Review

RNA-binding proteins in Mendelian disease

Alfredo Castello[1], Bernd Fischer[1], Matthias W. Hentze[1], and Thomas Preiss[2]

[1] European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, D-69117 Heidelberg, Germany
[2] Genome Biology Department, The John Curtin School of Medical Research (JCSMR), The Australian National University, Acton (Canberra), ACT 0200, Australia

RNA-binding proteins (RBPs) control all aspects of RNA fate, and defects in their function underlie a broad spectrum of human pathologies. We focus here on two recent studies that uncovered the *in vivo* mRNA interactomes of human cells, jointly implicating over 1100 proteins in RNA binding. Surprisingly, over 350 of these RBPs had no prior RNA binding-related annotation or domain homology. The datasets also contain many proteins that, when mutated, cause Mendelian diseases, prominently neurological, sensory, and muscular disorders and cancers. Disease mutations in these proteins occur throughout their domain architectures and many are found in non-classical RNA-binding domains and in disordered regions. In some cases, mutations might cause disease through perturbing previously unknown RNA-related protein functions. These studies have thus expanded our knowledge of RBPs and their role in genetic diseases. We also expect that mRNA interactome capture approaches will aid further exploration of RNA systems biology in varied physiological and pathophysiological settings.

Cellular functions of RBPs

RBPs are omnipresent partners of cellular RNA. Together they form dynamic ribonucleoprotein particles (RNPs), often in a highly combinatorial fashion, that affect virtually all aspects of the life of RNA from its genesis to its eventual demise. RBPs are critically important to RNA function in structural, regulatory, or catalytic capacities in the case of noncoding RNA (ncRNA), or for controlling mRNA as the template for protein synthesis. A wealth of literature focusing on mRNA in eukaryotic cells documents that RBPs, together with ncRNAs, such as micro RNAs (miRNAs), direct and regulate the post-transcriptional fate of mRNA in the nucleus and cytoplasm affecting its splicing and 3′ end formation, editing, localization, translation, and turnover, often in a dynamic and cell type-specific manner [1–3]. RBPs often interact with the untranslated regions (UTRs) of mRNAs, which are rich repositories of RBP binding sites with *cis*-acting regulatory functions. The importance of 3′UTRs as hubs of post-transcriptional regulation is further underscored by the recent discovery of widespread, regulated, alternative mRNA 3′-end formation in many cellular and disease contexts (e.g., [4–6]). This typically leads to the presence of multiple mRNA variants per gene that differ in 3′UTR length and thus in responsiveness to the cellular regulatory milieu of RBPs and miRNAs [7].

Much evidence implicates defective RBP expression or function in genetic disease, and the literature in this area has been expertly reviewed recently [8–10]. Box 1 outlines the molecular processes that might be affected in genetic diseases involving RBPs, and examples representing many of these can be found in the literature, particularly cases affecting pre-mRNA splicing [8,11,12]. Similarly, a spectrum of pathologies and syndromes are known to be caused by RBP defects, with a preponderance of published examples among neurological diseases, muscular atrophies, metabolic disorders, and cancer [9,10]. A fuller insight into the role of RBPs in genetic disease, however, requires a deeper knowledge of the repertoire of physiological RBPs and maps of their dynamic and intricate interactions with RNA targets [2]. Two recent studies have made significant progress in this direction by globally capturing and comprehensively identifying large sets of RBPs bound to mRNA in cultured human cells [13,14]. Here we describe the approaches developed in these studies, outline the data resources they generated, and provide an expanded analysis of the links between RBPs and genetic disease that they uncovered.

mRNA interactome capture

Methods for the unbiased identification of RBPs have long been employed in RNA research. For instance, two studies used hybridization of labeled mRNA preparations to protein arrays in global screens and identified about 200 proteins from budding yeast, including not only many established RBPs but also multiple novel and unexpected candidates [15,16]. Stable isotope labeling by amino acids in cell culture (SILAC) and mass spectrometry (MS) was used to identify RBPs associated with specific immobilized RNA probes *in vitro* [17]. Although the latter approach can yield much useful information, it cannot distinguish direct RNA–protein interactions from indirect protein–protein interactions with RBPs. Procedures that rely on establishing *in vitro* interactions also cannot discriminate between bona fide *in vivo* interactions from non-physiological RNA binding, for example through physicochemical properties of polypeptides. Recently, a new approach was taken to capture the mRNA interactome [13,14] that employed *in vivo* ultraviolet (UV) light-induced crosslinking of proteins

Corresponding author: Preiss, T. (thomas.preiss@anu.edu.au).
Keywords: disease genetics; mRNA metabolism; RNA-binding protein; interactome capture; RNA-binding domain; mass spectrometry; proteomics; gene set enrichment.

 0168-9525/$ – see front matter. http://dx.doi.org/10.1016/j.tig.2013.01.004

Box 1. Genetic disease and RBP function

A genetic lesion might cause heritable disease through affecting RBP function in several ways. Mutations might occur either in the RBP gene itself (Figure Ia) or in a gene expressing an RNA target (Figure Ib). In the latter case, excluding missense, nonsense or frameshift mutations, the lesion might affect RBP binding (loss of binding or gain of aberrant binding specificity) and consequently alter normal RBP function in the processing, utilization, or stability of that RNA. In addition, the mutated RNA might become 'toxic' to the cell, for instance by depleting RBPs or miRNAs, or function in aberrant signaling processes [73]. Mutations in the RBP gene itself might lead to RBP loss or expression of an aberrant variant. These mutated RBP can (i) display an abnormal subcellular localization [54,74], (ii) be defective in binding to RNA targets [51] or protein partners [48], (iii) harbor altered enzymatic activity [62], or (iv) form intracellular protein aggregates [75]. A mutation might interfere with post-translational modifications of the RBP and consequently its normal perception of intracellular signals. Where applicable, a mutation might also affect an enzymatic function of the RBP (e.g., as a kinase or an RNA-modifying enzyme). In all these cases loss of function, as well as gain of an aberrant function, are conceivable effects of the lesion.

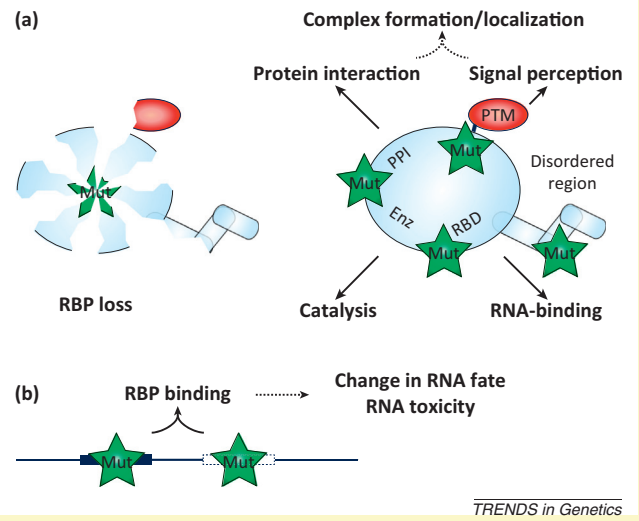

Figure I. Potential consequences of gene mutations for RBP function. Impairment of RNA–protein interaction can occur by mutations in **(a)** the RBP gene that may lead to RBP loss or otherwise affect its properties, or in **(b)** the RNA target that can generate or eliminate an RBP binding site. Enz, enzymatic activity; Mut, mutation; PTM, post-translational modification; PPI, protein–protein interaction; RBD, RNA-binding domain.

to RNA to 'freeze' protein–mRNA interactions in their native cellular context. This was followed by cell lysis, purification of polyadenylated RNA (mostly mRNA) on oligo(dT) beads, stringent washing to remove non-covalently associated proteins, and identification of copurifying proteins by quantitative MS, as summarized in Figure 1 [2,18]. One study used crosslinking aided by a photoactivatable-ribonucleoside {PAR-CL; in this case 4-thiouridine (4SU) or 6-thioguanosine incorporation into cellular RNAs and UV irradiation at 365 nm, based on [19]} and SILAC-MS [13], whereas the other study employed both conventional, short wavelength UV crosslinking (cCL, UV irradiation at 254 nm activating naturally photoreactive nucleosides) as well as PAR-CL and quantitative non-label-based MS [14]. Both studies yielded similar-sized sets of specifically enriched proteins referred to collectively as the mRNA interactome or mRNA-bound proteome (797 derived from HEK-293 [13] and 860 from HeLa cells [14], respectively), although the purified RBPs also potentially include those bound to non-coding polyadenylated RNAs.

