Gerstberger *et al. Genome Biology* 2013, **14**:130
http://genomebiology.com/2013/14/8/130

Genome Biology

RESEARCH HIGHLIGHT

Learning the language of post-transcriptional gene regulation

Stefanie Gerstberger†, Markus Hafner† and Thomas Tuschl*

Abstract

A large-scale RNA *in vitro* selection study systematically identified RNA recognition elements for 205 RNA-binding proteins belonging to families conserved in most eukaryotes.

The challenges of characterizing post-transcriptional gene regulation

Messenger RNAs (mRNAs) are regulated at every stage of their life cycle. All cellular RNA, including mRNA, is packaged into distinct ribonucleoprotein (RNP) complexes to orchestrate RNA maturation and turnover processes summarized as post-transcriptional gene regulation. The most relevant processes involving mRNAs include pre-mRNA splicing, 5' and 3' end modification, editing, transport, translation and degradation. Among the challenges for decoding post-transcriptional gene regulation is the elucidation of the mRNP composition, which changes as mRNAs mature or are translated. This is a prerequisite for understanding the consequences of dysregulation and/or mutation of RNA-binding proteins (RBPs) and/or their target RNA-binding sites in disease.

The human genome encodes 1,500 RBPs, and 600 microRNAs targeting mRNAs [1]. Most RBPs are composed of at least one, but frequently also combinations of multiple distinct RNA-binding domains (RBDs). At least 800 distinct RBDs are known [2]; among the most frequent in humans are the single-stranded-RNA-binding RRM, KH, zf-CCCH and zf-CCHC domains, and the double-stranded-RNA-binding DSRM domain. Recent proteomic analysis consolidated the number of mRBPs to 700 proteins and revealed at least 20 previously unknown RBDs [1,3].

Following or coinciding with the determination of the composition of mRNPs is the identification of the precise binding site(s) located within the mRNA targets of RBPs and the derivation of the underlying RNA recognition element(s) (RRE(s)). This task is non-trivial considering that RBDs generally recognize short and degenerate sequences of three to eight nucleotides, sometimes involving additional RNA secondary structure. In addition, *in vivo* binding is modulated by competition with other RBPs for the same or adjacent sites [1]. Since the implementation of high-throughput methods in RNA biology, various protocols for the experimental identification of RBP binding sites have been developed. A recent study by Ray *et al.* [4] used a single-cycle RNA *in vitro* selection approach to characterize the binding specificities for 205 recombinant RBPs and, in doing so, has brought us an important step closer to solving the post-transcriptional RBP regulatory code.

Experimental methods for determining RREs

RREs are traditionally determined by sequence comparison and/or conservational analysis from known RNA targets, and validated by biochemical interaction analysis (such as electrophoretic mobility shift assays, filter binding or surface plasmon resonance). For RBPs with unknown RNA-binding sites, *in vitro* evolutionary methods (primarily SELEX) that identify high-affinity RNA ligands within pools of randomized sequences have been employed with some success [5]. The RRE is then derived by comparing multiple independently sequenced RNA ligands. Alternatively, various crosslinking and immunoprecipitation (CLIP) methods have been introduced that rely on covalent crosslinking of an RBP to its RNA targets in live cells, followed by the isolation of crosslinked RBP-RNA segments (Figure 1) [1,6]. Coupled with deep sequencing of the crosslinked RNAs, CLIP methods allow for the comprehensive determination of *in vivo* RNA target sites and their underlying RRE. Until recently, knowledge of RREs was rather scant and experimental binding data from SELEX, CLIP and other methods were available for less than 10% of the known RBPs in humans [1,6,7].

To increase throughput and identify the highest-affinity RREs, Ray *et al.* [4,8] introduced a SELEX method termed RNAcompete (Figure 1). In contrast to random sequence

†Equal contributors
*Correspondence: ttuschl@rockefeller.edu
Howard Hughes Medical Institute and Laboratory for RNA Molecular Biology, The Rockefeller University, 1230 York Ave, New York, NY 10065, USA

