A census of human RNA-binding proteins

Article *in* Nature Reviews Genetics · November 2014

DOI: 10.1038/nrg3813 · Source: PubMed

3 authors, including:

Stefanie Gerstberger
The Rockefeller University
14 PUBLICATIONS **1,242** CITATIONS
SEE PROFILE

Markus Hafner
National Institutes of Health
232 PUBLICATIONS **9,804** CITATIONS
SEE PROFILE

ANALYSIS

A census of human RNA-binding proteins

Stefanie Gerstberger[1], Markus Hafner[2] and Thomas Tuschl[1]

Abstract | Post-transcriptional gene regulation (PTGR) concerns processes involved in the maturation, transport, stability and translation of coding and non-coding RNAs. RNA-binding proteins (RBPs) and ribonucleoproteins coordinate RNA processing and PTGR. The introduction of large-scale quantitative methods, such as next-generation sequencing and modern protein mass spectrometry, has renewed interest in the investigation of PTGR and the protein factors involved at a systems-biology level. Here, we present a census of 1,542 manually curated RBPs that we have analysed for their interactions with different classes of RNA, their evolutionary conservation, their abundance and their tissue-specific expression. Our analysis is a critical step towards the comprehensive characterization of proteins involved in human RNA metabolism.

Ribonucleoprotein
(RNP). Protein (or proteins) complexed with RNA as an obligate binding partner.

RNA-binding proteins
(RBPs). Proteins involved in the maturation, stability, transport and degradation of cellular RNAs. RBPs directly bind to RNA or are integral parts of macromolecular protein complexes that bind to RNA.

[1]*Howard Hughes Medical Institute and Laboratory for RNA Molecular Biology, The Rockefeller University, 1230 York Ave, New York 10065, USA.*
[2]*Laboratory of Muscle Stem Cells and Gene Regulation, National Institute of Arthritis and Musculoskeletal and Skin Disease, National Institutes of Health, Bethesda, Maryland 20892, USA.*
Correspondence to T.T.
e-mail:
ttuschl@rockefeller.edu
doi:10.1038/nrg3813
Published online
4 November 2014

Post-transcriptional gene regulation (PTGR) is essential to sustain cellular metabolism, coordinating maturation, transport, stability and degradation of all classes of RNAs (FIG. 1). Mechanistically, each of these events is regulated by the formation of different ribonucleoprotein (RNP) complexes with RNA-binding proteins (RBPs) at their core. Initially, it was thought that RNA mainly served either as the template, in the form of mRNA, or as an adaptor or a structural component during protein synthesis, provided by tRNAs and ribosomal RNAs. With the discovery of catalytic RNAs and a multitude of non-coding RNA (ncRNA) species, it was recognized that RNA is a highly versatile molecule that carries out many regulatory functions in the cell, either by acting as a guide to recognize RNA sequence motifs or RNA recognition elements present in their target RNAs, or by functioning as a scaffold and assembly platform for recruiting proteins to act synergistically[1]. The characterization of the proteins transiently or stably interacting with RNAs is a prerequisite for the dissection of RNA regulatory processes.

The recent development of large-scale quantitative methods, especially next-generation sequencing and modern protein mass spectrometry[2–6], facilitates genome-wide identification of RBPs, their protein cofactors and their RNA targets. Deep-sequencing approaches using immunoprecipitation of RBPs, with or without *in vivo* RNA–protein crosslinking (crosslinking and immunoprecipitation followed by sequencing (CLIP–seq) and RNA immunoprecipitation and sequencing (RIP-seq), respectively)[2,3], as well as *in vitro* evolution methods[7,8], revealed the binding ranges of RBPs and showed that many RBPs bind to thousands of transcripts in cells at defined binding sites.

Despite the growing amount of data collected on RBPs, many questions remain to be answered. Researchers still have an incomplete understanding of how binding specificity is achieved and how the regulatory function of an individual RBP is influenced by synergy and competition with other RBPs. We argue that a balanced approach of detailed biochemical and functional studies paired with complex systems-biology methods will ultimately lead to an understanding of the principles underlying PTGR networks.

Although much of the published research centres on mRNA-binding proteins (mRBPs) and messenger RNPs, PTGR is not limited to mRNA maturation and regulation; it also includes processes acting on ncRNAs. In this respect, it may not be surprising that, among the ~150 RBPs listed in the Online Mendelian Inheritance in Man (OMIM) database as being linked to human diseases, only one-third are described as directly binding mRNAs; the others target diverse ncRNAs[9].

Here, we present a census of 1,542 human RBPs that interact with all known classes of RNAs, detail their families and evolutionary conservation across species, and analyse their expression across tissues and their potential roles in developmental processes. This catalogue of RBPs will guide future analyses of RBPs and provide an overview of known RNA pathways and their protein components.

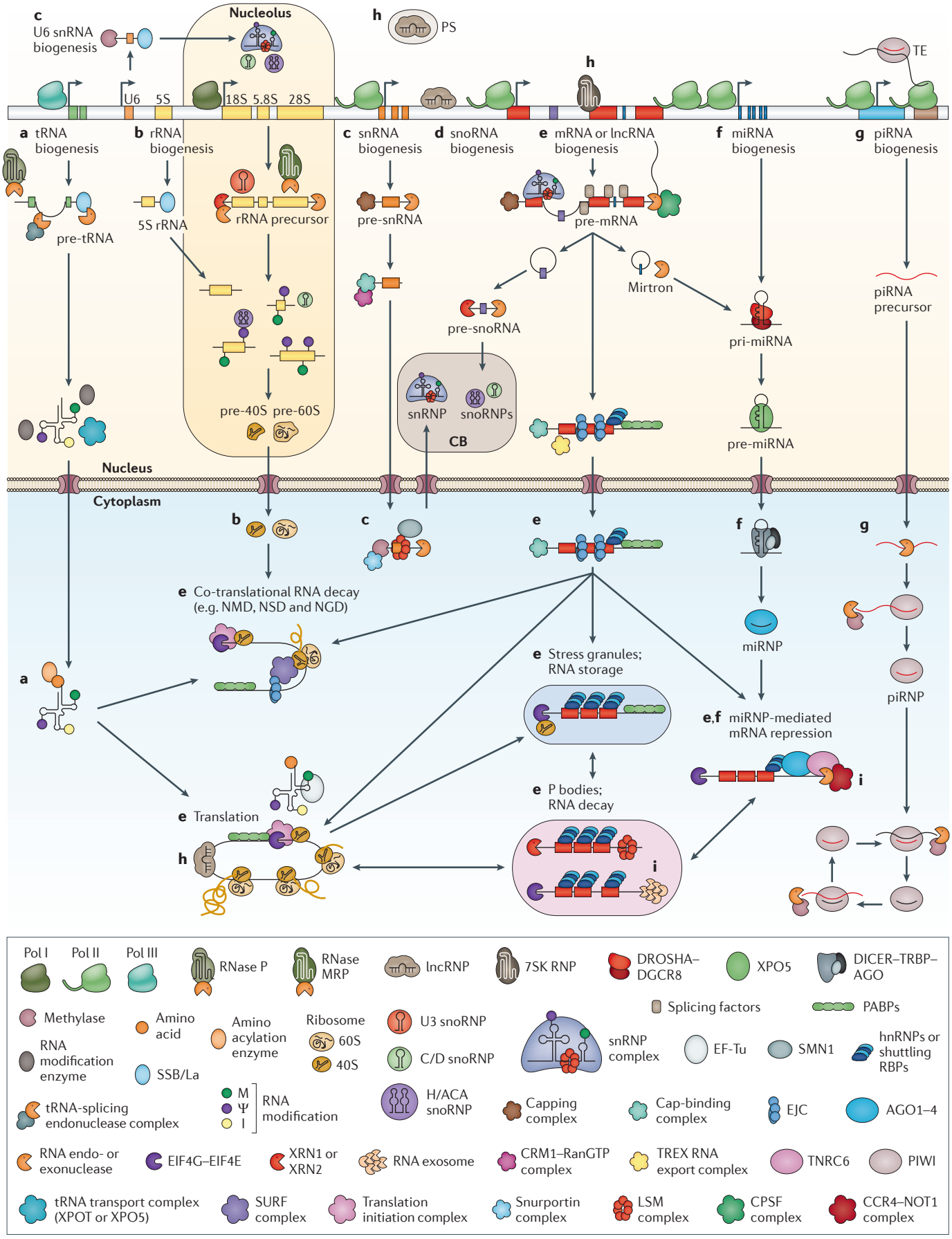


Figure 1 | **Overview of the main post-transcriptional gene regulation pathways in eukaryotes.** An overview is given for the biogenesis, decay and function of the most abundant RNAs: tRNAs, ribosomal RNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), mRNAs, microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs) and long non-coding RNAs (lncRNAs). Processes are described from left to right. Referenced gene names and complexes in the figure are listed in Supplementary information S3 (table) and within the listed references. **a** | tRNAs are transcribed by RNA polymerase III (Pol III); the 5′ leader and 3′ trailer sequences are removed, introns are spliced, and the ends are joined. CCA nucleotides are added to 3′ ends, and nucleotide modifications — such as methylation (M), pseudouridylation (ψ) and deamination of adenosines to inosines (I) — are introduced before tRNA aminoacylation[195]. **b** | The 5S rRNA is transcribed by Pol III, whereas 28S, 18S and 5.8S rRNAs are transcribed as one transcript by Pol I. The precursor is processed by RNA exonucleases, endonucleases and the ribonucleoprotein (RNP) RNase MRP, guided by U3 small nucleolar RNP (snoRNP). Nucleotide modifications are introduced by snoRNPs. rRNAs are assembled together with ribosomal proteins into ribosomal precursor complexes in the nucleus and transported to the cytoplasm, where they mature to functional ribosomes[92,196,197]. **c** | Most snRNAs are transcribed by Pol II, capped and processed in the nucleus. When exported to the cytoplasm, they undergo methylation and assemble with LSM proteins into small nuclear ribonucleic particles (snRNPs) in a process aided by the survival motor neuron 1 (SMN1). These snRNPs are re-imported into the Cajal body (CB) within the nucleus, where they undergo final maturation and snRNP assembly[81]. U6 and U6atac snRNAs are transcribed by Pol III and are alternatively processed in the nucleus and the nucleolus[198]. Mature snRNPs form the core of the spliceosome. **d** | snoRNAs and small Cajal body-specific RNAs (scaRNAs) are processed from mRNA introns, capped and modified before they assemble into snoRNPs or scaRNPs in the CB. snoRNPs and scaRNPs carry out methylation and pseudouridylation of rRNAs, snoRNAs and snRNAs, or function in rRNA processing (for example, processing of U3 snoRNA)[81]. **e** | mRNAs are transcribed by Pol II, capped, spliced, edited and polyadenylated in the nucleus. Correctly matured mRNAs are exported into the cytoplasm. Regulatory RNA-binding proteins (RBPs) control correct translation, monitor stability, decay and localization, and shuttle mRNAs between actively translating ribosomes, stress granules and P bodies[37,141,142,199–202]. **f** | miRNAs are either transcribed from separate genes by Pol II as long primary miRNA (pri-miRNA) transcripts or expressed from mRNA introns (mirtrons) and processed into hairpin pre-miRNAs in the nucleus. After transport into the cytoplasm, they are processed into 21-nucleotide-long double-stranded RNAs. One strand is incorporated into Argonaute (AGO) proteins (forming miRNA-containing RNPs (miRNPs)) and guides them to partially complementary target mRNAs to recruit deadenylases and repress translation[203]. **g** | piRNAs are ~28-nucleotides-long, germline-specific small RNAs. Primary piRNAs are directly processed and assembled from long, Pol II-transcribed precursor transcripts, whereas secondary piRNAs are generated in the 'ping pong' cycle by the cleavage of complementary transcripts by PIWI proteins. Mature piRNAs are 2′-O-methylated and incorporated into PIWI proteins. The piRNA–PIWI complexes (piRNPs) silence transposable elements (TEs) either by endonucleolytic cleavage in the cytoplasm or through transcriptional silencing at their genomic loci in the nucleus[107]. **h** | Most lncRNAs are transcribed and processed in a similar way to mRNAs. Nuclear lncRNAs play an active part in gene regulation by directing proteins to specific gene loci, where they recruit chromatin modification complexes and induce transcriptional silencing or activation[185]. Other non-coding RNAs (for example, 7SK RNA) regulate transcription elongation rates[204] or induce the formation of paraspeckles (PS)[205]. Cytoplasmic non-coding RNAs can modulate mRNA translation[206]. **i** | Incorrectly processed RNAs are recognized by several complexes in the nucleus and cytoplasm that initiate and execute their degradation[207,208]. CPSF, cleavage and polyadenylation specificity factor; EJC, exon junction complex; hnRNP, heterogeneous nuclear RNP; NGD, no-go decay; NMD, nonsense-mediated RNA decay; NSD, non-stop decay; PABP, poly(A)-binding protein.

Non-coding RNA
(ncRNA). An RNA that does not encode a protein. In this Analysis, ncRNA is also used to specifically group together all remaining ncRNAs that are not ribosomal RNAs, tRNAs, small nuclear RNAs, small nucleolar RNAs or small Cajal body-specific RNAs.

Establishing a census of RBPs

Identification of RBPs. RBPs were first characterized using biochemical methods, such as gel electrophoresis of ultraviolet (UV)-crosslinked nuclear extracts or RNA affinity purification coupled with mass spectrometry and/or immunodetection[10,11]. Further insights into the specific RNA targets of RNPs and their protein composition were gained using RIP coupled with cDNA array hybridization of recovered RNA and protein mass spectrometry, respectively [12–14]. The definition of RNA-binding domains (RBDs) in proteins was facilitated by the growing amount of experimental and structural data, as well as by the completion of sequenced genomes from all kingdoms of life, which allowed sequence alignments of multiple proteins and computational RBD predictions across organisms based on sequence information[15–21]. Early counts of RBPs using predictions of single-stranded RBDs (ssRBDs) estimated ~500 RBPs in humans and mice[22–24], and ~700 RBPs in humans when including additional RBDs involved in other aspects of RNA metabolism[25]. These approaches had already suggested the existence of a complex regulatory network controlled by RBPs and recognized that a census of RBPs was a prerequisite for the interpretation of synergistic and competitive action of RBPs on their targets.

