nature structural & molecular biology

# The RNA-binding protein repertoire of embryonic stem cells

S Chul Kwon[1,2], Hyerim Yi[1,2], Katrin Eichelbaum[3], Sophia Föhr[3], Bernd Fischer[3], Kwon Tae You[1,2], Alfredo Castello[3], Jeroen Krijgsveld[3], Matthias W Hentze[3,4] & V Narry Kim[1,2]

**RNA-binding proteins (RBPs) have essential roles in RNA-mediated gene regulation, and yet annotation of RBPs is limited mainly to those with known RNA-binding domains. To systematically identify the RBPs of embryonic stem cells (ESCs), we here employ interactome capture, which combines UV cross-linking of RBP to RNA in living cells, oligo(dT) capture and MS. From mouse ESCs (mESCs), we have defined 555 proteins constituting the mESC mRNA interactome, including 283 proteins not previously annotated as RBPs. Of these, 68 new RBP candidates are highly expressed in ESCs compared to differentiated cells, implicating a role in stem-cell physiology. Two well-known E3 ubiquitin ligases, Trim25 (also called Efp) and Trim71 (also called Lin41), are validated as RBPs, revealing a potential link between RNA biology and protein-modification pathways. Our study confirms and expands the atlas of RBPs, providing a useful resource for the study of the RNA-RBP network in stem cells.**



ESCs are derived from the inner cell mass of the blastocyst, and they retain the molecular identity and pluripotency of the preimplantation epiblast[1]. Their unlimited self-renewal capacity in culture allows biochemical dissection of the regulatory network in the stem-cell state[2]. Transcriptional networks governing pluripotency have in recent years been intensively characterized and used for genetic reprogramming[3–7]. The regulation and establishment of the stem cell at the post-transcriptional level is starting to be elucidated[8–11], paving the way for understanding combined regulatory networks at the DNA and RNA levels. RNAs in cellular environments typically act in a form of ribonucleoprotein complexes (RNPs) by interacting with RBPs. Thus RBPs constitute a major family of regulators by assisting and/or directing RNAs[12–15].

RBPs have been studied for decades, yet our understanding of RBPs is largely limited to proteins with known or predicted RNA-binding domains (RBDs). To overcome these limitations, several approaches have been undertaken to identify the complete repertoire of RBPs. For instance, protein microarrays representing 80% of the *Saccharomyces cerevisiae* proteome have been used to screen for proteins that bind to fluorescence-labeled RNA *in vitro*[16,17]. To complement this method, RNA-coated beads were incubated with whole-cell lysate, and the bound proteins were analyzed by LC-MS/MS[16]. Although these two approaches yielded candidates for novel RBPs, it is difficult to judge whether or not the interactions are physiologically relevant because the experiments were conducted *in vitro*. Recent studies have developed mRNA interactome capture, a method for systematic identification of the repertoire of RBPs in proliferating HeLa and HEK293 cells, respectively[18,19]. This method builds on previous work performed in the 1980s to identify heterogeneous nuclear RNPs[20] (hnRNPs). UV light introduces covalent bonds between RNA and binding proteins within subnanometer proximity, thereby enabling stringent purification and greatly reducing indirect or artificial interactions[18,19]. Because UV irradiation is carried out directly on living cells before cell lysis, only endogenous RNA-protein interactions are captured by cross-linking[18,19]. RNPs are subsequently recovered using oligo(dT) magnetic beads, and the associated proteins are analyzed by quantitative MS. A similar approach was taken recently to globally analyze yeast RBPs[21].

To investigate the nature, extent and functionality of RNA-protein interactions in stem cells, we here employ mRNA interactome capture. We identify 555 proteins that constitute the mESC mRNA interactome, including 283 novel RBP candidates. Among these, 68 are preferentially expressed in ESCs, as compared to differentiated cells. A substantial fraction of RBPs are transcriptionally regulated by the Myc complex. We further find that two well-known E3 ubiquitin ligases, Trim25 and Trim71, directly interact with RNA, which suggests a new mechanism coupling RNA biology and protein ubiquitination.

## RESULTS

### Identification of RBPs

To systematically and comprehensively identify RBPs in mESCs, we performed mRNA interactome capture analyses (**Fig. 1a**). Silver staining showed that UV cross-linking greatly enhanced protein isolation upon poly(A)$^+$ RNA capture (**Fig. 1b**, lanes 5 and 6). The protein pattern differed profoundly from that of the input sample (**Fig. 1b**, lanes 2 and 6). To test whether the eluted proteins were enriched for bona fide RBPs, we carried out western blotting to detect the well-known RNA binders Lin28a, hnRNP A and hnRNP C and found that they were present in UV-irradiated samples and absent from controls (**Fig. 1c**). In contrast, common contaminants such as histone H3 and α-tubulin were not detected in the precipitates (**Fig. 1c**). We also

[1]Center for RNA Research, Institute for Basic Science (IBS), Seoul, Korea. [2]School of Biological Sciences, Seoul National University, Seoul, Korea. [3]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [4]Molecular Genetics Division, Victor Chang Cardiac Research Institute, Sydney, Australia. Correspondence should be addressed to V.N.K (narrykim@snu.ac.kr).

Received 21 February; accepted 19 June; published online 4 August 2013; doi:10.1038/nsmb.2638

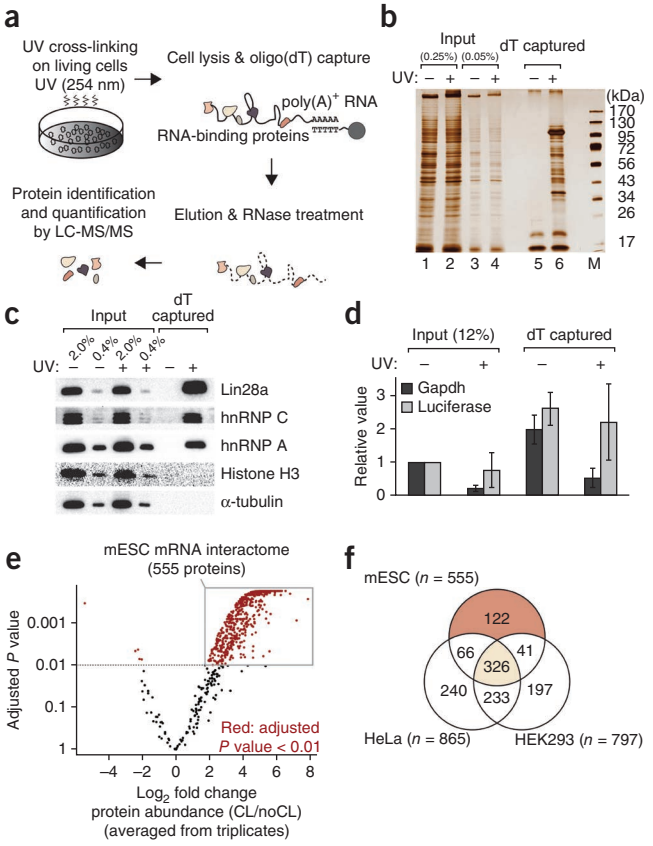



**Figure 1** Identification of RBPs in mESCs. (**a**) Schematic of the interactome-capture method. (**b**,**c**) Silver staining (**b**) and western blotting (**c**) of protein sample eluted from UV-irradiated (+) and nonirradiated (–) oligo(dT)-captured samples. (**d**) Quantitative RT-PCR to measure endogenous Gapdh mRNA levels and exogenous luciferase mRNA (spike-in) levels. *In vitro* transcribed, poly(A)-tailed luciferase mRNA was added to lysate after cell lysis. Error bars represent mean ± s.d. from three independent experiments. (**e**) The mESC mRNA interactome. Volcano plot displaying the average $\log_2$ ratios for all proteins quantified in at least two out of three biological replicates (*x* axis) and their adjusted *P* values (*y* axis). Proteins with $\log_2$ fold change that is statistically different from 0 ($P < 0.01$; empirical Bayes moderated *t*-test, adjusted by Benjamini-Hochberg method) are highlighted in red. Proteins ($n = 555$) that are significantly enriched in cross-linked samples (CL) compared to non–cross-linked samples (noCL) were included in the mRNA interactome (gray box). (**f**) Comparison of the mESC mRNA interactome and two human data sets[18,19].

investigated whether a comparable amount of RNA could be pulled down by quantitative reverse transcription PCR (RT-PCR) (**Fig. 1d**). Endogenous Gapdh mRNA and exogenous poly(A)-tailed luciferase mRNA (spike-in) that was mixed with lysate after cell lysis were pulled down from both the irradiated and nonirradiated samples. Of note, the level of endogenous Gapdh mRNA was reduced after UV exposure, possibly because cross-linked residues interfere with reverse transcription[22,23]. These results indicate that UV irradiation successfully induced cross-linking between RNAs and RBPs and that stringent washing with 0.5% lithium dodecyl sulfate and 500 mM lithium chloride was effective to remove contamination of abundant proteins lacking RNA-binding activity.

From three biologically independent experiments we identified 1,236 proteins, 720 of which were quantified from at least two replicates. Of these, 555 proteins were enriched in UV cross-linked samples, with a false-discovery rate (FDR) of 1% at the protein identification and quantification level (**Fig. 1e** and **Supplementary Table 1**). We refer to this collective set of candidate RBPs as the mESC mRNA interactome (**Fig. 1e** and **Supplementary Table 2**).

## Comparison to human mRNA interactome

Recent studies identified 860 human proteins (corresponding to 865 mouse homologs) from HeLa cells and 797 proteins (matching 797 mouse homologs) from HEK293 cells as candidate RBPs[18,19]. We compared the mESC data with the human data sets and found that orthologs of 326 proteins are common to all three data sets (**Fig. 1f** and **Supplementary Table 3**). Considering the differences in species, cell types and study protocols, the overlapping 326 proteins can be considered high-confidence RBPs of a core mammalian mRNA interactome. MS data for those proteins showed lower *P* values than for the proteins found only in one or two studies (**Supplementary Fig. 1a**). Forty percent (131/326) of the common proteins are not currently annotated as RBPs. As these 326 proteins were detected in three different cell types of distinct origins, many of them may have constitutive functions rather than tissue-specific roles.

