Advances in Experimental Medicine and Biology 825

Gene W. Yeo *Editor*

Systems Biology of RNA Binding Proteins

Springer

Advances in Experimental Medicine and Biology

More information about this series at http://www.springer.com/series/5584

Gene W. Yeo
Editor

Systems Biology of RNA Binding Proteins

Springer

Editor
Gene W. Yeo
UCSD Moores Cancer Center
Institute for Genomic Medicine UCSD Stem Cell Program
La Jolla, CA, USA

ISSN 0065-2598 ISSN 2214-8019 (electronic)
ISBN 978-1-4939-1220-9 ISBN 978-1-4939-1221-6 (eBook)
DOI 10.1007/978-1-4939-1221-6
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014947870

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Ribonucleic acid (RNA)-binding proteins are rapidly being recognized as a major class of approximately a thousand proteins that are widely conserved throughout eukaryotes and play key roles in almost every aspect of RNA metabolism. RNA-binding proteins interact with hundreds to thousands of RNA substrates, including coding transcripts and long and short noncoding RNAs via *cis*-regulatory sequences or guided by associated small RNAs. Defects in the regulation of RNA targets by mutations within the RNA regulatory proteins, RNA *cis*-elements, or changes in protein availability or expression lead to numerous diseases.

With the surge of high-throughput, massively parallel microarray and sequencing technologies in the past few years, there have been tremendous advances in genomics and systems-level approaches and computational methods to probe how RNA-binding proteins affect various aspects of the RNA processing life cycle and uncover key RNA substrates. Advances in this area have increased recognition of the relevance of RNA-binding proteins to neurological disease, heart and muscle abnormalities, germ-line defects, and other genetic diseases. Additional developments include new molecular engineering approaches that utilize RNA-binding proteins to control gene expression, computational models of splicing regulation, and successful therapeutic strategies to modify RNA-binding protein-RNA interactions.

Given the exciting onset of intersection between systems approaches and RNA processing, I felt that it was timely to assemble the first book focused on genome-wide and systems-level perspectives on the diverse roles that RNA-binding proteins play in development and disease. In particular, I wanted to bring to the forefront some of the major questions that are currently under intense investigation by the experts contributing to this book.

The content of this book surveys a wide range of genome-wide and systems approaches to studying RNA-binding proteins in a myriad of organisms, cells, and tissues. More importantly, many chapters illustrate major steps in the processing of RNA and development and diseases caused by defects in these steps. Lastly, many authors discuss open questions, anticipated answers, and potential new areas of research on posttranscriptional gene regulation.

Tuschl and colleagues present a comprehensive analysis of RNA-binding proteins in humans, their evolutionary conservation, structural domains and expression, and a survey of various classes of RNA-binding proteins implicated in human disease.

Lecuyer and Bergalet opine on the importance of and regulatory principles involved in RNA localization and review emerging genomic and imaging technologies that have provided insights into RNA localization and diseases associated with localization defects.

Tian and Zheng review the *cis* elements and *trans* factors involved in cleavage and polyadenylation, emphasizing the importance of alternative cleavage and polyadenylation in gene regulation and the contribution of transcriptome-wide technologies in identifying the diversity of alternative 3′ ends.

Zhao and Chen provide an overview of the exciting new field of long noncoding RNAs and describe the methodologies used to study protein and DNA interactions with long noncoding RNAs during development and in disease states.

Lau and Clark recount the systems-wide approaches that have contributed to our understanding of Piwi proteins and Piwi-interacting RNAs on germ-line genome regulation.

Wang and Choudhury discuss current advances in the molecular engineering of RNA-binding proteins, emphasizing design principles and their applications as therapeutic agents and basic tools in biology.

Fairbrother and colleagues present current practices in medical genetics, the principles behind biochemical binding and functional assays, and advances in scaling up assays in assessing how genetic variation impacts RNA-binding protein interaction.

Carstens and colleagues feature the impact of RNA-binding proteins on key posttranscriptional changes during the epithelial to mesenchymal transition.

Bennett and colleagues provide an inside view of antisense oligonucleotide-based therapies for diseases caused by defects in premature messenger RNA processing.

Swanson and Goodwin emphasize the importance of altering RNA-binding protein functions in microsatellite expansion diseases.

Cooper and Giudice discuss how activities of RNA-binding proteins affect heart development and cardiomyopathy and how these RNA-binding proteins modulate these processes.

Pasquinelli and Azoubel Lima highlight genome-wide methodologies utilized to find microRNAs and their mRNA targets in *C. elegans*.

Barash and Vaquero-Garcia survey the current state of developing regulatory models for splicing using machine learning-based approaches and an introduction to the AVISPA web tool.

The contributors of this book are internationally recognized leaders in the arena of RNA processing, and we envision that this book will serve as a valuable resource for both experts and non-experts. Advanced undergraduates, entering graduate students in biology, chemistry, molecular engineering, computer science, and bioinformatics, and medical students and postdoctoral fellows who are new to the arena of posttranscriptional gene regulation should find this book accessible. We hope that the chapters in this volume will stimulate interest and appreciation of the complexity and importance of posttranscriptional gene regulation to its readers and even lead

them to pose new solutions to the many challenges that lie ahead in understanding how RNA-binding proteins affect gene regulation.

I sincerely express my greatest gratitude to the contributors to this book, Yoseph Barash, Jorge Vaquero-Garcia, Thomas Bebee, Benjamin Cieply, Russ Carstens, Jimena Guidice, Thomas Cooper, Rachel Soemedi, Hugo Vega, Judson M Belmont, Sohini Ramachandran, William Fairbrother, Josef Clark, Nelson Lau, Julie Bergalet, Eric Lecuyer, Sarah Azoubel Lima, Amy Pasquinelli, Punit Seth, Frank Rigo, Frank Bennett, Marianne Goodwin, Maurice Swanson, Dinghai Zheng, Bin Tian, Stefanie Gerstberger, Markus Hafner, Manuel Ascano, Thomas Tuschl, Rajarshi Choudhury, Zefeng Wang, Ling-ling Chen, and Jing Crystal Zhao.

La Jolla, CA, USA Gene W. Yeo

Contents

Chapter 1
Evolutionary Conservation and Expression of Human RNA-Binding Proteins and Their Role in Human Genetic Disease

Stefanie Gerstberger, Markus Hafner, Manuel Ascano, and Thomas Tuschl

Abstract RNA-binding proteins (RBPs) are effectors and regulators of posttranscriptional gene regulation (PTGR). RBPs regulate stability, maturation, and turnover of all RNAs, often binding thousands of targets at many sites. The importance of RBPs is underscored by their dysregulation or mutations causing a variety of developmental and neurological diseases. This chapter globally discusses human RBPs and provides a brief introduction to their identification and RNA targets. We review RBPs based on common structural RNA-binding domains, study their evolutionary conservation and expression, and summarize disease associations of different RBP classes.

