sequences. Nonetheless, motifs determined for proteins known to prefer hairpin loops, such as yeast Vts1p and human SNRPA and LIN28A, were determined, and correspond to the single-stranded portions (e.g. the loop portion of a hairpin loop) of the structural motif. Significant preferences of many RBPs for ssRNA and, in a few cases, hairpin loops were identified [68]. An alternative approach may be provided by the RNA Bind-n-Seq, in which a library of 40 nt RNAs is incubated with varying concentrations of protein, and bound RNAs are isolated and sequenced [69]. The greater representation of secondary structures in this starting pool may support more sensitive inference of the impact of RNA secondary structure on binding.

To study the kinetics of RBP binding to RNA sequences in a high-throughput manner, a recent study from the Greenleaf laboratory adapted an Illumina sequencing machine to measure binding affinity of the MS2 coat protein to ~120 000 RNA sequences [70] representing a range of mutations to the consensus hairpin motif. Direct measurements of off-rates to >3000 sequences were observed, and the detailed data also enabled decomposition of the sequence and structural determinants of binding affinity at each base pair of the hairpin structure.

In vivo approaches

A series of developments over the past decade have revolutionized determination of the RNAs bound *in vivo* by an RBP. The first genome-wide analyses were microarray based, and involved immunoprecipitation of RBP–RNA complexes using an antibody to the endogenous protein or to an epitope tag (denoted as RNA immunoprecipitation followed by microarray analysis, RIP-chip or high-throughput

sequencing, RIP-seq). Tenenbaum et al were the first to apply RIP-chip to the determination of the RNA fraction bound to HuB, HuA/HuR, eIF-4 E and PABP [71]. Since then, RIP-chip/seq has been applied to determine the bound RNA complement of dozens of proteins in several species (collected in RBPDB [14] and reviewed in [72]), leading to the observation that RBP-to-mRNA interactions are, in general, many-to-many: each RBP interacts with a number of mRNAs, and each mRNA is regulated by at least several RBPs [73].

Although RIP allows identification of the target RNA molecules binding to an RBP, the data may include indirectly bound sequences, and precise locations of the binding site on the target mRNA may be difficult to determine. Additionally, RIP conditions must be calibrated to minimize reassociation of RBPs with mRNA *in vitro* after cell lysis, which has been observed under some conditions [74] but not others [75]. Cross-linking the RBP to the RNA using UV radiation before immunoprecipitation (CLIP) provides a way of both ensuring that *in vivo* contacts are maintained, as well as narrowing down the binding site [76]. UV cross-linking creates covalent bonds between proteins and RNA within a range of a few Ångstroms [77], and so background can be reduced by stringent purification protocols. Coupling of CLIP to high-throughput sequencing (HITS-CLIP or CLIP-seq) discloses RBP binding sites genome-wide. While resolution of CLIP-seq is generally limited to ~30–60 nt [78] based on the length of the cross-linked RNA molecules after fragmentation, digestion of the cross-linked protein leaves an amino acid 'tag' on the RNA sequence, which can occasionally cause the reverse transcriptase to skip or misread the cross-linked base, producing mutations that can be diagnostic of the binding site [78]. The behavior of reverse transcriptase at the cross-linked nucleotide is also exploited by iCLIP, which takes advantage of the fact that often the amino acid tag causes termination of reverse transcription at that site to determine the precise location of the termination (and thus the cross-linking) site [79]. Precise determinations of binding site locations are also enabled by PAR-CLIP [31]. In PAR-CLIP, cells are exposed to a modified nucleoside such as 4-thiouridine (4SU) or 6-thioguanosine (6SG), which cross-links more efficiently with proteins at 365 nm UV light (as compared with the 254 nm UV light used for basic CLIP). The reverse transcriptase misreads the modified uridine, causing T → C conversions in the sequenced reads that can be used to pinpoint binding sites.

CLIP-seq and its variants are not without biases. UV cross-linking preferentially bonds certain nucleotides and certain amino acids [80], and not all proteins will cross-link effectively, possibly because of the absence of aromatic amino acids close to the RNA binding site [35, 81]. A recent study quantified UV cross-linking sensitivity by calculating the ratio of RNase-sensitive radioactive signal to protein abundance for a number of yeast RBPs, and observed a wide range of cross-linking efficiencies, even for RRM domains [82]. The 365 nm UV light used in PAR-CLIP only creates bonds at the modified base, so PAR-CLIP tags will tend to be enriched at locations with several of that base. This can bias motif finding toward U-rich motifs (in the context of 4SU PAR-CLIP) [83], as opposed to motifs derived from *in vitro* affinity measurements [84]. UV cross-linking has been applied to a number of systems including suspensions of mouse brain cells [2], and whole *C. elegans* [85] animals, whereas PAR-CLIP is usually applied to cells in culture that can efficiently take up the modified nucleosides, although it has also been performed in the *C. elegans* germ line [86].