Notably, our comparison revealed 122 proteins that appeared only in the mESC interactome, indicating that the mammalian mRNA interactome has not reached saturation (**Fig. 1f**). Given that RNA pathways are generally conserved in mammals, the difference between the mESC and human interactomes may have arisen from features unique to ESCs rather than species-specific variation. About half of the 122 proteins were preferentially expressed in ESCs rather than differentiated cell types, and this preferential expression may partly explain the exclusive detection of these proteins in our experiments.

It is also plausible that some of the proteins are active in ESCs only in the presence of certain post-transcriptional modifications and/or specific substrates and cofactors. Technical differences among labs may also have contributed to the differences in the proteomes.

## The mESC interactome is enriched with known RBPs

We inspected the biochemical properties of the mESC mRNA interactome to exclude the possibility of technical bias in MS analysis. We did not observe a substantial bias in terms of protein length or hydrophobicity compared to the predicted mouse proteome (**Supplementary Fig. 1b**). As previously shown[18], we noticed a shift in the distribution of isoelectric points of the interactome proteins toward a basic pH, a characteristic feature of known RBPs (**Fig. 2a**).

To determine the representative features of the mRNA interactome, we also conducted analyses using Gene Ontology (GO). The term 'RNA-binding' was strongly over-represented in our data set (**Fig. 2b**). Other RNA-related terms, such as 'structural constituent of ribosome' or 'ATP-dependent helicase activity' were also enriched, suggesting that the interactome contains a large number of known RBPs (**Fig. 2b** and **Supplementary Table 4**). Consistently, InterPro domain analysis showed that most of the top represented domains comprise RBDs, of which RNA recognition motif had the smallest *P* value (**Fig. 2c**). Moreover, general RNA-related GO terms such as 'ribonucleoprotein complex biogenesis', 'cytoplasmic stress granule', 'polysome', 'translational initiation complex formation', 'mRNA splicing' and 'transport of mature transcript to cytoplasm' were highly enriched, whereas signaling-related and membrane-related GO terms were underrepresented (**Supplementary Fig. 1c**,**d** and **Supplementary Table 4**).

We confirmed that nonclassical RBDs such as ribosomal, WD40 and YTH are also enriched in the mESC mRNA interactome (at an FDR of 1%; parameters for definition as classical or nonclassical are described

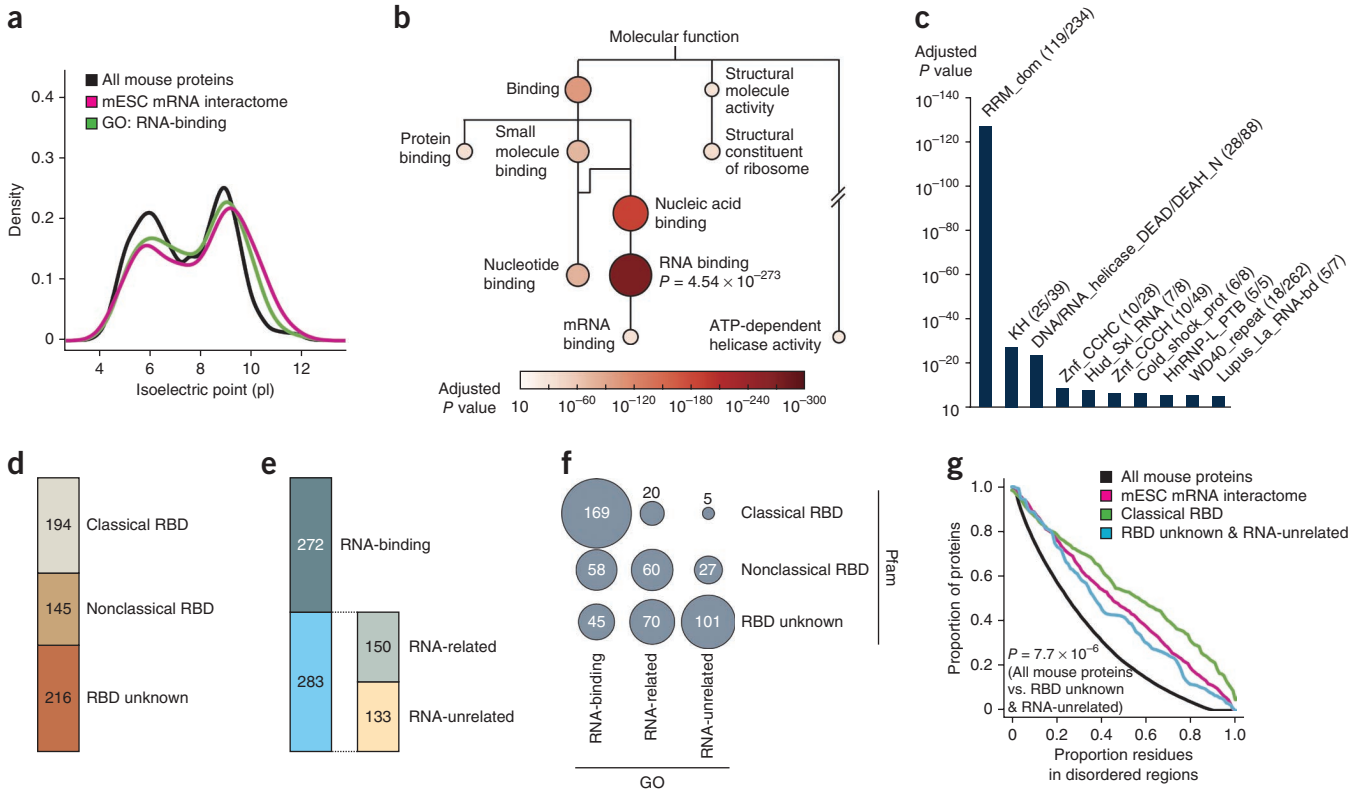



**Figure 2** Characteristics of the mESC mRNA interactome in comparison to known RBPs. (**a**) Density of isoelectric point. Shown are all mouse proteins (black), the mESC mRNA interactome (magenta) and annotated RBPs (green). (**b**) GO analysis. Top ten molecular function terms with the smallest *P* values are shown. The mESC mRNA interactome is compared with whole mouse proteome. *P* values were calculated by Fisher's exact test and corrected for multiple testing by Benjamini-Hochberg method. The color and the size of circles correspond to adjusted *P* values. Lines indicate hierarchical relationships. (**c**) Top ten protein domains with the smallest *P* values from InterPro. The numbers in parentheses indicate the numbers of mRNA-interactome proteins and of all proteins belonging to each domain, respectively. (**d**–**f**) Classification of the mESC mRNA interactome on the basis of Pfam (**d**), GO (**e**) and Pfam and GO (**f**). The numbers of proteins belonging to each category are indicated. (**g**) Distribution of predicted disordered regions for all mouse proteins (black), whole mESC mRNA interactome (magenta), mESC mRNA interactome proteins containing classical RBDs (green) and mESC mRNA interactome proteins without known RBDs and RNA-related GO terms (blue). *P* values were calculated by Kolmogorov–Smirnov test.



in Online Methods) (**Supplementary Fig. 2a**,**b** and **Supplementary Table 5a**,**b**). The WD40 domain is particularly noteworthy, as it is generally known as a protein-binding domain composed of tandem copies of WD40 repeats. But it was shown recently that the WD40 domain of Gemin5 can bind to RNA[24], and human mRNA interactome studies have revealed 25 proteins containing WD40 domains[18,19] (**Supplementary Table 5**). We found 16 WD40 proteins in the mESC interactome, of which 13 proteins had been identified in human interactomes as well, strongly supporting a wide usage of WD40 motifs as an RNA-binding platform. Notably, the RNA-binding WD40 proteins are enriched with basic amino acids compared to the global set of mouse WD40 proteins (239 in total), which is consistent with human data[18] (**Supplementary Fig. 2**). Therefore, the β-propeller structure of the WD40 domain may serve as both a protein- and an RNA-interaction platform, depending on the residues exposed on the interaction surface.

Our mESC mRNA interactome includes known RBPs as well as proteins not annotated as RNA-binding. According to Pfam domain annotation, 216 of the 555 mRNA interactome proteins (39%) do not carry any known RBDs (**Fig. 2d**). Among the 122 proteins unique to the mESC interactome (**Fig. 1f**), 66 (54%) were classified as 'RBD unknown', suggesting that a substantial fraction may be novel RBPs. We searched for domains or motifs that are enriched among RBD-unknown proteins (**Supplementary Fig. 2c** and **Supplementary Table 5c**) and found several poorly defined motifs, such as Pfam-B_3355, Pfam-B_969 and Pfam-B_2050, that may represent new RNA binding motifs.

Only 272 out of the 555 proteins were annotated with the GO term 'RNA-binding' (**Fig. 2e**). We examined the remaining 283 proteins manually and found that 150 proteins are associated with RNA-related GO terms and InterPro terms. We classified these proteins as 'RNA-related' (**Fig. 2e**). The rest (133 out of 555 (24%)) were neither RNA-binding nor RNA-related, so are classified here as 'RNA-unrelated' (**Fig. 2e**). When we intersected the domain-based grouping with the GO classifications (**Fig. 2f**), we found that 101 proteins (18% of the interactome) belong to the RBD-unknown and the RNA-unrelated group. Although it should be taken into account that GO annotation does not fully reflect the current knowledge in the literature, our results collectively indicate that a substantial fraction of the mESC mRNA interactome is made up of novel RBPs.