Automated functional annotations, including the Gene Ontology project, integrated literature reports, database entries and structural features, and arrived at ~1,900 human RBPs[26]. Although these gene groups are useful for gene set pathway analyses, they are not designed to establish a census of RBPs, as they include proteins that were falsely assigned owing to inferred participation in biological processes or exclude valid proteins owing to absence of annotation.

In an effort to experimentally validate RBPs across genomes or transcriptomes, four recent studies used a combination of *in vivo* UV RNA–protein crosslinking followed by poly(A)-RNA pulldown and protein mass spectrometry to characterize the mRNA-binding proteome in HEK293 and HeLa human cell lines, mouse embryonic stem cells and yeast cells[13,27–30]. A common set of ~600 crosslinked mRBPs was identified (see Supplementary information S1 (table)). Comparative large-scale studies isolating the many ncRNA-binding proteins have not yet been undertaken. Hence, although bioinformatic and experimental methods have advanced enough to allow the investigation of RBPs at a system-wide level, we currently do not have a clear understanding of the identity and number of genes involved in PTGR.

Defining the RBP repertoire. To annotate RBPs in the human genome, we defined proteins as RBPs if they contain domains known to directly interact with RNA or if they reside within well-characterized RNPs, even if they do not directly contact RNA in some structurally characterized conformations. Considering that most RNPs are dynamically assembled and disassembled, transient sequence-unspecific RNA contacts are still plausible. Examples include components of the exosome complex, the polyadenylation and cleavage complex, the spliceosome and the ribosome. We extracted 800 protein domains from the Protein families (Pfam) database (see Supplementary information S1 (table)) that are known to be RNA-binding or exclusively found in RNA-related proteins, and used the Pfam protein-domain hidden Markov models to search the human genome, comprising ~20,500 protein-coding genes, for proteins containing these RBDs[17,31]. From these candidates, we filtered out proteins with established RNA-unrelated functions, mostly DNA-binding zinc-finger proteins, and manually

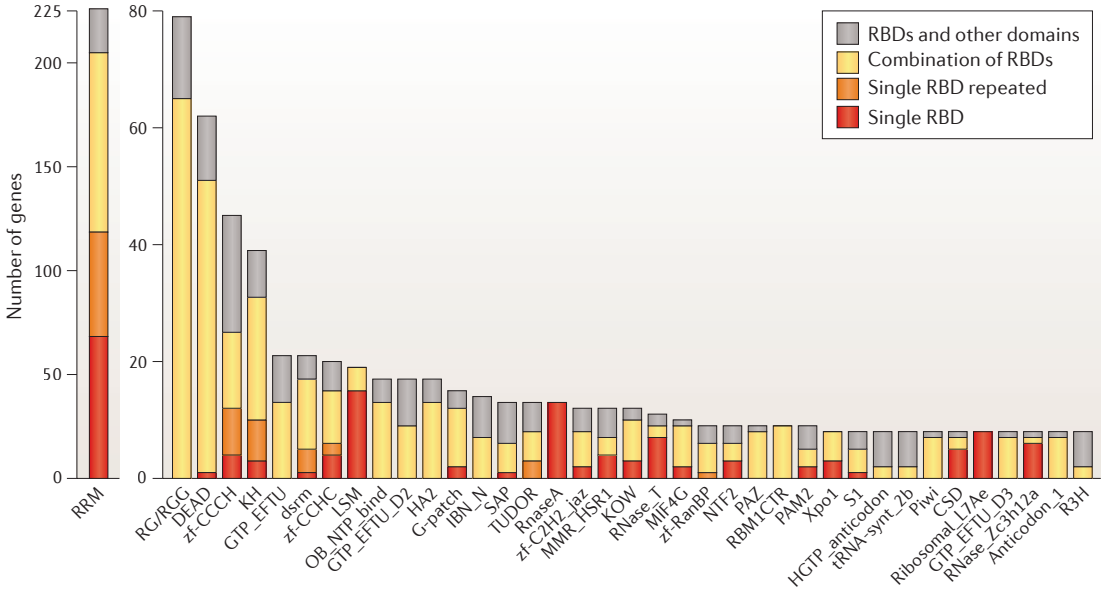


Figure 2 | **Single or repeated presence of frequent RBDs in human genes.** Counts of proteins with RNA-binding domains (RBDs) from the Protein families (Pfam) database with eight or more members in humans. Domain names are listed according to Pfam nomenclature; additional information can be found in Supplementary information S2 (box). In addition, low-complexity RG- or RGG-repeat regions — defined by at least three RG/RGG repeats spaced 10 amino acids or fewer apart — in RNA-binding proteins (RBPs) are shown. Counts are further subdivided to indicate the number of genes containing one RBD as the only structural domain in the encoded protein (red); repeats of the same class of RBD (orange); one or more RBDs in combination with RBDs of different classes (yellow); or combinations of the RBD with one or more domains unrelated to RNA metabolic function (grey), for example, protein kinase domains.

RNA recognition elements
Short (rarely more than 4–6-nucleotide-long) sequence elements within RNA targets that are recognized and bound by RNA-binding proteins.

Crosslinking and immunoprecipitation followed by sequencing
(CLIP–seq). An experimental method to map the binding sites of RNA-binding proteins (RBPs) on RNA targets transcriptome-wide. RBPs are ultraviolet-crosslinked to RNA *in vivo*, followed by partial RNase treatment of cell lysates, immunoprecipitation of RBPs, recovery of covalently bound RNA, and small RNA cDNA library preparation for deep sequencing of crosslinked RNA segments.

RNA immunoprecipitation and sequencing
(RIP-seq). An experimental method to identify enrichment and targets of RNA-binding proteins (RBPs). RBPs are immunoprecipitated, and bound RNAs are library-prepared for deep sequencing.

Small Cajal body-specific RNAs
(scaRNAs). Small RNAs that have a similar structure and sequence to small nucleolar RNAs (snoRNAs), localize to the Cajal body and are involved in the methylation and pseudouridylation of snoRNAs.

RNA-binding domains
(RBDs). Structural protein domains that directly bind to RNA. In this Analysis, RBD is also used to include structural domains found exclusively in RNA-binding proteins that are able to transiently contact RNA during ribonucleoprotein assembly or disassembly.

Hidden Markov models
Statistical probability models that assume a Markov chain with unobserved (hidden) states. In protein domain predictions, HMMs are calculated from protein sequence alignments and compute the probability of a specific protein sequence.

added RBPs that were missed by domain searches but that are clearly defined in the literature (see Supplementary information S2 (box)). This resulted in a final census of 1,542 RBPs (see Supplementary information S3 (table)), or 7.5% of all protein-coding genes in humans, which formed the basis of subsequent analyses described here. This catalogue provides a fresh starting point for future curation efforts but is likely to change as experimental studies uncover new RBPs or recognize that candidate RBPs containing established RBDs have evolved to adopt new functionalities that are unrelated to RNA binding.

Structural features of RBPs. RBPs are commonly classified based on their specific RBDs, as the structure and function of these RBDs provide some insights into the binding preferences and targets of the RBPs. Many excellent reviews have covered the different RBD families and their modes of RNA binding[25,32–55], and we will therefore limit the depth of our RBD discussion. RBDs are deeply conserved across bacteria, archaea and eukaryotes. The 1,542 human RBPs we identified contain a repertoire of ~600 structurally distinct RBDs. Among the RBD classes, only 20 have more than 10 human gene members; most of them have one or two members (FIG. 2). RBDs with the highest number of members are mostly found in mRBPs and mirror the rapid expansion of mRNA-related processes in the evolution of higher eukaryotes, such as alternative splicing and polyadenylation[56,57]. Of the estimated 692 mRBPs, 405 contain an RNA recognition motif (RRM), a K homology (KH) domain, a DEAD motif, a double-stranded RNA-binding motif (DSRM) or a zinc-finger domain (FIG. 2). By contrast, the 169 ribosomal proteins have 119 distinct domains that are exclusively found in ribosomal proteins. This diversity of RBDs complicates both the definition of a census and the *de novo* identification of RBPs, and it explains why earlier approximations based on the few large structural groups underestimated the number of proteins involved in PTGR.

A common feature of the abundant mRNA-binding domain classes is their frequent occurrence in multiple repeats or in combination with other RBDs. ssRBDs — such as RRMs, KH domains, zinc-finger domains or cold-shock domains — recognize degenerate 4–6-nucleotide segments and predominantly occur in combinations or repeats, which increase sequence specificity and affinity of RBPs[3,32]. Their modular design also enables the rapid evolutionary adaptation of proteins to new RNA targets[32], which in some cases poses interesting questions about the regulatory functions of RBPs and the evolution of their targets. For example, whereas most KH-domain-containing RBPs have one or two KH domains, the high-density lipoprotein-binding protein (HDLBP) expanded from 7 KH repeats in *Saccharomyces cerevisiae* to 14 KH repeats in humans. We are currently far from a comprehensive characterization of RBDs and RBPs. The biological functions of at least one-third of the 1,542 RBPs are unknown or have merely been inferred from those of homologues. Even among the established large classes of RBDs, many individual members have not been studied in detail, including RRM-containing proteins; helicases (DEAD box helicases and

Transcription factors
(TFs). Proteins that bind to specific DNA sequences at gene promoters, upstream and downstream elements, or within the gene body; they influence gene expression by enhancing or blocking transcription.

RPKM
(Reads per kilobase per million mapped reads). A measure for quantifying single-end read RNA-sequencing data per transcript or gene exon model; it normalizes the total number of mapped reads per transcript or gene exon model by the length of the transcript or gene exon model (in kilobases) and the library size (total number of reads mapped to the genome or transcriptome in million reads).

Small nuclear RNA
(snRNA). A type of short (~70–200-nucleotide) RNA found in the nucleus of eukaryotic cells. snRNAs associate with proteins of the spliceosome to form the spliceosomal core complexes.

helicase-associated domain-containing proteins); zinc-finger proteins (with zf-CCCH, zf-CCHC zf-C2H2_jaz motifs); RNA nucleases of type RNase A, RNase T and RNase ZC3H12A domain; and translation factor domain-containing proteins (for example, those of the GTP elongation factor family GTP-EFTU, which usually consist of three structural domains, two oligonucleotide-binding domains and a GTP-binding domain).

Abundance of RBPs and TFs across tissues. The importance of PTGR becomes evident when analysing RNA-sequencing (RNA-seq) expression levels of RBPs and transcription factors (TFs) relative to those of the residual proteome across 16 human tissues of the Illumina Body Map. The majority of RBPs were ubiquitously expressed, typically at higher levels than average cellular proteins (FIG. 3a), which is consistent with previous analyses[58,59]. Although RBPs and TFs were encoded by a similar number of genes (1,542 and 1,704 genes, respectively), the cumulative abundance of RPKM expression levels of RBPs contributed up to 20% of the expressed, protein-coding transcriptome, whereas TFs constituted only up to 3% by transcript abundance (FIG. 3b). The equivalent of 10–12% of the expressed transcriptome originates from the 169 ribosomal proteins but represents only approximately 0.8% of the protein-coding genes in the genome. By contrast, the transcripts encoded by the 692 mRBP genes, which represent 3% of protein-coding genes, accounted for 4–5% of the transcriptome, whereas all remaining RBPs (including tRNA-binding and pre-rRNA-binding proteins, as well as other RBPs) contributed the remaining 4–5%. These abundances illustrate the central importance of translation-related processes in the cell. Tumours and immortalized cell lines express mRBPs and ribosomal proteins at even higher levels than normal tissues. The increased demand for continuous protein production, the changes in the size of the nucleolus (the site of ribosome and rRNA biogenesis) and the modified translational activities of cells have long been considered as hallmarks of cancer[60–62]. More generally, altered translational activity has been observed in a wide range of human pathologies and has recently been connected to neurodegenerative diseases, including Parkinson's disease and Alzheimer's disease[63–66]. Given the central importance of protein translation and ribosome biogenesis for energy metabolism and cellular growth, understanding disease-related changes in PTGR pathways is of diagnostic and possibly prognostic value, and also enables new therapeutic approaches. Consequently, targeting of PTGR pathways has been explored in drug development of inhibitors to block translation initiation or ribosome biogenesis; for example, the anticancer drug silvestrol inhibits eukaryotic initiation factor 4A-I (EIF4A1) and EIF4A2 (REFS 67–70).