Recently, Friedersdorf and Keene demonstrated that a large fraction (up to 45%) of reads (including high-abundance sites) from published PAR-CLIP data sets overlap with binding regions observed in background data from FLAG-GFP immunoprecipitations [87]. Background sites included T → C conversions, albeit at a lower rate, and similar background profiles were observed in several published data sets, suggesting that the background is a result of cross-linking proteins other than the target RBP. PAR-CLIP data from novel RBPs or small data sets had a higher fraction of background overlap. Although motifs determined *in vitro* are enriched in CLIP-seq reads [68], often motifs are not extractable from the CLIP-seq data *de novo* [88]. Background subtraction as described by Friedersdorf and Keene could enrich for the presence of known motifs and improve motif finding [87].

CLIP/PAR-CLIP data analysis

Identification of protein-bound sites from CLIP-seq (and variants) data is nontrivial. Because transcript abundance varies, ascertainment of the protein's preference for a target RNA through quantitation of the number of CLIP-seq reads mapping to a

transcript sequence requires normalization to transcript abundance. Additionally, the choice of RNase and conditions chosen to fragment the cross-linked protein–RNA complex has significant impact on the base composition of the observed CLIP tags [83]. To overcome these challenges, several strategies have been used. CIMS [78] uses the diagnostic deletions present in CLIP-seq tags to help pinpoint the location of RBP binding site by clustering reads and identifying reproducibly occurring deletions. PARalyzer [89] uses the number of T → C conversions that are diagnostic of PAR-CLIP binding sites to define the limits of the binding site using a kernel density-based classifier, but does not use transcript abundance data and so generates a set of RBP-bound sites rather than RBP target preferences. Uren and colleagues [90] directly modeled the background distribution of read counts and account for mappability and transcript abundance (and, optionally, data from a contrasting experiment for comparison of differential binding). These methods, the production of additional data sets for comparison, and the incorporation of control background data [87] may allow for more precise quantitation of binding activity from CLIP data, which could reflect the stability or half-life of RBP-RNA interactions, and provide quantitative data to aid motif finding.

Proteome-wide approaches

While the majority of recent efforts have focused on identification of the RNAs bound to a given RBP, mass spectrometry has allowed complementary approaches to discover new RBPs and determine all the RBPs binding an RNA. In addition, modification of the *in vivo* methods discussed above to be protein-agnostic allows the identification of all the protein-bound sites in expressed transcripts.

Identification of RBPs proteome wide

Proteomics approaches have been applied to identify a more complete complement of RBPs in eukaryotic genomes, which is important because there are a number of RBPs without canonical RNA-binding domains [91, 92]. In yeast, probing protein microarrays with labeled RNA revealed a number of putative novel RBPs, many which already bear annotations as enzymes [93, 94]. More recently, two studies cross-linked proteins to RNA and applied an oligo (dT) pulldown and mass spectrometry to identify proteins contained in ribonucleoprotein complexes in human cells [35, 95]. Both studies identified ~800 RNA-associated proteins, several of which they validated by PAR-CLIP [95] and a fluorescence-based *in vivo* binding assay [35]. Surprisingly ~250–300 of these 800 do not contain classical RBDs or bear previous functional annotation as RBPs. A wide variety of proteins were identified in these studies, including several enzymes involved in intermediary metabolism. However, in contrast to the previous studies in yeast, metabolic enzymes were overall depleted in the set of RNA-associated proteins identified in both studies. Despite this, further *in vitro* or *in vivo* analyses are required to establish whether or not these novel RBPs simply bind RNA non-specifically as the PAR-CLIP data from many of the RBPs resembles background data derived from FLAG-GFP immunoprecipitations [87].

Application of mass spectrometry to determine proteins targeting a specific RNA

Complementary to the RBP-centric approaches described above, development of mass spectrometric methods for determining the protein complement of diverse mixtures has allowed the direct ascertainment of proteins bound to an RNA sequence. Most approaches have been performed *in vitro*, and involve tethering the target RNA molecule to a solid support by chemical modification of the RNA [96] or using an RNA aptamer to a protein that can then be either immunoprecipitated or attached to a solid support such as streptavidin [97] and incubating it with cellular lysate. Purification of *in vivo* assembled RNPs has been accomplished using oligonucleotides complementary to the RNA sequence [98], coexpression of MS2 coat protein and the RNA sequence harbouring an MS2 aptamer [99], and delivery of a peptide nucleic acid (PNA) complementary to the target RNA sequence and bonded to a photoactivatable reagent that will form cross-links between the PNA and nearby proteins [100]. Although quantitative mass spectrometry-based methods are less well developed than high-throughput sequencing approaches, these techniques have been used to identify RBP regulators of a telomere-associated noncoding RNA [101] and to classify a conserved RNA secondary structure predicted using a covariation approach as a putative internal ribosome entry site [97], demonstrating their value as assays to probe the functional significance of predicted RNA motifs. Interestingly, a SILAC-based mass spectrometry approach has been used to investigate binding sites determined using PAR-CLIP and confirmed many

of the binding sites, a direct demonstration of the complementarity of the approaches [102].