It was previously found that disordered regions, which lack stable structures, frequently appear in human RBPs[18]. Our mESC mRNA interactome is also enriched for disordered regions, indicating that they may constitute a general feature of RBPs[18,25–27] (**Fig. 2g**). Disordered regions are enriched with arginine, lysine, tyrosine and glycine, reflecting the presence of RGG, RS, YGG and polylysine motifs (**Supplementary Fig. 3a**,**b**). A recent report showed that low complexity peptide sequences can mediate the formation of RNA granules

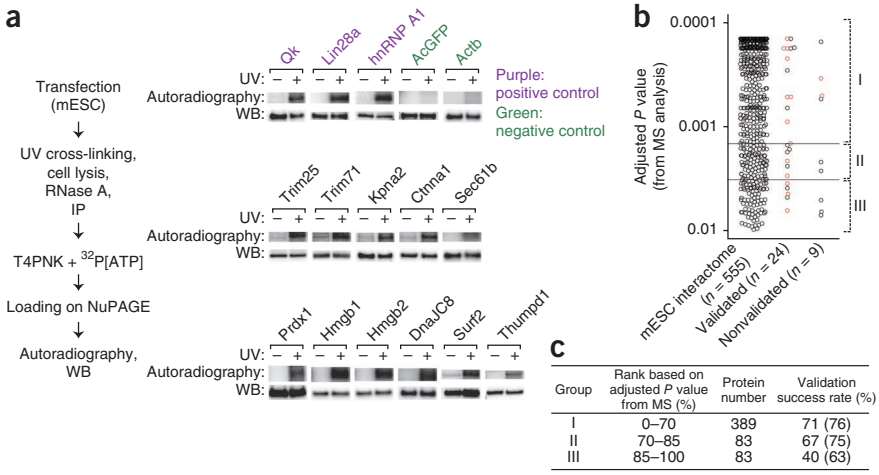
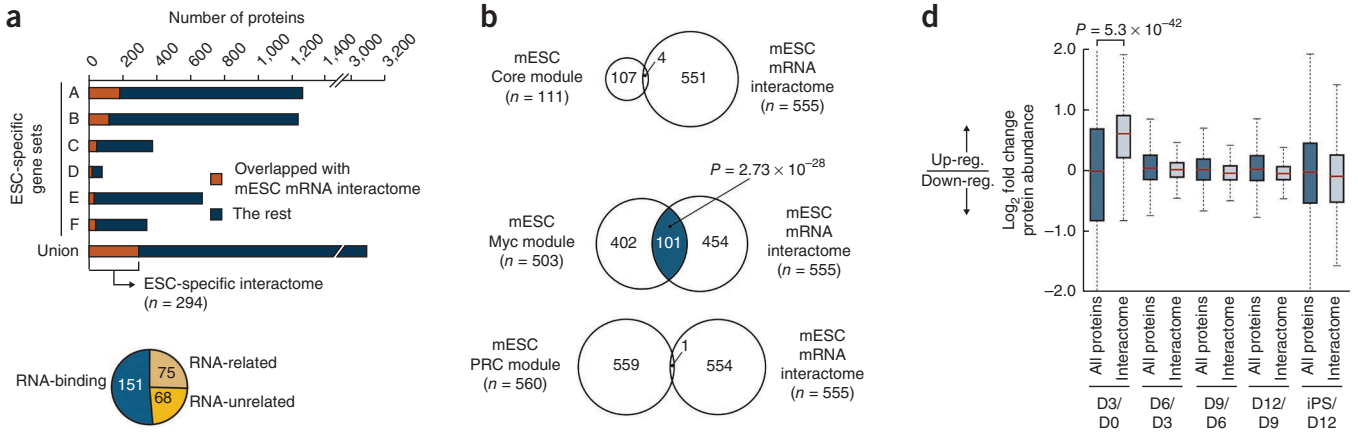



**Figure 3** Validation of RNA-binding activity of novel RBP candidates. (**a**) Schematic of validation method (left) and autoradiography and western blotting (WB; right). Samples are immunoprecipitates radiolabeled on the cross-linked RNA (see **Supplementary Fig. 4** for whole-blot images). (**b**) Summary of validation experiments. Distribution of adjusted *P* values (defined as in **Fig. 1e**) in each protein group is presented as a bee swarm plot. Open red circles indicate RBP candidates whose human homologs were tested in two recent papers[18,19] (**Supplementary Table 6**). (**c**) Validation success rates. Protein number, number of proteins that belong to each *P*-value group in the mESC mRNA interactome. Parenthetical values for validation success rates were calculated including RBPs tested by other groups[18,19] (**b**, open red circles).

*in vitro*[28,29]. In line with this finding, the interactome is enriched for low-complexity and repetitive sequences within disordered regions (see Online Methods), implying that they may participate in membrane-free, RNA-mediated cellular structures (**Supplementary Fig. 3c**,**d**). Taken together, these results indicate that the mESC mRNA interactome is enriched with known RBPs, disordered-region-containing proteins and many novel RBP candidates.

### Validation of RNA-binding activity of novel RBP candidates

We classified 555 proteins into three groups on the basis of the adjusted *P* value (calculated from the MS analysis), which may reflect different RNA-binding affinities or cross-linking efficiencies: 0–70% (group I, 389 proteins), 70–85% (group II, 83 proteins) and the remaining 85–100% (group III, 83 proteins). Eighteen novel RBP candidates with different adjusted *P* values were selected for validation (**Fig. 3a**). Eight candidates were chosen from those uniquely detected in our mESC interactome. The remaining ten candidates were identified from mouse- as well as human-interactome capture[18,19], but none except for Prdx1 have been validated previously (**Supplementary Table 6**). The selected candidates were ectopically expressed with an HA, Myc or Flag tag in mESCs. Cells were irradiated with UV before cell lysis to induce cross-linking. The cell lysate was then treated with RNase A



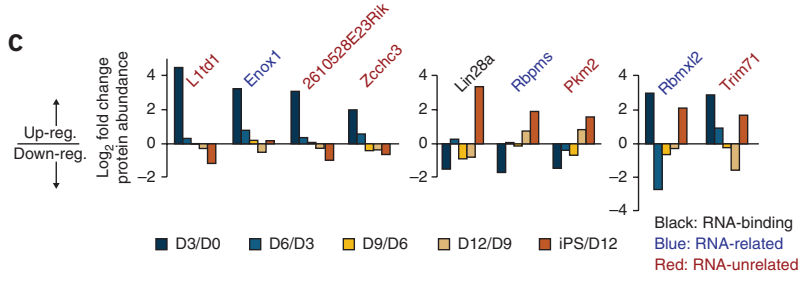




**Figure 4** Regulation of RBPs in stem cells. (**a**) The mESC mRNA interactome is compared with several ESC-specific gene sets: A, genes coordinately upregulated in a compendium of mESCs[33]; B, genes coordinately upregulated in a compendium of human ESCs[33]; C, genes overexpressed in human ESCs according to five or more out of 20 profiling studies[34]; D, genes upregulated in and common to six human ESC lines tested[35]; E, genes downregulated upon retinoic acid–induced differentiation[36]; F, genes downregulated upon embryoid body differentiation[37]; union, a combined set of ESC-specific gene sets. The ESC-specific interactome ($n = 294$) is composed of RBP candidates that are preferentially expressed in ESCs on the basis of at least one gene set (**Table 1** and **Supplementary Table 7a**,**b**). (**b**) Comparison of the mESC mRNA interactome with three major ESC regulatory modules: Core, Myc and PRC[41]. The intersection between the mESC mRNA interactome and the Myc module is statistically significant (one-tailed Fisher's exact test, $P = 2.73 \times 10^{-28}$; see also **Supplementary Table 8**). (**c**) Examples of novel RBP candidates that are dynamically regulated during iPSC reprogramming, represented as fold change over 3-d intervals[8]. See **Supplementary Table 9** for more information. (**d**) Comparison of the mESC mRNA interactome with proteome data during iPSC reprogramming[8]. Expression of mESC mRNA interactome proteins is significantly increased during the first 3 d of iPSC reprogramming (Wilcoxon rank-sum test, $P = 5.3 \times 10^{-42}$). Up-reg., upregulation; down-reg., downregulation.

to shorten the length of the cross-linked RNAs. After immunoprecipitation, the RNA fragments cross-linked to the protein were end-labeled for autoradiography. The residual RNA fragments typically increase the apparent size of RBP by ~7 kDa on polyacrylamide gels. This method was derived from cross-linking immunoprecipitation-sequencing (CLIP-seq) protocols and has been used for studies of RBPs and their directly interacting RNAs[19,30–32]. Known RBPs (Lin28a, quaking and hnRNP A1) were used as positive controls, and an engineered GFP (AcGFP) and β-actin served as negative controls.

The validation success rates roughly correlate with the *P* value–based grouping: 71% of proteins in group I, 67% of those in group II and 40% of those in group III passed the test (**Fig. 3b**,**c**). Eleven out of 18 proteins were validated (names used in the present study are in parentheses): Trim25, Trim71, importin subunit α-2 (Kpna2), catenin-α1 (Ctnna1), Sec61-β (Sec61b), Prdx1, Hmgb1, Hmgb2, DnaJC8, Surf2 and Thumpd1 (**Fig. 3a**,**b**, **Supplementary Fig. 4** and **Supplementary Table 6**). The human ortholog of Prdx1 was validated previously using a different cross-linking method[19] (photoactivatable ribonucleoside–enhanced cross-linking and immunoprecipitation (PAR-CLIP)). When we include the proteins whose human orthologs have been experimentally tested in other studies[18,19], the validation success rate reaches 76% (13/17 proteins) for group I, 75% (6/8 proteins) for group II and 63% (5/8 proteins) for group III (**Fig. 3c** and **Supplementary Table 6**). Of note, we cannot exclude the possibility that some of the proteins tested may have been falsely invalidated owing to limitations of the method. Transiently expressed tagged proteins may behave differently from their endogenous counterparts, as the tag may affect RNA interaction and/or subcellular localization. In addition, immunoprecipitation yielded nonspecific signals that sometimes overlapped with specific ones.