There is considerable overlap between the two human mRNA interactomes but, as anticipated given the distinct cellular origin and differences in experimental detail, there are also a number of RBPs unique to each study (545 of a total of 1106 proteins are common to both studies; Figure 2a). Gene set enrichment analyses confirmed expectations in that gene ontology (GO) terms related to RNA-binding are highly enriched in both interactomes, and members of all classical RBP domain families are abundantly represented (~50% of the interactome). Beyond that, the analyses revealed multiple new insights into modes of RNA binding and unexpected connections of RBPs to other cellular functions. For instance, prevalent links of RBPs to DNA damage responses are seen [13], and a high enrichment of repetitive disordered protein regions was noted among RBPs, suggesting a common involvement of such regions in RNA binding [14]. Many individual proteins with unrelated or under-represented GO terms were identified within the mRNA interactomes, including specific DNA-binding factors, kinases, and numerous metabolic enzymes. Unexpectedly, both studies identified as RBPs hundreds of proteins with no RNA-related ontology or domain homology (315 [14] and 245 cases [13], respectively; 352 distinct cases in total). These novel RBPs display enrichment for a range of recognized protein domains that in the light of this evidence warrant testing for putative RNA-binding function (for a comprehensive listing see Table 1 in [18]).

Global protein occupancy profiling

One of the studies also globally identified mRNA regions that interact with RBPs using an approach termed protein occupancy profiling [13]. The authors used PAR-CL followed by oligo(dT) selection; however, they then identified RNA sites protected from mild RNase digestion by crosslinked RBPs using next-generation sequencing. Peaks of mapped reads, as well as diagnostic T to C transitions at crosslinked sites generated during reverse transcription and sequencing of PAR-CL derived RNA fragments [19], were then used to call RBP-bound regions for each detectable transcript. Overall, this analysis identified widespread occupancy by RBPs in all regions of mRNAs.

Crosslinked regions show heightened evolutionary conservation across 44 vertebrate species (based on PhyloP conservation scores [20]) and, where available, good concordance with footprints of individual RBPs obtained by the related PAR-CLIP method [19]. RBP occupancy regions overall exhibited reduced SNP frequencies, suggesting conservation of functionally important sites. Nevertheless, 28 known disease single-nucleotide polymorphisms (SNPs) lie in close proximity (±10 nt) to crosslinking sites. For example, two of these are situated within the 3′UTRs of HOXB5 (homeobox B5) and ZNRD1 (zinc ribbon domain-containing 1) mRNAs and are implicated in childhood obesity and AIDS progression, respectively [21,22]. Extensive RBP occupancy maps generated in different cell types and under different physiological conditions will no doubt

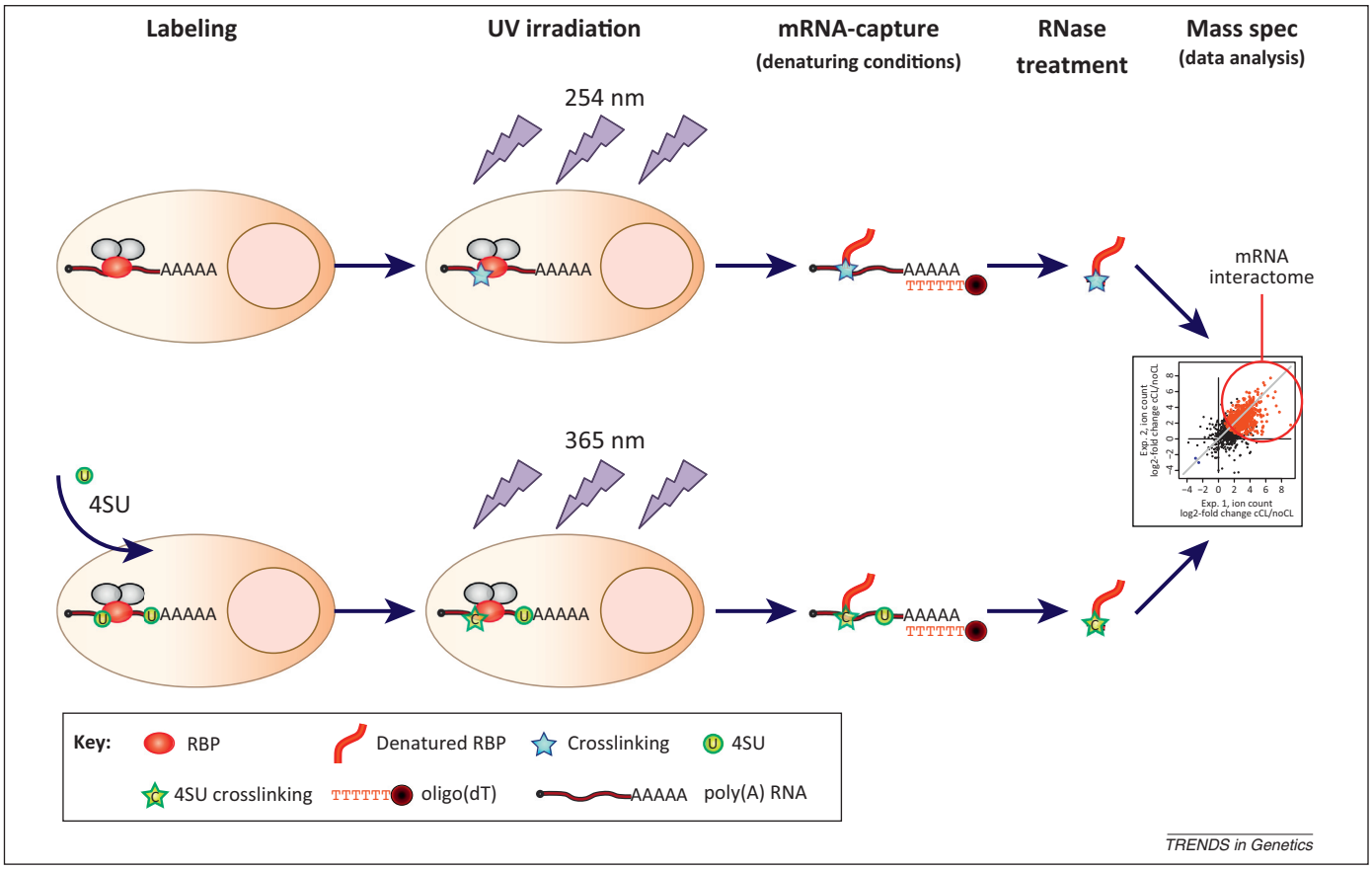

Figure 1. Schematic of the (m)RNA interactome capture workflow. mRNA–protein interactions are preserved in cultured cells employing either the cCL (top) or PAR-CL (bottom) UV crosslinking approach. mRNA–protein complexes are captured on oligo(dT) magnetic beads and stringently washed. Bound proteins are released by RNase treatment and identified by quantitative mass spectrometry (Mass spec).

broaden our view of disease mutations affecting RBP binding sites within target mRNAs.