Protein occupancy profiling

Several methods have been developed for discovering protein-bound sites on RNAs agnostic of the RBP, denoted protein occupancy (or interaction) profiling. These methods function similarly to PAR-CLIP, except that instead of immunoprecipitation of a target RBP, RNPs are purified using oligo (dT) beads [95] or chemical biotinylation of proteins [103]. High-throughput sequencing identifies bound sites. Silverman and colleagues applied a slightly different approach: cross-linking with formaldehyde, digesting RNA using RNases, reversing the cross-links and sequencing the resulting RNA [104]. This method uncovered fewer binding sites overall but was not limited to processed mRNAs and so observed sites on introns and non-polyadenylated RNAs. The method of Baltz and colleagues has also been applied to MCF7 cells [105], allowing observation of overall differences in binding site occupancy by RBPs between cell types. Interestingly, binding sites with increased occupancy in MCF7 cells contained predicted binding sites for the ELAV family of ARE-binding proteins, regulators of mRNA stability that have been previously implicated with carcinogenesis and poor prognosis [106, 107], and mRNAs with differentially occupied sites had longer half-lives in MCF7 cells, suggesting a widespread role for this protein family in post-transcriptional regulation in cancer cells [105].

Computational methods for examining protein-RNA interactions

The goal of complete understanding of the functions of RBPs in post-transcriptional regulation requires computational analysis of RBP-RNA interactions to interpret experimental data and model how RBPs find and bind to their targets. A general strategy for predicting RBP binding involves (1) motif finding and (2) prediction of binding sites using the discovered motifs. Online databases collecting RBP motifs and *in vivo* data will also be described.

Motif finding

Learning RBP binding motifs can be accomplished by applying DNA motif finders, which only consider the RNA sequence, or by considering RNA secondary structure either explicitly or as a layer on top of sequence preference. The basic approaches are summarized in Table 1. The DNA-based methods MEME, PhyloGibbs and cERMIT have been used to identify motifs from RIP-chip (both raw, and filtered to include only sequences with enriched hexamers), CLIP-seq and PAR-CLIP data [64, 84, 108–110]. Despite ignoring secondary structure, these methods are often successful, presumably because many RBPs fundamentally bind short ssRNA sequences (5–10 nt), often without variable gaps between bound segments that can confound standard position weight matrix (PWM) based methods.

Specialized methods that incorporate RNA secondary structure generally break down into two camps: the first is based on determining the linear structural context around a sequence motif. MEMERIS is an extension of the popular MEME algorithm that uses RNA accessibility as a prior probability to guide motif finding to single-stranded regions [122]. Similarly, Li and colleagues applied accessibility to select motifs that best distinguish bound and unbound transcripts [121]. MatrixREDUCE input is filtered to include only possible hairpin loop structures as part of StructRED [115]. Finally, RNAcontext models the probability that each base in a motif is in a particular secondary structure context (e.g. a hairpin loop) and learns the weights for each position from a set of input sequences annotated with relative binding affinity [117]. RNAcontext is also available on the RBPmotif Web server [123].

The second approach considers RNA secondary structure explicitly and includes stochastic context-free grammar (SCFG) and graphical approaches. CMfinder [119] and RNApromo [120] both start with a set of structures predicted using thermodynamic methods. RNApromo was used to predict motifs in sets of RNAs bound by RBPs in yeast [120]. Maticzka and colleagues developed the graph kernel-based GraphProt, and applied it to learn motifs from CLIP-seq data [118], producing motifs that were highly predictive of binding: the certainty of predicted motifs for PTB correlated with measured RBP affinity.

The fact that many RBPs have multiple RNA binding domains raises the interesting possibility of RBPs binding to bipartite or complex motifs. Gapped motifs have been described for PTB (4 RRM domains, [124]) and the STAR family of RBPs (1 KH domain, but the proteins bind as a dimer, [125]). Motif finding algorithms that