**Table 1 Examples of ESC-specific novel RBP candidates in the mESC mRNA interactome**

| Gene name | Other names | Encoded protein | Group[a] | ESC-specific gene sets |
|---|---|---|---|---|
| *Rcc2* | *Kiaa1470* | Protein RCC2 | I | A,B,C,D,F |
| *Nasp* | *5033430J04Rik* | Nuclear autoantigenic sperm protein | I | A,B,C,D |
| *Pdia4* | *Cai*, *Erp72* | Protein disulfide-isomerase A4 | I | A,B |
| *Ctnna1* | *Catna1* | Catenin-α1 | II | A,B |
| *Prdx1* | *Msp23*, *Paga*, *Tdpx2* | Peroxiredoxin-1 | II | A,B |
| *Nusap1* | *2610201A12Rik* | Nucleolar and spindle-associated protein 1 | III | A,B |
| *Kpna2* | *Rch1* | Importin subunit α-2 | II | A,C,D |
| *Wdr46* | *Bing4* | WD repeat–containing protein 46 | I | A,E,F |
| *Zcchc3* | *2810406K24Rik* | Zinc finger, CCHC domain–containing-3 | I | A,F |
| *Knop1* | *Tsg118*, *2310008H09Rik* | Testis-specific gene 118 protein | I | A |
| *Utp3* | *Crl1*, *Crlz1*, *Sas10* | Something about silencing protein 10 | I | A |
| *Usp10* | *Kiaa0190*, *Ode-1*, *Uchrp* | Deubiquitinating enzyme 10 | I | A |
| *Ubap2* | | Ubiquitin-associated protein 2 | I | A |
| *Zmat3* | *Pag608*, *Wig1* | Zinc finger, matrin-type 3 | II | A |
| *Ipo5* | *Kpnb3*, *Ranbp5* | Importin-5 | III | A |
| *Pkm* | *Pkm2*, *Pk3*, *Pykm* | Pyruvate kinase isozymes M1/M2 | III | A |
| *Tnpo1* | *Kpnb2* | Transportin-1 | I | B |
| *Tma16* | *1810029B16Rik* | UPF0534 protein C4orf43 homolog | III | B |
| *L1td1* | *Ecat11* | LINE-1-type transposase domain–containing 1 | I | C,F |
| *Pak1ip1* | *5830431I15Rik* | PAK1-interacting protein 1 | I | C |
| *8430410A17Rik* | *C85376* | UPF0361 protein C3orf37 homolog | II | E,F |
| *Tbrg4* | *Kiaa0948* | Transforming growth factor-β regulator 4 | I | E |
| *Cmss1* | *2610528E23Rik* | Uncharacterized protein C3orf26 homolog | I | E |
| *Trim25* | *Efp*, *Zfp147*, *Znf147* | Trim25 | III | E |
| *Trim71* | *Gm1127*, *Lin41* | Trim71 | I | F |

[a]Rank based on adjusted *P* value calculated from the initial MS analysis (**Fig. 1e**): Group I (0–70%), Group II (70–85%) and Group III (85–100%). A, Wong_MOUSE_ESC-LIKE[33]; B, Wong_HUMAN_ESC-LIKE[33]; C, Benporath_ES_1 (ref. 34); D, Bhattacharya_EMBRYONIC_STEM_CELL[35]; E, Ivanova_RA_DIFFERENTIATION[36]; F, Liu_EB_DIFFERENTIATION[37] (**Fig. 4a**).

## RBPs that are preferentially expressed in ESCs

Next, we used published data sets to investigate the expression patterns of the mESC mRNA interactome proteins. We included four collections of genes that are expressed at higher levels in ESCs than in differentiated cells[33–35] (**Fig. 4a**, sets A–D) and two that are expressed in ESCs and either downregulated upon retinoic acid–induced differentiation[36] (**Fig. 4a**, set E) or during differentiation into embryoid bodies[37] (**Fig. 4a**, set F). Many (294 of 555 interactome proteins) are shown in at least one data set to be preferentially expressed in ESCs. We refer to these 294 proteins as ESC-specific RBPs. Among ESC-specific RBPs, 68 proteins are not currently annotated as RNA-binding or RNA-related (**Table 1** and **Supplementary Table 7a**). Notably, several known ESC marker proteins are present in the list of ESC-specific RBPs. For example, L1td1 (or Ecat11) is rapidly downregulated upon differentiation[38]. In human ESCs, L1td1 is required for self-renewal[39], although it may be dispensable in mice[40]. It is not yet annotated as an RBP in the GO database, but it was recently reported to interact with RNA[39]. Additionally, seven proteins (Kpna2, Hmgb1, Hmgb2, Ctnna1, Prdx1, Trim25 and Trim71) validated in this study are preferentially expressed in ESCs (**Supplementary Table 7a**). Because their known biochemical activities are not directly related to RNA biology, it will be important to uncover the link between their known functions and their newly identified RNA-binding activity in ESCs.

## RBPs constitute a substantial portion of the ESC Myc module

The ESC transcription program is composed of several distinct modules that make important contributions to stem-cell maintenance and cell reprogramming. Three major modules (Core, Myc and Polycomb repressive complex (PRC)) have been defined on the basis of interactions between transcription factors and their promoter co-occupancy[41]. The genes encoding the Core module are controlled by transcription factors important for early development (such as Oct4 and Nanog). The genes encoding the Myc module are also associated with Myc and Myc-interacting proteins. The Myc module is activated widely in cancer cells and stem cells, and this is largely why stem cells and cancer cells share many features, such as rapid proliferation owing to facilitation of the G1/S transition. The genes encoding the PRC module are generally repressed in ESCs and are involved in cell differentiation and lineage-specific processes. We compared the mESC mRNA interactome with genes in the three modules (**Supplementary Table 8**). Notably, the mESC interactome intersects significantly with the Myc module ($P = 2.73 \times 10^{-28}$), but not with the Core module

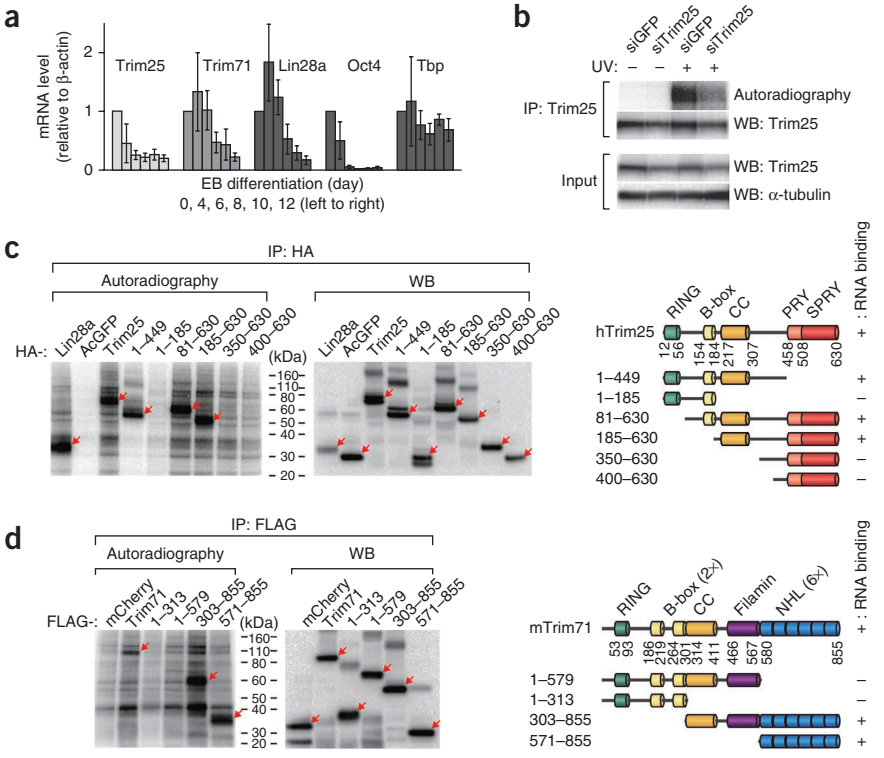



**Figure 5** Trim25 and Trim71 are RNA-binding E3 ubiquitin ligases. (**a**) Quantitative RT-PCR confirms that Trim25 and Trim71 are downregulated upon embryoid body (EB) differentiation. Lin28a and Oct4 are positive controls for ESC-specific genes; Tbp represents constitutively expressed genes. Error bars represent mean ± s.d. from three independent experiments. (**b**) Autoradiography and western blotting (WB) to confirm RNA-binding activity of endogenous human Trim25 (hTrim25) in MCF-7 cells. si-, short interfering RNA against the indicated gene products. α-tubulin, loading control. (**c**) Mapping of the RBD of hTrim25. Plasmids encoding HA-tagged hTrim25 deletion mutants were transfected into HEK293T cells, and RNA-binding activity was tested. Lin28a, positive control; AcGFP, negative control. (**d**) Mapping of the RBD of mouse Trim71 (mTrim71). Plasmids encoding FLAG-tagged mTrim71 deletion mutants were transfected into HEK293T cells, and RNA-binding activity was tested. mCherry, negative control. Red arrows (**c**,**d**) indicate specific signals from the RNA-protein complex indicated.

($P = 0.914$) or the PRC module ($P = 1$; one-tailed Fisher's exact test) (**Fig. 4b**). We found that not only known RBPs but also many newly identified RBPs belong to this intersected group (**Supplementary Table 8**). This suggests that Myc may depend strongly on RBPs to exert its regulatory function and that a substantial proportion of the mRNA interactome may be involved in Myc-directed control.

### RBPs are dynamically regulated during iPSC reprogramming

Recent proteomic and transcriptomic studies have delineated changes in gene expression during reprogramming of induced pluripotent stem cells[8,9] (iPSCs). The initial wave of changes immediately follows overexpression of the reprogramming factors Oct4, Klf4, Sox2 and c-Myc and mainly involves genes that function in cell cycle, DNA repair, chromatin modification and RNA processing. The second major wave of changes arrives in the final stages of reprogramming and is chiefly linked to epigenetic remodeling and stem-cell circuitry[8,9]. When we examined expression profiles of the mESC mRNA interactome using published proteome data[8], we found that many RBPs are dynamically regulated during reprogramming (**Fig. 4c** and **Supplementary Table 9**). For example, L1td1 and Zcchc3 are induced early and remain constant, whereas Lin28a and Pkm2 are not induced until the final phase.