Cellular roles of disease-associated RBPs

Both mRNA interactome datasets contain numerous proteins with known links to human Mendelian diseases [based on the Online Mendelian Inheritance in Man (OMIM) database; see Box 2 for a list of online resources]. The HeLa cell data revealed 86 such 'OMIM-RBPs', 48 of which were not previously known to bind RNA [14]; in HEK-293 cells, 59 OMIM-RBPs were found, and of these 13 had not been annotated as RNA-binding before [13]. To further explore the roles of RBPs in genetic disease we integrated the datasets from both studies and supplemented them with RBPs annotated with the gene ontology (GO) term 'RNA-binding' (through literature and by domain homology) that were not found by either study. This joint RBP set comprises a total of 1502 RBPs (Figure 2a and Table S1, worksheet 1, in the supplementary material online), including 157 OMIM-RBPs, 63 of which are newly identified by one or both of the mRNA interactome studies (Figure 2b,c). In the following, we use this joint OMIM-RBP set to highlight links between RBPs, their cellular functions, and disease.

The GO terms most highly represented among the OMIM-RBP set relate to the metabolism of mRNA, rRNA, and tRNA, with the molecular functions of tRNA aminoacyl ligase, nuclease, and helicase being most common along with terms relating to the cellular components of the spliceosome and ribosome (Figure 2d–g and Table S1, worksheet 2, in the supplementary material online). Because splicing and translation are highly regulated, defects in their control might alter the level or the function of proteins involved in cell differentiation, cell division, integrity checkpoints, or cellular responses to stimuli, all processes where accurate regulation is essential. Indeed, the importance of alternative splicing to human disease and development is well recognized [8,11,12]. A prominent example is the case of cell-specific alternative splicing of FAS (TNF receptor superfamily, member 6) mRNA that has been linked to cancer predisposition [23]. Several RBPs such as PTB (polypyrimidine tract binding protein), HuR (Hu antigen R), and TIA-1 (T cell intracellular antigen-1-related/like protein) promote inclusion or skipping of exon 6 in FAS mRNA, resulting in either a pro-apoptotic transmembrane form or an anti-apoptotic secreted from of FAS protein [24–26]. Genetic alterations in components of the translational apparatus are linked to cancer and a heterogeneous family of inherited syndromes, known as 'ribosomopathies'. Surprisingly, ribosomopathies present with a high degree of cell and tissue specificity, rather than systemic symptoms [27]. This suggests that ribosomal proteins might regulate protein synthesis in a cell- and tissue-specific manner, or have extra-ribosomal roles in post-transcriptional regulation of gene expression as shown for RPL13a (ribosomal protein L13a) [28]. Owing

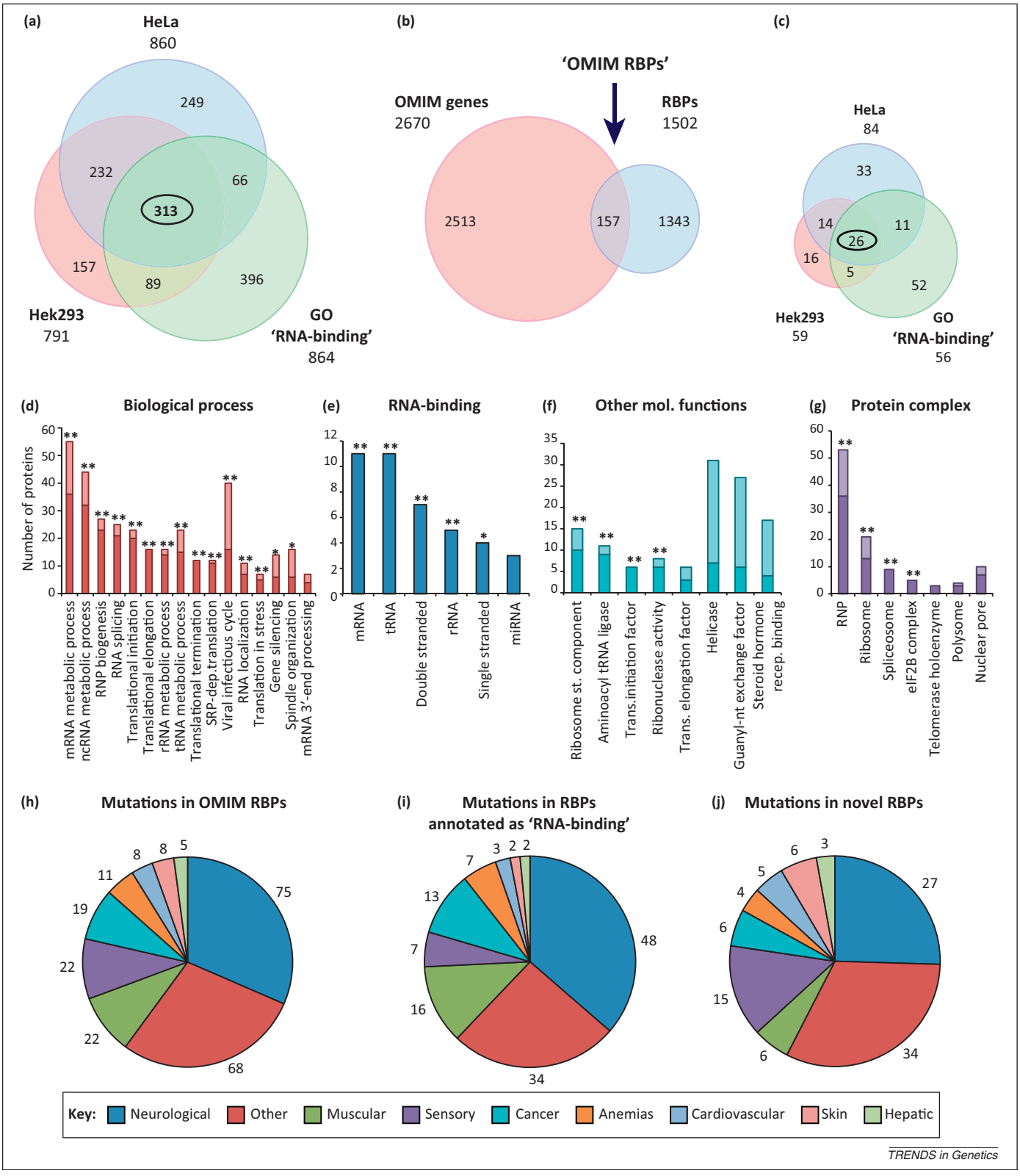

Figure 2. RBPs and Mendelian disease. **(a)** Venn diagram comparison of RBPs identified in HeLa and HEK293 cell interactome data [13,14] with proteins annotated with the GO term 'RNA-binding'. Overlap of groups is significant ($P < 10^{-16}$, Fisher's exact test). **(b)** Overlap of the union set from (a) with proteins listed in OMIM defines the 'OMIM-RBP set'. **(c)** Breakdown of the OMIM-RBPs by original identification. Overlap of groups is significant ($P < 10^{-16}$, Fisher's exact test). **(d–g)** Analysis of GO term enrichment within the OMIM-RBP set (dark color bars) versus all other OMIM proteins annotated with the same GO. GO and Interpro annotations were downloaded from ENSEMBLE (version 68). Enrichment of categories was tested for the OMIM RBPs compared to all proteins annotated in OMIM. *P* values were calculated by Fisher's exact test and corrected for multiple testing by the method of Benjamini–Hochberg; **, $P < 0.01$; * $P < 0.05$. Significantly enriched, non-redundant GO terms are shown. Number of mutations in OMIM-RBPs causing hereditary diseases **(h)**, in RBPs annotated as 'RNA-binding' **(i)** and in novel RBPs identified in mRNA interactome data **(j)** [13,14]. Note that several mutations can occur within the same RBP and that mutations within the same protein can be involved in different diseases. The number of OMIM-RBPs involved in different diseases is shown in Table S2 in the supplementary material online. Abbreviations: dep, dependent; GO, gene ontology; mol, molecular; OMIM, Online Mendelian Inheritance in Man; RBP, RNA-binding protein; st, structural; trans, translation.