Conservation of RBPs and their families

RBP subclasses. A classification of RBPs by interacting RNA targets is useful, as it isolates individual PTGR pathways and can also explain similar phenotypes for RBP genes in human diseases. We grouped the set of 1,542 human RBPs based on literature reports (FIG. 4A) into mRNA-binding, pre-rRNA-binding, tRNA-binding, small nuclear RNA (snRNA)-binding and small nucleolar RNA (snoRNA)-binding proteins, as well as a residual

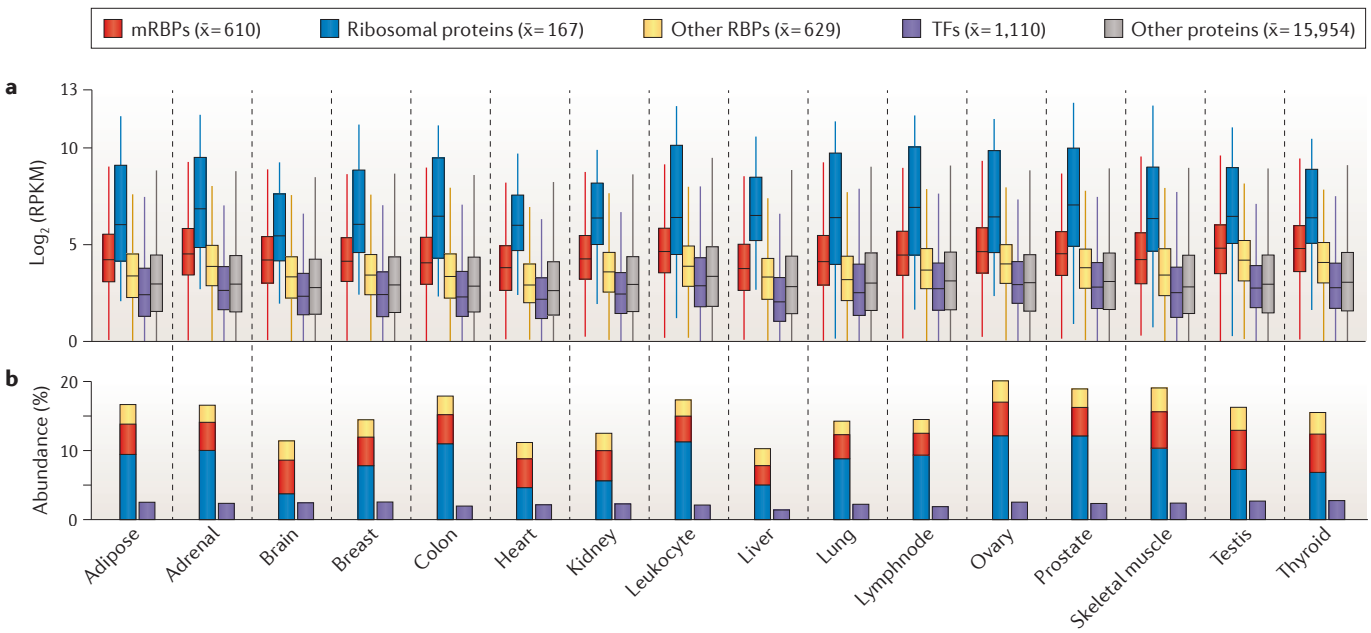


Figure 3 | **Transcript abundance of RBPs and TFs across 16 different human tissues.** **a** | Distribution of gene expression levels of protein-coding genes, measured by RNA sequencing (RNA-seq) with RPKM (reads per kilobase per million mapped reads) expression values ≥1, is displayed. Shown as subgroups are mRNA-binding proteins (mRBPs), ribosomal proteins, the remaining RNA-binding proteins (RBPs), transcription factors (TFs) and the residual protein-coding transcriptome. For each group, the mean number of expressed proteins across the tissues is shown in parentheses. **b** | Cumulative abundance of RBPs and TFs as percentages of all RNA-seq reads is shown.

ncRNA-binding category (see Supplementary information S3 (table)). Further categories were introduced to define protein components of the ribosome, diverse RBPs that interact indiscriminately with many types of RNAs (such as the RNA exosome in general RNA turnover) and RBPs with unknown targets (that is, proteins with known RBDs or some experimental data on RNA binding that lacked information on specific targets). RBPs with more than one target class were also found, and emerging transcriptome-wide binding studies reveal that it may be common for certain RBPs to interact with and regulate multiple classes of RNAs; for example, LIN28 proteins bind to let-7 pre-microRNA (miRNA), mRNAs and snoRNAs, and microprocessor complex subunit DGCR8 binds to double-stranded RNA regions within pre-miRNAs and mRNAs[71–74]. For simplification, we grouped RBPs by the class of RNA with which they predominantly interact, when known.

Almost all categories of RBPs are directly or indirectly invested in the process of protein synthesis: 692 proteins are mRNA-binding, 169 are ribosomal proteins, and 130 proteins are involved in the biogenesis and delivery of charged tRNAs to the ribosome. Another 90 proteins are involved in the biogenesis of snRNAs or the formation of small nuclear ribonucleic particles (snRNPs; snRNA–protein complexes); 122 and 41 RBPs take part in the biogenesis of rRNA and snoRNA, respectively, which may be an underestimation given that rRNA biogenesis and the functional roles of some 'orphan' snoRNAs have yet to be fully characterized[75–78]. Another group comprised 122 RBPs that interact with the remaining ncRNAs, including all remaining ncRNA categories, such as miRNAs, PIWI-interacting RNAs (piRNAs), long ncRNAs (lncRNAs) and the RNA components of MRP and RNase P (FIG. 1; see Supplementary information S2, S3 (box, table)). These proteins associate with a range of ncRNAs, some of which are involved in gene regulation (for example, miRNAs, piRNAs and lncRNAs) and control gene expression through post-transcriptional RNA degradation, transcriptional silencing or activation of gene loci, and translational repression or activation. Others act as structural and catalytic components of RNP complexes (including RNase MRP, RNase P and telomerase RNP) or form RNP complexes of unknown functions (for example, vault RNAs and Y RNAs). The group with diverse RNA targets includes 47 RBPs, mostly RNA nucleases involved in general RNA turnover. Surprisingly, the functions of many rRNA and tRNA biogenesis factors in humans are mostly inferred from distant homologues and domain relationships[79–82]. As these processes are essential for cellular life and are highly conserved, core functions often remain the same. However, with the increasing complexity of organisms, protein factors, their family members and/or the range of target RNAs of pathway components evolved, grew in size and diverged. The recently discovered new roles of tRNA methylases in mRNA and ncRNA metabolism may be an example of this[83,84].

Conservation of RBP and TF families. Phylogenetic relationships of RBPs reveal the creation of gene families during evolution, a process that in principle allows diversification of RNA metabolic pathways. However, in most instances, human paralogues are functionally overlapping, with similar or even identical binding sites[3,8,85,86], and may therefore represent an alternative to increasing protein synthesis and/or facilitating regulation across cell types. Understanding family relationships and examining protein families together, instead of singular members, is therefore a valuable approach. Furthermore, PTGR functions of RBPs may be redundant, and structurally unrelated proteins are frequently found to act in the same pathway or on the same target. For example, during rRNA biogenesis, distinct processing factors can carry out the same processing steps for both 3′-end and 5′-end maturations[87]. In these cases, considering family relationships together with structural information and localization of RBPs is useful for isolating components of regulatory pathways.

To define redundant properties of RBPs, we compared the evolutionary characteristics of RBPs with those of TFs, which constitute the other main group of gene regulatory factors. We used the phylogenetic homology classification as already defined by the Ensembl Compara database[88] and further grouped together paralogues with even closer homology. Most RBP paralogues share 20–70% sequence identity and, by this criteria, RBPs that are known to be functionally related are grouped together, such as members of the cytoplasmic polyadenylation element-binding protein family, which share ~25% sequence identity. We refer to these subgroups as paralogous RBP families throughout this Analysis. Lower cutoffs included functionally unrelated, more distant paralogues, whereas higher cutoffs missed family members of known functional similarity. Consistent with their high structural diversity, the 1,542 RBPs formed 1,111 families, in which individual members within RBP families generally interact with the same RNA classes (FIG. 4B,Ca; see Supplementary information S3 (table)). By contrast, the 1,704 human TFs, which diverged more recently than the RBPs (reviewed in REF. 58), formed only 554 protein families by the homology criteria above (FIG. 4Cb; see Supplementary information S4 (table)). RBDs and DNA-binding domains often originate from common superfamily folds, such as the oligonucleotide-binding fold (OB-fold), nucleotidyltransferase domains, zinc-finger domains, and DNA and RNA helicase domains[34,43,89,90]. However, although TFs and RBPs have a shared evolutionary history, TFs expanded late in evolution into large families by multiple gene duplications[58], with up to 50 members per family and 2.5 members on average. By contrast, RBPs diversified early, and paralogous families have 1.3 members on average, with the largest RBP families including up to 10 members. Paralogous RBP families are well conserved across eukaryotes, and 50% of the human RBP families are also present in *S. cerevisiae* (FIG. 4Ca; see Supplementary information S5 (table)). This finding is consistent with previous observations that at least 200 distinct RBPs are present in the lowest common ancestor of animals, and that 80 orthologous groups of RBPs are traceable even to the lowest universal common ancestor[25,91]. By striking

Small nucleolar RNA
(snoRNA). A type of short (~50–200-nucleotide) RNA that is localized to the nucleolus and that guides methylation or pseudouridylation of ribosomal RNAs and small nuclear RNAs.

MicroRNA
(miRNA). A type of small (~21-nucleotide) non-coding RNA involved in post-transcriptional gene silencing. miRNAs form ribonucleoprotein complexes with Argonaute proteins to repress mRNA stability and protein expression by recruiting RNA deadenylation and degradation complexes to their RNA targets.

PIWI-interacting RNAs
(piRNAs). Small (~28-nucleotide) non-coding RNAs involved in post-transcriptional gene silencing that are expressed in the germ line; they form ribonucleoprotein complexes with PIWI proteins, and protect the genome from genomic instability by transcriptional and post-transcriptional repression of transposons.

Long ncRNAs
(lncRNAs). RNAs that do not encode proteins and are >200 nucleotides long; they are found as structural components in nuclear and cytoplasmic ribonucleoprotein complexes and are transcribed by RNA polymerase II, similarly to mRNAs. Less abundant lncRNAs may influence the gene expression of neighbouring genes (in *cis*) at the transcriptional, post-transcriptional and translational levels.

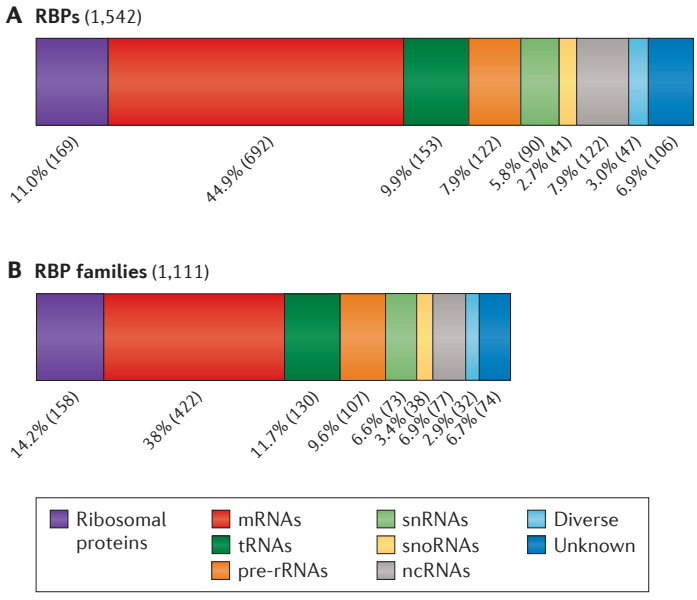
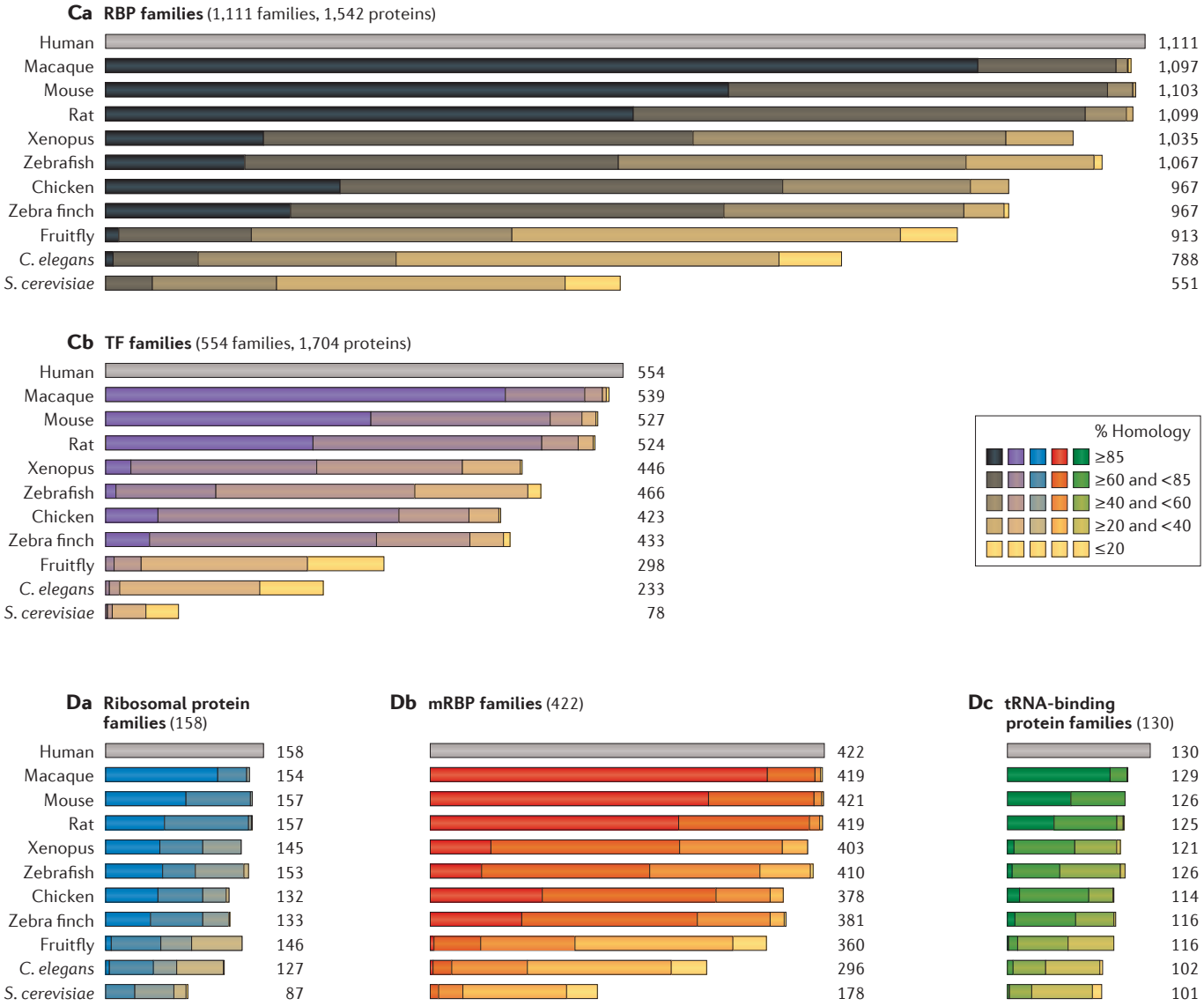