Table 1: Motif-finding algorithms used for analyzing RBP-RNA interaction data

Algorithm	Input	Type of motif generated	Considers secondary structure?	Reference
MEME	Positive (and optionally, negative) sequences	PWM	No	[111]
PhyloGibbs	Positive (and optionally, negative) sequences	PWM	No	[112]
REFINE	Positive sequences	N/A, Filtering procedure to only consider sequences containing three enriched hexamers; filtered sequences are then submitted to another motif finding algorithm	No	[64]
cERMIT	Rank ordered sequences	PWM	No	[113]
DRIMUST	Rank ordered sequences	IUPAC motif, possibly gapped	No	[114]
StructuRED	Positive and negative sequences	PWM in a hairpin loop	Yes, considers possible hairpin loops up to 7 bases with at least 3 paired bases	[115]
TEISER	Sequences and scores (e.g., stability scores)	PWM in a hairpin loop	Yes, considers possible hairpin loops with stems 4-7 bases long and loop sizes of 4-9 bases	[116]
RNAcontext	Sequences and affinity scores	PWM with structural context scores	Yes, learns the preferred structural context of each base in a motif	[117]
GraphProt	Positive and negative sequences	graph-based sequence and structure motifs, can be visualized with logos	Yes, models RNA structure using a graph-based encoding	[118]
CMfinder	Positive sequences	structured sequence	Yes, SCFG-based, examines the most stable structures in the input	[119]
RNApromo	Positive sequences	structured sequence	Yes, SCFG-based, optimizes a motif from an initial set of substructures generated from the input	[120]
#ATS	Positive and negative sequences	IUPAC	Yes, scores candidate binding sites by accessibility	[121]
MEMERIS	Positive and negative sequences	PWM	Yes, uses accessibility as prior knowledge to guide motif finding toward single-stranded regions	[122]

incorporate gapped positions have been developed [126] but have not been extensively applied to RNA–protein interaction data. Exceptions include Leibovich and Yakhini, who applied the DRIMUST algorithm to a RIP-chip data set of yeast PUM-HD proteins, several of which displayed gapped motifs [114], and hidden Markov model-based approaches for detecting clustered binding sites in PTB [127] and Nova/Mbnl [128] data.

Databases

Online databases that collect RNA–protein interactions are summarized in Table 2. The RBP DataBase (RBPDB) stores low- and high-throughput experimental evidence of RNA-binding for metazoan RBPs [14]. The Catalogue of Inferred Sequence Binding Preference of RNA binding proteins (CISBP-RNA) focuses on sequence motifs, and includes inference of motifs for RBPs homologous to a studied protein [68]. *In vivo* RBP binding sites are catalogued in three databases, starBase [129], doRiNA [130], and CLIPz [131]. These databases integrate high-throughput CLIP, PAR-CLIP, and iCLIP data with other data such as miRNA binding sites, and offer tools for online analysis.

OUTLOOK

High-throughput identification of protein–RNA interactions has improved understanding of the targets of RBPs in diverse cellular contexts, and focus is

Table 2: Databases that collect RNA-protein interactions

Database	URL	Features	Reference
RBPDB	http://rbpdb.ccbr.utoronto.ca/	Direct observations of protein-RNA interactions in metazoans, both low- and high-throughput	[14]
CISBP-RNA	http://cisbp-rna.ccbr.utoronto.ca/	Directly observed and predicted (by homology with known proteins) motifs. Tools for scanning sequences and comparing motifs	[68]
starBase	http://starbase.sysu.edu.cn/	RBP-RNA and miRNA-RNA interactions from CLIP data	[129]
doRiNa	http://dorina.mdc-berlin.de/	mRNA-centric or RBP-centric search of CLIP data including combinatorial search	[130]
CLIPz	http://www.clipz.unibas.ch/	Storage and analysis (mapping reads, extracting clusters, mapping T→C conversions) of CLIP data	[131]

shifting from cataloging RNA–protein interactions to understanding the individual and combined effects of RBPs on global RNA metabolism and gene expression. Methods to produce extensive functional data sets of RNA splicing [132], stability [133] and translation [134] have been developed, and more data are likely on the way. High-throughput methods to assay the effect of sequence features in 3′ and 5′ UTRs on RNA expression levels will help generate training datasets to learn subtle features of complex or combinatorial regulation [135–137]. Computational analysis and modeling will undoubtedly play a primary role in understanding RNA biology in the near future.

Key points

- RBPs apply different strategies for binding RNA.
- For certain RBPs, RNA secondary structure is a key component of RNA binding, but widespread prediction and measurement of RNA secondary structure remains difficult especially due to the potential impact of other RBPs on mRNA structure.
- *In vitro* binding specificity of RBPs can be determined using microarray-based methods (such as RNAcompete) or high-throughput sequencing-based methods (such as RNA Bind-n-Seq).
- *In vivo* targets of RBPs are determined using CLIP-seq/HITS-clip, and precise binding sites are more easily defined using the PAR-CLIP and iCLIP variants.
- A general strategy for predicting RBP binding involves (1) motif finding and (2) prediction of binding sites using the discovered motifs.

FUNDING

K.B.C. is supported by an NSERC Alexander Graham Bell Canada Graduate Scholarship. This work was funded by Canadian Institute for Health Research operating grant to Q.D.M. and T.R.H. MOP-125894.