Box 2. Online resources

The Gene Ontology (GO) Project	http://www.geneontology.org/
OMIM (Online Mendelian Inheritance in Man) database	http://www.ncbi.nlm.nih.gov/omim
UniProt: Human polymorphisms and disease mutations	http://www.uniprot.org/docs/humsavar
Ensembl Genes 68 (WTSI, UK)	http://www.ensembl.org
Saccharomyces Genome Deletion Project	http://www-sequence.stanford.edu:16080/group/yeast_deletion_project/
STRING (search tool for the retrieval of interacting genes/proteins)	http://string-db.org/
RBPDB, RNA-binding protein database	http://rbpdb.ccbr.utoronto.ca/
DoRiNA Database of Post-transcriptional Regulatory Elements	http://dorina.mdc-berlin.de/rbp_browser/dorina.html
mRNA Interactome Database	http://www.embl.de/mRNAinteractome

to the high requirement for protein synthesis in proliferative cells, eukaryotic initiation factors (eIFs) play an important role in cancer establishment and progression. Increased levels of key eIFs are often found in transformed cells, including the cap-binder eIF4E and the adapter protein eIF4G, which recruits the small ribosomal subunit to mRNA via multiple protein–protein interactions. Multiple efforts are currently being undertaken to develop therapeutic approaches to specifically inhibit translation initiation by targeting eIF4E (via 4E-binding proteins or the mTOR pathway, which controls eIF4E activity) [29].

OMIM-RBPs are also involved in other RNA metabolic processes that have been previously linked to disease, such as host–virus interactions [30], RNA transport [31], gene silencing (via RNA) [32], and mRNA 3′-end processing [33,34] (Figure 2d). Also present in the OMIM-RBP set are proteins from the telomerase complex (see below).

Diseases linked to RBP mutations

In total, the 157 OMIM-RBPs are linked to 221 Mendelian diseases, with a spectrum of pathologies including neuropathies, muscular atrophies, sensorial disorders, and cancer [9,10]. Importantly, the proteins known and annotated as 'RNA-binding', as well as the RBPs newly identified in mRNA interactome datasets, are implicated in a similar spectrum of genetic diseases, suggesting that they are involved in analogous biological functions (Figure 2h–j and Table S2 in the supplementary material online) [14]. In some cases the same or similar diseases are caused by mutation of both known and novel RBPs (e.g., retinitis pigmentosa, familial cirrhosis, Bardet–Biedl syndrome, prostate cancer, Parkinson's disease, amyotrophic lateral sclerosis, Charcot–Marie–Tooth disease) (Table S2 in the supplementary material online), suggesting hitherto unknown links between these proteins. This is further corroborated by analyses with STRING (search tool for the retrieval of interacting genes/proteins) [35,36] indicating high connectivity between novel and previously known RBPs involved in retinitis pigmentosa and other sensorial disorders (Figure 3a). Thus, mRNA interactome studies can reveal wider RBP networks featuring physical and functional connections. Dysfunction of any of the proteins of such a network might cause similar phenotypes and syndromes.

Neurological disorders are the most prominent group of diseases caused by RBP mutations (Figure 2h–j). Of the 59 RBPs linked to hereditary neurological disorders, 18 were newly identified by mRNA interactome capture (Table S2 in the supplementary material online). Response to chemical stimulus ($P = 2 \times 10^{-4}$), neurological system process ($P = 2 \times 10^{-04}$), nervous system development ($P = 2.8 \times 10^{-4}$), heat response (6.32×10^{-4}), regulation of membrane potential ($P = 6.32 \times 10^{-4}$) as well as aminoacyl-tRNA biosynthesis ($P = 2 \times 10^{-4}$), eukaryotic translation initiation factor 2B complex ($P = 0.002$) and guanyl-nucleotide exchange factor activity ($P = 0.015$) are the most enriched biological process GO terms for these proteins when compared to the total joint RBP set (Figure 2a). Analyses with STRING revealed two clusters for RBPs involved in neuropathies: one corresponding to eIF2B and the other to aminoacyl-tRNA biosynthesis (Figure 3b). This resonates with previous reports showing that control of translation plays a key role in memory consolidation and neuronal plasticity [37,38]. Mutations in all five subunits of the eIF2B complex have been shown to be involved in leukoencephalopathy with vanishing white matter [39,40]. The eIF2B complex is a guanine nucleotide exchange factor (GEF) that specifically recycles inactive eIF2–GDP into eIF2–GTP, the active form required for translation initiation [41]. Protein synthesis is inhibited when the α subunit of eIF2 is phosphorylated by kinases in response to stress, such as viral infection or nutrient starvation [42]. Phospho-eIF2α strongly binds to eIF2B, blocking this rate-limiting GEF and preventing the recycling of the growing pool of eIF2–GDP [41]. Recent findings revealed that eIF2α and its regulation by the kinase GCN2 mediate the switch from short to long-term synaptic plasticity and memory [43,44]. Similarly, defects in eIF2B function might imbalance this delicate system and promote deregulation of protein synthesis [40].

RBP mutations and domain architecture

Conventional RBPs are built through combinations of a small number of classical RNA-binding domains (RBDs), a strategy that allows for the modular expansion of RNA-binding affinities and specificities [45]. RNA recognition motifs (RRM), heterogeneous nuclear ribonucleoprotein K-homology domains (KH), and zinc fingers (Znf) are the most frequent RBDs found in RBPs. Despite their general prevalence in the joint RBP set, only a few of the OMIM RBPs harbor these domains (13 of 221 for RRM, 3 of 38 for KH, and 0 of 48 for CCCH Znf; Table S1, worksheet 3, in the supplementary material online). Similarly, common enzymatic activities such as DEAD- and DEAH-box helicases are also under-represented (3 of 86). One plausible explanation for this is that most proteins harboring these classical RBDs play essential roles in RNA metabolism, and their aberrant expression or activity might be lethal. Indeed, conserved genes coding for RBD-containing proteins