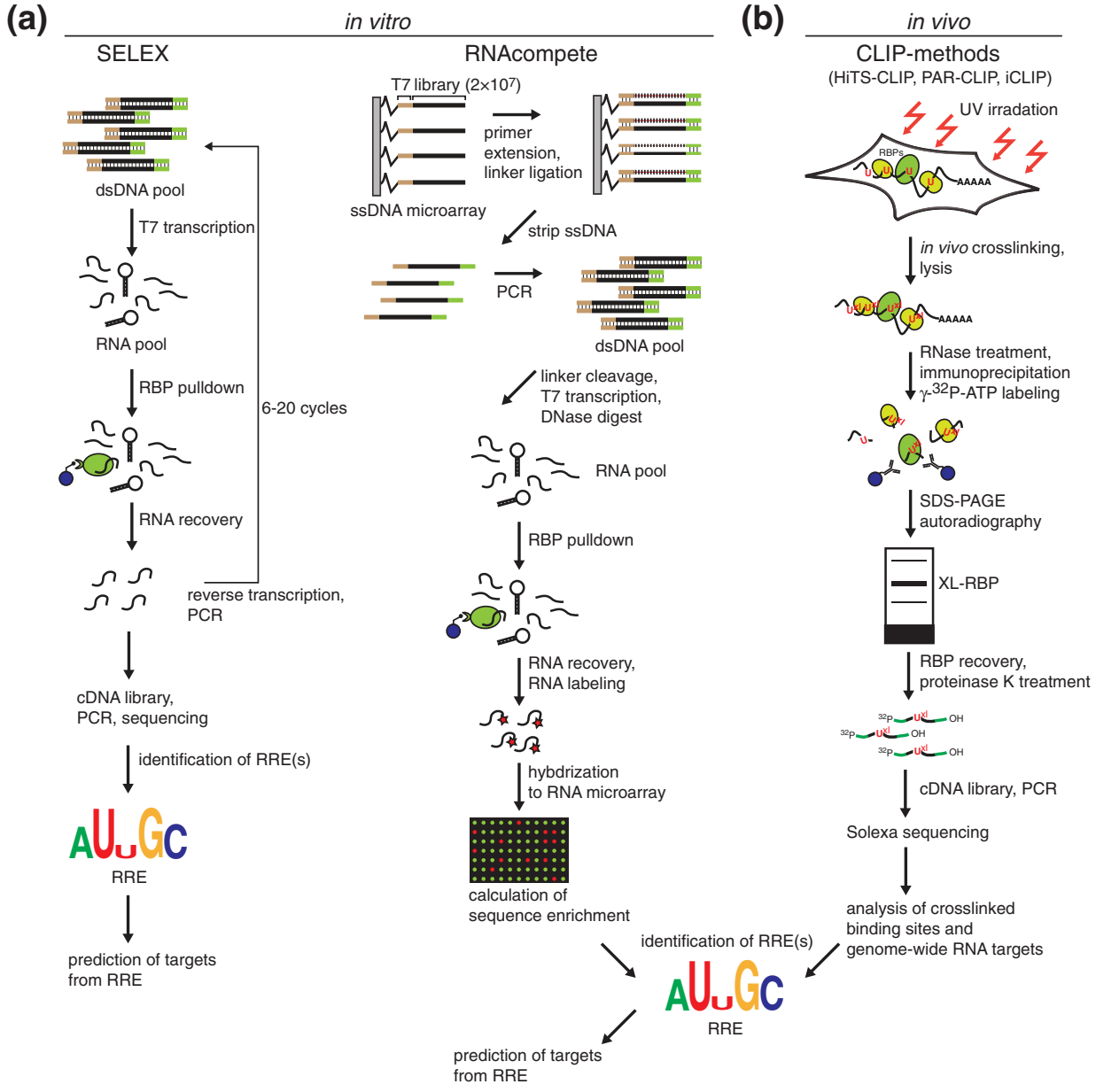

Figure 1. Overview of *in vitro* and *in vivo* methods for RRE determination. (a) SELEX and RNAcompete start with the preparation of a diverse DNA sequence pool, which is *in vitro* transcribed into RNA. The protein of interest is incubated with the random sequence RNA pool, followed by RBP pulldown and recovery of the bound RNA. In SELEX, high-affinity ligands are enriched by several rounds of reverse transcription, (mutagenic) PCR and selection, before sequencing of the RNA ligands. In RNAcompete, the recovered RNA is directly quantified on a microarray, rather than sequenced, and enrichment for each individual sequence over the initial pool is calculated. Enrichment scores, which directly correlate with the binding affinity of the RNA sequence, are used to derive the RRE, which serve as input for computational prediction of *in vivo* RNA targets. **(b)** CLIP-based methods use *in vivo* crosslinking to covalently link RBPs to their RNA targets by UV light. After cell lysis, limited RNase treatment and immunoprecipitation of the RBP, the crosslinked RNA segments are recovered, converted into cDNA libraries and deep sequenced. CLIP methods directly identify *in vivo* RNA targets and binding sites, and motif finding algorithms are used to deduce the RRE from the crosslinked RNA sequences. dsDNA, double-stranded DNA; HITS-CLIP, high-throughput sequencing of RNA isolated by CLIP; iCLIP, individual-nucleotide resolution CLIP; PAR-CLIP, photoactivatable-ribonucleoside-enhanced CLIP; ssDNA, single-stranded DNA; XL-RBP, crosslinked RBP.

pools used in SELEX, which contain up to 10^{14} different molecules of 20- to 80-nucleotide random sequence flanked by constant primer binding sites, RNAcompete pools were designed to contain only 240,000 different sequences of 30 to 40 nucleotides in length. These RNA sequences were predicted to be only weakly structured,

with each possible 9-mer represented at least 16 times in the RNAcompete sequence pool. To prepare this RNA sequence pool, oligodeoxynucleotides printed on a microarray were amplified, transcribed into RNA, and subsequently incubated with a recombinantly expressed, affinity-tagged RBP of interest. The RNA pool was then incubated with 75-fold molar excess over protein to ensure efficient competition between the various sequences during binding, so that at equilibrium the proportion of each sequence bound to the RBP reflected its affinity. The incubated protein was recovered and the enrichment of bound RNAs over the initial pool RNA was quantified on microarrays. In contrast to SELEX, the bound RNA was directly analyzed after the first competitive binding reaction without further cycles of amplification and mutagenesis. The RRE for the protein was inferred by combining the calculated Z- and E-values for each possible 7-mer.

Evolutionary insights and global patterns in protein-RNA sequence recognition