We found that the majority of the mESC mRNA interactome is selectively upregulated during the first 3 days of iPSC reprogramming and remains relatively constant in the later stages (**Fig. 4d**). This result is not due to a bias related to expression levels because the interactome proteins are upregulated during this period more significantly ($P = 8.7 \times 10^{-8}$; Wilcoxon rank-sum test) than the proteins expressed highly in ESCs (top 500) (**Supplementary Fig. 5a**). This observation is consistent with a previous report that proteins with RNA-related GO terms such as 'RNA splicing', 'mRNA processing' and 'ribosome biogenesis' tend to be upregulated early in reprogramming[8]. Here we extend previous studies by finding that other RBPs, including newly identified RBP candidates (RNA-unrelated), are also induced during the first wave (**Supplementary Fig. 5b**). This result implicates that post-transcriptional regulation may have an extensive role during the initial stage of iPSC reprogramming.

### Trim25 and Trim71 interact directly with RNA

We identified two members of the TRIM family of E3 ubiquitin ligases (**Table 1** and **Supplementary Table 7a**). Trim25 is encoded by an ESC Core module target gene (**Supplementary Table 8**). Both Trim25 and Trim71 are expressed abundantly in ESCs and are downregulated during differentiation to embryoid bodies (**Fig. 5a** and **Supplementary Table 7b**). This expression pattern is similar to that of other stem-cell factors, such as Lin28a and Oct4. This result is consistent with previous high-throughput data[36,37] and a study on Trim71 (ref. 42).

We next treated the RNA-protein complexes with RNase after the polynucleotide kinase reaction, to further confirm their RNA-binding activity. The radioactive signal diminished after RNase treatment (**Supplementary Fig. 5c**), indicating that the signal was RNA-dependent rather than DNA-dependent. We also confirmed that endogenous Trim25 interacts with RNA using antibodies to human Trim25 in MCF-7 human breast cancer cell lines (**Fig. 5b**). The RNA signal decreased upon Trim25 knockdown, indicating that endogenous Trim25 was indeed responsible for the RNA pulldown (**Fig. 5b**). TRIM proteins share a common domain structure at the N terminus that includes a RING finger motif, one or two B-boxes and a coiled coil domain, but the middle and C-terminal parts are more variable among subtypes. To delineate the RBDs within Trim25 and Trim71, we generated a series of deletion mutants. Trim25 lost RNA-binding activity upon deletion of the middle part that contains the coiled coil domain (**Fig. 5c**), indicating that the middle portion is required for RNA binding. For Trim71, NHL repeats (named after NCL-1, HT2A and Lin-41) are sufficient for RNA binding (**Fig. 5d**). TRIM proteins containing NHL domains were previously shown to control microRNA function through a mechanism that is not yet known[43]. Recent studies have also found that Trim71 can associate with mRNAs and reduce their expression level; however, it has been unclear whether or not Trim71 interacts with mRNA directly[42,44]. Our results now demonstrate that Trim71 binds RNAs in a direct manner (**Fig. 5d**). As Trim25 and Trim71 use different domains for RNA interaction, they are likely to have distinct sets of RNA partners.

To further understand whether RNA binding is a common feature of all TRIM proteins, we tested the RNA-binding activity of Triml1, which belongs to the same subfamily as Trim25 (ref. 45). Triml1 is expressed in mESCs[46] but was not identified in the mESC-interactome capture. Triml1 does not exhibit RNA-binding activity (**Supplementary Fig. 5d**), suggesting that only a subset of TRIM proteins can interact with RNA.

## DISCUSSION

We have captured the *in vivo* poly(A)$^+$ RNA interactome of mESCs and uncovered numerous mouse proteins as novel RBPs. Only about 61% of the mESC mRNA interactome proteins contain known RBDs (339 of 555 proteins), suggesting the existence of unknown RNA-binding architectures. Indeed, 10 out of 11 proteins that were experimentally validated in our study (Hmgb1, Hmgb2, Trim71, Trim25, Ctnna1, Kpna2, Surf2, Sec61b, Dnajc8 and Prdx1) do not harbor known RBDs. Thus, interactome capture allows unbiased discovery of RBPs without previous knowledge of domains. Further analysis of the other candidates reported here is likely to substantially expand the repertoire of RBPs and provide a valuable resource for studies of RNA-protein networks in stem cells and mammalian systems.

There are interesting RBP candidates for future studies, as they may contribute to ESC pluripotency. When we compared the mESC interactome proteins to previous genome-wide RNA-interference (RNAi) screens for ESC pluripotency[47–49], we identified 28 proteins as candidates for pluripotency regulators (**Supplementary Table 10**). The results from the three RNAi studies do not overlap (reflecting the complex nature of ESC physiology), but two RBPs (Cnot1 and Fip1l1) were found in two independent studies[47–49] (**Supplementary Table 10**). Eight of the 28 proteins are not currently annotated as RBPs in GO, so further investigation of these proteins is needed. For example, Zc3h18 contains one CCCH-type zinc-finger domain that is possibly responsible for RNA binding. In trypanosomes, Zc3h18 has been shown to be involved in differentiation[50]; its mutation in humans is associated with progression from severe congenital neutropenia to acute myeloid leukemia[51]; and it mediates canonical nuclear factor-κB activation by performing crucial functions in IκB kinase activation[52]. Cnot1 is the protein-binding platform in the CCR4–NOT mRNA deadenylation complex[53]; therefore, we did not expect to identify it as an RBP candidate in our mESC interactome as well as in two other studies in human cells[18,19]. Thus, Cnot1 may function as both a protein-binding platform and an RBP. A recent study showed that Cnot1, Cnot2 and Cnot3 are required for the maintenance of ESC identity by repressing extra-embryonic differentiation[54]. The authors suggested that this function may not come from general deadenylation activity, because knockdown of other members of the CCR4–NOT complex did not affect the Oct4 reporter expression[54]. Endothelial differentiation-related factor 1 (Edf1, also called Mbf1) is a transcriptional coactivator that links TATA box–binding protein and several transcription factors, such as Gcn4, steroidogenic factor 1 and liver receptor homolog 1 (refs. 55–58). Because the suppression of human Edf1 expression affects endothelial cell differentiation[59], it would be of interest to examine the role of Edf1 in ESC differentiation as an RNA-binding transcription regulator.

Other RBPs that are highly expressed in undifferentiated ESCs are also attractive subjects for further investigation (**Supplementary Table 7a**). For instance, Kpna2 (also known as importin subunit α-2, pendulin or Rch1) belongs the nuclear transport receptor family. In mESCs knockdown of Kpna2 induces neural differentiation, and overexpression of Kpna2 suppresses it, suggesting a role in stem-cell maintenance[60]. Ctnna1 is another unexpected RBP highly expressed in ESCs. It is an essential component of the cadherin–catenin complex that constitutes adherens junctions, but it is also known to modulate several signaling pathways. Overall, investigation of ESC-specific RBPs may help us to uncover new regulatory pathways governing pluripotent stem cells.

Tripartite motif (TRIM) proteins belong to one of the subfamilies of the RING-type E3 ubiquitin ligases. In mammals, about 70 proteins belong to the TRIM family, and they are involved in many biological processes, including proliferation, differentiation, apoptosis, oncogenesis and antiviral immunity[45,61]. Trim25 is one of the best-characterized TRIM proteins because of its crucial roles in cancer progression and antiviral pathways[62,63]. Trim25 ubiquitinates and destabilizes 14-3-3σ protein (a cell-cycle regulator that leads to G2 arrest), thereby promoting cell growth[62], so it is not surprising that Trim25 has been linked to breast, ovarian and endometrial cancer[45,64–66]. Trim25 was also shown to play a crucial part in antiviral pathways by ubiquitinating RIG-I, a host sensor of viral replication, in response to RNA virus infection[63]. The function of Trim25 in ESCs remains unknown. Trim71 was first identified as one of the heterochronic genes in *Caenorhabditis elegans*[67,68]. In mice, *Trim71* knockout results in a defect in neural-tube closure and lethality during embryonic development[68,69]. Expression of Trim71 is suppressed by microRNA let-7 in differentiated tissues[70–72]. In mESC, Trim71 represses the expression of Cdkn1, a cell-cycle regulator that inhibits the G1/S transition, thereby promoting rapid proliferation[42].

It is notable that protein-modifying enzymes have RNA-binding activity. A recent computational study predicted that more than 15 proteins carry both known RBDs and E3 ubiquitin ligase domains[73]. Several E3 ligases with RBDs have been detected in mouse and human interactomes: Roquin (also called Rc3h1), Rc3h2, Cnot4, Mex3a, Mex3b, Mex3c, Mex3d, Mkrn1 and Mkrn2. Only a few of these proteins have been studied, but some have interesting properties. For example, Roquin, which harbors a RING E3 ligase domain, also contains a CCCH zinc-finger domain and promotes mRNA decay, which is required for the prevention of autoimmune responses in mice[74–76]. However, Roquin's E3 ubiquitin ligase activity has not been demonstrated, and the target proteins for ubiquitination remain to be determined. Trim25 and Trim71 would be important additions to the newly emerging class of RNA-binding E3 ligases and may serve as good models because their ubiquitination substrates are known. Two other TRIM proteins (Trim56 and Trim28) have been detected in human interactome studies, but they have not yet been validated. It would be important to investigate which RNAs these enzymes are bound to and what the roles of the interacting RNAs are. One can imagine that the RNAs help the enzymes to find their substrates by serving as a 'ribo-bridge'. The RNAs may also modulate the activity of E3 ligases or drive them to specific subcellular localizations (for example, RNA granules). Further investigation of TRIM proteins would offer the opportunity to reveal the cross-talk between RNA pathways and protein ubiquitination.