References

1. Glisovic T, Bachorik JL, Yong J, *et al*. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 2008;**582**:1977–86.
2. Licatalosi DD, Mele A, Fak JJ, *et al*. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 2008;**456**:464–9.
3. Wang ET, Cody NAL, Jog S, *et al*. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell* 2012;**150**:710–24.
4. Sawicka K, Bushell M, Spriggs KA, *et al*. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem Soc Trans* 2008;**36**:641–7.
5. Markus MA, Morris BJ. RBM4: a multifunctional RNA-binding protein. *Int J Biochem Cell Biol* 2009;**41**:740–3.
6. Lukong KE, Chang K, Khandjian EW, *et al*. RNA-binding proteins in human genetic disease. *Trends Genet* 2008;**24**: 416–25.
7. Auweter SD, Oberstrass FC, Allain FHT. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res* 2006;**34**:4943–59.
8. Aviv T, Lin Z, Ben-Ari G, *et al*. Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. *Nat Struct Mol Biol* 2006;**13**:168–76.
9. Battle DJ, Doudna JA. The stem-loop binding protein forms a highly stable and specific complex with the 3′ stem-loop of histone mRNAs. *RNA* 2001;**7**:123–32.
10. Adam SA, Nakagawa T, Swanson MS, *et al*. mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. *Mol Cell Biol* 1986;**6**:2932–43.
11. Bandziulis RJ, Swanson MS, Dreyfuss G. RNA-binding proteins as developmental regulators. *Genes Dev* 1989;**3**: 431–7.
12. Maris C, Dominguez C, Allain FH-T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* 2005;**272**: 2118–31.
13. Daubner GM, Cléry A, Allain FH-T. RRM-RNA recognition: NMR or crystallography...and new findings. *Curr Opin Struct Biol* 2013;**23**:100–8.
14. Cook KB, Kazan H, Zuberi K, *et al*. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* 2011;**39**: D301–8.

15. Deo RC, Bonanno JB, Sonenberg N, *et al*. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* 1999;**98**:835–45.
16. Handa N, Nureki O, Kurimoto K, *et al*. Structural basis for recognition of the TRA mRNA precursor by the Sex-lethal protein. *Nature* 1999;**398**:579–85.
17. Oberstrass FC, Auweter SD, Erat M, *et al*. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* 2005;**309**:2054–7.
18. Allain FH, Howe PW, Neuhaus D, *et al*. Structural basis of the RNA-binding specificity of human U1A protein. *EMBO J* 1997;**16**:5764–72.
19. Siomi H, Matunis MJ, Michael WM, *et al*. The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. *Nucleic Acids Res* 1993;**21**:1193–8.
20. Grishin NV. KH domain: one motif, two folds. *Nucleic Acids Res* 2001;**29**:638–43.
21. Valverde R, Edwards L, Regan L. Structure and function of KH domains. *FEBS J* 2008;**275**:2712–26.
22. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* 2010;**79**:321–49.
23. Gregory RI, Yan K-P, Amuthan G, *et al*. The Microprocessor complex mediates the genesis of microRNAs. *Nature* 2004;**432**:235–40.
24. Ramos A, Grünert S, Adams J, *et al*. RNA recognition by a Staufen double-stranded RNA-binding domain. *EMBO J* 2000;**19**:997–1009.
25. Laver JD, Li X, Ancevicius K, *et al*. Genome-wide analysis of Staufen-associated mRNAs identifies secondary structures that confer target specificity. *Nucleic Acids Res* 2013; **41**:9438–60.
26. Krovat BC, Jantsch MF. Comparative mutational analysis of the double-stranded RNA binding domains of Xenopus laevis RNA-binding protein A. *J Biol Chem* 1996;**271**: 28112–9.
27. Masliah G, Barraud P, Allain FH-T. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cell Mol Life Sci* 2013;**70**:1875–95.
28. A Bloomfield VM, Crothers D, Tinoco I. *Nucleic Acids: Structures, Properties, and Functions*. Sausalito, California: University Science Books, 2000.
29. Stefl R, Oberstrass FC, Hood JL, *et al*. The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove. *Cell* 2010;**143**: 225–37.
30. Chen Y, Varani G. Finding the missing code of RNA recognition by PUF proteins. *Chem Biol* 2011;**18**: 821–3.
31. Cheong C-G, Hall TMT. Engineering RNA sequence specificity of Pumilio repeats. *Proc Natl Acad Sci USA* 2006;**103**: 13635–9.
32. Dong S, Wang Y, Cassidy-Amstutz C, *et al*. Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. *J Biol Chem* 2011;**286**: 26732–42.
33. Filipovska A, Razif MFM, Nygård KKA, *et al*. A universal code for RNA recognition by PUF proteins. *Nat Chem Biol* 2011;**7**:425–7.
34. Cavaloc Y, Bourgeois CF, Kister L, *et al*. The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA* 1999;**5**:468–83.
35. Castello A, Fischer B, Eichelbaum K, *et al*. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 2012;**149**:1393–406.
36. Hall TMT. Multiple modes of RNA recognition by zinc finger proteins. *Curr Opin Struct Biol* 2005;**15**:367–73.
37. Ziehler WA, Engelke DR. Probing RNA structure with chemical reagents and enzymes. *Curr Protoc Nucleic Acid Chem* 2001 Chapter 6:Unit 6.1.
38. Underwood JG, Uzilov AV, Katzman S, *et al*. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods* 2010;**7**:995–1001.
39. Kertesz M, Wan Y, Mazor E, *et al*. Genome-wide measurement of RNA secondary structure in yeast. *Nature* 2010; **467**:103–7.
40. Ding Y, Tang Y, Kwok CK, *et al*. *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 2014;**505**:696–700.
41. Rouskin S, Zubradt M, Washietl S, *et al*. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature* 2014;**505**:701–5.
42. Wan Y, Qu K, Zhang QC, *et al*. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 2014;**505**:706–9.
43. Shabalina SA, Ogurtsov AY, Spiridonov NA. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res* 2006;**34**:2428–37.
44. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003;**31**: 3406–15.
45. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 1990;**29**:1105–19.
46. Wuchty S, Fontana W, Hofacker IL, *et al*. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 1999;**49**:145–65.
47. Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* 2003; **31**:7280–301.
48. Hofacker IL, Fontana W, Stadler PF, *et al*. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Mon* 1994;**125**:167–88.
49. Bellaousov S, Reuter JS, Seetin MG, *et al*. RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res* 2013;**41**:W471–4.
50. Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 2004;**5**:140.
51. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math* 1985;**45**:810–25.
52. Griffiths-Jones S. Annotating noncoding RNA genes. *Annu Rev Genomics Hum Genet* 2007;**8**:279–98.
53. Schwartz A, Myers E, Pachter L. Alignment metric accuracy. *arXiv Prepr q-bio/0510052* 2005.
54. Babak T, Blencowe BJ, Hughes TR. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics* 2007;**8**:33.
55. Eddy SR. *User's Guide for COVE—Covariance Models of RNA Sequence Families*. Available at: ftp://selab.janelia.org/pub/software/tRNAscan-SE/tRNAscan-SE-1.23/Cove/Guide.tex.

56. Bernhart SH, Hofacker IL, Stadler PF. Local RNA base pairing probabilities in large sequences. *Bioinformatics* 2006; **22**:614–5.
57. Lange SJ, Maticzka D, Möhl M, *et al*. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res* 2012;**40**:5215–26.
58. Tafer H, Ameres SL, Obernosterer G, *et al*. The impact of target site accessibility on the design of effective siRNAs. *Nat Biotechnol* 2008;**26**:578–83.
59. Lu ZJ, Mathews DH. Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res* 2008;**36**: 640–7.
60. Li X, Kazan H, Lipshitz HD, *et al*. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip Rev RNA* 2014;**5**: 111–30.
61. Ouyang Z, Snyder MP, Chang HY. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res* 2013;**23**: 377–87.
62. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. *Nature* 1990;**346**:818–22.
63. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 1990;**249**:505–10.
64. Riordan DP, Herschlag D, Brown PO. Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome. *Nucleic Acids Res* 2011;**39**:1501–9.
65. Campbell ZT, Bhimsaria D, Valley CT, *et al*. Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell Rep* 2012;**1**:570–81.
66. Jolma A, Kivioja T, Toivonen J, *et al*. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* 2010;**20**: 861–73.
67. Ray D, Kazan H, Chan ET, *et al*. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* 2009;**27**:667–70.
68. Ray D, Kazan H, Cook KB, *et al*. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 2013; **499**:172–7.
69. Lambert N, Robertson A, Jangi M, *et al*. RNA bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell* 2014;**1**:1–14.
70. Buenrostro JD, Araya CL, Chircus LM, *et al*. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat Biotechnol* 2014;**32**:562–8.
71. Tenenbaum SA, Carson CC, Lager PJ, *et al*. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci USA* 2000;**97**: 14085–90.
72. Morris AR, Mukherjee N, Keene JD. Systematic analysis of posttranscriptional gene expression. *Wiley Interdiscip Rev Syst Biol Med* 2010;**2**:162–80.
73. Hogan DJ, Riordan DP, Gerber AP, *et al*. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol* 2008;**6**:e255.
74. Mili S, Steitz JA. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA* 2004;**10**:1692–4.
75. Penalva LOF, Burdick MD, Lin SM, *et al*. RNA-binding proteins to assess gene expression states of co-cultivated cells in response to tumor cells. *Mol Cancer* 2004;**3**:24.
76. Ule J, Jensen KB, Ruggiu M, *et al*. CLIP identifies Nova-regulated RNA networks in the brain. *Science* 2003;**302**: 1212–5.
77. Hockensmith JW, Kubasek WL, Vorachek WR, *et al*. Laser cross-linking of nucleic acids to proteins. Methodology and first applications to the phage T4 DNA replication system. *J Biol Chem* 1986;**261**:3512–8.
78. Zhang C, Darnell RB. Mapping *in vivo* protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol* 2011;**29**:607–14.
79. König J, Zarnack K, Rot G, *et al*. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* 2010;**17**:909–15.
80. Ule J, Jensen K, Mele A, *et al*. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* 2005;**37**:376–86.
81. Lu Z, Guan X, Schmidt CA, *et al*. RIP-seq analysis of eukaryotic Sm proteins identifies three major categories of Sm-containing ribonucleoproteins. *Genome Biol* 2014; **15**:R7.
82. Klass DM, Scheibe M, Butter F, *et al*. Quantitative proteomic analysis reveals concurrent RNA-protein interactions and identifies new RNA-binding proteins in *Saccharomyces cerevisiae*. *Genome Res* 2013;**23**:1028–38.
83. Kishore S, Jaskiewicz L, Burger L, *et al*. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods* 2011;**8**:559–64.
84. Brümmer A, Kishore S, Subasic D, *et al*. Modeling the binding specificity of the RNA-binding protein GLD-1 suggests a function of coding region-located sites in translational repression. *RNA* 2013;**19**:1317–26.
85. Zisoulis DG, Lovci MT, Wilbert ML, *et al*. Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nat Struct Mol Biol* 2010;**17**:173–9.
86. Jungkamp A-C, Stoeckius M, Mecenas D, *et al*. *In vivo* and transcriptome-wide identification of RNA binding protein target sites. *Mol Cell* 2011;**44**:828–40.
87. Friedersdorf MB, Keene JD. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol* 2014;**15**:R2.
88. Uniacke J, Holterman CE, Lachance G, *et al*. An oxygen-regulated switch in the protein synthesis machinery. *EMBO J* 2012;**486**:126–9.
89. Corcoran DL, Georgiev S, Mukherjee N, *et al*. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol* 2011;**12**:R79.
90. Uren PJ, Bahrami-Samani E, Burns SC, *et al*. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* 2012;**28**:3013–20.
91. Pandey NB, Sun JH, Marzluff WF. Different complexes are formed on the 3′ end of histone mRNA with nuclear and polyribosomal proteins. *Nucleic Acids Res* 1991;**19**: 5653–9.
92. Smibert CA, Wilson JE, Kerr K, *et al*. smaug protein represses translation of unlocalized nanos mRNA in the *Drosophila* embryo. *Genes Dev* 1996;**10**:2600–9.

93. Tsvetanova NG, Klass DM, Salzman J, *et al*. Proteome-wide search reveals unexpected RNA-binding proteins in *Saccharomyces cerevisiae*. *PLoS One* 2010;**5**:pii: e12671.
94. Scherrer T, Mittal N, Janga SC, *et al*. A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. *PLoS One* 2010;**5**:e15499.
95. Baltz AG, Munschauer M, Schwanhäusser B, *et al*. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell* 2012;**46**:674–90.
96. Miniard AC, Middleton LM, Budiman ME, *et al*. Nucleolin binds to a subset of selenoprotein mRNAs and regulates their expression. *Nucleic Acids Res* 2010;**38**: 4807–20.
97. Butter F, Scheibe M, Mörl M, *et al*. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc Natl Acad Sci USA* 2009;**106**:10626–31.
98. Soeno Y, Taya Y, Stasyk T, *et al*. Identification of novel ribonucleo-protein complexes from the brain-specific snoRNA MBII-52. *RNA* 2010;**16**:1293–300.
99. Tsai BP, Wang X, Huang L, *et al*. Quantitative profiling of *in vivo*-assembled RNA-protein complexes using a novel integrated proteomic approach. *Mol Cell Proteomics* 2011;**10**: M110.007385.
100. Zielinski J, Kilk K, Peritz T, *et al*. *In vivo* identification of ribonucleoprotein-RNA interactions. *Proc Natl Acad Sci USA* 2006;**103**:1557–62.
101. Scheibe M, Arnoult N, Kappei D, *et al*. Quantitative interaction screen of telomeric repeat-containing RNA reveals novel TERRA regulators. *Genome Res* 2013;**23**: 2149–57.
102. Scheibe M, Butter F, Hafner M, *et al*. Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions. *Nucleic Acids Res* 2012;**40**:9897–902.
103. Freeberg MA, Han T, Moresco JJ, *et al*. Pervasive and dynamic protein binding sites of the mRNA transcriptome in *Saccharomyces cerevisiae*. *Genome Biol* 2013;**14**:R13.
104. Silverman IM, Li F, Alexander A, *et al*. RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biol* 2014; **15**:R3.
105. Schueler M, Munschauer M, Gregersen LH, *et al*. Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biol* 2014;**15**:R15.
106. López de Silanes I, Fan J, Yang X, *et al*. Role of the RNA-binding protein HuR in colon carcinogenesis. *Oncogene* 2003;**22**:7146–54.
107. Heinonen M, Bono P, Narko K, *et al*. Cytoplasmic HuR expression is a prognostic factor in invasive ductal breast carcinoma. *Cancer Res* 2005;**65**:2157–61.
108. Gerber AP, Luschnig S, Krasnow MA, *et al*. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 2006;**103**:4487–92.
109. Wright JE, Gaidatzis D, Senften M, *et al*. A quantitative RNA code for mRNA target selection by the Germline fate determinant GLD-1. *EMBO J* 2010;**30**: 533–45.
110. Mukherjee N, Corcoran DL, Nusbaum JD, *et al*. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol Cell* 2011;**43**:327–39.
111. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994;**2**:28–36.
112. Siddharthan R, Siggia ED, van Nimwegen E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 2005;**1**:e67.
113. Georgiev S, Boyle AP, Jayasurya K, *et al*. Evidence-ranked motif identification. *Genome Biol* 2010;**11**:R19.
114. Leibovich L, Yakhini Z. Efficient motif search in ranked lists and applications to variable gap motifs. *Nucleic Acids Res* 2012;**40**:5832–47.
115. Foat BC, Stormo GD. Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. *Mol Syst Biol* 2009;**5**:268.
116. Goodarzi H, Najafabadi HS, Oikonomou P, *et al*. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* 2012; **485**:264–8.
117. Kazan H, Ray D, Chan ET, *et al*. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol* 2010;**6**:e1000832.
118. Maticzka D, Lange SJ, Costa F, *et al*. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol* 2014;**15**:R17.
119. Yao Z, Weinberg Z, Ruzzo WL. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* 2006;**22**:445–52.
120. Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci USA* 2008;**105**: 14885–90.
121. Li X, Quon G, Lipshitz HD, *et al*. Predicting *in vivo* binding sites of RNA-binding proteins using mRNA secondary structure. *RNA* 2010;**16**:1096–107.
122. Hiller M, Pudimat R, Busch A, *et al*. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res* 2006; **34**:e117.
123. Kazan H, Morris Q. RBPmotif: a web server for the discovery of sequence and structure preferences of RNA-binding proteins. *Nucleic Acids Res* 2013;**41**:W180–6.
124. Lamichhane R, Daubner GM, Thomas-Crusells J, *et al*. RNA looping by PTB: Evidence using FRET and NMR spectroscopy for a role in splicing repression. *Proc Natl Acad Sci USA* 2010;**107**:4105–10.
125. Galarneau A, Richard S. The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs. *BMC Mol Biol* 2009;**10**:47.
126. Frith MC, Saunders NFW, Kobe B, *et al*. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol* 2008;**4**:e1000071.
127. Han A, Stoilov P, Linares AJ, *et al*. De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. *PLoS Comput Biol* 2014;**10**:e1003442.
128. Zhang C, Lee KY, Swanson MS, *et al*. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Res* 2013;**41**:6793–807.
129. Li J-H, Liu S, Zhou H, *et al*. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA

interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2014;**42**:D92–7.

130. Anders G, Mackowiak SD, Jens M, *et al*. doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* 2012;**40**:D180–6.
131. Khorshid M, Rodak C, Zavolan M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res* 2011; **39**:D245–52.
132. Witten JT, Ule J. Understanding splicing regulation through RNA splicing maps. *Trends Genet* 2011;**27**:89–97.
133. Geisberg J V, Moqtaderi Z, Fan X, *et al*. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* 2014;**156**:812–24.
134. Ingolia NT, Ghaemmaghami S, Newman JRS, *et al*. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* 2009;**324**: 218–23.
135. Shalem O, Carey L, Zeevi D, *et al*. Measurements of the impact of 3′ end sequences on gene expression reveal wide range and sequence dependent effects. *PLoS Comput Biol* 2013;**9**:e1002934.
136. Dvir S, Velten L, Sharon E, *et al*. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci USA* 2013;**110**:E2792–801.
137. Zhao W, Pollack JL, Blagev DP, *et al*. Massively parallel functional annotation of 3′ untranslated regions. *Nat Biotechnol* 2014;**32**:387–91.