In their recent study, Ray *et al.* [4] applied RNAcompete to determine RREs for a collection of 205 different RBPs distributed across 24 species and representing approximately 60 conserved families of RBPs. The parallel processing of samples using a single method facilitated comparison of the RREs and specificities of various RBPs. Most RBPs were expressed in truncated forms comprising all constituent RBD(s) with 30 to 50 flanking amino acid residues to enhance solubility. The selected RBPs contained at least one of nine well-characterized RBDs (RRM, KH, S1, YTH, Pumilio repeats (PUF), zf-CCCH, zf-CCHC, zf-RanBP and SAM), whereby the majority of RBPs contained multiple RBDs. Approximately 90% of the RBPs tested recognized five to seven nucleotide-long sequence motifs and did not require structured RNA for binding, which is expected based on the inclusion of predominantly single-strand-specific RBDs in this study.

For 52 proteins, RNAcompete RREs were compared with RREs previously determined by CLIP or other methods. Of these, 35 were highly similar, 6 matched partially and 11 were dissimilar to RNAcompete RRE. For example, for PUM1/2 or ELAVL1/HuR the RREs agreed perfectly, while for proteins such as FMR1 only one of two established RREs were identified. The discrepancies may mirror technical differences between the methods or differences between *in vivo* and *in vitro* specificities of RBPs. Enrichment of an RRE by RNAcompete is dependent on affinity, and for multi-RBD proteins affinities of individual RBDs for RNA can vary by orders of magnitude, and contributions of weaker binding RBDs, which can be detected in *in vivo* data, may be potentially overlooked. In addition, *in vivo*, the highest affinity sites may not always be accessible due to competition with other RBPs, the cell-type- and subcellular-compartment-dependent concentration of RBP and RNA targets, modulation of RNA affinities by protein cofactors, and the secondary structure of RNA.

Of importance was the validation of the intuitive notion that RBPs with high sequence identity bind to similar RREs. It was found that RBPs with 70% sequence identity have close to identical RREs, and RBPs with 50% identity share related binding specificities. Based on this notion, the authors predicted RREs for a total of 8,056 RBPs in humans and other metazoans, as well as in plants and protists. Specifically, this number amounted to 159 RBPs in human belonging to 62 protein families, of which approximately 90% were putative or experimentally validated mRNA-binding proteins (mRBPs). Estimating that 700 of the 1,500 RBPs are mRNA-binding, this study elucidated RRE motifs for 20% of all human mRBPs, and 53% of proteins containing canonical single-strand RBDs (Figure 2). The results are available as a public database and represent a valuable resource for researchers interested in prediction of RBP binding sites.

Conservation of motifs and functional implications

RNAcompete-derived RREs demonstrated predictive power for anticipating regulatory functions of RBPs [4]. Evolutionary conservation analysis showed that sequence elements containing these RREs were frequently under positive selection pressure in 5' UTRs, coding regions, 3' UTRs and intronic regions flanking alternative exons. The location of conserved RREs correlated well with previously elucidated RBP binding patterns, with a few surprising twists; for example, conserved RREs for several splicing factors were unexpectedly frequent in the 3' UTR of mRNAs. RNA sequencing experiments from diverse cell lines and tissues with different RBP expression levels allowed correlation of RBP levels with predicted target RNA levels or splicing patterns. This analysis confirmed known RBP functions in some cases (ELAVL1/HuR, RBM4), but also hinted at unanticipated roles for others (PUM1/2, RBFOX1). A study of RNA knockdown data confirmed that RBFOX1, a splicing regulator, also had a positive effect on RNA stability of putative targets with predicted RBFOX1 sites in the 3' UTR, confirming previous reports that some RBPs may have multiple functions in post-transcriptional gene regulation.

Some of the regulatory effects predicted by the evolutionary conservation analysis of RNAcompete RREs, however, are difficult to reconcile with other available data, such as the implied negative effect of the FMR1 protein on target mRNA levels. An effect of FMR1 on RNA abundance was explicitly ruled out in two recent studies, although FMR1 was shown at the same time to

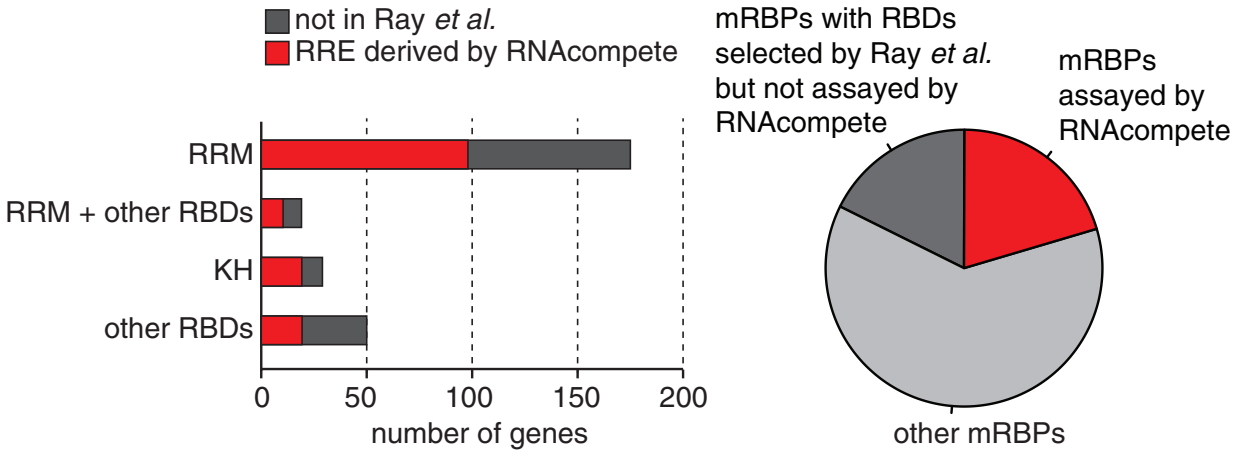

Figure 2. Summary of human mRBPs characterized by RNAcompete. The left panel shows, categorized by selected domain types, the number of RBPs for which RREs were characterized by RNAcompete (red) or remain to be studied (dark gray). RBD categories are RRM, RRM + other (combination of RRM with KH or zinc finger), KH and other RBDs selected in Ray *et al.* (zinc finger, PUF, CSD, YTH, S1 and SAM). The right panel shows the fraction of human mRBPs studied by RNAcompete. Of all known/putative mRBPs (n = 717; reviewed by Ascano *et al.* [1]), 20% were studied by RNAcompete (red, n = 146). Another 18% of mRBPs contain at least one or more of the selected RBDs but remain to be studied (dark gray, n = 127). Sixty-two percent of mRBPs contained RBDs (for example, DSRM, La, PWI, and so on) not studied by RNAcompete (light gray, n = 444).