Figure 4 | **Target RNA classification and evolutionary conservation of RBP and TF paralogous families.** **A,B** | RNA-binding proteins (RBPs) and RBP families are grouped by their respective targets: ribosomal proteins, mRNA, tRNA, pre-ribosomal RNA, small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), non-coding RNA (ncRNA); diverse targets and unknown targets are also indicated. Percentage and counts (in parentheses) of RBPs in the category are shown. In part **B**, RBP paralogues are grouped into families defined by 20% sequence identity according to the Ensembl Compara database[88]. **C** | The number of RBP families conserved across 11 species and their percentage identity score for human RBP families (part **Ca**) and transcription factor (TF) families (part **Cb**) are shown. The number of families with different degrees of conservation are grouped into the following five categories: ≥85% homology; ≥60% and <85% homology; ≥40% and <60% homology; ≥20% and <40% homology; and ≤20% homology. **D** | The number of paralogous families and degree of conservation are shown for human ribosomal proteins (part **Da**), mRNA-binding proteins (mRBPs; part **Db**) and tRNA-binding proteins (part **Dc**). *C. elegans*, *Caenorhabditis elegans*; *S. cerevisiae*, *Saccharomyces cerevisiae*.



contrast, few TFs are conserved throughout evolution, and only 14% of the human TF families are found in *S. cerevisiae* (FIG. 4Cb; see Supplementary information S6 (table)). Although the expansion of TFs traced the increase in organismal complexity, which possibly enabled the development of new functionalities[58], evolutionary stability of RBPs went along with the early establishment of core RNA metabolic processes in all cellular systems[25].

Conservation of RBP families interacting with different RNA classes. With the growing number of RBP families in higher eukaryotes, the relative size of the various RBP families committed to different RNA targets has remained constant across phylogenies (38% mRBPs, 12% tRNA-binding proteins and 14% ribosomal proteins; FIG. 4D). However, RBP families in different target groups display varying levels of evolutionary conservation. Most RBP classes — including families in the rRNA, tRNA, snRNA and snoRNA pathways — display average homologies of 31% between humans and yeast (for example, the tRNA-binding proteins in FIG. 4Dc). Ribosomal proteins are among the most conserved, with an average homology of 51%, in contrast to the factors involved in maturation and processing of rRNAs, which are conserved up to only 31%, reflecting the increasing divergence of rRNA biogenesis factors in higher eukaryotes[82,92]. ncRNA-binding protein and mRBP families display the lowest conservation (20% and 27%, respectively), and only one-fifth of all ncRNA-binding protein families have homologous families in yeast.

Tissue specificity of RBPs. Tissue-specific expression, phyletic age and cellular functions of proteins often correlate. Whereas ancient and highly conserved genes are widely expressed and support basic cellular functions, more recently evolved genes, such as TFs and secreted proteins, are expressed in a species- and tissue-specific manner[93–95]. Accordingly, we investigated the mRNA expression of 1,441 RBPs and 1,463 TFs profiled in a microarray study that measured transcript levels of 16,867 protein-coding genes across 31 human tissues[96], assuming that transcript abundance approximates protein abundance in the cell[97,98]. Tissue-specific variation of RBP isoforms owing to alternative splicing and to alternative cleavage and polyadenylation are not well understood and were not considered for this level of analysis. We favoured microarray studies over RNA-seq studies because they profiled larger collections of different human tissues and organs. We calculated a tissue-specificity score (S) for every gene on the array across the profiled samples, which ranged from 0 for ubiquitously expressed proteins to 5 for highly tissue-specific proteins (see Supplementary information S2 (box)). Based on the score for established tissue-specific RBPs, such as the germline-specific PIWI family ($S = 1.7–2.3$) or neuronal members of the ELAV-like (ELAVL) family ($S = 3.2$)[99,100], we set a cut-off score of 1 for referring to tissue-specific genes, at which 6% of RBPs and 13% of TFs showed some level of tissue-specific expression (FIG. 5; see Supplementary information S7 (table)).

As expected, ribosomal proteins (FIG. 5a,b), as well as general components of the spliceosome, RNA transport and turnover machineries (FIG. 5d), were ubiquitously expressed across tissues. Furthermore, although tissue-specific variation has been reported for some tRNAs and snoRNAs[101–103], the biogenesis factors of tRNAs and snoRNAs, as well as components of the snRNA and rRNA maturation pathways, were generally uniformly expressed across tissues (FIG. 5a–c). The majority of tissue-specific RBPs consisted of mRBPs and ncRNA-binding proteins, as well as a range of RNA nucleases with diverse target specificity (FIG. 5d). Perhaps unexpectedly, some tissue-specific outliers were observed among rRNA biogenesis factors and ribosomal proteins, including ribosomal protein L3-like (RPL3L), a homologue of RPL3 (FIG. 5d). These selectively expressed RBPs may reflect extra-ribosomal roles[104] or tissue-specific adaptations in the composition of ribosomes that regulate translation of subsets of proteins[105].

The majority of profiled RBPs, including mRBPs, displayed no tissue specificity: 79% of all RBPs (1,144 of 1,441 RBPs) had $S < 0.3$. Among paralogous RBP families, 808 of 1,049 families (77%), or 277 of 409 mRBP families (68%), were ubiquitously expressed with $S < 0.3$ (FIG. 5e; see Supplementary information S7 (table)). Few tissues contained specialized RBPs, and 90% of the 82 tissue-specific RBPs were identified in germline, brain, muscle, bone marrow or liver cells. The largest fraction (47 proteins) was enriched in adult testis, where they contribute to gametogenesis and fertility through regulation of transposon silencing, mitosis, meiosis, stem cell maintenance and differentiation[106–111]. Many of these proteins were also expressed during fetal ovary development, at the stage of mitotic and meiotic cell divisions and germ cell expansion[112]. We found that 2% of all RBP families were exclusively tissue-specific, and most of these families were expressed in the germ line, such as the deleted in azoospermia (DAZ) and PIWI families, and other RBPs involved in the piRNA pathway[99,107,113,114] (FIG. 5e).

Rather than being part of families in which all paralogues are expressed in a tissue-specific manner, most tissue-specific RBPs belonged to protein families with at least one ubiquitously expressed member. Overall, 5% of RBP families were broadly expressed with one or more highly tissue-specific member. Examples included several helicase families with germline-specific paralogues, such as the tissue-specific DDX4 protein with its ubiquitously expressed family members DDX3X and DDX3Y; the MOV10L1 helicase and its ubiquitously present paralogue MOV10; and the tissue-specific helicase DDX25 with its ubiquitous paralogues DDX19A and DDX19B[115–118]. Moreover, most members of the secreted, vertebrate-specific RNase A family — including angiogenin (ANG), RNase 2 (RNASE2) and RNASE3 — displayed high tissue specificity and were expressed in the liver and bone marrow cells, where they have a role in angiogenesis and the innate immune response, respectively[119]. Other families in this group were the mRNA splicing and regulatory families ELAVL and insulin-like growth factor 2-binding protein

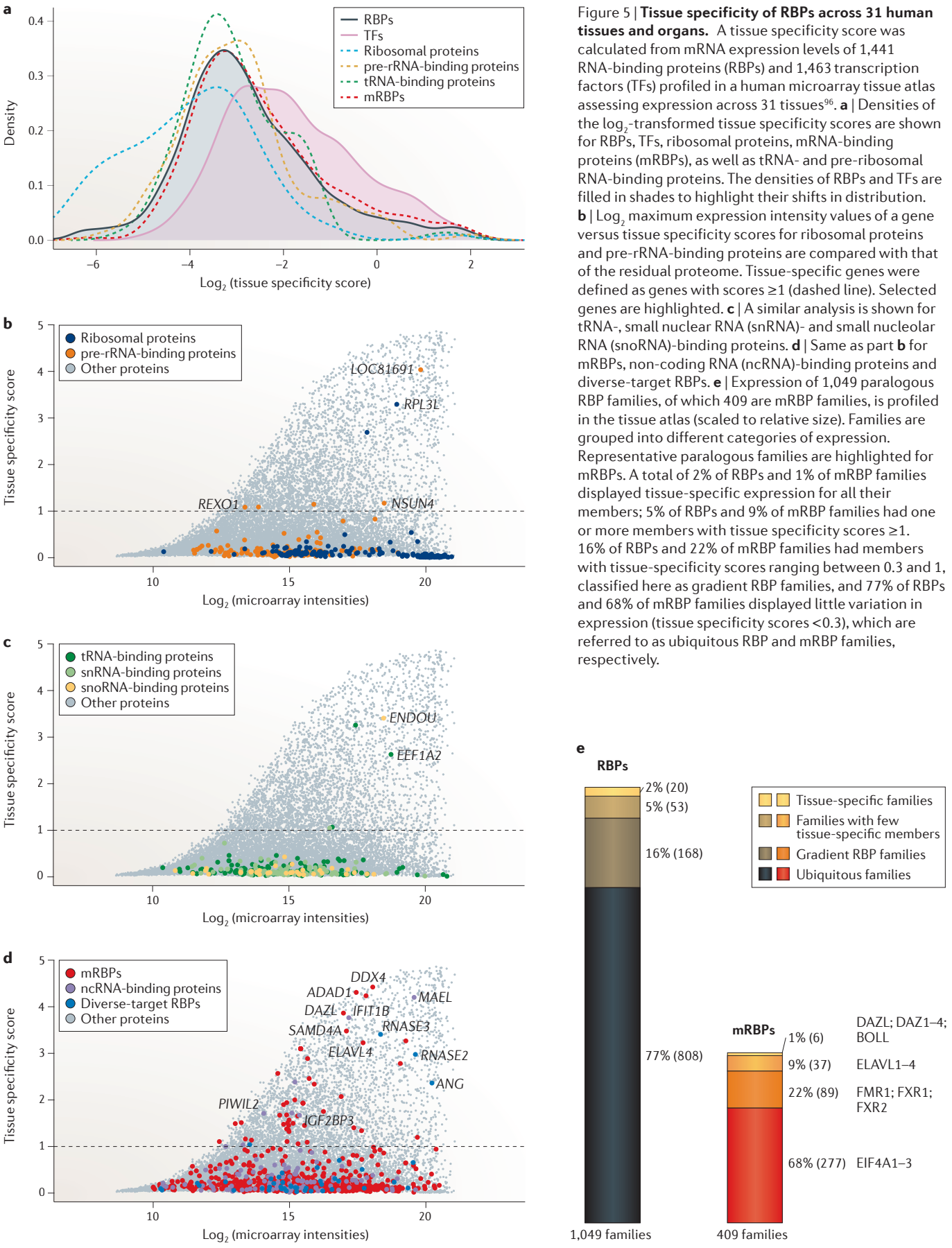


Figure 5 | **Tissue specificity of RBPs across 31 human tissues and organs.** A tissue specificity score was calculated from mRNA expression levels of 1,441 RNA-binding proteins (RBPs) and 1,463 transcription factors (TFs) profiled in a human microarray tissue atlas assessing expression across 31 tissues[96]. **a** | Densities of the $\log_2$-transformed tissue specificity scores are shown for RBPs, TFs, ribosomal proteins, mRNA-binding proteins (mRBPs), as well as tRNA- and pre-ribosomal RNA-binding proteins. The densities of RBPs and TFs are filled in shades to highlight their shifts in distribution. **b** | $\log_2$ maximum expression intensity values of a gene versus tissue specificity scores for ribosomal proteins and pre-rRNA-binding proteins are compared with that of the residual proteome. Tissue-specific genes were defined as genes with scores ≥1 (dashed line). Selected genes are highlighted. **c** | A similar analysis is shown for tRNA-, small nuclear RNA (snRNA)- and small nucleolar RNA (snoRNA)-binding proteins. **d** | Same as part **b** for mRBPs, non-coding RNA (ncRNA)-binding proteins and diverse-target RBPs. **e** | Expression of 1,049 paralogous RBP families, of which 409 are mRBP families, is profiled in the tissue atlas (scaled to relative size). Families are grouped into different categories of expression. Representative paralogous families are highlighted for mRBPs. A total of 2% of RBPs and 1% of mRBP families displayed tissue-specific expression for all their members; 5% of RBPs and 9% of mRBP families had one or more members with tissue specificity scores ≥1. 16% of RBPs and 22% of mRBP families had members with tissue-specificity scores ranging between 0.3 and 1, classified here as gradient RBP families, and 77% of RBPs and 68% of mRBP families displayed little variation in expression (tissue specificity scores <0.3), which are referred to as ubiquitous RBP and mRBP families, respectively.

(IGF2BP) family, which had ubiquitously expressed paralogues (ELAVL1 and IGF2BP2) and highly tissue-specific members (ELAVL3, ELAVL4, IGF2BP1 and IGF2BP3) in the brain, germ line and liver[100,120,121] (FIG. 5e). For 16% of RBP families, here named gradient RBP families, individual members were ubiquitously expressed with tissue specificity scores below 1 but displayed some degree of differential expression across tissues without being tissue-specific (FIG. 5e). Loss-of-function of proteins in these families often affects the tissue of highest expression most strongly. A prominent example of RBPs with this expression pattern is the fragile X mental retardation 1 (FMR1) family, which comprises three ubiquitously expressed members (FMR1, fragile X mental retardation syndrome-related 1 (FXR1) and FXR2) with redundant target specificities[122]. Of these proteins, FMR1 has the highest expression levels in the brain, thyroid and gonads, and FXR1 and FXR2 are most abundant in the skeletal muscle and testes. Thus, even though activity of this protein family is necessary in every tissue, loss-of-function of FMR1 mainly affects the brain and gonads, and causes mental retardation and premature ovarian insufficiency in fragile X syndrome or fragile X-associated ataxia syndrome[123], whereas mouse knockout models of FXR1 are embryonically lethal owing to skeletal muscle defects[124]. For families with some tissue-specific variation, the closely related paralogues often bind to the same sites on target RNAs with similar affinities, such as the members of the FMR1 and ELAVL families, which have identical binding sites in cell culture models[3,121,122]. The redundant functions of the ubiquitous paralogues can therefore complicate the dissection of the role of the tissue-specific proteins and may require technically challenging experimental designs, such as the generation of animals with tissue-specific knock-in or knockout of family members.