To understand the molecular functions of the RBPs identified here, it will be crucial to systematically identify the RNA partners of the RBPs. CLIP-seq (also known as HITS-CLIP) or related approaches (such as PAR-CLIP) will be useful methods to identify bound RNAs and map their binding sites *in vivo*. CLIP-seq combines UV irradiation, immunoprecipitation of RBPs of interest and deep sequencing of the co-precipitated RNAs. Because interactome capture and CLIP-seq use virtually identical cross-linking conditions, we expect the RBPs uncovered in this study to be easily amenable to CLIP-seq protocols. These 'CLIPable' RBPs will allow us to probe RNA-RBP networks in pluripotent stem cells. We also note that because our method enriches

poly(A)$^+$ RNAs, it should precipitate proteins bound not only to mRNAs but also to any polyadenylated long noncoding RNAs. It will be interesting to see whether the RBPs found in our study interact with selective sets of regulatory noncoding RNAs.

## METHODS

Methods and any associated references are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*


### ACKNOWLEDGMENTS

We thank H. Chang for teaching and sharing Python codes, A. Cho, Y.C. Chang and H. Kim for technical help, and all members of our laboratories for helpful discussion. We gratefully acknowledge the EMBL Proteomics Core Facility for technical support. We are grateful to G. Dreyfuss (University of Pennsylvania), F.G. Wulczyn (Charité–Universitätsmedizin Berlin), S.H. Baek (Seoul National University), H.-Y. Kao (Case Western Reserve University), D.-E. Zhang (The Scripps Research Institute), D. Rimm (Yale School of Medicine), K. Helin (Københavns Universitet) and Y. Kawakami (University of Minnesota) for valuable plasmids and antibodies. This work was supported by the Research Center Program (EM1202) of the Institute for Basic Science (S.C.K., H.Y., K.T.Y. and V.N.K.) and the BK21 Research Fellowships (S.C.K. and H.Y.) from the Ministry of Education, Science and Technology of Korea. Work in the group of M.W.H. was funded by an ERC Advanced grant (ERC-2011-ADG_20110310) to M.W.H.


### AUTHOR CONTRIBUTIONS

S.C.K., A.C., M.W.H. and V.N.K. designed the project; S.C.K., H.Y. and K.T.Y. performed biochemical experiments and analyzed ESC-related data; K.E., S.F. and J.K. performed and analyzed MS experiments; B.F., A.C. and M.W.H. analyzed the mRNA interactome; S.C.K. and V.N.K. wrote the manuscript.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.




1. Smith, A.G. Embryo-derived stem cells: of mice and men. *Annu. Rev. Cell Dev. Biol.* **17**, 435–462 (2001).
2. Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4**, 487–492 (2009).
3. Loh, Y.-H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440 (2006).
4. Orkin, S.H. & Hochedlinger, K. Chromatin connections to pluripotency and cellular reprogramming. *Cell* **145**, 835–850 (2011).
5. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S.H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049–1061 (2008).
6. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
7. Young, R.A. Control of the embryonic stem cell state. *Cell* **144**, 940–954 (2011).
8. Hansson, J. *et al.* Highly coordinated proteome dynamics during reprogramming of somatic cells to pluripotency. *Cell Rep.* **2**, 1579–1592 (2012).
9. Polo, J.M. *et al.* A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151**, 1617–1632 (2012).
10. Sampath, P. *et al.* A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **2**, 448–460 (2008).
11. Lu, R. *et al.* Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature* **462**, 358–362 (2009).
12. Gebauer, F. & Hentze, M.W. Molecular mechanisms of translational control. *Nat. Rev. Mol. Cell Biol.* **5**, 827–835 (2004).
13. Abaza, I. & Gebauer, F. Trading translation with RNA-binding proteins. *RNA* **14**, 404–409 (2008).
14. Glisovic, T., Bachorik, J.L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).
15. Kishore, S., Luber, S. & Zavolan, M. Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Brief. Funct. Genomics* **9**, 391–404 (2010).
16. Tsvetanova, N.G., Klass, D.M., Salzman, J. & Brown, P.O. Proteome-wide search reveals unexpected RNA-binding proteins in *Saccharomyces cerevisiae*. *PLoS ONE* **5**, e12671 (2010).
17. Scherrer, T., Mittal, N., Janga, S.C. & Gerber, A.P. A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. *PLoS ONE* **5**, e15499 (2010).
18. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
19. Baltz, A.G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690 (2012).
20. Choi, Y.D. & Dreyfuss, G. Isolation of the heterogeneous nuclear RNA-ribonucleoprotein complex (hnRNP): a unique supramolecular assembly. *Proc. Natl. Acad. Sci. USA* **81**, 7471–7475 (1984).
21. Mitchell, S.F., Jain, S., She, M. & Parker, R. Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol.* **20**, 127–133 (2013).
22. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
23. Sugimoto, Y. *et al.* Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.* **13**, R67 (2012).
24. Lau, C.-K., Bachorik, J.L. & Dreyfuss, G. Gemin5-snRNA interaction reveals an RNA binding function for WD repeat domains. *Nat. Struct. Mol. Biol.* **16**, 486–491 (2009).
25. Radivojac, P. *et al.* Intrinsic disorder and functional proteomics. *Biophys. J.* **92**, 1439–1456 (2007).
26. Tompa, P. & Csermely, P. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.* **18**, 1169–1175 (2004).
27. Dyson, H.J. & Wright, P.E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
28. Han, T.W. *et al.* Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell* **149**, 768–779 (2012).
29. Kato, M. *et al.* Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012).
30. Chi, S.W., Zang, J.B., Mele, A. & Darnell, R.B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**, 479–486 (2009).
31. Ule, J. *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**, 1212–1215 (2003).
32. Cho, J. *et al.* LIN28A is a suppressor of ER-associated translation in embryonic stem cells. *Cell* **151**, 765–777 (2012).
33. Wong, D.J. *et al.* Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell* **2**, 333–344 (2008).
34. Ben-Porath, I. *et al.* An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).
35. Bhattacharya, B. *et al.* Gene expression in human embryonic stem cell lines: unique molecular signature. *Blood* **103**, 2956–2964 (2004).
36. Ivanova, N. *et al.* Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**, 533–538 (2006).
37. Liu, Z., Scannell, D.R., Eisen, M.B. & Tjian, R. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* **146**, 720–731 (2011).
38. Wong, R.C.-B. *et al.* L1TD1 is a marker for undifferentiated human embryonic stem cells. *PLoS ONE* **6**, e19355 (2011).
39. Närvä, E. *et al.* RNA-binding protein L1TD1 interacts with LIN28 via RNA and is required for human embryonic stem cell self-renewal and cancer cell proliferation. *Stem Cells* **30**, 452–460 (2012).
40. Iwabuchi, K.A. *et al.* ECAT11/L1td1 is enriched in ESCs and rapidly activated during iPSC generation, but it is dispensable for the maintenance and induction of pluripotency. *PLoS ONE* **6**, e20461 (2011).
41. Kim, J. *et al.* A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. *Cell* **143**, 313–324 (2010).
42. Chang, H.-M. *et al.* Trim71 cooperates with microRNAs to repress Cdkn1a expression and promote embryonic stem cell proliferation. *Nat. Commun.* **3**, 923 (2012).
43. Loedige, I. & Filipowicz, W. TRIM-NHL proteins take on miRNA regulation. *Cell* **136**, 818–820 (2009).
44. Loedige, I., Gaidatzis, D., Sack, R., Meister, G. & Filipowicz, W. The mammalian TRIM-NHL protein TRIM71/LIN-41 is a repressor of mRNA function. *Nucleic Acids Res.* **41**, 518–532 (2013).
45. Hatakeyama, S. TRIM proteins and cancer. *Nat. Rev. Cancer* **11**, 792–804 (2011).
46. Tian, L. *et al.* Characterization and potential function of a novel pre-implantation embryo-specific RING finger protein: TRIML1. *Mol. Reprod. Dev.* **76**, 656–664 (2009).
47. Ding, L. *et al.* A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. *Cell Stem Cell* **4**, 403–415 (2009).
48. Hu, G. *et al.* A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes Dev.* **23**, 837–848 (2009).
49. Chia, N.-Y. *et al.* A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* **468**, 316–320 (2010).
50. Benz, C., Mulindwa, J., Ouna, B. & Clayton, C. The *Trypanosoma brucei* zinc-finger protein ZC3H18 is involved in differentiation. *Mol. Biochem. Parasitol.* **177**, 148–151 (2011).
51. Beekman, R. *et al.* Sequential gain of mutations in severe congenital neutropenia progressing to acute myeloid leukemia. *Blood* **119**, 5071–5077 (2012).
52. Gewurz, B.E. *et al.* Genome-wide siRNA screen for mediators of NF-κB activation. *Proc. Natl. Acad. Sci. USA* **109**, 2467–2472 (2012).
53. Denis, C.L. & Chen, J. The CCR4–NOT complex plays diverse roles in mRNA metabolism. *Prog. Nucleic Acid Res. Mol. Biol.* **73**, 221–250 (2003).

54. Zheng, X. *et al.* Cnot1, Cnot2, and Cnot3 maintain mouse and human ESC identity and inhibit extraembryonic differentiation. *Stem Cells* **30**, 910–922 (2012).
55. Kabe, Y. *et al.* The role of human MBF1 as a transcriptional coactivator. *J. Biol. Chem.* **274**, 34196–34202 (1999).
56. Takemaru, K.-i., Li, F.Q., Ueda, H. & Hirose, S. Multiprotein bridging factor 1 (MBF1) is an evolutionarily conserved transcriptional coactivator that connects a regulatory factor and TATA element-binding protein. *Proc. Natl. Acad. Sci. USA* **94**, 7251–7256 (1997).
57. Takemaru, K.-i., Harashima, S., Ueda, H. & Hirose, S. Yeast coactivator MBF1 mediates GCN4-dependent transcriptional activation. *Mol. Cell Biol.* **18**, 4971–4976 (1998).
58. Brendel, C., Gelman, L. & Auwerx, J. Multiprotein bridging factor-1 (MBF-1) is a cofactor for nuclear receptors that regulate lipid metabolism. *Mol. Endocrinol.* **16**, 1367–1377 (2002).
59. Dragoni, I., Mariotti, M., Consalez, G.G., Soria, M.R. & Maier, J.a. EDF-1, a novel gene product down-regulated in human endothelial cell differentiation. *J. Biol. Chem.* **273**, 31119–31124 (1998).
60. Yasuhara, N. *et al.* Triggering neural differentiation of ES cells by subtype switching of importin-α. *Nat. Cell Biol.* **9**, 72–79 (2007).
61. Nisole, S., Stoye, J.P. & Saïb, A. TRIM family proteins: retroviral restriction and antiviral defence. *Nat. Rev. Microbiol.* **3**, 799–808 (2005).
62. Urano, T. *et al.* Efp targets 14–3-3σ for proteolysis and promotes breast tumour growth. *Nature* **417**, 871–875 (2002).
63. Gack, M.U. *et al.* TRIM25 RING-finger E3 ubiquitin ligase is essential for RIG-I-mediated antiviral activity. *Nature* **446**, 916–920 (2007).
64. Suzuki, T. *et al.* Estrogen-responsive finger protein as a new potential biomarker for breast cancer. *Clin. Cancer Res.* **11**, 6148–6154 (2005).
65. Sakuma, M. *et al.* Expression of estrogen-responsive finger protein (Efp) is associated with advanced disease in human epithelial ovarian cancer. *Gynecol. Oncol.* **99**, 664–670 (2005).
66. Nakayama, H., Sano, T., Motegi, A., Oyama, T. & Nakajima, T. Increasing 14-3-3 sigma expression with declining estrogen receptor alpha and estrogen-responsive finger protein expression defines malignant progression of endometrial carcinoma. *Pathol. Int.* **55**, 707–715 (2005).
67. Rybak, A. *et al.* The let-7 target gene mouse *lin-41* is a stem cell specific E3 ubiquitin ligase for the miRNA pathway protein Ago2. *Nat. Cell Biol.* **11**, 1411–1420 (2009).
68. Chen, J., Lai, F. & Niswander, L. The ubiquitin ligase mLin41 temporally promotes neural progenitor cell maintenance through FGF signaling. *Genes Dev.* **26**, 803–815 (2012).
69. Maller Schulman, B.R., Liang, X. & Stahlhut, C. The let-7 microRNA target gene, *Mlin41/Trim71* is required for mouse embryonic survival and neural tube closure. *Cell Cycle* **7**, 3935–3942 (2008).
70. Slack, F.J. *et al.* The *lin-41* RBCC gene acts in the *C. elegans* heterochronic pathway between the *let-7* regulatory RNA and the LIN-29 transcription factor. *Mol. Cell* **5**, 659–669 (2000).
71. Kanamoto, T., Terada, K., Yoshikawa, H. & Furukawa, T. Cloning and regulation of the vertebrate homologue of *lin-41* that functions as a heterochronic gene in *Caenorhabditis elegans*. *Dev. Dyn.* **235**, 1142–1149 (2006).
72. Lin, Y.-C. *et al.* Human TRIM71 and its nematode homologue are targets of *let-7* microRNA and its zebrafish orthologue is essential for development. *Mol. Biol. Evol.* **24**, 2525–2534 (2007).
73. Cano, F., Miranda-Saavedra, D. & Lehner, P.J. RNA-binding E3 ubiquitin ligases: novel players in nucleic acid regulation. *Biochem. Soc. Trans.* **38**, 1621–1626 (2010).
74. Vinuesa, C.G. *et al.* A RING-type ubiquitin ligase family member required to repress follicular helper T cells and autoimmunity. *Nature* **435**, 452–458 (2005).
75. Yu, D. *et al.* Roquin represses autoimmunity by limiting inducible T-cell co-stimulator messenger RNA. *Nature* **450**, 299–303 (2007).
76. Glasmacher, E. *et al.* Roquin binds inducible costimulator mRNA and effectors of mRNA decay to induce microRNA-independent post-transcriptional repression. *Nat. Immunol.* **11**, 725–733 (2010).

## ONLINE METHODS

**Cell culture and *in vitro* differentiation of ESCs.** R1 mouse embryonic stem cells were purchased from ATCC (catalog no. SCRC-1011) and maintained on mouse embryonic fibroblasts (MEFs, CF-1 strain, MCTT 001-CF1-V) treated with mitomycin C (Sigma, M4287) in GlutaMax DMEM (Gibco, 10569) supplemented with 10% FBS (Gibco, 16000-044), nonessential amino acids (Gibco, 11140-050), 100 µM 2-mercaptoethanol (Amresco) and 1,000 U/ml leukemia inhibitory factor (LIF) (Millipore, ESG1107). Prior to interactome capture experiments, R1 cells were purified from MEFs and cultured on 0.1% gelatin-coated dishes. For *in vitro* differentiation of ESCs into embryoid bodies, $1 \times 10^6$ R1 cells were plated on 90-mm Petri dishes in ESC medium without LIF. Medium was renewed every other day. HEK293T cells were grown in DMEM supplemented with 10% FBS (WelGENE). MCF-7 cells were maintained in RPMI1640 supplemented with 10% FBS (WelGENE).

**Interactome capture.** *In vivo* isolation of mESC (R1) RBPs was carried out as previously described[18,77]. Briefly, $6 \times 10^8$ R1 cells (per condition) were irradiated with 0.15 J/cm$^2$ UV light at 254 nm, lysed in lysis buffer (20 mM Tris (pH 7.5), 500 mM LiCl, 0.5% lithium dodecyl sulfate, 1 mM EDTA, 5 mM DTT) and supplemented with 30 ng of *in vitro* transcribed, poly(A)-tailed luciferase mRNA. Oligo $dT_{25}$ beads (NEB, S1419S) were used to capture poly(A)$^+$ RNAs, and covalently cross-linked proteins were released from captured RNAs by treatment with RNase A/T1 mix, as described elsewhere[77]. Eluted proteins were identified and quantified by LC-MS/MS. For quantitative RT-PCR of pulled-down RNAs (**Fig. 1d**), equal volumes of each lysate were treated with proteinase K (0.4 mg/ml, Roche) for 1 h at 50 °C, and RNAs were extracted by addition of Trizol to the lysate. The same volumes of RNA were used for subsequent reverse transcription.

**Sample preparation for mass spectrometry.** Sample preparation for mass spectrometry including alkylation, buffer exchange and digestion was performed as reported with minor modifications[18,77]. Samples were adjusted to 5 mM DTT, incubated for 1 h at 56 °C and concentrated using Amicon Ultra Centrifugal Filters (0.5 ml, 3-kDa cut-off) (Millipore). After addition of 200 µl 55 mM iodoacetamide, the samples were mixed at 600 r.p.m. for 30 min in the dark and then concentrated. The buffer was exchanged by adding 100 µl 50 mM $NH_4HCO_3$, and the sample was then concentrated for three successive rounds. After addition of 0.6 µg trypsin in 40 µl 50 mM $NH_4HCO_3$, the filter units were incubated at room temperature overnight. The derived peptides were collected by centrifugation of the filter units, washed with 50 µl of 0.5 M NaCl and centrifuged again. Reductive dimethylation for stable isotope labeling of the samples was performed according to a previously described protocol[78]. Labeled peptide samples retrieved from the pulldown and the according controls were combined and fractionated into 12 fractions on an Agilent 3100 OFFGEL Fractionator according to the manufacturer's protocol using Immobiline DryStrips (pH 3–10 NL, 13 cm; GE Healthcare). Focusing was performed at a constant current of 50 mA with a maximum voltage of 8,000 V. After reaching 20 kVh the samples were collected, acidified with $CF_3COOH$ and desalted using StageTips[79]. The dried samples were dissolved in 10 µl 4% acetonitrile and 0.1% formic acid.

**Liquid chromatography–tandem mass spectrometry.** Peptides were separated using the nanoAcquity UPLC system (Waters) fitted with a trapping column (nanoAcquity Symmetry $C_{18}$, 5 µm, 180 µm × 20 mm) and an analytical column (nanoAcquity BEH $C_{18}$, 1.7 µm, 75 µm × 200 mm). The outlet of the analytical column was coupled directly to an LTQ Orbitrap Velos Pro (Thermo Fisher Scientific) using a Proxeon nanospray source. Solvent A was water and 0.1% formic acid and solvent B was acetonitrile and 0.1% formic acid. The samples (5 µl) were loaded with a constant flow of solvent A at 15 µl/min onto the trapping column. Trapping time was 1 min. Peptides were eluted via the analytical column at a constant flow of 0.3 µl/min. During the elution step, the percentage of solvent B increased in a linear fashion from 3% to 40% in 45 min. The peptides were introduced into the mass spectrometer via a PicoTip Emitter (360 µm OD × 20 µm ID; 10 µm tip; New Objective) and a spray voltage of 2.1 kV was applied. The capillary temperature was set at 200 °C. Full scan MS spectra with mass range of 300–1,700 *m/z* were acquired in profile mode in the orbitrap with a resolution of 30,000. The filling time was set at a maximum of 300 ms with limitation of $10^6$ ions. The most intense ions (up to 15) from the full scan MS were selected for sequencing in the LTQ. A normalized collision energy of 40% was used, and the fragmentation was performed after accumulation of $3 \times 10^4$ ions or after a filling time of 100 ms for each precursor ion, whichever occurred first. MS/MS data were acquired in centroid mode. Only multiply charged (2+, 3+) precursor ions were selected for MS/MS. The dynamic exclusion list was restricted to 500 entries with maximum retention periods of 30 s and a relative mass window of 10 p.p.m. To improve the mass accuracy, a lock mass correction using a background ion (*m/z* 445.12003) was applied.

**Mass spectrometric data analysis.** The mass spectrometric raw data was processed using MaxQuant version 1.2.2.5 (ref. 80) and MS/MS spectra were searched using the Andromeda search engine[81] against mouse proteins in UniProt[82] (53,623 entries, downloaded on 21 June 2011), to which 265 frequently observed contaminants as well as reversed sequences of all entries had been added. Enzyme specificity was set to trypsin/P and a maximum of two missed cleavages was allowed. Cysteine carbamidomethylation was used as fixed modification, and methionine oxidation and protein N-terminal acetylation were used as variable modifications. The minimal peptide length was set to six amino acids. Initial maximal allowed mass tolerance was set to 20 p.p.m. for peptide masses, followed by 6 p.p.m. in the main search and 0.5 Da for fragment ion masses. False discovery rates for peptide and protein identification were set to 1%. At least one unique peptide was required for protein identification. The protein identification was reported as an indistinguishable protein group if no peptide sequence unique to a single database entry was identified.

For protein quantification a minimum of two ratio counts was set and the 'requantify' and 'match between runs' function enabled. Protein groups assigned to contaminants or reverse sequences were removed. Average protein ratios were reported if they were quantified in at least two biological replicates on the basis of at least two ratio counts per replicate.

**Statistical analysis of mass spectrometry data.** Statistical analysis was performed using the Limma package in R/Bioconductor[83,84]. After fitting a linear model to the data, an empirical Bayes moderated *t*-test was used, and *P* values were adjusted for multiple testing with the Benjamini-Hochberg method. Proteins that had an adjusted *P* value < 0.01 and were more abundant in the cross-linked sample than in the control sample were assigned to the mRNA interactome (584 protein groups). Correlations between replicates were calculated in R using Pearson correlation. The 584 protein groups mapped to 555 unique genes that were used for subsequent processing.

**Gene set enrichment analysis.** InterPro and GO annotations were downloaded from Ensembl release 67. For gene-set enrichment analysis, categories with at least three genes were considered. Enrichment and under-representation was tested for statistical significance by Fisher's exact test. *P* values were corrected for multiple testing by the Benjamini-Hochberg method. For data shown in **Figure 2b**, Filtered Gene Ontology (OBO v1.2; cross-products, inter-ontology links and has_part relationships removed) was downloaded (http://www.geneontology.org/), and top ten terms from the mRNA interactome were selected with their parent terms and visualized using OBO-Edit version 2.3 (http://oboedit.org/).

**Human orthologs.** Human orthologous genes were downloaded from Ensembl release 67. A mouse gene is considered present (as ortholog) in the human mRNA interactome if at least one of its orthologous human genes is contained in the human mRNA interactome.

**Sequence features and RBDs.** Classical and nonclassical RBDs were selected from all Pfam domains as previously described[18]. Briefly, the RBDs listed in a recent review paper[85] are considered classical RBDs. Protein domains with experimental evidence for RNA-binding in at least one case are classified as nonclassical RBDs (see list in **Supplementary Table 5a**,**b**). All classical Pfam RBDs were tested by Fisher's exact test for enrichment in the mESC mRNA interactome compared to all mouse proteins. Proteins not containing a classical RBD were tested for enrichment of nonclassical RBDs, and all proteins containing neither a classical nor a nonclassical RBD were tested for enrichment of all other domains. *P* values were corrected for multiple testing by the Benjamini-Hochberg method.

Amino acids with a score >0.4 for intrinsic disorder (IUPred)[86] were considered disordered. The fraction of residues in disordered regions was computed for each protein. An amino acid was defined to be in a low-complex region when




doi:10.1038/nsmb.2638

the Shannon entropy of the amino acid distribution within a window of 10 residues before and after the position was <3 bits. The proportion of low-complexity residues within the set of disordered residues was computed for each protein. For each pair of amino acids the probability of occurrence as neighboring amino acids (and with one or two wild-card amino acids in between) was computed for all mouse proteins. Over-representation of an amino acid pair in the disordered regions was tested by a binomial test. *P* values were corrected for multiple testing by the Benjamini-Hochberg method. Dimers with a *P* value <0.01 were regarded as repetitive within disordered regions. For each of the three measures, a distribution was plotted for the proteins in the mESC mRNA interactome, for all mouse proteins, for mRNA interactome proteins containing a classical RBD and for novel RBPs not annotated as RNA-related (GO) and that do not contain a nonclassical RBD.

**Plasmid construction.** To generate a vector that expresses 3XMyc-tagged proteins, we first inserted new multi cloning sites (MCS) between EcoRI and BamHI sites of pLVX-EF1a-AcGFP1-N1 (Clontech, 631983). Next, we removed coding sequences of AcGFP from the above vector by using XbaI sites, and inserted new sequences comprised of Kozak-ATG-3XMyc between BamHI and EcoRI sites. The following plasmids were kindly provided as gifts: hActn4 (H.-K. Kao), hHnrnpa1 (G. Dryfuss), and mTrim71 (F.G. Wulczyn). We also obtained plasmids from Addgene: hTrim25 (no. 12452), hCtnna1 (no. 24194), hEed (no. 24231), hHmgb1 (no. 31609) and mHmgb2 (no. 31610). Other plasmids used in validation experiments were generated by using Infusion cloning method (Enzynomics, EZ-Cloning core kit) or pGEM-T easy cloning system (Promega). For oligonucleotide sequences and other detailed information, see **Supplementary Table 11**.

**Validation of RNA-binding activity of novel RBP candidates.** Coding sequences of novel RBP candidates were cloned into a mammalian expression vector with an N-terminal or C-terminal tag (HA, Myc or Flag), then ectopically expressed in mESCs by using Lipofectamine 2000 (Invitrogen). One or two days later, cells were washed twice with cold PBS and irradiated with UV light at 254 nm (160 mJ/cm$^2$) without PBS, and lysed in CLIP Wash buffer[30] (1× PBS, 0.1% SDS, 0.5% sodium deoxycholate, 0.5% NP-40). Lysate was passed through a 21-gauge needle ten times to reduce viscosity and then incubated with 100 U/ml of DNase I (Takara, 2270A) for 5 min at 37 °C. To reduce the size of RNAs bound to RBPs, lysate was incubated with 4 ng/ul of RNase A for 10 min at 37 °C. After centrifugation at 16,000 × *g* for 20 min at 4 °C, lysate was incubated with proper antibodies (anti-HA, Santacruz, sc-7392, 4 μg/ml; anti-HA, Abcam, ab9110, 4 μg/ml; anti-FLAG, Sigma, A2220, 1.6 μl/ml per sample; anti-Myc, 9E10 ascites fluids, 6 μg/ml) and protein A–Sepharose beads (GE Healthcare, 17-5138- 01) for 1–2 h. Immunopurified proteins were washed four times with wash buffer and two times with PNK buffer[30] (50 mM Tris (pH 7.4), 10 mM MgCl2, 0.5% NP-40), incubated with T4 polynucleotide kinase (Takara) and [γ-$^{32}$P]rATP for 15 min at 37 °C to label 5′ end of cross-linked RNAs, loaded onto 4–12% NuPAGE gel (Invitrogen, NP0321BOX) and transferred to the nitrocellulose membrane (Amersham, RNP303E). Autoradiography signals were detected by BAS-2500 (GE Healthcare), and corresponding proteins were detected by western blotting (antibody to HA (anti-HA), Santa Cruz, sc-7392, 1:1,000; anti-HA, Invitrogen, 71-5500, 1:2,000; anti-FLAG, Sigma, F7425, 1:2,000; anti-Myc, Abcam, ab9106, 1:5,000). For the following highly conserved genes, we used human proteins instead of mouse proteins: Ctnna1 (99.45% identity), Hmgb1 (99.07%), Actn4 (98.47%), Eed (100%). (When similar amino acids are considered as functionally identical residues, the identities are 100% for Ctnna1, 100% for Hmgb1 and 99.8% for Actn4.) Whole-blot images are shown in **Supplementary Figure 4**.

**Western blots.** Anti-Lin28a (2J3 mouse monoclonal antibody[32], 1:500), anti-hnRNP A (4B10, G. Dreyfuss, 1:3,000), anti-hnRNP C (4F4, G. Dreyfuss, 1:500), anti-histone H3 (Cell signaling, 9715, 1:2,400), anti-α-tubulin (Abcam, ab52866, 1:500), anti-Trim25 (BD biosciences, 610570, 1:1,000) and anti-Gapdh (Santa Cruz, sc-32233, 1:1,000) were used for western blotting.

**RNA-seq analysis.** RNA-seq data during embryoid body differentiation were obtained from GEO (GSM774956, GSM774958, GSM774960, GSM774962, GSM774964 and GSM774966)[37]. For data presented in **Supplementary Figure 5a**, we used GSM521650 (ref. 87). Gene-level expression was calculated by following a previously described protocol[88] (quantification of reference annotation only (no gene/transcript discovery)). For data shown in **Supplementary Figure 5a**, we selected genes that have CDS information and FPKM (fragments per kb of transcript per million mapped reads) >8 (9,933 genes) and then sorted them by FPKM.

**RNA preparation and quantitative RT-PCR.** Total RNA was extracted using Trizol (Invitrogen, 15596-018) and treated with DNase I (Takara). cDNAs were generated by using RevertAid reverse transcriptase (Fermentas, EP0442), amplified by using power SYBR green PCR master mix (Applied Biosystems, 4368702) and detected by StepOnePlus real-time PCR system (Applied Biosystems). Relative expression level of each gene was calculated using the comparative Ct method with *Actb* as a normalization control. Primer sequences are provided in **Supplementary Table 11**.

**RNA interference.** MCF-7 cells (1 × 10$^6$) were seeded on 100-mm dishes 1 d before transfection and transfected with 30 nM of short interfering RNAs (siRNAs) using Lipofectamine 2000 (Invitrogen). Forty-eight hours after transfection, cells were harvested for further experiment. The siRNA sequences are provided in **Supplementary Table 11**.


77. Castello, A. *et al.* System-wide identification of RNA-binding proteins by interactome capture. *Nat. Protoc.* **8**, 491–500 (2013).
78. Boersema, P.J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A.J.R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).
79. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
80. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
81. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
82. Jain, E. *et al.* Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10**, 136 (2009).
83. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
84. Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).
85. Lunde, B.M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479–490 (2007).
86. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005).
87. Guttman, M. *et al. Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
88. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).