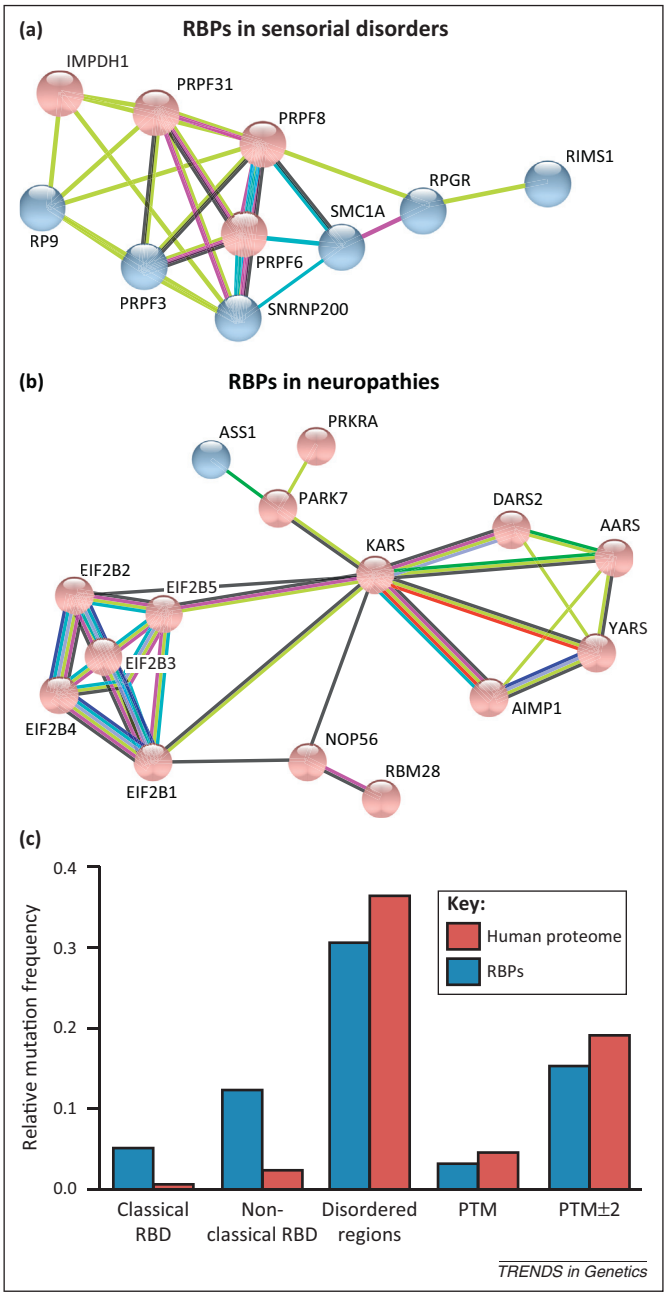

Figure 3. Functional implications of RBP mutations. Protein network connections between newly identified (blue spheres) and previously known (pink spheres) RBPs were explored using STRING. Colored lines indicate the type of evidence: green, genomic neighborhood; red, gene fusion; dark blue, co-occurrence; brown, coexpression; magenta, experiments; light blue, databases; yellow, text mining; light grey, homology. Shown are OMIM RBPs involved in **(a)** retinitis pigmentosa and other ocular disorders, or **(b)** neuropathies. The latter reveals two clusters with functionalities in translation initiation control (via eIF2 guanyl-nucleotide exchange factor activity) and aminoacyl-tRNA biosynthesis. The protein network shows a high connectivity between the two groups of proteins. **(c)** The graph displays the frequency of mutations affecting classical and non-classical RBDs, disordered regions, sites of post-translational modification (PTM), and PTM ± 2. Mutations within RBDs might affect the RNA-binding properties of the RBP. Disordered regions exert important biological functions including RNA binding, subcellular localization, hydrogel formation, etc., and OMIM mutations in these region probably alter their activities. Mutations in or near PTM sites might induce RBP deregulation. Mutations in RBDs are more prevalent in OMIM-RBPs than in the rest of the OMIM proteins, whereas the incidence of mutations in disordered regions and at PTM sites is similar in both groups of OMIM proteins. Abbreviations: RBP, RNA-binding proteins; RBD, RNA-binding domain; OMIM, Online Mendelian Inheritance in Man; STRING, search tool for the retrieval of interacting genes/proteins.

are often essential in yeast ($P < 0.003$, Fisher's exact test; using data from the *Saccharomyces* Genome Deletion Project).

Known disease mutations are not randomly distributed among protein domains in the OMIM-RBP set. For instance, four of the eleven human RBPs harboring Tudor domains are associated with Mendelian disease. This domain usually occurs together with classical RBDs in a given RBP. In the SMN (survival motor neuron) protein the Tudor domain recognizes symmetrically dimethylated arginine residues found in the arginine/glycine-rich C-terminal tails of the Sm proteins, a family of RBPs essential for the RNA splicing machinery [46]. Mutations in *SMN1*, the gene encoding SMN, cause autosomal recessive spinal muscular atrophy (SMA) and four of these mutations map to the Tudor domain. The SMN complex (Box 3) performs an essential role in the maturation of small nuclear ribonucleoproteins (snRNPs), and aberrant function of the Tudor domain in some instances is correlated with decreased recruitment of Sm proteins [47]. Indeed, mutations in other SMN protein–protein interaction domains {binding sites for Gemin-2 [gem (nuclear organelle)-associated protein 2]; SM protein B; and SYNCRIP (synaptotagmin binding, cytoplasmic RNA interacting protein)} have been associated with SMA [48]. This illustrates that aberrant activity of RBPs can be induced by defects in protein–protein binding interfaces, promoting the assembly of functionally impaired complexes on targeted RNAs. Therefore, by furthering our understanding of the protein–protein interaction networks of RBPs we might uncover important clues about disease etiology.

Only nine disease-associated mutations are located within classical RBDs found among OMIM-RBPs (Figure 3c and Table S3, worksheet 1, in the supplementary material online); 7 of these reside in the PUA (pseudouridine synthase and archeosine transglycosylase) RNA-binding domain of dyskerin (dyskeratosis congenita 1 or *DKC1*). Dyskerin is a subunit of the telomerase complex which maintains telomeres at the ends of chromosomes that would otherwise be gradually lost during replication (Figure 2g; Box 3). *DKC1* mutations cause X-linked dyskeratosis congenita, which is associated with defects in telomerase activity. Another explanation for the small number of OMIM mutations mapping to classical RBDs could be their structural properties. RRM, KH, and Znf domains establish multiple interactions with the RNA, and these domains are often found in multiple copies per protein. Therefore, a single point mutation is unlikely to abolish classical RBD activities completely. Abrogation of protein–RNA interaction might require accumulation of multiple point mutations to interfere with RNA binding, and combinatorial mutations are difficult to detect in association studies. Nevertheless, two OMIM mutations were identified in the RRM of RBM28 (RNA-binding motif protein 28) and TARDBP (TAR DNA-binding protein), causing amyotrophic lateral sclerosis or alopecia, neurological defects, and endocrinopathy syndrome, respectively. Mutations in TARDBP (D169G) and RBM28 (L351P) cause aberrant function of these RBPs (Table S3, worksheet 1, in the supplementary material online) [49,50], although it is unknown whether the RNA-binding properties or other

Box 3. Multi-subunit RBP complexes

The human telomerase holoenzyme complex (Figure Ia) is composed of the human telomerase reverse transcriptase (hTERT), human telomeric RNA (hTR), dyskerin (DKC1), and additional proteins such as NOP10 (nucleolar protein family A member 3), NHP2 (nucleolar protein family A member 2), GAR1 (nucleolar protein family A member 1), and WRAP53 (WD repeat-containing, antisense to TP53), which are important in telomerase assembly, stability, localization, and function [54,76]. Because telomeric DNA is gradually lost during DNA replication, the telomerase holoenzyme is essential to maintain telomere integrity and its deregulation has been linked to disease [77,78]. Interactome capture identified GAR1, DKC1, and NHP2, suggesting that they might also bind poly(A)$^+$ RNAs. The schematic in Figure Ib shows the human survival motor neuron (SMN) complex, where the SMN protein (blue) interacts directly and indirectly with members of the Gemini body (gem, nuclear organelle)-associated protein family (gemins 2 to 8) and STRAP (serine/threonine kinase receptor-associated protein; also named UNRIP). Red circles represent gemins probed to interact with SMN proteins by at least two independent assays. Yellow circles represent proteins reported to bind SMN proteins by a single assay. Grey circles represent proteins that do not interact directly with SMN proteins. Modified from [79]. The (SMN) complex acts in the cytoplasm as an 'assemblyosome' of ribonucleoproteins, recruiting the proteins of the Smith (Sm) family to form a ring around the small nuclear RNAs (snRNAs) [80]. Assembled snRNPs are imported to the nucleus and function in splicing. snRNAs interact with the survival motor neuron proteins (SMN1 and SMN2), forming the SMN complex. In addition, the SMN complex might also be involved in mRNP complex localization, perhaps explaining the isolation of some of its components (gemin 5 and STRAP) by mRNA interactome capture.