In conclusion, we found that 98% of paralogous RBP families were ubiquitously expressed, and their deep evolutionary conservation supports their predominant basic cellular function. Of these, 20% are families with tissue-specific and ubiquitous paralogues or ubiquitous members that are enriched in some tissues. Only 2% of families are tissue-specific for all paralogues, suggesting a strictly cell type-specific contribution to PTGR pathways, which is similar to the evolution of TFs. Cell type-specific expression levels of an RBP and its paralogues must be considered when choosing a system to study regulatory networks and targets.

RBPs and human diseases

Disease phenotypes of RBPs may correlate with tissue-specific expression; for example, loss-of-function of germline-specific proteins causes infertility[114], and loss of FMR1 causes more severe phenotypes in the tissues in which it is usually most enriched[123]. However, highly tissue-specific pathologies are often observed for loss-of-function of RBPs with no specificity in expression at all. These tissue-specific phenotypes may be explained either by tissue-specific expression of critical RNA targets and cofactors of the RBP or by a greater sensitivity to expression changes of PTGR components in general for the affected tissue.

RBPs that bind to the same RNA class often affect the same tissues and display similar pathologies. For example, ribosomopathies, such as Diamond–Blackfan anaemia and Shwachman–Diamond syndrome, are caused by defects in ribosomal proteins and rRNA biogenesis factors, and severely affect the bone marrow and skin[125]. By contrast, mutations in mRBPs are found in multiple neurodegenerative and neuromuscular diseases that affect mRNA metabolism in neurons, particularly in motor neurons[66,126,127]. In these cases, mutations in mRBPs or their RNA targets impair RNA transport and translation, often leading to protein or RNA aggregation and inefficient clearance of neuronal RNA or protein granules, which cause a range of neuropathological diseases[66,126–132]. Examples include the neurological diseases amyotrophic lateral sclerosis (ALS) — which is caused by mutations in the heterogeneous nuclear RNPs (hnRNPs) TAR DNA-binding protein (TARDBP; also known as TDP43), FUS, hnRNP A2/B1 (HNRNPA2B1) and hnRNP A1 (HNRNPA1)[131,132] — and spinocerebellar ataxia, which is caused by polyglutamine expansions in ataxin proteins (ATXN1 and ATXN2)[133,134]. Both diseases show toxic prion-like protein aggregations and accumulation of RNA-protein granules in cerebellar and/or motor neurons. Although FMR1 and RBP fox-1 homologue 1 (RBFOX1) are involved in different processes (mRNA transport and splicing, respectively), loss-of-function of these proteins results in a spectrum of mental retardation and autism[123,135]. mRNA-repeat expansions lead to sequestration of RBPs, often splicing factors, and have typically been linked to muscular diseases caused by dysregulated splicing (such as myotonic dystrophies), mental retardation and ataxia[136]. Loss-of-function of the snRNP assembly factor survival motor neuron 1 (SMN1) directly affects the spliceosome and causes spinal muscular atrophy[127]. Loss-of-function mutations in tRNA splicing components and aminoacyl tRNA synthetases typically cause encephalopathies and the neuropathy Charcot–Marie–Tooth disease[66,137,138]. Several RNA and DNA nucleases that are crucial for nucleic acid clearance have been implicated in the autoimmune disease Aicardi–Goutières syndrome[139,140]. These characteristic patterns for RBPs show that, instead of RBP expression, the interacting RNA targets are often a better predictor for the disease pathologies of RBPs.

Expression dynamics of RBPs

Dynamic complexes of RBPs. RBPs assemble into dynamic RNP complexes that mature, process, regulate or transport RNAs. In addition to remodelling RNA structures to keep RNAs accessible to other RBPs and enzymatic RNA-processing complexes, RBPs and RNPs act as RNA chaperones; prevent aggregation, misfolding and incomplete processing; and facilitate movement of RNA targets to required locations in the cell across cellular compartments. As a consequence, the abundance of RBPs and their constituents differentially affects RNA regulation[37,141–144]. For example, the abundance of various splicing factors can influence mRNA splicing

patterns[57,145–148], whereas U1 snRNP levels control alternative polyadenylation sites[149,150]. The competition among RBPs with similar or overlapping target specificity can also define regulatory modes. For example, ELAVL1 antagonizes miRNA regulation on numerous mRNA targets[151], LIN28 competes with the miRNA-processing machinery to suppress pre-let-7 miRNA processing[72–74], and Pumilio homologue (PUM) proteins synergize with miR-221 and/or miR-222 to destabilize the cyclin-dependent kinase inhibitor 1B (*CDKN1B*) mRNA[152]. Similarly, multiple RBPs are involved in the localization and transport of RNA to distinct RNP granules, which contain highly concentrated subsets of RNAs and RBPs and act in the storage and/or degradation of mRNAs[153]. The importance of PTGR regulatory networks in cellular processes is evident from genetic knockouts of RBPs, which are often lethal or affect all tissues, and is consistent with their high conservation, number of targets and low tissue specificity. Selective expression of a single RBP typically does not result in differentiation or dedifferentiation into distinct cell types, in contrast to the selective expression of TFs[154]. Instead, the interactions of many RBPs in regulatory complexes determine specificity of PTGR processes. Hence, groups of RBPs in common PTGR pathways are often co-expressed and can drive coordinated expression of targets in cells and tissues, as well as across developmental processes[22,155,156]. Specific expression of RBPs can be used to deduce their putative roles and to identify novel components of regulatory pathways. In the next two sections, we explore these concepts and use available gene expression data across several developmental stages in human ovary and brain to identify the correlated expression of groups of RBPs that potentially act in the same developmental pathways.

Co-expression of RBPs required for ovarian development. The germ line presents a unique system for functional studies of process-specific RBPs, as it has a highly specialized RNA metabolism. At least 50 tissue-specific RBPs contribute to the differentiation and maintenance of germ cells[108], and many are involved in germline-specific piRNA-induced transposon silencing, alternative polyadenylation and translational regulation that affects hundreds of mRNA targets[107,108,157–160]. Between 8 and 20 weeks of gestation, oogonia proliferate and their numbers increase from 0.6 million to 6 million cells[161]. By 20 weeks of gestation, primordial oocytes enter meiosis and arrest in the diplotene stage of meiosis I prophase I until oogenesis resumes in puberty. We examined RBP expression in a microarray study of 9–18-week-old human fetal ovaries[112]. Expression of germline-specific RBPs peaked at 14, 16.4, 16.9 and 18 weeks (FIG. 6), and displayed highly correlated expression dynamics, reaching Pearson coefficients close to 1 (FIG. 6b; see Supplementary information S7 (table)). Although some of the expression changes may be attributed to changing percentages of tissue composition of germline and somatic cells, the increase in expression for known germline-specific RBPs was evident and also correlated with their high tissue-specific expression in the adult testis, confirming a role in germline development for both sexes. The expression dynamics were clearly distinct from differentially expressed somatic RBPs with functions unrelated to germline development. For example, the IGF2BPs, which are required during embryogenesis and organ development, were highly expressed at week 9, before expression levels rapidly decreased[120,162]. All constituents of the piRNA pathway[107] were upregulated in the course of germline development, including piRNA biogenesis factors, such as the RNA endonucleases maelstrom and phospholipase D6 (PLD6, the *Drosophila melanogaster* Zucchini homologue); the RNA helicases DDX4, DDX39A (the UAP56 homologue) and MOV10L1; and most members of the Tudor domain-containing protein family (TDRD1–9 and RING finger 17)[107] (FIG. 6a,c). In addition, we observed coordinated expression of the established germline-specific translational regulators DAZ1–4, DAZL and boule-like (BOLL), which are also essential regulators of gametogenesis[113,163]. The expression dynamics and patterns of groups of RBPs during ovarian development mirror their role in germline development derived from genetic experiments. Our clustering of expression profiles of RBPs across developmental stages also points to novel regulatory roles of RBPs that have not been previously studied in germline development, including RBM46, PIH1 domain-containing protein 2 (PIH1D2), adenosine deaminase domain-containing protein 1 (ADAD1) and poly(A)-specific ribonuclease PARN-like domain-containing protein 1 (PNLDC1) (FIG. 6a,c).

Co-expression of RBPs in brain development. Neurons demonstrate unique alternative splicing and polyadenylation of mRNAs[57,100,157–159]. Furthermore, the considerable length of neuronal projections make mRNA transport and local translation at neuronal dendrites indispensable for development, synaptic plasticity and long-term memory[164]. Unsurprisingly, many RBPs regulating splicing, RNA transport, storage and translation are critical for neuroplasticity[165–167]. To capture brain-specific PTGR networks, we examined the expression dynamics of RBPs at different fetal and postnatal stages in human hippocampus development using RNA-seq data from the BrainSpan database. Although more than 75% of all protein-coding genes were reported to be expressed in the brain[168], the expression of thousands of these genes was found to be either restricted to a particular cell type or temporally regulated[168,169]. Concordantly, distinct groups of RBPs were upregulated at different developmental stages, which is consistent with previous studies[22,170].

The largest cluster of differentially expressed RBPs in the hippocampus contained ~100 RBPs that are highly expressed during early development and rapidly downregulated at later stages (FIG. 7, group I; see Supplementary information S7 (table)). This cluster included proteins required for the regulation of developmentally relevant pathways, such as the IGF2BP family[120,162] and LIN28 (REF. 171), as well as general factors involved in translation, mRNA splicing, transport and rRNA biogenesis. Another distinct group of ~20 proteins displayed low expression in the first fetal

Post-conception week (PCW). A time measurement used to describe stages of human development in prenatal weeks. PCW records the time elapsed since the day of conception. Also commonly used is gestation week, which counts from the day of the last menstrual period. Assuming a normal 28-day menstrual cycle, PCW is 2 weeks less than gestation week.

3′ untranslated regions (3′UTRs). 3′ ends of mRNAs, specifically the region between the stop codon and the poly(A) tail. 3′UTRs are targets of post-transcriptional regulation by many ribonucleoprotein and RNA-binding protein complexes, which determine their stability, translation and turnover.

post-conception week (PCW) and rapidly increased expression during the period coinciding with hippocampal development. This group included multiple splicing regulators required for neuronal function, such as the RBFOX family, which contributes to the characteristic splicing pattern of many neuronal transcripts[172] (FIG. 7, group II).

A separate group of ~20 RBPs, which were enriched in RNA nucleases and mRNA regulatory proteins involved in inflammatory and innate immune responses, was highly expressed at 37 PCWs (FIG. 7, group III), coinciding with the maturation wave of pyramidal neurons and synaptogenesis at 34–36 PCWs[173]. Mutations in the RNA nucleases ANG and RNase T2 (RNASET2) were found in patients with the neurological diseases ALS and cystic leukoencephalopathy, respectively, which is consistent with these proteins having a neuronal function[174–177]. Strikingly, the mRNA regulatory protein tristetraprolin (TTP; also known as ZFP36) was the most specifically upregulated RBP in the hippocampus by more than 200-fold. ZFP36 is known to destabilize mRNAs that encode cytokines and other inflammatory immune genes by recruiting the CCR4–NOT1 complex to AU-rich elements in the 3′ untranslated regions (3′UTRs) of targets, leading to deadenylation and subsequent degradation of mRNAs[178–180]. Recently, it was found that cytokines and other immune-regulatory proteins are expressed in the developing and adult nervous system, in which they are required for normal brain development and synaptic plasticity[181,182]. Indeed, coinciding with TTP, immune regulatory genes such as the members of major histocompatibility complex class I (for example, human leukocyte antigen A (HLA-A), HLA-B, HLA-C, HLA-E and HLA-F) were selectively expressed at 37 PCWs[183] (FIG. 7c, group III). Whether the molecular function of TTP remains the same in neurons is unknown, but the coordinated expression of these regulatory RBPs may imply a biological role during neural development and synaptic plasticity.