Keywords RNA-binding domains, overview • RNA-binding proteins, tissue specificity • RNA-binding proteins, abundance • RNA-binding proteins, genetic diseases

1 Principles of Posttranscriptional Gene Regulation

RNA is an essential constituent of all living organisms and central to decoding the genetic information of every cell. Recent advances in RNA sequencing technologies have facilitated the discovery of novel transcripts and we will soon know the precise composition of most cellular transcriptomes. While functional annotation for many RNAs is still in progress, the major classes of RNAs have now been described (Table 1.1). The most abundant RNAs, constituting 90 % of cellular RNAs by copy number, are shared by all organisms and required for protein synthesis: rRNAs, tRNAs, and mRNAs (Table 1.1). The remaining 10 % are noncoding RNAs (ncRNAs) that mainly serve as guides or molecular scaffolds in a variety of processes including RNA splicing, RNA modification, and RNA silencing. The structure, length, and composition of these RNAs and their

S. Gerstberger • M. Hafner • M. Ascano • T. Tuschl (✉)
e-mail: ttuschl@mail.rockefeller.edu

G.W. Yeo (ed.), *Systems Biology of RNA Binding Proteins*, Advances in Experimental Medicine and Biology 825, DOI 10.1007/978-1-4939-1221-6_1

Table 1.1 Functional description of the main RNA classes in humans and their length distribution

RNA class	Size (nt)	Biological role (additional reviews on function and biogenesis)
Messenger RNA (mRNA)	~200–100,000	Encodes the information for protein-coding genes, translated by ribosomes (Dreyfuss et al. 2002; Glisovic et al. 2008; Müller-McNicoll and Neugebauer 2013)
Transfer RNA (tRNA)	~70–95	RNA adaptor molecule, transports amino acids to ribosome and recognizes specific triplet codons on mRNA (Suzuki et al. 2011; Maraia and Lamichhane 2011; Simos and Hurt 1999)
Ribosomal RNA (rRNA)	121–5,072	Structural component of ribosomes (Boisvert et al. 2007; Ciganda and Williams 2011; Granneman and Baserga 2004)
Small nuclear RNA (snRNA)	~70–190	snRNAs U1, U2, U4, U5, U6, U11, U12, U4atac, and U6atac are core components of the spliceosome; U7 snRNA functions in 3′end maturation of histone RNAs (Kiss 2004; Matera et al. 2007)
Small nucleolar RNA (snoRNA) and small Cajal-body-specific RNA (scaRNA)	~50–450	Guide chemical modifications (methylation and pseudouridylation) of rRNAs, snRNAs, and snoRNAs (Filipowicz and Pogacić 2002; Kiss et al. 2006; Matera et al. 2007)
microRNA (miRNA) and small interfering RNA (siRNA)	21–22	Associate with AGO proteins, guide them to target sequences predominantly in the 3′UTRs of mRNAs, induce degradation and translational repression (Bartel 2009; Kim et al. 2009)
piwi-interacting RNA (piRNA)	~28–32	Associates with PIWI proteins; PIWI RNP complexes induce ribonucleolytic cleavage and epigenetic silencing of transposable elements (Kim et al. 2009; Siomi et al. 2011)
Long intervening noncoding RNA (lincRNA), 7SK RNA	>200	Recruits chromatin modifiers and remodeling complexes, modulates transcription by recruitment of protein cofactors to transcription starts sites and enhancers, functions as molecular scaffolds for nuclear RBPs (Batista and Chang 2013; Ulitsky and Bartel 2013); 7SK RNA regulates transcription elongation (Peterlin et al. 2011)
Ribonuclease P/(RNase P) and mitochondrial RNA-processing endonuclease (MRP RNase)	~260–340	Ribonucleolytic RNP complexes that carry out processing of precursor tRNAs, rRNAs, snRNAs, and other noncoding RNAs (Xiao et al. 2002; Jarrous 2002; Ellis and Brown 2009; Esakova and Krasilnikov 2010)
Y RNA	~80–110	Small noncoding RNAs that form an RNP complex with TROVE2 (Ro60) protein and act as RNA chaperones, have a role in DNA replication and immune response (Hall et al. 2013; Köhn et al. 2013)

(continued)

Table 1.1 (continued)

RNA class	Size (nt)	Biological role (additional reviews on function and biogenesis)
Signal recognition particle RNA (7SL/SRP RNA)	~300	RNA of the signal recognition particle; the complex recognizes signal sequences of newly synthesized peptides and targets them to the endoplasmatic reticulum (Akopian et al. 2013)
Vault-associated RNA (vtRNA)	~80–120	Small noncoding RNAs, part of the vault RNP complex, involved in drug resistance, downregulate mRNA targets through posttranscriptional gene silencing (Berger et al. 2008)
Telomerase RNA (telRNA)	~450	RNA component of the telomerase complex TERC, which acts as reverse transcriptase and elongates telomerase repeats, TERC is structurally related to box H/ACA snoRNAs (Egan and Collins 2012)

Additional reviews on biogenesis pathways and RBP components interacting with each class of RNA are referenced

ribonucleoprotein particles (RNP) are distinct and allow their integration into diverse functions and layers of regulation to control target RNAs and their many functions.

Posttranscriptional gene regulation (PTGR) is a term that refers to the cellular processes that control gene expression at the level of RNA; it encompasses RNA maturation, modification, transport, and degradation. Consequently, every RNA molecule independent of its ultimate function is at some level subject to PTGR. RNA-binding proteins (RBPs) are central players of PTGR, as they directly bind to RNAs to form RNPs. In many cases, the RNP is the most basic unit, comprising a complex of obligate RNA and protein partners (e.g., snRNPs, snoRNPs, RNase P, ribosome subunits), which elicits its respective function. However, many other types of RNAs, particularly mRNAs and tRNAs, only transiently associate with RBPs, whose functions are necessary for their proper maturation, localization, and turnover (Dreyfuss et al. 2002; Granneman and Baserga 2004; Phizicky and Hopper 2010; Müller-McNicoll and Neugebauer 2013). Indeed most mRNA-binding proteins (mRBPs) have thousands of targets they regulate (Ascano et al. 2011). Hence the proper assembly and function of RNA-protein complexes are critical for development and maintenance of all cells and organisms. For a large fraction of RBPs, we are only starting to understand the complexity of their basic molecular roles, modes of recognition, and global targets.

In this chapter we review the current state of knowledge of the protein components involved in PTGR in humans. We discuss common patterns found among RBPs, based on targets, evolutionary conservation, shared structural domains, and cell-type-specific or ubiquitous expression. We then examine various classes of RBPs commonly implicated in human disease.

2 Human RBPs

2.1 Experimental and Bioinformatic Approaches Leading Towards a Census of RBPs

A complete catalogue of the proteins involved in PTGR is an important goal. Historically, different strategies have been employed towards the identification of RBPs (Ascano et al. 2013). Common approaches used RNA pull-down assays to recover associated proteins in cell lysates, followed by their mass spectrometric identification, or candidate proteins were recombinantly expressed and interrogated for their RNA-binding properties in vitro. These RNA-centric approaches identified the interactome of subsets of RNAs but did not capture the whole RBP proteome, and were not suitably of high throughput.

The first genome-wide approaches for the identification of proteins involved in PTGR utilized predictive methods and searched for the presence of protein domains conferring RNA binding. Early studies on the protein components of heteronuclear RNPs (hnRNPs) led to the identification of the first conserved, canonical RNA-binding domain (RBD) within RBPs (Burd and Dreyfuss 1994). Following these initial discoveries, and facilitated by advances in genome sequencing and the acquisition of protein structures, more precise classification of structural and functional protein domains followed rapidly (Henikoff et al. 1997). Computational prediction algorithms that use probability matrices from multiple sequence alignments enabled the detection of structural domains in uncharacterized protein sequences across organisms. The results of these predictions are publically available in a number of databases such as Interpro, Pfam, SCOP, SMART, or CDD (Murzin et al. 1995; Apweiler et al. 2001; Marchler-Bauer et al. 2003; Letunic et al. 2009; Finn et al. 2010). Among these domain classifications, at least 600 can be found with annotation referring to involvement in RNA-related processes.

Predicting the number of RBPs encoded in various genomes has remained a challenge. RBPs were defined by the presence of one or more canonical RBDs, such as RRM, KH, CSD, zinc fingers, and PUF domains (Lunde et al. 2007). Selecting these predominantly mRNA-binding RBDs, the number of RBPs was initially estimated to ~400–500 in human and mouse (McKee et al. 2005; Galante et al. 2009; Cook et al. 2011), ~300 in *D. melanogaster* (Lasko 2000; Gamberi et al. 2006), ~250–500 in *C. elegans* (Lasko 2000; Lee and Schedl 2006; Tamburino et al. 2013), and ~500 RBPs in *S. cerevisiae* (Hogan et al. 2008). Inclusion of RNA-processing domains involved in RNA metabolism of every known type of RNA leads to numbers near ~700 RBPs in humans (Anantharaman et al. 2002).