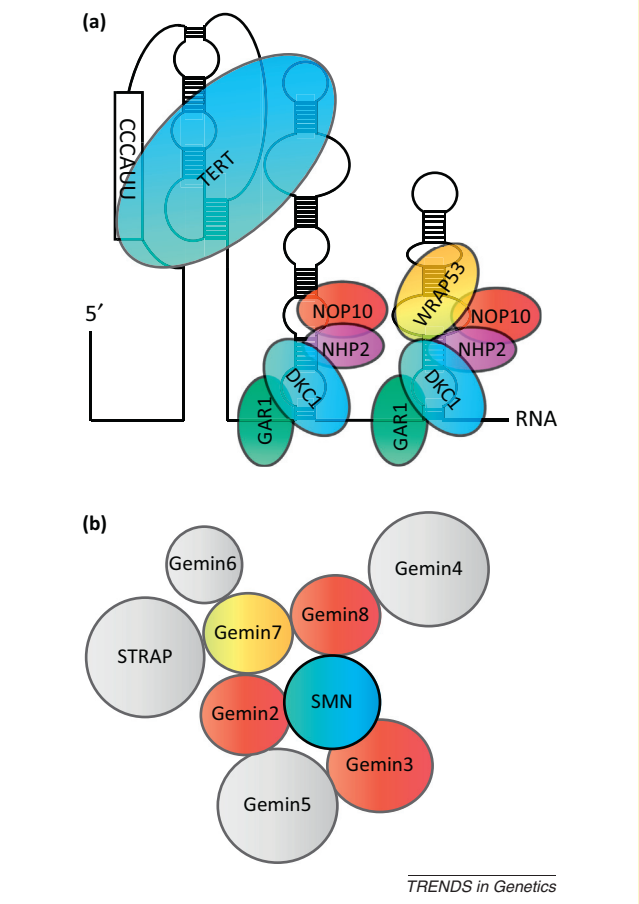

Figure I. Schematic representation of two multi-subunit RBP complexes. **(a)** The human telomerase holoenzyme complex. **(b)** The human survival motor neuron (SMN) complex.

biological roles are affected. The functional role of a mutation in the second KH domain (I340N) of FMR1 (fragile X mental retardation 1 protein) is better understood. It abolishes RNA binding, and this has been proven in a mouse model [51]. However, because this mutation is not annotated in the OMIM database, it was not included in Table S3 in the supplementary material online.

Apart from the RBDs listed in [45], any other protein domain proven to bind RNA in biochemical or structural studies in at least one well-characterized example can be referred to as a non-classical RBD [14]. A larger number of disease-associated mutations have been found in OMIM-RBPs with non-classical RBDs, affecting 13 different RBPs (Figure 3c and Table S3, worksheet 1, in the supplementary material online). One such non-classical RBD is the WD40 domain that usually acts as a protein–protein interface [52]. However, it promotes RNA binding in the SMN complex protein Gemin-5 and 23 proteins harboring WD40 domains were found in the HeLa mRNA interactome [14,53]. Two mutations occur in the WD40 domain of the telomerase component WRAP53 (WD repeat containing, antisense to TP53, also known as telomerase Cajal body protein 1), which consists of repeats of a 31–60 residue conserved motif (the WD40 motif) that form β-propeller structures. In contrast to the mutations in other telomerase components described above, these mutations cause dyskeratosis congenita without affecting telomerase activity. Instead, the subcellular distribution of the complex is altered from Cajal bodies to nucleoli [54]. How WRAP53 mutations affect WD40 domain function and telomerase relocalization is still unknown and calls for further exploration.

Role of disordered protein regions

Large portions of RBPs identified by mRNA interactome capture are intrinsically disordered and lack stable 3D structure under native conditions [14]. These disordered regions might undergo 'induced fit' transitions following interactions with binding partners and are frequently endowed with high functional density, containing multiple interaction interfaces, including facilitation of RNA folding as RNA chaperones [55,56], hydrogel formation [57,58], and RNA binding [59]. Disease-associated mutations are often found in disordered regions of the human proteome (Figure 3c), 55 of them affecting OMIM-RBPs (Table S3, worksheet 2, in the supplementary material online). In the latter cases, the two amino acids most frequently mutated are arginine (R, 17 cases) and glycine (G, 10 cases), which often co-occur in disordered regions of RBPs, forming a repetitive motif known as an RGG-box. FMRP, encoded by

the *FMR1* gene, binds to guanine-quadruplex-forming sequences via the RGG-box [60], where the $R_{534}GGGGR_{539}$ peptide is positioned along the major groove of the RNA duplex and forms a sharp turn at the duplex–quadruplex junction. Mutations in any of these R residues or in the poly(G) spacer impair the RNA-binding activity of the RGG-box [59]. These findings suggest that RGG-boxes strongly rely on their primary sequence and mutations might affect their RNA-binding properties. The FUS (fused in sarcoma) protein harbors large disordered regions containing RGG-boxes. The R244C mutation disrupts an RGG box (G.**R**.GGGRGGRGG > G.**C**.GGGRGGRGG), causing amyotrophic lateral sclerosis type 6 (Table S3, worksheet 2, in the supplementary material online) [61]. Although the molecular consequences of this mutation are still unknown, it might alter the RNA-binding specificity or/and affinity of FUS. Further efforts should be undertaken to understand the role of RBP low-complexity sequences in RNA metabolism and human diseases.

Regulation of RBP activity by post-translational modifications (PTMs)

Disease-associated mutations are also found within OMIM-RBPs at PTM sites or more frequently within two amino acids upstream or downstream of such sites (PTM ± 2) (Figure 3c, Table S3, worksheet 3, in the supplementary material online). Because PTMs usually control protein activity, localization, or turnover, mutated RBPs might behave aberrantly, generating a pathological environment. The F1127L mutation of telomerase reverse transcriptase, which is in close proximity to a phosphoserine, causes dyskeratosis congenita; in this case shortened telomeres are found in patients even though telomerase activity *per se* is not affected. Replacement of an aromatic for an aliphatic residue might alter the recognition of the phosphosite, affecting important properties of this protein such as localization and stability [62]. Change of phosphoserine to tyrosine at amino acid 1217 of another OMIM-RBP, BRCA1 (breast cancer 1 early-onset protein), promotes increased predisposition to breast and ovarian cancer development (Table S3, worksheet 3, in the supplementary material online) [63]. BRCA1 is an E3 ligase that has been associated with DNA damage response [64], but its specific role in RNA biology is thus far unknown. Because this mutation occurs at a phosphosite, the lack of phosphorylation at this residue might impact upon the regulation of the protein.