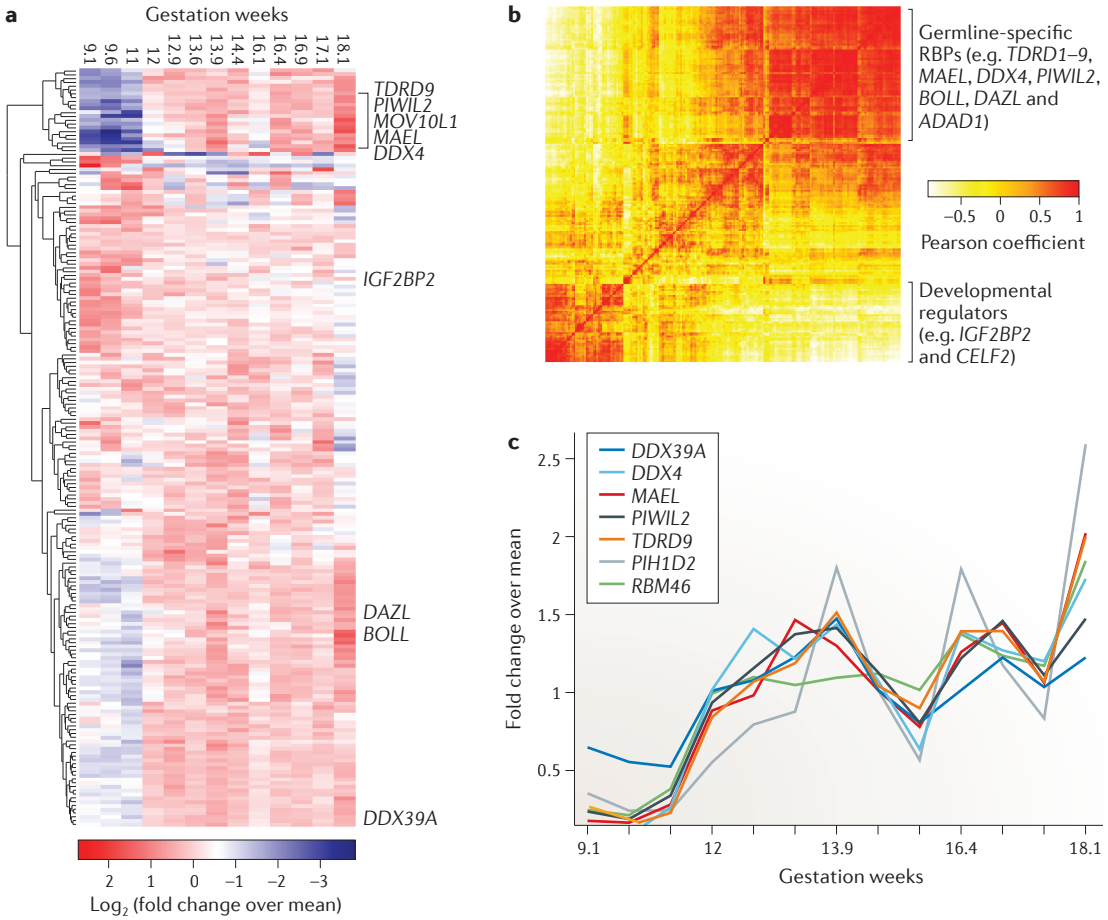


Figure 6 | **Expression of RBPs across nine gestational stages of human fetal ovarian development.** The top 200 most differentially expressed RNA-binding proteins (RBPs) from a microarray study profiling human fetal gonad development are shown[112]. For each gene microarray, intensity values were normalized to relative fold changes by dividing the expression value by the mean expression value across developmental stages. **a** | The heatmap shows the $\log_2$-transformed relative fold changes of the RBPs sorted by unsupervised clustering. Some gonad-specific RBPs are indicated. **b** | The Pearson correlation map indicates correlated expression changes of the 200 selected RBPs. Functionally related RBPs in gonad development cluster into a distinct expression group. **c** | The plot shows the normalized expression changes of selected genes relevant in gonad development.

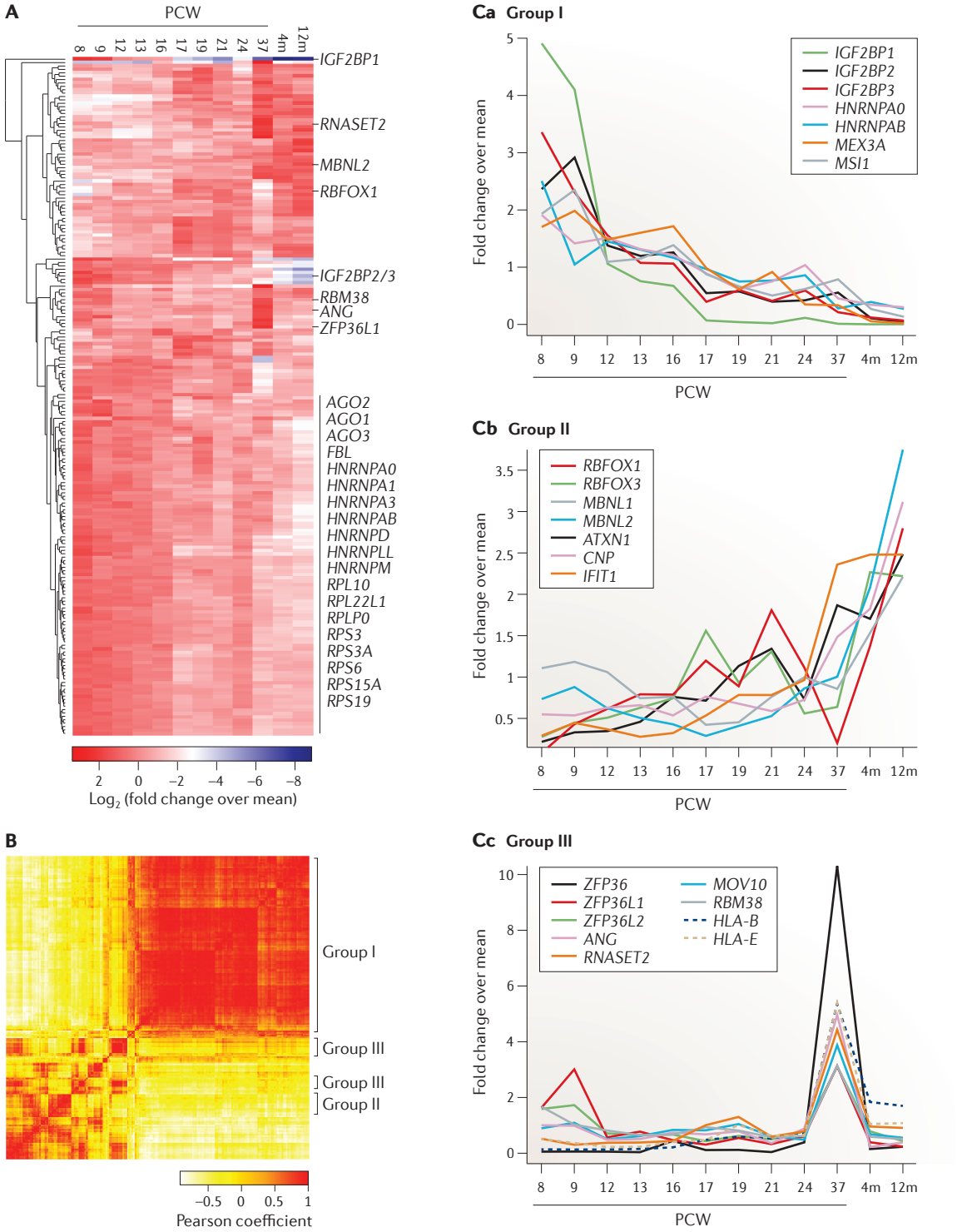


Figure 7 | **Expression of RBPs across human fetal hippocampus development.** The top 200 most differentially expressed RNA-binding proteins (RBPs) are shown across 12 stages of human hippocampus development ranging from post-conception week (PCW) 8 up to 12 months (12m) after birth, as profiled by RNA-sequencing (data from the BrainSpan database). For each gene, RPKM (reads per kilobase per million mapped reads) values were normalized to relative fold changes by dividing the expression value by the mean expression value across developmental stages. **A** | The heatmap shows the $\log_2$-transformed relative fold changes of the RBPs sorted by unsupervised clustering. **B** | The Pearson correlation map indicates correlated expression changes of the 200 selected RBPs. **C** | Characteristic expression fold changes across developmental stages are shown for genes in the three different groups. Group I includes genes with high expression levels at early PCWs, which rapidly decrease at later stages (part **Ca**). Group II includes genes with low expression levels at early PCWs and rapidly increasing levels at late PCWs and postnatal stages (part **Cb**). Group III includes genes with a single high-expression peak at 37 PCWs (part **Cc**).

Conclusions

A census of human RBPs is essential for organizing our current molecular and genetic understanding of the role of RNA in general gene expression and PTGR. This catalogue provides researchers with a newly curated resource to guide their investigations of PTGR processes and to systematically study RBPs. An analogous catalogue that assesses the abundance of all expressed RNAs (that is, the RBP targets) and that classifies them across tissues and cell types is still missing. Such a catalogue would be a useful complementary document to this census.

Of the ~20,500 protein-coding genes in humans, we determined that 7.5% are directly involved in RNA metabolism by binding to and/or processing RNA, or by constituting essential components of RNPs. RBPs are structurally diverse and include many distinct classes of RBDs. Indeed, whereas the three most abundant DNA-binding domains account for 80% of all TFs[58], the three most abundant RBDs accounted for only 20% of all RBPs in our census. Based on target-RNA categorization, we found that nearly 50% of RBPs acted in mRNA metabolic pathways and 11% constituted ribosomal proteins, while the rest were involved in the diverse number of ncRNA metabolic processes. The target-based categorization of RBPs can assist interpretation of disease phenotypes and mutations emerging from rapidly increasing patient genome sequencing, and may guide future functional studies. When considering abundances, we found that ribosomal proteins and mRBPs were the most abundant RBPs in the cell. Nevertheless, most RBPs were ubiquitously expressed at higher levels than the residual protein-coding transcriptome, and up to 20% of the total expressed protein-coding transcripts encoded RBPs. Therefore, not only is RNA metabolism one of the most conserved cellular processes, but it also has one of the highest protein copy number demands.

Many details of PTGR remain to be revealed, including the dissection of newly discovered RNA regulatory processes[1,184,185]. The investigation of PTGR networks is aided by the rapid development of next-generation sequencing-based methods, such as RIP- and CLIP-based methods[2,3,14], ribosome profiling[186], *in vivo* RNA secondary structure profiling[187–189], small and long RNA-seq[6,190,191], and 3′-end sequencing methods that profile alternative polyadenylation sites and poly(A) tail lengths[158,192–194]. These studies reveal an unanticipated complexity in RBP binding and targeting, and highlight the need to experimentally dissect PTGR networks in various cellular systems.

1. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution — trashing old rules to forge new ones. *Cell* **157**, 77–94 (2014).
This is a concise overview of the different RNA classes in bacteria, archaea and eukaryotes, highlighting their discovery and regulatory roles.
2. Konig, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein–RNA interactions: new genomic technologies and perspectives. *Nature Rev. Genet.* **13**, 77–83 (2011).
3. Ascano, M., Hafner, M., Cekan, P., Gerstberger, S. & Tuschl, T. Identification of RNA–protein interaction networks using PAR-CLIP. *Wiley Interdiscip. Rev. RNA***3**, 159–177 (2011).
4. Gerstberger, S., Hafner, M. & Tuschl, T. Learning the language of post-transcriptional gene regulation. *Genome Biol.* **14**, 130 (2013).
5. Mann, M. Functional and quantitative proteomics using SILAC. *Nature Rev. Mol. Cell. Biol.* **7**, 952–958 (2006).
6. Wang, Z., Gerstein, M. & Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
7. Stoltenburg, R., Reinemann, C. & Strehlitz, B. SELEX — a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol. Engineer.* **24**, 381–403 (2007).
8. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
9. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
10. Dreyfuss, G., Choi, Y. D. & Adam, S. A. Characterization of heterogeneous nuclear RNA–protein complexes *in vivo* with monoclonal antibodies. *Mol. Cell. Biol.* **4**, 1104–1114 (1984).
11. Pinol-Roma, S., Choi, Y. D., Matunis, M. J. & Dreyfuss, G. Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA-binding proteins. *Genes Dev.* **2**, 215–227 (1988).
12. Tenenbaum, S. A., Carson, C. C., Lager, P. J. & Keene, J. D. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl Acad. Sci. USA* **97**, 14085–14090 (2000).
13. Ascano, M., Gerstberger, S. & Tuschl, T. Multi-disciplinary methods to define RNA–protein interactions and regulatory networks. *Curr. Opin. Genet. Dev.* **23**, 20–28 (2013).
14. McHugh, C. A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNA–protein interactions. *Genome Biol.* **15**, 203 (2014).
15. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
16. Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res.* **37**, D229–D232 (2009).
17. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
18. Wilson, D. *et al.* SUPERFAMILY — sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37**, D380–D386 (2009).
19. Marchler-Bauer, A. *et al.* CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* **41**, D348–D352 (2013).
20. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
21. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29**, 41–43 (2001).
22. McKee, A. E. *et al.* A genome-wide *in situ* hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. *BMC Dev. Biol.* **5**, 14 (2005).
23. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* **39**, D301–D308 (2011).
24. Galante, P. A. F. *et al.* A comprehensive *in silico* expression analysis of RNA binding proteins in normal and tumor tissue: Identification of potential players in tumor formation. *RNA Biol.* **6**, 426–433 (2009).
25. Anantharaman, V., Koonin, E. V. & Aravind, L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.* **30**, 1427–1464 (2002).
This is one of the first genome-wide comparative studies profiling the proteins involved in RNA metabolism, which concluded that RNA metabolism is the most evolutionary conserved of all cellular systems. It gives a detailed account of the structural, functional and phylogenetic relationships of protein domains in RNA metabolism, and analyses the number of genes containing RBDs across 30 different organisms.
26. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
27. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
28. Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690 (2012).
References 27 and 28 describe the first large-scale crosslinking studies combined with quantitative mass spectrometry for the proteome-wide identification of poly(A)-RBPs.
29. Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nature Struct. Mol. Biol.* **20**, 1122–1130 (2013).
30. Mitchell, S. F., Jain, S., She, M. & Parker, R. Global analysis of yeast mRNPs. *Nature Struct. Mol. Biol.* **20**, 127–133 (2013).
31. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
32. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nature Rev. Mol. Cell. Biol.* **8**, 479–490 (2007).
This review summarizes the most commonly found RBDs and gives an overview of their structural characteristics and binding modes.
33. Burd, C. G. & Dreyfuss, G. Conserved structures and diversity of functions of RNA-binding proteins. *Science* **265**, 615–621 (1994).
34. Arcus, V. OB-fold domains: a snapshot of the evolution of sequence, structure and function. *Curr. Opin. Struct. Biol.* **12**, 794–801 (2002).
35. Kim, C. A. & Bowie, J. U. SAM domains: uniform structure, diversity of function. *Trends Biochem. Sci.* **28**, 625–628 (2003).
36. Rajkowitsch, L. *et al.* RNA chaperones, RNA annealers and RNA helicases. *RNA Biol.* **4**, 118–130 (2007).
37. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).
38. Sommerville, J. Activities of cold-shock domain proteins in translation control. *Bioessays* **21**, 319–325 (1999).
39. Mihailovich, M., Militti, C., Gabaldón, T. & Gebauer, F. Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression. *Bioessays* **32**, 109–118 (2010).