Other predictive approaches such as the Kyoto Encyclopedia of Genes and Genomes database (KEGG) (Kanehisa and Goto 2000) and the Gene Ontology project (GO) (Ashburner et al. 2000) integrate domain annotations, protein homologies, and searches of scientific literature statements. These estimate the number of human proteins with RNA-related functions to ~1,800 proteins (Fig. 1.1). However, these methods are often not reliable due to false classifications of proteins, leading to a large number false positives and false negatives.

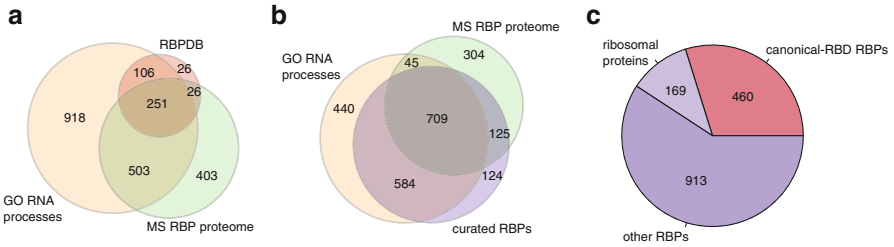

Fig. 1.1 Different approaches to define the catalogue of human RBPs. (**a**) Venn diagram showing the overlap of proteins with RNA-related Gene Ontology (GO) categories (Ashburner et al. 2000) (*orange*), the human RNA-binding proteome identified by RNA-cross-linking and mass spectrometry studies (MS RBP proteome, *green*) (Baltz et al. 2012; Castello et al. 2012; Kwon et al. 2013), and the RBPDB database of human RBPs with canonical RBDs (Cook et al. 2011) (*red*). (**b**) Venn diagram showing the overlaps of GO RBPs (*orange*), MS RBP proteome (*green*), and the curated RBP list based on analysis of RNA-binding domains and experimental evidence of RNA binding found in the literature (*violet*). (**c**) Composition of RBPs in the curated RBP list: Canonical-RBD RBPs (containing canonical RBDs (Lunde et al. 2007; Cook et al. 2011), *red*), ribosomal proteins (*bright violet*), other RBPs (*dark violet*)

In parallel, experimental proteome-wide methods were employed to identify the number of known and novel RBPs such as the development of protein microarrays, which allowed increased throughput for probing the RNA-binding capabilities of a fraction of the proteome in vitro, using RNA probes of defined sequence (Scherrer et al. 2010; Tsvetanova et al. 2010; Siprashvili et al. 2012). In an attempt to comprehensively identify existing and novel RBPs in human at large scale with a singular approach, cross-linking-based methods were recently introduced. In these methods, RBPs were covalently cross-linked to endogenous RNAs using in vivo UV cross-linking, followed by polyA selection of mRNAs, and subsequent identification of interacting proteins by mass spectrometry. These approaches identified ~800 mRBPs in human HEK293 and HeLa cell lines, respectively (Baltz et al. 2012; Castello et al. 2012), 555 in mouse embryonic stem cells (mESCs) (Kwon et al. 2013), and 200 mRBPs in yeast (Mitchell et al. 2013). Together, 1,100 of known and putative human mRBPs were experimentally defined and, assuming homologous function between mouse and human proteins, an additional ~80 proteins may be added (Fig. 1.1a, b). A significant portion of these (64 %) overlapped with known GO-classified RBPs (Fig. 1.1a, b). Many of the residual mRBP candidates did not contain previously described RBDs and require further experimental validation, while other known and expressed RBPs were missed due to the sensitivity of the experiments. However, in comparison to earlier predictive counts of the number of mRBPs (Cook et al. 2011), this approach expanded the mRBP proteome from ~400 to ~1,200 proteins and may, with increasing sensitivity, represent the most suitable method to identify novel RBPs in proteome-wide experiments in different cell types.

Here, we describe our attempt in generation of a curated and comprehensive list of RBPs involved in PTGR processes to guide us in their study of molecular and cellular function and definition of all RNA-related processes.

2.2 Generation of a Curated List of Human RBPs

Our approach selects RBDs involved in RNA-related processes as defined by Pfam (Finn et al. 2010) and searches the human genome for any protein-coding gene that contains at least one of the selected domains, but remains overall unbiased to the putative function of the gene and its RNA targets. Arriving at a list of 2,130 candidates, we added known RBPs from literature searches with unclassified RBDs and additionally screened proteins defined as RBPs by GO (Ashburner et al. 2000) and proteome-wide mass spectrometry datasets (Baltz et al. 2012; Castello et al. 2012; Kwon et al. 2013) for literature-based evidence of their involvement in PTGR.

The list of RBPs was finalized according to the following main criteria: (1) the proteins possessed defined RNA-binding or RNA-enzymatic domains, (2) the proteins were experimentally shown to be part of RNP complexes and thereby involved in RNA metabolic pathways, or (3) they possessed high sequence identity to homologs and paralogs involved in PTGR. Some of the candidate RBPs identified in the recent cross-linking-mass spectrometry studies were not considered as RBPs, if their RNA-binding activity could not be confirmed independently in other published datasets or their domain structure, family members, and homologs were not indicative of an RBP. We furthermore disregarded proteins containing putative RBDs if they showed strong evidence for exclusive roles in RNA-unrelated pathways, such as the majority of C2H2 zinc finger transcription factors of which only a small subset are RNA-binding, e.g., TF3A binding to 5S rRNA (Brown 2005). We included proteins, which are components of well-defined, large RBP complexes, such as the ribosome or the spliceosome, as it is difficult to establish with certainty which proteins interact with RNA directly or indirectly in these large RNPs. This approach is supported by recent proteome-wide cross-linking studies, and the RNA-binding properties of scaffold proteins CNOT1 and TDRD3 have for example emerged through this process (Thomson and Lasko 2005; Siomi et al. 2010; Baltz et al. 2012; Castello et al. 2012; Kwon et al. 2013). Through this curation process we reduced the union of ~3,700 proteins, derived from domain annotation, mass spectrometry datasets, literature search, and GO annotation, to arrive at a final of 1,542 proteins (Fig. 1.1b). The resulting curated list of RBPs contains proteins interacting with all RNA classes. A comparison to the conventionally named “canonical RBPs” (Lunde et al. 2007; Cook et al. 2011) shows that canonical RBPs only represent one-third of RBPs in this set and the majority of RBPs in our set would not be considered using currently available datasets (Fig. 1.1c). The following sections discuss abundance, evolution, expression, and RBPs in human disease based on this curated set of RBPs.

3 Quantitative Aspects of Proteins in RNA Metabolism

The dynamics of complex assembly and composition of RNPs, their targets, and protein cofactors are extremely sensitive to the quantitative relationship between the abundance of RBPs and their targets (Dreyfuss et al. 2002; Müller-McNicoll and

Neugebauer 2013). RBPs are in constant competition for binding to frequently occurring short and degenerated RNA sequence elements and thus the cellular compartment concentration of RNA and RBPs will affect the equilibrium of dynamic RNP formation and disassembly. Processes such as pre-mRNA splicing and alternative polyadenylation, where the choice of alternative splice sites or 3′UTR lengths is dependent on the abundance of splicing enhancers, silencers, or U1 snRNP (Smith and Valcarcel 2000; Kaida et al. 2010; Berg et al. 2012; Kornblihtt et al. 2013), emphasize the importance of determining precise RBP levels.