RNA binding and the link to metabolism

The REM (RNA–enzyme–metabolite) hypothesis proposes the existence of regulatory links between gene expression and intermediary metabolism mediated by bifunctional RNA-binding metabolic enzymes [65]. Metabolites (substrates or cofactors) might regulate the RNA-binding and catalytic activity of the bifunctional enzyme/RBP. Sporadic reports accumulated over several decades have supported the notion of RNA binding by multiple metabolic enzymes, reviewed in [66,67], as have recent *in vitro* system-wide screens for yeast RBPs [15,16]. In most cases a physiological role is not yet known, a notable exception being cytoplasmic aconitase (ACO1; better known as iron regulatory protein 1, IRP1) [68]. The HeLa mRNA interactome revealed that (at least) 17 enzymes in central metabolic pathways bind to RNA in living cells [14], to which the HEK293 interactome adds further examples [13]. Some of these RNA-binding metabolic enzymes have been linked to hereditary diseases (Table S2 in the supplementary material online). Interestingly, in the cases of IMPDH1 (inosine 5′-monophosphate dehydrogenase 1) and HSD17B10 (hydroxysteroid 17β-dehydrogenase 10), the severity of the disease caused by mutations does not correlate with impairment of the catalytic activity [69,70], suggesting that other protein functions, such as RNA binding, might be affected in these pathological contexts. In particular, most of the IMPDH1 mutations identified in autosomal dominant retinitis pigmentosa prevent single-chain nucleic acid binding [71], thus affecting the RBP properties of the protein. Therefore, the existing mRNA interactome datasets already support the REM hypothesis in that they demonstrate the existence of an RNA–enzyme axis *in vivo* [13,14]. Further mRNA interactome analyses in different cellular contexts and under different metabolic conditions might expose additional enzymes with RBP properties and uncover the role of metabolites in regulating interactions in REM networks.

Concluding remarks

mRNA interactome capture was developed to generate comprehensive surveys of the (m)RNA-binding proteins of living cells. It builds on, and complements, methods for global identification of the RNA targets of a given RBP, such as crosslink/immunoprecipitation (CLIP) protocols, that have recently come to the fore. These approaches can be deployed in a highly synergistic fashion to survey networks of protein–RNA interactions in different cellular contexts, for example, by focusing on the aberrant pathophysiological cellular conditions associated with common diseases such as cancer, cardiac disease, diabetes, and infection, or responses to drugs. Interactome capture will detect disease-associated changes in the RBP profile of cells and CLIP will then identify the RNA targets and *cis*-regulatory binding sites of RBPs of interest. Together, the two approaches are destined to uncover new avenues for therapy.

Box 4. Outstanding questions

- Why is it that among the many steps of RNA metabolism, splicing and translation are most prominently linked to hereditary diseases?
- Considering that around half of the proteins within the mRNA interactome datasets do not harbor canonical RBDs, what is the real incidence of disease-associated mutations within RNA-binding architectures?
- Because the severity of the disease does not correlate with changes in the enzymatic activity of metabolic enzymes that also act as RBPs, is the RNA-binding activity of these proteins affected in pathological states?
- Because disease-associated mutations are more frequently found in disordered regions than in globular RBDs, what is the biological role of those motifs?
- PTM sites within RBPs are also mutated in disease. What are the roles of these PTMs in RBP function, localization, and expression?

The mRNA interactome studies to date already offer informative systems-wide views on the mRNA interactomes of human cells and substantiate the established disease links of particular aspects of RNA metabolism, as well as the prevalence of specific disease spectra resulting from mutations in RBP-coding genes. Importantly, many among the pool of newly identified RBPs were encoded by known disease genes. This raises the prospect that a subset of disease mutations might affect RNA binding or other previously unknown RNA-related functions of the encoded proteins (Box 4). These and the many other exciting implications of these new resources now await exploration by future research.

Update

A study describing the mRNA interactome of *S. cerevisiae* under glucose deprivation stress using similar *in vivo* capture methodology has just appeared [72]. It identifies 120 mRNA-binding proteins, of which 92 have human orthologs; 72 of these in turn were also identified as RBPs by the human studies detailed above [13,14]. Among 17 proteins new to RNA binding were kinases, proteins involved in DNA biology, and several metabolic enzymes, again extending trends seen in human cells to yeast.

Acknowledgments

We thank Markus Landthaler and Yalin Liao for their suggestions on this manuscript. M.W.H. is supported by a European Research Council Advanced Grant and by the Virtual Liver Network of the German Ministry for Science and Education. T.P. acknowledges grant support from the National Health and Medical Research Council of Australia and the Australian Research Council.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.tig.2013.01.004.

References

1 Glisovic, T. *et al.* (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* 582, 1977–1986
2 Gebauer, F. *et al.* (2012) From cis-regulatory elements to complex RNPs and back. *Cold Spring Harb. Perspect. Biol.* 4, a012245
3 Keene, J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.* 8, 533–543
4 Mayr, C. and Bartel, D.P. (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138, 673–684
5 Wang, E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476
6 Danckwardt, S. *et al.* (2011) p38 MAPK controls prothrombin expression by regulated RNA 3′ end processing. *Mol. Cell* 41, 298–310
7 Hynes, C. *et al.* (2012) miRNAs in cardiac disease: sitting duck or moving target? *IUBMB Life* 64, 872–878
8 Cooper, T.A. *et al.* (2009) RNA and disease. *Cell* 136, 777–793
9 Lukong, K.E. *et al.* (2008) RNA-binding proteins in human genetic disease. *Trends Genet.* 24, 416–425
10 Darnell, R.B. (2010) RNA regulation in neurologic disease and cancer. *Cancer Res. Treat.* 42, 125–129
11 Norris, A.D. and Calarco, J.A. (2012) Emerging roles of alternative pre-mRNA splicing regulation in neuronal development and function. *Front. Neurosci.* 6, 122
12 Padgett, R.A. (2012) New connections between splicing and human disease. *Trends Genet.* 28, 147–154
13 Baltz, A.G. *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* 46, 674–690
14 Castelló, A. *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 149, 1393–1406
15 Scherrer, T. *et al.* (2010) A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. *PLoS ONE* 5, e15499
16 Tsvetanova, N.G. *et al.* (2010) Proteome-wide search reveals unexpected RNA-binding proteins in *Saccharomyces cerevisiae*. *PLoS ONE* 5, e12671
17 Butter, F. *et al.* (2009) Unbiased RNA–protein interaction screen by quantitative proteomics. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10626–10631
18 Sibley, C.R. *et al.* (2012) The greatest catch: big game fishing for mRNA-bound proteins. *Genome Biol.* 13, 163
19 Hafner, M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129–141
20 Pollard, K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20, 110–121
21 Bradfield, J.P. *et al.* (2012) A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat. Genet.* 44, 526–531
22 Limou, S. *et al.* (2009) Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J. Infect. Dis.* 199, 419–426
23 Lima, L. *et al.* (2008) Association between FAS polymorphism and prostate cancer development. *Prostate Cancer Prostatic Dis.* 11, 94–98
24 Izquierdo, J.M. (2008) Hu antigen R (HuR) functions as an alternative pre-mRNA splicing regulator of Fas apoptosis-promoting receptor on exon definition. *J. Biol. Chem.* 283, 19077–19084
25 Izquierdo, J.M. (2010) Cell-specific regulation of Fas exon 6 splicing mediated by Hu antigen R. *Biochem. Biophys. Res. Commun.* 402, 324–328
26 Izquierdo, J.M. *et al.* (2005) Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol. Cell* 19, 475–484
27 Ruggero, D. (2012) Translational control in cancer etiology. *Cold Spring Harb. Perspect. Biol.* 4, a012336
28 Kapasi, P. *et al.* (2007) L13a blocks 48S assembly: role of a general initiation factor in mRNA-specific translational control. *Mol. Cell* 25, 113–126
29 Malina, A. *et al.* (2012) Emerging therapeutics targeting mRNA translation. *Cold Spring Harb. Perspect. Biol.* 4, a012377
30 Brass, A.L. *et al.* (2008) Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 319, 921–926
31 Dictenberg, J.B. *et al.* (2008) A direct role for FMRP in activity-dependent dendritic mRNA transport links filopodial-spine morphogenesis to fragile X syndrome. *Dev. Cell* 14, 926–939
32 Osman, A. (2012) MicroRNAs in health and disease – basic science and clinical applications. *Clin. Lab.* 58, 393–402
33 Danckwardt, S. *et al.* (2008) 3′ end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J.* 27, 482–498
34 Danckwardt, S. *et al.* (2006) 3′ end processing of the prothrombin mRNA in thrombophilia. *Acta Haematol.* 115, 192–197
35 Szklarczyk, D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39, D561–D568
36 von Mering, C. *et al.* (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 33, D433–D437
37 Costa-Mattioli, M. *et al.* (2009) Translational control of long-lasting synaptic plasticity and memory. *Neuron* 61, 10–26
38 Gkogkas, C. *et al.* (2010) Translational control mechanisms in long-lasting synaptic plasticity and memory. *J. Biol. Chem.* 285, 31913–31917
39 Bugiani, M. *et al.* (2010) Leukoencephalopathy with vanishing white matter: a review. *J. Neuropathol. Exp. Neurol.* 69, 987–996
40 Pavitt, G.D. and Proud, C.G. (2009) Protein synthesis and its control in neuronal cells with a focus on vanishing white matter disease. *Biochem. Soc. Trans.* 37, 1298–1310
41 Mohammad-Qureshi, S.S. *et al.* (2008) Clues to the mechanism of action of eIF2B, the guanine-nucleotide-exchange factor for translation initiation. *Biochem. Soc. Trans.* 36, 658–664
42 Garcia, M.A. *et al.* (2007) The dsRNA protein kinase PKR: virus and cell control. *Biochimie* 89, 799–811
43 Costa-Mattioli, M. *et al.* (2005) Translational control of hippocampal synaptic plasticity and memory by the eIF2alpha kinase GCN2. *Nature* 436, 1166–1173

44 Costa-Mattioli, M. *et al.* (2007) eIF2alpha phosphorylation bidirectionally regulates the switch from short- to long-term synaptic plasticity and memory. *Cell* 129, 195–206
45 Lunde, B.M. *et al.* (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* 8, 479–490
46 Sprangers, R. *et al.* (2003) High-resolution X-ray and NMR structures of the SMN Tudor domain: conformational variation in the binding site for symmetrically dimethylated arginine residues. *J. Mol. Biol.* 327, 507–520
47 Selenko, P. *et al.* (2001) SMN tudor domain structure and its interaction with the Sm proteins. *Nat. Struct. Biol.* 8, 27–31
48 Sun, Y. *et al.* (2005) Molecular and functional analysis of intragenic SMN1 mutations in patients with spinal muscular atrophy. *Hum. Mutat.* 25, 64–71
49 Kabashi, E. *et al.* (2008) TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat. Genet.* 40, 572–574
50 Nousbeck, J. *et al.* (2008) Alopecia, neurological defects, and endocrinopathy syndrome caused by decreased expression of RBM28, a nucleolar protein associated with ribosome biogenesis. *Am. J. Hum. Genet.* 82, 1114–1121
51 Zang, J.B. *et al.* (2009) A mouse model of the human fragile X syndrome I304N mutation. *PLoS Genet.* 5, e1000758
52 Stirnimann, C.U. *et al.* (2010) WD40 proteins propel cellular networks. *Trends Biochem. Sci.* 35, 565–574
53 Lau, C.K. *et al.* (2009) Gemin5–snRNA interaction reveals an RNA binding function for WD repeat domains. *Nat. Struct. Mol. Biol.* 16, 486–491
54 Batista, L.F. *et al.* (2011) Telomere shortening and loss of self-renewal in dyskeratosis congenita induced pluripotent stem cells. *Nature* 474, 399–402
55 Dyson, H.J. and Wright, P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6, 197–208
56 Tompa, P. and Csermely, P. (2004) The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.* 18, 1169–1175
57 Han, T.W. *et al.* (2012) Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell* 149, 768–779
58 Kato, M. *et al.* (2012) Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* 149, 753–767
59 Phan, A.T. *et al.* (2011) Structure–function studies of FMRP RGG peptide recognition of an RNA duplex–quadruplex junction. *Nat. Struct. Mol. Biol.* 18, 796–804
60 Darnell, J.C. *et al.* (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell* 107, 489–499
61 Vance, C. *et al.* (2009) Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science* 323, 1208–1211
62 Xin, Z.T. *et al.* (2007) Functional characterization of natural telomerase mutations found in patients with hematologic disorders. *Blood* 109, 524–532
63 Easton, D.F. *et al.* (2007) A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am. J. Hum. Genet.* 81, 873–883
64 Aressy, B. and Greenberg, R.A. (2012) DNA damage: placing BRCA1 in the proper context. *Curr. Biol.* 22, R806–R808
65 Hentze, M.W. and Preiss, T. (2010) The REM phase of gene regulation. *Trends Biochem. Sci.* 35, 423–426
66 Ciesla, J. (2006) Metabolic enzymes that bind RNA: yet another level of cellular regulatory network? *Acta Biochim. Pol.* 53, 11–32
67 Hentze, M.W. (1994) Enzymes as RNA-binding proteins: a role for (di)nucleotide-binding domains? *Trends Biochem. Sci.* 19, 101–103
68 Muckenthaler, M.U. *et al.* (2008) Systemic iron homeostasis and the iron-responsive element/iron-regulatory protein (IRE/IRP) regulatory network. *Annu. Rev. Nutr.* 28, 197–213
69 Hedstrom, L. (2008) IMP dehydrogenase-linked retinitis pigmentosa. *Nucleosides Nucleotides Nucleic Acids* 27, 839–849
70 Rauschenberger, K. *et al.* (2010) A non-enzymatic function of 17beta-hydroxysteroid dehydrogenase type 10 is required for mitochondrial integrity and cell survival. *EMBO Mol. Med.* 2, 51–62
71 Mortimer, S.E. and Hedstrom, L. (2005) Autosomal dominant retinitis pigmentosa mutations in inosine 5′-monophosphate dehydrogenase type I disrupt nucleic acid binding. *Biochem. J.* 390, 41–47
72 Mitchell, S.F. *et al.* (2013) Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol.* 20, 127–133
73 Wang, G.S. and Cooper, T.A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* 8, 749–761
74 Kwiatkowski, T.J., Jr *et al.* (2009) Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* 323, 1205–1208
75 Keller, B.A. *et al.* (2012) Co-aggregation of RNA binding proteins in ALS spinal motor neurons: evidence of a common pathogenic mechanism. *Acta Neuropathol.* 124, 733–747
76 Cohen, S.B. *et al.* (2007) Protein composition of catalytically active human telomerase from immortal cells. *Science* 315, 1850–1853
77 Blasco, M.A. (2005) Telomeres and human disease: ageing, cancer and beyond. *Nat. Rev. Genet.* 6, 611–622
78 Shay, J.W. and Wright, W.E. (2011) Role of telomeres and telomerase in cancer. *Semin. Cancer Biol.* 21, 349–353
79 Cauchi, R.J. (2010) SMN and Gemins: 'we are family'... or are we?: insights into the partnership between Gemins and the spinal muscular atrophy disease protein SMN. *Bioessays* 32, 1077–1089
80 Paushkin, S. *et al.* (2002) The SMN complex, an assemblyosome of ribonucleoproteins. *Curr. Opin. Cell Biol.* 14, 305–312