40. Curry, S., Kotik-Kogan, O., Conte, M. R. & Brick, P. Getting to the end of RNA: structural analysis of protein recognition of 5′ and 3′ termini. *Biochim. Biophys. Acta.* **1789**, 653–666 (2009).
41. Auweter, S. D., Oberstrass, F. C. & Allain, F. H. T. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res.* **34**, 4943–4959 (2006).
This is a highly detailed review on the structural determinants of RNA binding for ssRBDs.
42. Singh, R. & Valcarcel, J. Building specificity with nonspecific RNA-binding proteins. *Nature Struct. Mol. Biol.* **12**, 645–653 (2005).
43. Kuchta, K., Knizewski, L., Wyrwicz, L. S., Rychlewski, L. & Ginalski, K. Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human. *Nucleic Acids Res.* **37**, 7701–7714 (2009).
44. Valverde, R., Edwards, L. & Regan, L. Structure and function of KH domains. *FEBS J.* **275**, 2712–2726 (2008).
45. Masliah, G., Barraud, P. & Allain, F. H. T. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cell. Mol. Life Sci.* **70**, 1875–1895 (2013).
46. Chang, K.-Y. & Ramos, A. The double-stranded RNA-binding motif, a versatile macromolecular docking platform. *FEBS J.* **272**, 2109–2117 (2005).
47. Wilusz, C. J. & Wilusz, J. Eukaryotic Lsm proteins: lessons from bacteria. *Nature Struct. Mol. Biol.* **12**, 1031–1036 (2005).
48. Tharun, S. Roles of eukaryotic Lsm proteins in the regulation of mRNA function. *Int. Rev. Cell. Mol. Biol.* **272**, 149–189 (2009).
49. Wang, X., McLachlan, J., Zamore, P. D. & Hall, T. M. T. Modular recognition of RNA by a human pumilio-homology domain. *Cell* **110**, 501–512 (2002).
50. Linder, P. & Jankowsky, E. From unwinding to clamping — the DEAD box RNA helicase family. *Nature Rev. Mol. Cell. Biol.* **12**, 505–516 (2011).
51. Jankowsky, E. RNA helicases at work: binding and rearranging. *Trends Biochem. Sci.* **36**, 19–29 (2011).
52. Tanner, N. K. & Linder, P. DExD/H box RNA helicases: from generic motors to specific dissociation functions. *Mol. Cell* **8**, 251–262 (2001).
53. Rocak, S. & Linder, P. DEAD-box proteins: the driving forces behind RNA metabolism. *Nature Rev. Mol. Cell. Biol.* **5**, 232–241 (2004).
54. Meister, G. Argonaute proteins: functional insights and emerging roles. *Nature Rev. Genet.* **14**, 447–459 (2013).
55. Draper, D. E. & Reynaldo, L. P. RNA binding strategies of ribosomal proteins. *Nucleic Acids Res.* **27**, 381–388 (1999).
56. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nature Rev. Genet.* **11**, 345–355 (2010).
57. Chen, M. & Manley, J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature Rev. Mol. Cell. Biol.* **10**, 741–754 (2009).
58. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Rev. Genet.* **10**, 252–263 (2009).
Analogous to this Analysis, this article presents a catalogue for curated human TFs. It describes a census of ~1,400 TFs and gives an overview of common structural domains, tissue-specific expression and evolutionary conservation.
59. Kechavarzi, B. & Janga, S. C. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol.* **15**, R14 (2014).
60. Boisvert, F.-M., van Koningsbruggen, S., Navascués, J. & Lamond, A. I. The multifunctional nucleolus. *Nature Rev. Mol. Cell. Biol.* **8**, 574–585 (2007).
61. Montanaro, L., Trerè, D. & Derenzini, M. Nucleolus, ribosomes, and cancer. *Am. J. Pathol.* **173**, 301–310 (2008).
62. Ruggero, D. & Pandolfi, P. P. Does the ribosome translate cancer? *Nature Rev. Cancer* **3**, 179–192 (2003).
63. Ma, T. *et al.* Suppression of eIF2α kinases alleviates Alzheimer's disease-related plasticity and memory deficits. *Nature Neurosci.* **16**, 1299–1305 (2013).
64. Martin, I. *et al.* Ribosomal protein s15 phosphorylation mediates LRRK2 neurodegeneration in Parkinson's disease. *Cell* **157**, 472–485 (2014).
65. Klein, C. & Westenberger, A. Genetics of Parkinson's disease. *Cold Spring Harb. Perspect. Med.* **2**, a008888 (2012).
66. Scheper, G. C., van der Knaap, M. S. & Proud, C. G. Translation matters: protein synthesis defects in inherited disease. *Nature Rev. Genet.* **8**, 711–723 (2007).
This is a comprehensive review of mRNA-binding, tRNA-binding and ribosomal proteins involved in translation, genetic mutations of which cause human diseases.
67. Silvera, D., Formenti, S. C. & Schneider, R. J. Translational control in cancer. *Nature Rev. Cancer* **10**, 254–266 (2010).
This article discusses dysregulation of translation in human cancers and the factors involved, the loss or increased expression of which are found in different cancers, as well as the relevant druggable targets.
68. Hein, N., Hannan, K. M., George, A. J., Sanij, E. & Hannan, R. D. The nucleolus: an emerging target for cancer therapy. *Trends Mol. Med.* **19**, 643–654 (2013).
69. Skrtić, M. *et al.* Inhibition of mitochondrial translation as a therapeutic strategy for human acute myeloid leukemia. *Cancer Cell* **20**, 674–688 (2011).
70. Grzmil, M. & Hemmings, B. A. Translation regulation as a therapeutic target in cancer. *Cancer Res.* **72**, 3891–3900 (2012).
This paper describes different druggable targets for regulating aberrant protein translation in diseases such as cancers.
71. Macias, S. *et al.* DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. *Nature Struct. Mol. Biol.* **19**, 760–766 (2012).
72. Hafner, M. *et al.* Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition. *RNA* **19**, 613–626 (2013).
73. Wilbert, M. L. *et al.* LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. *Mol. Cell* **48**, 195–206 (2012).
74. Cho, J. *et al.* LIN28A is a suppressor of ER-associated translation in embryonic stem cells. *Cell* **151**, 765–777 (2012).
75. Tafforeau, L. *et al.* The complexity of human ribosome biogenesis revealed by systematic nucleolar screening of pre-rRNA processing factors. *Mol. Cell* **51**, 539–551 (2013).
76. Henras, A. K. *et al.* The post-transcriptional steps of eukaryotic ribosome biogenesis. *Cell. Mol. Life Sci.* **65**, 2334–2359 (2008).
77. Bratkovič, T. & Rogelj, B. The many faces of small nucleolar RNAs. *Biochim. Biophys. Acta.* **1839**, 438–443 (2014).
78. Yin, Q.-F. *et al.* Long noncoding RNAs with snoRNA ends. *Mol. Cell* **48**, 219–230 (2012).
79. Phizicky, E. M. & Hopper, A. K. tRNA biology charges to the front. *Genes Dev.* **24**, 1832–1860 (2010).
80. Hopper, A. K., Pai, D. A. & Engelke, D. R. Cellular dynamics of tRNAs and their genes. *FEBS Lett.* **584**, 310–317 (2010).
81. Kiss, T. Biogenesis of small nuclear RNPs. *J. Cell Sci.* **117**, 5949–5951 (2004).
82. Phipps, K. R., Charette, J. M. & Baserga, S. J. The small subunit processome in ribosome biogenesis — progress and prospects. *Wiley Interdiscip. Rev. RNA* **2**, 1–21 (2011).
83. Hussain, S. *et al.* NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep.* **4**, 255–261 (2013).
84. Sibbritt, T., Patel, H. R. & Preiss, T. Mapping and significance of the mRNA methylome. *Wiley Interdiscip. Rev. RNA* **4**, 397–422 (2013).
85. Spencer, C. M. *et al.* Exaggerated behavioral phenotypes in *Fmr1/Fxr2* double knockout mice reveal a functional genetic interaction between fragile X-related proteins. *Hum. Mol. Genet.* **15**, 1984–1994 (2006).
86. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143 (2001).
87. Woolford, J. L. & Baserga, S. J. Ribosome biogenesis in the yeast *Saccharomyces cerevisiae*. *Genetics* **195**, 643–681 (2013).
88. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
89. Fairman-Williams, M. E., Guenther, U.-P. & Jankowsky, E. SF1 and SF2 helicases: family matters. *Curr. Opin. Struct. Biol.* **20**, 313–324 (2010).
90. Krishna, S. S., Majumdar, I. & Grishin, N. V. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* **31**, 532–550 (2003).
91. Kerner, P., Degnan, S. M., Marchand, L., Degnan, B. M. & Vervoort, M. Evolution of RNA-binding proteins in animals: insights from genome-wide analysis in the sponge *Amphimedon queenslandica*. *Mol. Biol. Evol.* **28**, 2289–2303 (2011).
92. Granneman, S. & Baserga, S. J. Ribosome biogenesis: of knobs and RNA processing. *Exp. Cell Res.* **296**, 43–50 (2004).
93. Winter, E. E., Goodstadt, L. & Ponting, C. P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* **14**, 54–61 (2004).
94. Freilich, S. *et al.* Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol.* **6**, R56 (2005).
95. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5**, e1000598 (2009).
This is one of the first RNA-seq studies to investigate the tissue specificity of genes based on mRNA expression levels in 16 human tissues and cell types.
96. Dezso, Z. *et al.* A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol.* **6**, 49 (2008).
97. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840 (2010).
98. Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
99. Thomson, T. & Lin, H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu. Rev. Cell Dev. Biol.* **25**, 355–376 (2009).
100. Li, Q., Lee, J.-A. & Black, D. L. Neuronal regulation of alternative pre-mRNA splicing. *Nature Rev. Neurosci.* **8**, 819–831 (2007).
101. Castle, J. C. *et al.* Digital genome-wide ncRNA expression, including snoRNAs, across 11 human tissues using polyA-neutral amplification. *PLoS ONE* **5**, e11779 (2010).
102. Dittmar, K. A., Goodenbour, J. M. & Pan, T. Tissue-specific differences in human transfer RNA expression. *PLoS Genet.* **2**, e221 (2006).
103. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nature Rev. Genet.* **12**, 32–42 (2011).
104. Warner, J. R. & McIntosh, K. B. How common are extraribosomal functions of ribosomal proteins? *Mol. Cell* **34**, 3–11 (2009).
105. Xue, S. & Barna, M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nature Rev. Mol. Cell. Biol.* **13**, 355–369 (2012).
106. Luteijn, M. J. & Ketting, R. F. PIWI-interacting RNAs: from generation to transgenerational epigenetics. *Nature Rev. Genet.* **14**, 523–534 (2013).
107. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Rev. Mol. Cell. Biol.* **12**, 246–258 (2011).
108. Seydoux, G. & Braun, R. E. Pathway to totipotency: Lessons from germ cells. *Cell* **127**, 891–904 (2006).
109. Kotaja, N. & Sassone-Corsi, P. The chromatoid body: a germ-cell-specific RNA-processing centre. *Nature Rev. Mol. Cell. Biol.* **8**, 85–90 (2007).
110. Voronina, E., Seydoux, G., Sassone-Corsi, P. & Nagamori, I. RNA granules in germ cells. *Cold Spring Harb. Perspect. Biol.* **3**, a002774 (2011).
111. Kang, M. K. & Han, S. J. Post-transcriptional and post-translational regulation during mouse oocyte maturation. *BMB Rep.* **44**, 147–157 (2011).
112. Houmard, B. *et al.* Global gene expression in the human fetal testis and ovary. *Biol. Reprod.* **81**, 438–443 (2009).
113. Brook, M., Smith, J. W. S. & Gray, N. K. The DAZL and PABP families: RNA-binding proteins with interrelated roles in translational control in oocytes. *Reproduction* **137**, 595–617 (2009).
114. Reynolds, N. & Cooke, H. J. Role of the *DAZ* genes in male fertility. *Reprod. Biomed. Online* **10**, 72–80 (2005).
115. Lasko, P. The DEAD-box helicase Vasa: evidence for a multiplicity of functions in RNA processes and developmental biology. *Biochim. Biophys. Acta.* **1829**, 810–816 (2013).