negatively regulate protein abundance of targeted mRNAs [9,10]. As discussed above, these discrepancies may reflect differences between *in vivo* and *in vitro* preferences of multi-RBD proteins, including FMR1. Analysis of CLIP-derived motifs showed that the FMR1 RG-rich region bound WGGA with higher affinity than its KH domains bound ACUK [9]. The RNAcompete motifs GACAAG and ANGGAC more likely reflected contributions of the RG-rich region to binding. The implicit assumption that the highest-affinity RRE also reflects the optimal *in vivo* RRE may prove inaccurate in some cases, because of varying accessibility of a motif.

From RRE identification to elucidation of post-transcriptional gene regulatory network

The systematic analysis and identification of RREs, together with *in vivo* RNA targets of regulatory proteins, will remain one of the main focuses in post-transcriptional gene regulation research. Ray *et al.* have compiled the largest catalog of experimentally derived RREs at present and this resource may be used to understand evolutionary relationships between RBPs. It also allows researchers to find putative binding sites for RBPs of interest and gives computational biologists the opportunity to integrate RREs as predictors into statistical learning methods to model, in concert with microRNA binding sites, transcription factor recognition elements and epigenetic marks, the transcriptional and post-transcriptional control of gene expression.

To capture the physiological role of RBPs, we still need to dissect the target gene network for each RBP individually in various cellular contexts, and then integrate the knowledge into computational approaches that are able to recapitulate quantitatively the regulatory effects of RBPs. This includes understanding protein and target RNA levels in different cell types and tissues, insights into RRE occupancy, competition among RBPs and accounting for redundancies in protein families or regulatory pathways. Efforts such as SELEX- or CLIP-based methods increase the growing compendium of RREs and contribute to this goal to characterize post-transcriptional regulation in a comprehensive manner.

Abbreviations

CLIP, crosslinking and immunoprecipitation; mRNA, messenger RNA; RBD, RNA-binding domain; RBP, RNA-binding protein; RNP, ribonucleoprotein; SELEX, systematic evolution of ligands by exponential enrichment.

Competing interests

The authors declare that they have no competing interests

Published: 23 August 2013

References

1. Ascano M, Gerstberger S, Tuschl T: **Multi-disciplinary methods to define RNA-protein interactions and regulatory networks.** *Curr Opin Genet Dev* 2013, **23:**20-28.
2. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38:**D211-D222.
3. Sibley CR, Attig J, Ule J: **The greatest catch: big game fishing for mRNA-bound proteins.** *Genome Biol* 2012, **13:**163.
4. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LOF, *et al.*: **A compendium of RNA-binding motifs for decoding gene**

regulation. *Nature* 2013, **499:**172-177.
5. Stoltenburg R, Reinemann C, Strehlitz B: **SELEX - A (r)evolutionary method to generate high-affinity nucleic acid ligands.** *Biomol Eng* 2007, **24:**381-403.
6. Konig J, Zarnack K, Luscombe NM, Ule J: **Protein-RNA interactions: new genomic technologies and perspectives.** *Nat Rev Genet* 2011, **13:**77-83.
7. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR: **RBPDB: a database of RNA-binding specificities.** *Nucleic Acids Res* 2011, **39:**D301-D308.
8. Ray D, Kazan H, Chan ET, Castillo LP, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR: **Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins.** *Nat Biotechnol* 2009, **27:**667-670.
9. Ascano M, Mukherjee N, Bandaru P, Miller JB, Nusbaum JD, Corcoran DL, Langlois C, Munschauer M, Dewell S, Hafner M, Williams Z, Ohler U, Tuschl T: **FMRP targets distinct mRNA sequence elements to regulate protein expression.** *Nature* 2012, **492:**382-386.
10. Darnell JC, Van Driesche SJ, Zhang C, Hung KYS, Mele A, Fraser CE, Stone EF, Chen C, Fak JJ, Chi SW, Licatalosi DD, Richter JD, Darnell RB: **FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism.** *Cell* 2011, **146:**247-261.

doi:10.1186/gb-2013-14-8-130
Cite this article as: Gerstberger S, *et al.*: Learning the language of post-transcriptional gene regulation. *Genome Biology* 2013, **14:**130.