Approximately 7 % of all protein-coding genes are committed to PTGR, but their contribution to the pool of expressed proteins in cells is much higher. We analyzed expression level of ubiquitous RBPs based on RNA-seq data in HEK293 cells (Teplova et al. 2013) (Fig. 1.2). In this cell line, RBPs represented 9 % (1,364 genes) of the ~16,300 expressed genes (expressed with RPKM > 1), but their corresponding transcripts represent more than 25 % of total cellular mRNA, including 7 % mRBPs and 14 % ribosomal proteins (RBP categorization discussed in next section), stressing the abundance of mRNA metabolism and the central role of protein translation (Fig. 1.2). In contrast, transcription factors and cytoskeletal proteins were not

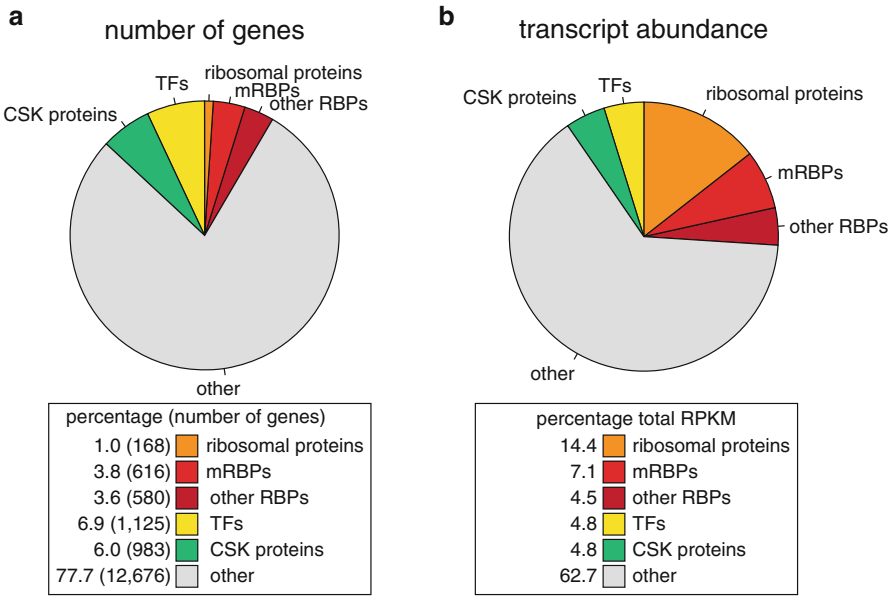

Fig. 1.2 Analysis of RBP abundance. Curated RBPs are subclassified into ribosomal proteins (*orange*), mRBPs (*bright red*), and other RBPs (*dark red*). Percentages of RBPs are compared to a set of GO-defined transcription factors (TFs, *yellow*), a set of GO-defined cytoskeletal proteins (CSK proteins, *green*) (Ashburner et al. 2000), and all other expressed genes (*grey*). (**a**) Count of expressed genes with RPKM > 1 in RNA-seq data from HEK293 cells (Teplova et al. 2013). (**b**) Relative abundance of each gene group is given by the summation of expression levels of genes in each category

overrepresented in the transcriptome of HEK293 cells. In summary, about a quarter of the transcriptome is committed to RNA metabolism, highlighting its fundamental role in the cell.

4 Paralogous RBP Families and their Targets

Determining the evolutionary relationships and the conservation of gene families has been critical for understanding gene function and emphasized the utility of model organisms for the study of fundamental biological processes (Henikoff et al. 1997). To account for redundancies among RBPs it is therefore beneficial to consider the RBP family as the smallest functional unit. Grouping the 1,542 RBPs into paralogous gene families with at least 20 % homology gives 1,113 RBP families with one or two members on average. The large number of families reflects a high diversity of RBPs in human.

Here, we categorized RBPs and RBP families based on their reported natural targets and examined their distribution and evolutionary relationships among different classes. Although RBPs often show some degree of interaction with a range of target RNAs in vivo, most of them are committed to one subtype of RNA (Hafner et al. 2010; Wang et al. 2012; Ascano et al. 2012; Hussain et al. 2013; Lovci et al. 2013; Wang et al. 2013) (Table 1.1). Some exceptions remain, such as RNA nucleases, and RBPs acting at the interface of two different RNA classes, such as spliceosomal proteins, XPO5, or EEF1A, recognizing snRNA/mRNAs, pre-miRNAs/mRNAs, and tRNAs/mRNAs, respectively (Liu et al. 2002; Lund 2004; Mickleburgh et al. 2006; Bennasser et al. 2011; Dever and Green 2012). For the RBPs with multiple targets, we either classified them as diverse in target preference or counted the proteins towards the predominant group of targets based on available literature. The resulting distribution of RBPs and RBP families across all RNA targets in human and their conserved homologs in yeast is shown in Fig. 1.3a–c.

Our analysis shows that mRBPs form the largest group among RBPs comprising 45 % of all human RBPs (~700 proteins). mRBPs frequently represent families of RBPs with more than two members. Ribosomal proteins constitute the next larger group of RBPs with ~170 proteins of the cytosolic and mitochondrial ribosomes. The next smaller groups of RBPs are committed to tRNA (~150) and rRNA (~120) biogenesis pathways, followed by proteins involved in snRNA, snoRNA, and other ncRNA pathways (Fig. 1.3a, b).

Of the ~1,100 human RBP families, ~550 have homologs in yeast with on average 30 % homology. Different clades of RBPs display varied degrees of conservation. Cytosolic ribosomal proteins are the most conserved with ~57 % homology, while proteins associating with mRNAs or ncRNAs are least conserved, 27 % and 20 %, respectively, and also have the least number of conserved homologs, 45 % and 21 %, respectively (Fig. 1.3c). Nevertheless, despite the gene expansions within protein

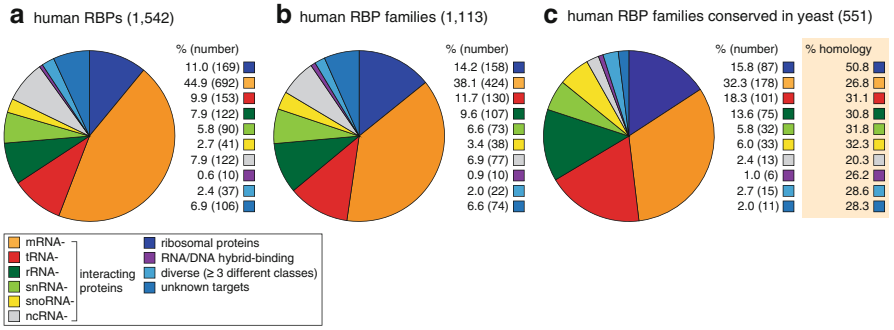

Fig. 1.3 Number of RBPs involved in different RNA pathways. Curated RBPs are categorized into the following groups: (*1*) Ribosomal proteins and RBP-interacting proteins (e.g., TUDOR proteins, RBP transport proteins) (*dark blue*), (*2*) mRNA-binding proteins (*orange*), tRNA-binding proteins (*red*), rRNA-binding proteins (*dark green*), snRNA-binding proteins (*bright green*), snoRNA-binding proteins (*yellow*), ncRNA-binding proteins (ncRNAs defined as miRNA, piRNA, MRP, 7SL, XIST, lincRNAs, telRNA, etc.) (*light grey*), RNA/DNA-hybrid-interacting proteins (*violet*), RBPs interacting unselectively with a range of RNA targets (*light blue*), RBPs with unknown RNA targets (*marine blue*). Distribution into the listed categories of the (**a**) 1,542 curated human RBPs, (**b**) 1,113 human paralogous RBP families, and (**c**) conserved paralogous RBP families in *S. cerevisiae* and their average conservation score (*orange box*)

families at later evolutionary stages (Venter 2001; Van de Peer et al. 2009), the relative ratios of paralogous RBPs families invested in the different RNA pathways remain approximately the same across evolution as seen for the distribution between human and yeast RBP families (Fig. 1.3b, c). This breadth of PTGR factors agrees with an earlier analysis of 32 RBP domain classes of canonical RBDs (including RRM, KH, dsrm, DEAD, PUF, Piwi, PAZ, zinc finger, LSM) showing that the large diversity of RBPs found in contemporary metazoans was already established in the last common ancestor (LCA) of animals, and which possessed an estimated total number of 88 RRM, 15 KH, 49 DEAD box, 9 dsrm, and 38 other RBD proteins (Kerner et al. 2011). Thus the complexity of PTGR was present at the earliest stages of evolution, reflecting that RNA metabolism lies at the heart of eukaryotic gene regulation.

Visualization of the evolutionary relationships of RBP families facilitates systems biology approaches to dissect their regulatory roles. Phylogenetic trees give an intuitive graphic representation of the conservation of proteins, highlight closely related homologs, and thereby provide a glimpse into function of uncharacterized RBPs if function has been already established for a relative. Phylogenetic comparison of the predominantly mRNA-binding KH-domain-containing proteins and the proteins of the small subunit of the cytosolic ribosome illustrates the differences in their evolutionary trajectory (Fig. 1.4).

KH proteins experienced multiple gene expansions, as noted earlier for mRBPs, and evolved new RBP families at the later metazoan stages, thereby

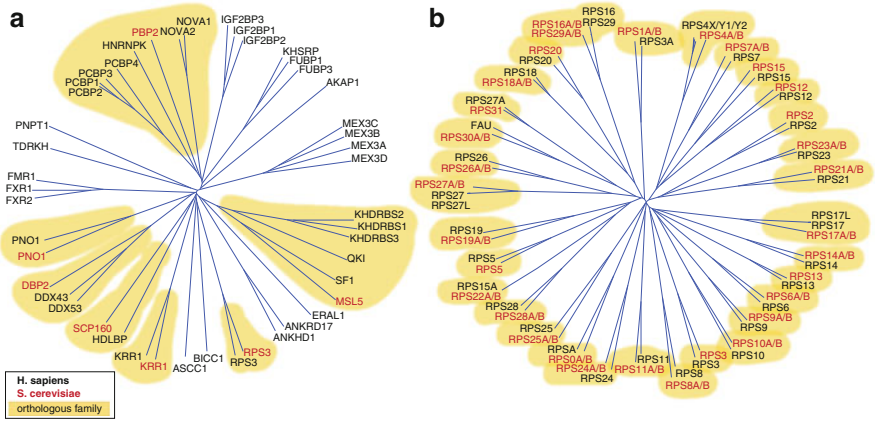

Fig. 1.4 Phylogenetic trees of (**a**) KH-containing proteins and (**b**) ribosomal proteins of the small subunit. Branch lengths are scaled to the sequence identity of the proteins. *S. cerevisiae* proteins are marked in *red*, human proteins in *black*, homologous families with conserved members in *S. cerevisiae* highlighted in *yellow*

expanding and diversifying components involved in various regulatory pathways, such as mRNA splicing, translational regulation, and transport. KH protein families contain between one and four members in human, and possess generally one distantly related homolog in yeast (Fig. 1.4a). Multiple family members often have redundant biological functions and RNA target spectra. For example, members of the FMR1 family (FMR1, FXR1, FXR2) or the IGF2BP1 family (IGF2BP1, 2, and 3) show >90 % identical RNA-binding specificities (Hafner et al. 2010; Ascano et al. 2012).

In contrast, cytosolic ribosomal proteins display an unusually high conservation, not too surprising, given that the process of protein translation is conserved to such a high degree between prokaryotes and all clades of eukaryotes that functional details of translation determined in bacteria are almost identical to higher systems (Wool et al. 1995; Melnikov et al. 2012; Dever and Green 2012). The ~90 human ribosomal cytosolic proteins are highly similar in structure and function between yeast and human and show late divergence in evolution, as illustrated for the phylogenetic tree of small ribosomal subunit proteins (Fig. 1.4b). With on average 57 % protein identity, all human cytosolic ribosomal proteins have direct one-to-one, or due to a whole-genome duplication in yeast, one-to-two or two-two matching homologs (Wool 1979; Wool et al. 1995; Anger et al. 2013). In contrast, the majority (80 %) of the ~80 human mitochondrial ribosomal proteins (Matthews et al. 1982) have no homologs in yeast, and the few that are conserved have comparatively low homology (22 % identity), reflecting that mitochondrial ribosomes, acquired through eubacterial endosymbiosis, rapidly evolved independently across species and that major remodeling events happened later in evolution (Cavdar Koc et al. 2001; O'Brien 2003).

5 Structural Analysis of RNA-Binding Domains

Analysis of structural features in proteins and the grouping of proteins into domain classes can help to understand their biological function (Henikoff et al. 1997; Anantharaman et al. 2002). Structural domains can predict how RBPs recognize and bind RNAs. It can also uncover redundancies to other RBPs in target recognition, as well as highlight families of RBPs that remain to be characterized. The size of a domain class mirrors its diversity and evolutionary adaptation to biological pathways.

Structure-guided searches can be valuable to place proteins into biological pathways and, for instance, DICER1 and DROSHA were identified as the endonucleases responsible for double-stranded RNA (dsRNA) processing in microRNA (miRNA) maturation based on known structure and substrate preferences of the dsRNA-processing bacterial and yeast RNase III enzymes (Hammond et al. 2000) (Bernstein et al. 2001; Lee et al. 2003). Similarly, the structural similarity of AGO proteins and the germline-specific PIWI proteins sparked the search for PIWI-interacting small RNAs with similar features as miRNAs, now known as piRNAs (Girard et al. 2006; Grivna 2006; Aravin et al. 2006).

For a review of the structural features of RBPs, we analyzed characteristic domain combinations of RBD classes (Fig. 1.5a) and will give here a brief overview over the abundant RBD classes and their modes of RNA-binding, natural targets, and the processes they are involved in. For excluded classes, we refer to a number of excellent review articles (Burd and Dreyfuss 1994; Sommerville 1999; Aravind and Koonin 2001a; Aravind and Koonin 2001b; Anantharaman et al. 2002; Arcus 2002; Szymczyna et al. 2003; Kim and Bowie 2003; Maraia and Bayfield 2006; Lunde et al. 2007; Rajkowitsch et al. 2007; Glisovic et al. 2008; Curry et al. 2009; Mihailovich et al. 2010; Zhang et al. 2010). To give additional insight into the structural properties of RBPs, we distinguished between RBDs with only RNA-binding properties (nonenzymatic RBDs), and RBDs that also contain enzymatic functions (enzymatic RBDs), such as RNA helicases, ATPases, polymerases, editing enzymes, and nucleases.

5.1 Modes of RNA Interaction by RBPs and their Domain Organization

Prototypical single-stranded RNA (ssRNA)-binding domains interact with their targets in a nucleobase-sequence-specific manner typically binding between 4 and 8 nucleotides (Singh and Valcarcel 2005; Lunde et al. 2007; Glisovic et al. 2008). Specificity is introduced mainly by hydrogen bonding and van der Waals interactions of the nucleobases with the protein side chains or the carbonyl and amide groups of the main chain (Auweter et al. 2006), often leaving the RNA phosphate backbone exposed to the solvent. Additional base stacking interactions with aromatic amino acids or positively charged residues in cationic π interactions serve to

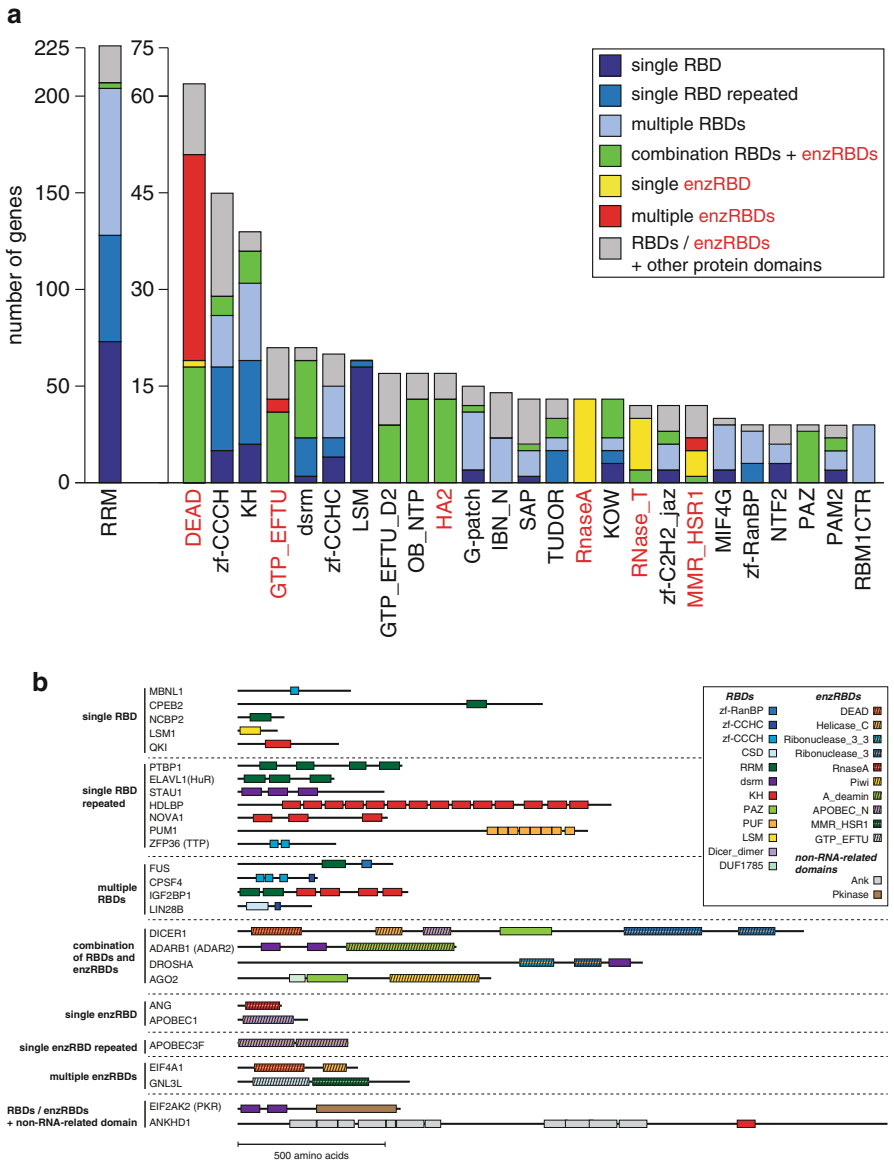

Fig. 1.5 (**a**) Analysis of structural patterns of the most abundant RNA-binding domains (with ≥ 9 members) in humans. Shown are the counts for the number of genes containing the listed RNA-binding domains (with ≥ 9 members) named by its Pfam abbreviation (Finn et al. 2010). RNA-binding domains are categorized into those binding RNA without additional enzymatic activity (RBD) (*black*) and those with additional enzymatic activity (enzRBD) (*red*). The RBD category was broadly defined to include protein-protein interacting domains known to interact with RBPs, such as those found in TUDOR family proteins (TUDOR) or ribosomal proteins. The following structural patterns are counted: (*1*) singular occurrence of an RNA-binding domain (RBD—*dark blue*, enzRBD—*yellow*), (*2*) single RBD repeated (RBD—*marine blue*), (*3*) multiple RBDs (RBD—*light blue*, enzRBD—*red*), (*4*) combinations of RBDs and enzRBDs (*green*), and (*5*) combination of at least one RBD/enzRBD with at least one other, non-RNA-related protein domain (*grey*). (**b**) Scheme of domain structure organization of representative RBPs, categorized into the domain combination classes listed in (**a**)

increase affinity. Double-stranded RNA (dsRNA)-binding proteins achieve specificity through recognition of shape of RNA secondary structure, such as stem–loops (Masliah et al. 2013). Non-sequence-specific RBDs generally interact with the negatively charged phosphate backbone, leaving the bases exposed to the solvent. To achieve specificity, these RBPs can interact with cofactors recruiting them to specific targets, as has been observed for many RNA helicases (Rocak and Linder 2004; Auweter et al. 2006).

Many RBDs (but also DNA-binding domains) derive from a few common superfamily folds, such as the oligonucleotidyl transferase fold and the oligosaccharide-binding fold (OB-fold). Oligonucleotidyl transferase fold proteins include enzymatic RBPs such as TUTases, polyA polymerases, RNA ligases, tRNA CCA-adding enzymes, and immune-stimulatory 2′,5′-oligoadenylate synthases (Kuchta et al. 2009). RBDs of the OB-fold superfamily are the S1, PAZ, and CSD domains (Murzin 1993; Arcus 2002; Lunde et al. 2007). However, RBDs largely diversified throughout evolution and most RBD classes have only one member, while only 4 % of all RBD classes found in human have more than eight members. Members of the 26 most abundant RBD classes (with 9 and more members) constitute a third of the 1,542 curated RBPs (Fig. 1.5a); most of them are mRBPs. Some of the highly studied RBDs (Lunde et al. 2007), such as the PUF (two proteins), S1, CSD, and PIWI domains (eight members each), define smaller RBD classes in humans. Particularly, ribosomal structural components and proteins involved in processes related to ribosome maturation are unique and thus cannot be classified into large families of related structural organization (Korobeinikova et al. 2012).

More than half of all RBPs contain only one RBD; mRBPs, however, form a notable exception and often have multiple RBDs, either one repeated RBD or multiple RBDs in combination. This modular design allows flexibility and versatility for target recognition and, as RBDs usually recognize relatively short stretches of RNA, increases the affinity and specificity for RNA targets by extending the RNA recognition element (RRE) of the protein (Lunde et al. 2007). A few RBDs are exclusively found in combination with other conserved protein domains, such as RBM1CTR in hnRNPs or the PAZ domain found in the Argonaute proteins and DICER1.

5.2 Abundant Nonenzymatic RBDs

5.2.1 RNA-Recognition Motif

The ssRNA-recognition motif (RRM) is the most frequently found RNA-binding domain in eukaryotes, and has 226 members in humans. The ~90 amino-acid-long domain adopts a βαββαβ topology and is composed of two RNP consensus motifs that recognize 4–6 nucleotides (nts) by stacking interactions of the bases with three conserved aromatic amino acids in the β-sheets (Auweter et al. 2006; Lunde et al. 2007; Cléry et al. 2008). Its small size and modular organization yield flexibility to adaptive change and allowed the RRM domain to vastly expand during evolution (Anantharaman et al. 2002).

Deviations from this canonical binding mode, including N- and C-terminal extensions of the domain, as well as usage of the linker regions and other regions outside of the β-sheet, have been characterized and allow for recognition of up to 8 nt (Maris et al. 2005; Auweter et al. 2006; Lunde et al. 2007; Cléry et al. 2008; Muto and Yokoyama 2012).

Sixty-one RRM proteins only comprise a single isolated RRM domain; examples include the polyA-binding protein CPEB family and the nuclear cap-binding protein NCBP2 (CBP20) (Fig. 1.5b). Sixty RBPs occur with several repeated RRMs; among them are the PTBP and the ELAVL families, regulators of mRNA splicing, stability, and localization (Sawicka et al. 2008; Simone and Keene 2013). Another 68 RRM proteins are found in combination with other RBDs; prominent examples in this group include the IGF2BP1 and FUS families, involved in translational regulation, mRNA transport, and splicing (Tan and Manley 2009; Bell et al. 2013) (Fig. 1.5b). Hence, RRM proteins are found in a variety of biological pathways, the majority of which involve mRNA-related processes, such as regulation of mRNA stability, splicing, translation, and transport. RRM domains are generally a diagnostic indicator for ssRNA binding. In a few cases, however, protein-binding partners have been shown to occlude interaction of the RRM domain with RNA, as seen for the RRM domain of the exon junction complex protein RBM8A (Y14), which binds to the protein cofactor MAGOH (Maris et al. 2005; Glisovic et al. 2008).

5.2.2 K-Homology Domain

The heterogeneous nuclear RNP K-homology (KH) domain binds to ssRNA and ssDNA, and has 39 members in humans. KH domains are ~70 amino acids long and characterized by a hydrophobic core domain with an (I/L/V)IGXXGXX(I/L/V) consensus sequence in the center. Structurally, all KH domains form a three-stranded β-sheet packed against three α-helices and belong either to the eukaryotic type I, of βααβα topology, or the prokaryotic type II, of αββααβ topology (Grishin 2001; Lunde et al. 2007; Valverde et al. 2008). KH domains typically recognize 4-nt ssRNA sequences through electrostatic interactions. Signal transduction and activation of RNA (STAR) proteins, such as SAM68 and QKI, contain just a single KH domain sandwiched between two short signaling motifs, which modulate the protein activity through posttranslational modifications in response to intracellular signaling pathways (Lasko 2003; Chénard and Richard 2008). However, most KH proteins contain combinations of RBDs including the IGF2BP1 family, with four KH and two RRM motifs (Bell et al. 2013), and the brain-specific NOVA splicing family with three repeated KH domains (Li et al. 2007) (Fig. 1.5b). The most extreme example of multiplication of RBDs is found in HDLBP, conserved from yeast to humans, which has 14 repeated KH domains in human (Fig. 1.5b). Analogous to RRM proteins, KH domain proteins predominantly interact with mRNAs and are found in posttranscriptional processes, such as mRNA splicing (PCBP and NOVA family), transport, and translation

(IGF2BP1, FMR1, and the MEX family). The two, highly conserved, KH domain proteins PNO1 and RPS3 represent an exception in their target specificity and interact with rRNA during ribosome biogenesis (Vanrobays et al. 2004; Anger et al. 2013).

5.2.3 Double-Stranded RNA-Binding Motif

The dsRNA-binding motif (dsrm) has 21 members and 4 dsrm-like members in humans. Dsrm domains are ~70 amino acids long and adopt an αββββα topology, in which the two α-helices are packed against the three β-sheets. This facilitates nonspecific, shape-dependent contacts with the RNA backbone along the minor and major grooves of A-form dsRNA helix, as well as base contacts along the minor groove and the apical loop (Chang and Ramos 2005; Lunde et al. 2007; Masliah et al. 2013). Dsrm domains are rarely found alone; 24 of the 25 of human dsrm proteins contain multiple dsrm domains or other enzymatic and nonenzymatic RBDs that modulate their function. While the best known dsrm-containing proteins are the Staufen family (STAU1, STAU2) of mRNA stability and transport regulators (Miki et al. 2005; Park and Maquat 2013), this domain type is not confined to mRBPs; instead, most dsrm proteins interact with a range of RNA substrates and are commonly found in RNA enzymes. Members include the adenine-to-inosine RNA-editing ADAR family, processing stem–loops or double strands in mRNAs, viral RNAs, and miRNA precursors (Savva et al. 2012), the two miRNA-processing endonucleases DROSHA and DICER1 (Kim et al. 2009; Wilson and Doudna 2013), as well as the interferon-inducible protein kinase EIF2AK2 (PKR), which, upon binding dsRNA, activates its kinase domain (Saunders and Barber 2003; Raven and Koromilas 2008) (Fig. 1.5b).

5.2.4 CCCH and CCHC Zinc Fingers

The two ssRNA-binding zinc fingers (zf), zf-CCCH and zf-CCHC, form rigid structures by coordination of a Zn^{2+} ion with three cysteine (C) and one histidine (H) residues. In humans, 45 genes contain the zf-CCCH (C-×8-C-×5-C-×3-H type) and 21 contain the zf-CCHC (C-×2-C-×4-H-×4-C) motif, also known as zinc knuckle. Zf proteins form sequence-specific interaction with RNAs through hydrogen bonding and van der Waals interactions of the protein backbone (Lunde et al. 2007; Kaymak et al. 2010), and use stacking interactions of aromatic side chains with the bases to increase RNA-binding affinity. In contrast to other ssRNA-binding domains, the rigidity and shape of the protein structure are the key determinant for specificity of zinc-finger proteins to their target RNAs. The domains generally occur in repeats or in combination with other RBDs. While for most of

the CCHC and CCCH zf proteins the molecular function remains unclear, characterized zf proteins are predominantly involved in regulation of mRNA-related processes. Classic examples of zf-CCCH proteins are the AU-rich-binding ZFP36 (TTP) proteins, which participate in rapid degradation of mRNAs transcribed after immune stimulation (Sandler and Stoecklin 2008; Brooks and Blackshear 2013), and the muscleblind (MBNL1,2,3) family, which regulates alternative splicing during muscle differentiation (Pascual et al. 2006; Cooper et al. 2009). Characterized zf-CCHC proteins include the CPSF4 mRNA polyadenylation and cleavage factor (Colgan and Manley 1997; Shatkin and Manley 2000; Proudfoot and O'Sullivan 2002; Proudfoot 2004) (Fig. 1.5b), and the ZCCHC7 (AIR1) protein, member of the nuclear polyadenylation TRAMP complex, required for the degradation of aberrant nuclear ncRNAs (Anderson and Wang 2009). Another prominent member in this class is the LIN28 family (Fig. 1.5b), which posttranscriptionally maintains pluripotency in early embryonic development by inhibiting maturation of miRNA let-7 family precursors and increasing stability and translation of mRNA targets (Thornton and Gregory 2012; Wilbert et al. 2012; Cho et al. 2012; Hafner et al. 2013).

5.2.5 LSM Domain

The LSM domain is found in 19 proteins in humans. First discovered in Sm proteins, it later was re-named LSM ("like-Sm") to include proteins outside of its founding members. The LSM fold is a bipartite domain that stretches along a region of ~65 amino acids and folds into an N-terminal α-helix followed by a twisted five-stranded β-strand (Wilusz and Wilusz 2005; Tharun 2009). The two motifs, motif I, ~22 residues long, and motif II, ~16 residues long, are connected by a variable linker region. LSM proteins form hexa- and heptameric-ring-shaped complexes around RNA. Binding to short, internal polyU- and polyA-rich stretches, they generally associate with snRNAs and some snoRNAs to form stable snRNP complexes (Achsel et al. 2001; Khusial et al. 2005). These RNP complexes commonly act as RNA-RNA and RNA-protein chaperones (Wilusz and Wilusz 2005; Tharun 2009).

The originally characterized Sm proteins (SNRPB, SNRBPD1, SNRBPD2, SNRBPD3, SNRBPE, SNRBPF, SNRBPG) form the core protein complex around the snRNAs of the major and minor spliceosome (U1, U2, U4, U5, U11, U12, U4atac) (Tharun 2009). Other combinations of LSM-fold proteins and their RNA components lead to functional complexes (Beggs 2005; Wilusz and Wilusz 2005; Tharun 2009), such as the nuclear LSM2-8 protein complex (LSM2, LSM3, LSM4, LSM5, LSM6, LSM7, LSM8), which forms around U6 and U6atac snRNA and participates in mRNA splicing. Association of the same LSM proteins around the C/D box U8 snoRNA results in an RNP complex functioning in rRNA maturation (Pannone et al. 2001; Tomasevic and Peculis 2002). In addi-

tion, binding of LSM2-8 to nuclear polyadenylated mRNAs promotes mRNA decapping (Kufel et al. 2004). Other LSM complexes include the U7 snRNP complex (SNRPB, LSM10, SNRPD3, LSM11, SNRPE, SNRPF, SNRPG), which is essential for histone 3′ end processing (Pillai et al. 2001), and the cytoplasmic LSM1-7 complex (LSM1, LSM2, LSM3, LSM4, LSM5, LSM6, LSM7), which localizes to P bodies and facilitates mRNA decapping after deadenylation (Tharun et al. 2000; Coller and Parker 2004; Parker and Song 2004; Parker and Sheth 2007). We can expect the identification of novel functions and target specificities in PTGR by as yet uncharacterized variations in the composition of LSM complexes (Wilusz and Wilusz 2005).

5.2.6 PIWI and PAZ Domain

The combination of PIWI, PAZ, and MID domains characterizes the Argonaute RBP family, a clade wirh four AGO and four PIWI protein members in humans (Peters and Meister 2007; Kim et al. 2009). Proteins of this clade bind miRNAs, siRNAs, and piRNAs by anchoring the 5′phosphate in the MID-domain pocket and the 3′end in the PAZ domain, while the PIWI domain interacts with the RNA backbone (Song 2004; Song and Joshua-Tor 2006; Wang et al. 2008b; Tian et al. 2011; Simon et al. 2011; Schirle and MacRae 2012). The PAZ RBD is also a structural component of the miRNA-processing endonuclease DICER1 (Zhang et al. 2004).

The 110 amino-acid-long PAZ domain consists of a β-barrel followed by an αβ-domain, and is structurally related to OB folds, S1, and LSM domains (Lunde et al. 2007). Forming a clamp-like structure, the PAZ domain selectively binds the two-nucleotide overhangs of small RNA duplexes at the 3′end, thereby acting as an anchor to position small RNAs for cleavage (Jinek and Doudna 2009). The PIWI domain is structurally similar to the RNase H endonuclease domain; however, in mammals, only PIWI proteins (Siomi et al. 2011) and AGO2 (Meister et al. 2004; Liu 2004) display nuclease activity, while in other AGO proteins subtle changes in the active site or the N-terminal regulatory domain prevent catalytic activity (Hauptmann et al. 2013; Faehnle et al. 2013; Nakanishi et al. 2013). AGO proteins initiate, guided by miRNAs and siRNAs, posttranscriptional silencing of mRNAs (Hutvagner and Simard 2008), and PIWI proteins, guided by piRNAs, silence transposons at the posttranscriptional and epigenetic levels (Kim et al. 2009; Siomi et al. 2011). Given the variability of possible guide RNA sequences, Argonaute proteins are tremendously versatile and by using different endogenously expressed guide RNA sequences they can form hundreds of distinct RNP complexes in vivo. Capable of targeting virtually any given cytosolic RNA sequence in a specific manner, they are used extensively as a tool in biotechnological applications (Dorsett and Tuschl 2004).

5.2.7 PUF Repeat

In humans, PUF repeats are only found in the two members of the Pumilio family; however, the structure and RNA-recognition mechanism of this domain are highly conserved and probably the best understood among all RBDs. PUF domains are ~40 amino acids long and consist of three α-helices that pack together into a half-ring structure. Each PUF domain recognizes only one nucleotide, but multiple repeats additively increase the number of bases recognized, and Pumilio proteins contain multiple PUF repeats that recognize highly sequence-specifically stretches within mRNAs (Wang et al. 2001). The extremely high specificity is achieved by hydrogen bonding interactions of two residues per repeat, while aromatic side chains wrap the bases into a tight fit. Human Pumilio proteins (PUM1, PUM2) contain eight PUF repeats, which together recognize the sequence UGUANAUA frequently located within the 3′UTRs of its targets to regulate mRNA stability and translation (Wickens et al. 2002; Wang et al. 2002; Hafner et al. 2010). Compared to other RBDs recognizing short and often degenerated RNA sequences, the RRE of Pumilio repeats is highly predictive for identifying Pumilio protein targets. Indeed, predictions of conserved RREs within 3′UTRs of mRNAs mainly identified, next to miRNA, Pumilio protein-binding sites (Xie et al. 2005), highlighting the exceptionally high information content of the Pumilio RRE. The high molecular specificity of the interaction has allowed engineering of RNA-binding specificity of Pumilio proteins to recognize different sequences (Cheong and Hall 2006).

5.3 Predominant Enzymatic RNA-Binding Domains

5.3.1 DExD/H helicases

DExD/H helicases, comprising DEAD and DEAH box helicases, are ATP-dependent enzymes that are involved in RNA-protein remodeling in the cell. They form the second largest class of RBPs comprising 73 members in humans, of which 62 interact specifically with RNA, and the remaining with DNA. The majority of the human RNA-binding DexD/H helicases, 42 members, belong to the DEAD box class, while the others are DEAH and DExH Ski-like helicases (named after its founding member Ski2p) (la Cruz et al. 1999). DExD/H RNA helicases belong to the SF2 helicase superfamily and contain NTPase characteristic Walker A and B motifs; their seven helicase signature motifs extend over ~400 amino acids (Tanner and Linder 2001; Rocak and Linder 2004; Pyle 2008; Jankowsky and Fairman-Williams 2010; Fairman-Williams et al. 2010). The helicases are differentiated by their catalytic core residues Asp-Glu-Ala-Asp for DEAD box helicases, and Asp-Glu-Ala/x-His for the related DEAH box and Ski2-like helicases. The enzymatic core arranges into two discrete domains connected by a linker that forms a cleft, in which an ATP can bind (Tanner and Linder 2001), whose hydrolysis provides the energy for unwinding RNA secondary structures or reorganizing RNPs in either a directional (DEAH helicases) or a bidirectional (DEAD helicases) manner. DExD/H RNA

helicases generally lack substrate specificity, or even affinity, towards RNA and DNA. This allows them to promiscuously unwind and remodel a broad range of targets, but also requires their association with cofactors that give specificity and affinity for their targets (Rocak and Linder 2004; Jankowsky and Fairman-Williams 2010). While most members of DExD/H helicases are involved in mRNA-related processes, in particular splicing, they play essential roles in diverse PTGR pathways such as transcriptional regulation, rRNA and tRNA maturation, viral defense, miRNA RISC loading, translation initiation, RNA export, and degradation (Rocak and Linder 2004; Fukuda et al. 2007; Pyle 2008; Jankowsky 2011; Linder and Jankowsky 2011; Martin et al. 2013; Schmidt and Butler 2013; Fullam and Schröder 2013). Next to RNA-RNA and RNA-protein remodeling, DExD/H helicases are also important in RNA-protein complex disassembly and facilitate removal of protein interactors from their targets during RNA export (Linder and Jankowsky 2011).

Paralogs within one RBP family can function in highly diverse roles and pathways, but even one helicase can assume a variety of different biological functions depending on its associated cofactors. For instance, EIF4A1, the first DEAD box helicase for which remodeling and unwinding was mechanistically characterized, forms the EIF4F translation initiation complex, together with the cap-binding protein EIF4E and the scaffolding protein EIF4G (Gingras et al. 1999; Andreou and Klostermeier 2013). Complexed with EIF4H or EIF4B cofactors, EIF4A1 unwinds secondary structures in the 5′UTR, allowing binding of the 43S ribosome complex for AUG start codon scanning. In contrast, although structurally very similar (65 %), the family member EIF4A3 is a core component of the exon junction complex (EJC), in which it acts as an RNA clamp to assist correct positioning of the EJC 20–24 nt upstream of mRNA exon-exon junctions (Linder and Fuller-Pace 2013).

5.3.2 EF-Tu GTP-Binding Domain

The EF-Tu GTP-binding domain (GTP_EFTU), named after its prokaryotic founding member EF-Tu, is a highly conserved domain across all kingdoms of life, and shared by 21 genes in humans. The domain is typically found in GTP-binding translation elongation factors, which are composed of three structural domains, the GTP-binding domain, and two β-barrel nucleotide-binding domains, D2 and D3, which bind to aminoacylated tRNAs (Nissen et al. 1995; Wang et al. 1997; Negrutskii and El'skaya 1998). Eukaryotic EF-1α (human ortholog EEF1A1) has also been shown to interact with higher molecular weight G/U-rich RNAs and rRNAs at a tRNA-independent binding site (Negrutskii and El'skaya 1998). Translation elongation factors are essential for protein synthesis; they bind aminoacyl-tRNAs in a GTP-dependent manner and direct them to the A-site of the ribosome where, upon codon recognition by the tRNA, GTP is hydrolyzed and the factor released (Dever and Green 2012). Furthermore, the GTP_EFTU domains are not only found in combination with D2 and D3 in various translation initiation and release factors, but also alone in GTPases involved in mRNA splicing (EFTUD2) (Fabrizio et al. 1997) and