116. Frost, R. J. A. *et al.* MOV10L1 is necessary for protection of spermatocytes against retrotransposons by PIWI-interacting RNAs. *Proc. Natl Acad. Sci. USA* **107**, 11847–11852 (2010).
117. Zheng, K. *et al.* Mouse MOV10L1 associates with PIWI proteins and is an essential component of the PIWI-interacting RNA (piRNA) pathway. *Proc. Natl Acad. Sci. USA* **107**, 11841–11846 (2010).
118. Dufau, M. L. & Tsai-Morris, C.-H. Gonadotropin-regulated testicular helicase (GRTH/DDX25): an essential regulator of spermatogenesis. *Trends Endocrinol. Metab.* **18**, 314–320 (2007).
119. Rosenberg, H. F. in *Ribonucleases* Ch. 2 (ed. Nicholson, A. W.) 35–53 (Springer, 2011).
120. Yisraeli, J. K. VICKZ proteins: a multi-talented family of regulatory RNA-binding proteins. *Biol. Cell* **97**, 87–96 (2005).
121. Simone, L. E. & Keene, J. D. Mechanisms coordinating ELAV/Hu mRNA regulons. *Curr. Opin. Genet. Dev.* **23**, 35–43 (2013).
122. Ascano, M. *et al.* FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492**, 382–386 (2012).
References 2, 14 and 122 give comprehensive and balanced accounts of different methods developed for the genome-wide identification of RBPs and RBP-binding sites.
123. Wang, T., Bray, S. M. & Warren, S. T. New perspectives on the biology of fragile X syndrome. *Curr. Opin. Genet. Dev.* **22**, 256–263 (2012).
124. Mientjes, E. J. *et al. Fxr1* knockout mice show a striated muscle phenotype: implications for Fxr1p function *in vivo*. *Hum. Mol. Genet.* **13**, 1291–1302 (2004).
125. Narla, A. & Ebert, B. L. Ribosomopathies: human disorders of ribosome dysfunction. *Blood* **115**, 3196–3205 (2010).
126. Lukong, K. E., Chang, K. W., Khandjian, E. W. & Richard, S. RNA-binding proteins in human genetic disease. *Trends Genet.* **24**, 416–425 (2008).
127. Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793 (2009).
This is a comprehensive overview of RNA- and RBP-based genetic diseases caused by mutations in RNAs and RBPs, and highlights the most prominent examples.
128. Ramaswami, M., Taylor, J. P. & Parker, R. Altered ribostasis: RNA–protein granules in degenerative disorders. *Cell* **154**, 727–736 (2013).
This paper highlights prion-like RBP aggregation in human diseases caused by mutations in RBPs.
129. Buchan, J. R., Kolaitis, R.-M., Taylor, J. P. & Parker, R. Eukaryotic stress granules are cleared by autophagy and Cdc48/VCP function. *Cell* **153**, 1461–1474 (2013).
130. Liu-Yesucevitz, L. *et al.* Local RNA translation at the synapse and in disease. *J. Neurosci.* **31**, 16086–16093 (2011).
131. Lagier-Tourenne, C., Polymenidou, M. & Cleveland, D. W. TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Hum. Mol. Genet.* **19**, R46–R64 (2010).
132. Kim, H. J. *et al.* Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature* **495**, 467–473 (2013).
133. Orr, H. T. *et al.* Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature Genet.* **4**, 221–226 (1993).
134. Banfi, S. *et al.* Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nature Genet.* **7**, 513–520 (1994).
135. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
136. Echeverria, G. V. & Cooper, T. A. RNA-binding proteins in microsatellite expansion disorders: mediators of RNA toxicity. *Brain Res.* **1462**, 100–111 (2012).
137. Budde, B. S. *et al.* tRNA splicing endonuclease mutations cause pontocerebellar hypoplasia. *Nature Genet.* **40**, 1113–1118 (2008).
138. Yao, P. & Fox, P. L. Aminoacyl-tRNA synthetases in medicine and disease. *EMBO Mol. Med.* **5**, 332–343 (2013).
139. Rice, G. I. *et al.* Mutations involved in Aicardi–Goutières syndrome implicate SAMHD1 as regulator of the innate immune response. *Nature Genet.* **41**, 829–832 (2009).
140. Crow, Y. J. *et al.* Mutations in genes encoding ribonuclease H2 subunits cause Aicardi–Goutières syndrome and mimic congenital viral brain infection. *Nature Genet.* **38**, 910–916 (2006).
141. Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nature Rev. Mol. Cell. Biol.* **3**, 195–205 (2002).
142. Müller-McNicoll, M. & Neugebauer, K. M. How cells get the message: dynamic assembly and function of mRNA–protein complexes. *Nature Rev. Genet.* **14**, 275–287 (2013).
143. Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nature Rev. Genet.* **8**, 533–543 (2007).
144. Mitchell, S. F. & Parker, R. Principles and properties of eukaryotic mRNPs. *Mol. Cell* **54**, 547–558 (2014).
145. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Rev. Mol. Cell. Biol.* **14**, 153–165 (2013).
146. Smith, C. W. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* **25**, 381–388 (2000).
147. Wahl, M. C., Will, C. L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701–718 (2009).
148. Kalsotra, A. & Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nature Rev. Genet.* **12**, 715–729 (2011).
149. Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010).
150. Berg, M. G. *et al.* U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**, 53–64 (2012).
151. Mukherjee, N. *et al.* Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell* **43**, 327–339 (2011).
152. Kedde, M. *et al.* A Pumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. *Nature Cell Biol.* **12**, 1014–1020 (2010).
153. Anderson, P. & Kedersha, N. RNA granules: post-transcriptional and epigenetic modulators of gene expression. *Nature Rev. Mol. Cell. Biol.* **10**, 430–436 (2009).
154. Hanna, J. H., Saha, K. & Jaenisch, R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell* **143**, 508–525 (2010).
155. Cirillo, D. *et al.* Constitutive patterns of gene expression regulated by RNA-binding proteins. *Genome Biol.* **15**, R13 (2014).
156. Mittal, N., Roy, N., Babu, M. M. & Janga, S. C. Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proc. Natl Acad. Sci. USA* **106**, 20300–20305 (2009).
157. Norbury, C. J. Cytoplasmic RNA: a case of the tail wagging the dog. *Nature Rev. Cancer* **13**, 643–653 (2013).
158. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**, 2380–2396 (2013).
This is a detailed study profiling genome-wide alternative polyadenylation sites in mRNAs across 12 human cell lines and tissues. The authors conclude that genes with multiple 3′UTRs tend to vary 3′UTR ratios across tissues, whereas genes with single 3′UTRs vary mRNA expression levels.
159. Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and consequences of alternative polyadenylation. *Mol. Cell* **43**, 853–866 (2011).
160. MacDonald, C. C. & McMahon, K. W. Tissue-specific mechanisms of alternative polyadenylation: testis, brain, and beyond. *Wiley Interdiscip. Rev. RNA***1**, 494–501 (2010).
161. Oktem, O. & Urman, B. Understanding follicle growth *in vivo*. *Hum. Reprod.* **25**, 2944–2954 (2010).
162. Bell, J. L. *et al.* Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell. Mol. Life Sci.* **70**, 2657–2675 (2013).
163. Kee, K., Angeles, V. T., Flores, M., Nguyen, H. N. & Reijo Pera, R. A. Human *DAZL*, *DAZ* and *BOULE* genes modulate primordial germ-cell and haploid gamete formation. *Nature* **462**, 222–225 (2009).
164. Bramham, C. R. & Wells, D. G. Dendritic mRNA: transport, translation and function. *Nature Rev. Neurosci.* **8**, 776–789 (2007).
165. Jung, H., Gkogkas, C. G., Sonenberg, N. & Holt, C. E. Remote control of gene function by local translation. *Cell* **157**, 26–40 (2014).
166. Kandel, E. R., Dudai, Y. & Mayford, M. R. The molecular and systems biology of memory. *Cell* **157**, 163–186 (2014).
167. Sutton, M. A. & Schuman, E. M. Dendritic protein synthesis, synaptic plasticity, and memory. *Cell* **127**, 49–58 (2006).
168. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
169. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
170. Mody, M. *et al.* Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc. Natl Acad. Sci. USA* **98**, 8862–8867 (2001).
171. Thornton, J. E. & Gregory, R. I. How does Lin28 let-7 control development and disease? *Trends Cell Biol.* **22**, 474–482 (2012).
172. Gehman, L. T. *et al.* The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nature Genet.* **43**, 706–711 (2011).
173. Arnold, S. E. & Trojanowski, J. Q. Human fetal hippocampal development: I. Cytoarchitecture, myeloarchitecture, and neuronal morphologic features. *J. Comp. Neurol.* **367**, 274–292 (1996).
174. Greenway, M. J. *et al. ANG* mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nature Genet.* **38**, 411–413 (2006).
175. Henneke, M. *et al.* RNASET2-deficient cystic leukoencephalopathy resembles congenital cytomegalovirus brain infection. *Nature Genet.* **41**, 773–775 (2009).
176. Thiyagarajan, N., Ferguson, R., Subramanian, V. & Acharya, K. R. Structural and molecular insights into the mechanism of action of human angiogenin-ALS variants in neurons. *Nature Commun.* **3**, 1121 (2012).
177. Skorupa, A. *et al.* Motoneurons secrete angiogenin to induce RNA cleavage in astroglia. *J. Neurosci.* **32**, 5024–5038 (2012).
178. Mukherjee, N. *et al.* Global target mRNA specification and regulation by the RNA-binding protein ZFP36. *Genome Biol.* **15**, R12 (2014).
179. Fabian, M. R. *et al.* miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4–NOT. *Nature Struct. Mol. Biol.* **18**, 1211–1217 (2011).
180. Brooks, S. A. & Blackshear, P. J. Tristetraprolin (TTP): Interactions with mRNA and proteins, and current thoughts on mechanisms of action. *Biochim. Biophys. Acta.* **1829**, 666–679 (2013).
181. Boulanger, L. M. Immune proteins in brain development and synaptic plasticity. *Neuron* **64**, 93–109 (2009).
182. Deverman, B. E. & Patterson, P. H. Cytokines and CNS development. *Neuron* **64**, 61–78 (2009).
183. Zhang, A. *et al.* The spatio-temporal expression of MHC class I molecules during human hippocampal formation development. *Brain Res.* **1529**, 26–38 (2013).
184. Meyer, K. D. & Jaffrey, S. R. The dynamic epitranscriptome: N^6-methyladenosine and gene expression control. *Nature Rev. Mol. Cell. Biol.* **15**, 313–326 (2014).
185. Ulitsky, I. & Bartel, D. P. lincRNAs: Genomics, evolution, and mechanisms. *Cell* **154**, 26–46 (2013).
This is a detailed review on the emerging roles of lncRNAs in gene regulation.
186. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Rev. Genet.* **15**, 205–213 (2014).
This article gives an overview of ribosome profiling, which is a method to measure actively translating RNAs genome-wide. Next to mass spectrometry, ribosome profiling allows the quantification of expressed proteins in the cell and also the measurement of translation rates of mRNAs.
187. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).
188. Ding, Y. *et al. In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).
189. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature* **505**, 701–705 (2014).
190. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature Rev. Genet.* **12**, 87–98 (2011).

191. Pritchard, C. C., Cheng, H. H. & Tewari, M. MicroRNA profiling: approaches and considerations. *Nature Rev. Genet.* **13**, 358–369 (2012). **References 189 and 191 describe transcriptome-wide methods for determining RNA structures *in vivo*, which give insights into RNA accessibility and regulation.**
192. Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* **469**, 97–101 (2011). **This study describes one of the first RNA-seq methods to accurately profile alternative polyadenylation sites genome-wide.**
193. Chang, H., Lim, J., Ha, M. & Kim, V. N. TAIL-seq: Genome-wide determination of poly(A) tail length and 3′ end modifications. *Mol. Cell* **53**, 1044–1052 (2014).
194. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**, 66–71 (2014). **This paper details a protocol to map genome-wide mRNA poly(A) tail length *in vivo*.**
195. Maraia, R. J. & Lamichhane, T. N. 3′ processing of eukaryotic precursor tRNAs. *Wiley Interdiscip. Rev. RNA* **2**, 362–375 (2011).
196. Thomson, E., Ferreira-Cerca, S. & Hurt, E. Eukaryotic ribosome biogenesis at a glance. *J. Cell Sci.* **126**, 4815–4821 (2013).
197. Lafontaine, D. L. & Tollervey, D. The function and synthesis of ribosomes. *Nature Rev. Mol. Cell. Biol.* **2**, 514–520 (2001).
198. Mroczek, S. & Dziembowski, A. U6 RNA biogenesis and disease association. *Wiley Interdiscip. Rev. RNA* **4**, 581–592 (2013).
199. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Rev. Mol. Cell. Biol.* **11**, 113–127 (2010).
200. Buchan, J. R. & Parker, R. Eukaryotic stress granules: The ins and outs of translation. *Mol. Cell* **36**, 932–941 (2009).
201. Parker, R. & Sheth, U. P bodies and the control of mRNA translation and degradation. *Mol. Cell* **25**, 635–646 (2007).
202. Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay. *Nature Rev. Mol. Cell. Biol.* **8**, 113–126 (2007).
203. Kim, V. N., Han, J. & Siomi, M. C. Biogenesis of small RNAs in animals. *Nature Rev. Mol. Cell. Biol.* **10**, 126–139 (2009).
204. Peterlin, B. M., Brogie, J. E. & Price, D. H. 7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription. *Wiley Interdiscip. Rev. RNA* **3**, 92–103 (2011).
205. Fox, A. H. & Lamond, A. I. Paraspeckles. *Cold Spring Harb. Perspect. Biol.* **2**, a000687 (2010).
206. Yoon, J.-H. *et al.* LincRNA-p21 suppresses target mRNA translation. *Mol. Cell* **47**, 648–655 (2012).
207. Doma, M. K. & Parker, R. RNA quality control in eukaryotes. *Cell* **131**, 660–668 (2007).
208. Houseley, J., LaCava, J. & Tollervey, D. RNA-quality control by the exosome. *Nature Rev. Mol. Cell. Biol.* **7**, 529–539 (2006).

Acknowledgements
The body map data were kindly provided by the Gene Expression Applications research group at Illumina. The authors thank P. Morozov, M. Carty, M. Brown, R. Kim and S. Lianoglou for discussions on the computational methods, as well as Z. Ozair, A. D. Haase and all laboratory members for comments on the manuscript. S.G. was supported by a Ph.D. fellowship from the Boehringer Ingelheim Fonds. M.H. is supported by the US National Institute of Arthritis and Musculoskeletal and Skin Diseases Intramural Research Program. T.T. is an Investigator of the Howard Hughes Medical Institute.

Competing interests statement
The authors declare competing interests: see Web version for details.

FURTHER INFORMATION
BrainSpan: http://www.brainspan.org
Illumina Body Atlas: http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/

SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (box) | S3 (table) | S4 (table) | S5 (table) | S6 (table) | S7 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF