Transcriptome-scale RNase-footprinting of RNA-protein complexes

Zhe Ji[1,2], Ruisheng Song[1], Hailiang Huang[2,3], Aviv Regev[2,4,5] & Kevin Struhl[1]

Ribosome profiling is widely used to study translation *in vivo*, but not all sequence reads correspond to ribosome-protected RNA. Here we describe Rfoot, a computational pipeline that analyzes ribosomal profiling data and identifies native, nonribosomal RNA-protein complexes. We use Rfoot to precisely map RNase-protected regions within small nucleolar RNAs, spliceosomal RNAs, microRNAs, tRNAs, long noncoding (lnc)RNAs and 3′ untranslated regions of mRNAs in human cells. We show that RNAs of the same class can show differential complex association. Although only a subset of lncRNAs show RNase footprints, many of these have multiple footprints, and the protected regions are evolutionarily conserved, suggestive of biological functions.

Target sites for individual RNA-binding proteins have been identified on a transcriptome scale using CLIP-seq (cross-linking and immunoprecipitation-sequencing) or PAR-CLIP (photoactivatable ribonucleoside-enhanced CLIP) techniques[1,2]. Two transcriptome-scale methods for more comprehensive identification of RNA-protein interactions *in vivo* have been described. One approach uses UV cross-linking of cells grown in the presence of 4-thiouridine[3,4], but this is limited to short-range interactions of appropriate stereochemistry to permit UV cross-linking. The other approach involves RNase footprinting of RNA cross-linked with formaldehyde[5]. Both transcriptome-scale approaches map the regions of RNA bound by proteins in the context of the RNA-protein complex, but they do not identify the specific proteins involved. In addition, both methods identify bound regions on a population basis, not at the levels of individual molecules, and hence cannot distinguish between different complexes associated with the same region of RNA.

Sequencing of ribosome-protected RNA, known as ribosome profiling, has been used widely to examine translation *in vivo*[6]. In this procedure, cell extracts are treated with RNase I to degrade all non-protected RNA, and the resulting material is subjected to velocity sedimentation through sucrose to enrich for material >7–10S (corresponds to a 100–200 kDa globular protein) while removing degraded RNA and other low-molecular-weight material. In the course of ribosome profiling experiments, we and others noted that many sequencing reads do not correspond to translated regions. Ribosomes are not specifically selected during the biochemical isolation procedure, and therefore nonribosomal RNA-protein complexes should also be present. In ribosome profiling, sequencing reads correspond to ribosomes that span the entire translated region and show 3-nt periodicity (**Fig. 1a**). In contrast, sequencing reads corresponding to RNase footprints of nonribosomal RNA-protein complexes should be highly localized (**Fig. 1a**,**b**). Each RNA species has a percentage of maximum entity (PME) value that reflects the degree of localization of sequence reads within this RNA (0 represents highly localized and 1 represents uniform distribution across the gene), and different types of RNA-protein complexes have different PME values (**Fig. 1b**).

Based on these considerations, we developed a computational pipeline, Rfoot (**Supplementary Software**), to systematically identify RNA regions protected by nonribosomal RNA-protein complexes. Specifically, Rfoot searches for protected RNA regions with at least ten sequencing reads that are highly localized and do not show 3-nt periodicity. Rfoot is distinct from standard peak-detecting methods in chromatin immunoprecipitation (ChIP)-seq and CLIP-seq analyses that identify, respectively, DNA or RNA regions bound by proteins. Rfoot considers read distribution patterns and distinguishes between RNA protected by ribosomes, which represent the majority of sequence reads, from RNA protected by nonribosomal complexes. Unlike analyses of ChIP-seq and CLIP-seq data that require peak detection methods to map bound regions from a population of molecules of varying size with endpoints having varying distances from the protected region, each sequencing read in Rfoot analysis corresponds directly to the fully protected region of an individual RNA-protein complex.

Rfoot analysis of our previous ribosome profiling data[7] from two isogenic human cancer cell models (Src-inducible mammary epithelial and Ras-dependent fibroblast)[8] revealed that 11.3% of the sequencing reads correspond to nonribosomal RNA-protein complexes. Protected RNA regions, and presumably RNA-protein complexes, were observed for virtually all types of cytoplasmic and nuclear RNAs: mRNAs (3′ untranslated regions (UTRs)), lncRNAs, small nucleolar (sno) RNAs, spliceosomal RNAs, microRNAs (miRNAs) and tRNAs. Detection of a given RNA-protein complex depends on the abundance of the RNA, the fraction of RNA stably bound by proteins throughout the experimental procedure and the total number of sequencing reads. Although the sequencing depth used here is sufficient to identify RNA-protein complexes from all

[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Analytic and Translation Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [4]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. Correspondence should be addressed to K.S. (kevin@hms.harvard.edu) or A.R. (aregev@broadinstitute.org).

Received 25 April 2015; accepted 25 November 2015; published online 22 February 2016; doi:10.1038/nbt.3441

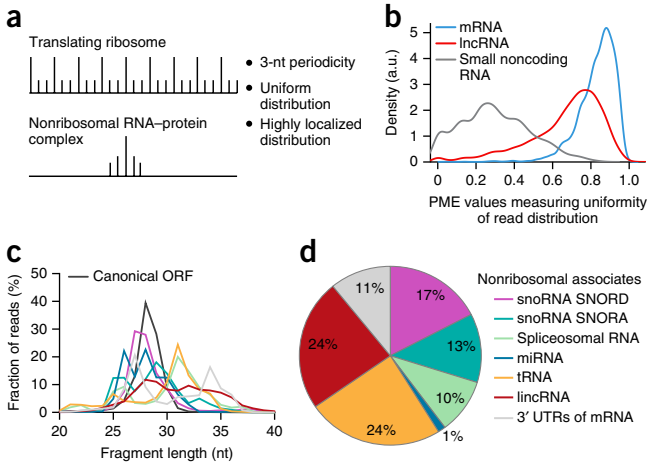

Figure 1 Identifying nonribosomal RNA-protein–associated footprints. (**a**) Read distribution pattern in translated ORFs and nonribosomal RNA-protein complexes. (**b**) Distribution of PME values across transcripts (60-nt window). a.u., arbitrary units. (**c**) Read fragment length (nt) of RNase footprints in types of transcripts. (**d**) Fraction (in percent) of the various types of RNA-protein complexes.

RNA classes, greater sequencing depth would likely reveal additional complexes involving mRNAs, miRNAs or lncRNAs that are poorly expressed. As expected, different types of RNA-protein complexes protected different lengths of RNAs (**Fig. 1c**,**d**), and the same complexes were observed when translation was inhibited by either cycloheximide or harringtonine.

Small nucleolar (sno)RNAs are primarily nuclear, with the 'C/D box' class of snoRNAs guiding methylation and the 'H/ACA box' class guiding pseudouridylation of other RNAs[9]. We identified RNase footprints for 112 C/D box RNAs and 68 H/ACA box RNAs (**Supplementary Table 1**), which represent almost all expressed snoRNAs. The protected region of C/D type snoRNAs covers the stem loop structure between the C motif (UGAUGA) and D motif (CUGA) (**Fig. 2a**,**b**). The region between C and D motifs forms an RNA duplex with the methylation site of the target RNA[10], and is bound by C/D ribonucleoproteins[9]. Notably, although C/D box snoRNAs can form symmetric stem loop structures (**Fig. 2a**), the protected region covers the left arm of the snoRNA SNORD105, the right arm of SNORD110, both arms of SNORD113–9, and the middle D and C motifs from different arms of SNORD87 (**Fig. 2b**). For H/ACA type snoRNAs, the protected regions flank the H box (ANANNA), the single-stranded region linking two stem loop structures, and the ACA box located in the tail region (**Fig. 2c**,**d**). These motifs are bound by the H/ACA ribonucleoproteins[9]. Interestingly, although C/D box snoRNAs can form symmetric stem loop structures (**Fig. 2a**), the protected region covers the left arm of SNORD105, the right arm of SNORD110, both arms for SNORD113–9, and the middle D and C motifs from different arms of SNORD87 (**Fig. 2b**). For H/ACA type snoRNAs, the protected regions flank the H box (ANANNA), the single-stranded region linking two stem loop structures, and the ACA box located in the tail region (**Fig. 2c**,**d**). Reads in SNORA23 are mostly in the H box (**Fig. 2d**), whereas reads in SNORA3 are associated more with the ACA box (**Fig. 2d**). Thus, it appears that RNA-protein complexes within an individual snoRNA class can have different stabilities or conformations.

Spliceosomal RNAs associate with spliceosomal proteins to form small nuclear ribonucleic particles (snRNPs) that are critical for RNA splicing[11], and we detected RNase footprints for all types of spliceosomal RNAs (**Supplementary Table 1**). For snRNP RNU11, the protected region is mainly associated with the Sm site (**Fig. 2e**), a conserved sequence (consensus AUUUGUGG) bound by the SMN complex[12]. For RNU12, protected regions are observed both for the Sm site and the 5′ hairpin structure (**Fig. 2f**) that interacts with branch points of pre-mRNA[12].

We detected RNase footprints for almost all expressed tRNAs (157 in **Supplementary Table 1**). The protected regions are located in the D loop and TΨC loop. The D loop is recognized by aminoacyl-tRNA synthases[13], whereas the TΨC loop is important for ribosome binding[14]. The read distribution between these loops varies among tRNAs. For example, more sequencing reads were observed for the D loop of tRNA9 on chromosome 1 (**Fig. 2g** and **Supplementary Fig. 1a**), or the TΨC loop of tRNA2 on chromosome 12 (**Fig. 2h** and **Supplementary Fig. 1b**). Thus, as observed for snoRNAs, tRNA-protein complexes can have different stabilities or conformations.

We detected RNase-protected regions for 12 miRNAs (**Supplementary Table 1**) that cover the mature miRNA (**Supplementary Fig. 2a**,**b**). If one transcript encodes two mature miRNAs (e.g., miR21 and miR21*), sequence reads were observed over both mature miRNAs (**Supplementary Fig. 2c**). The RNA-induced silencing complex may bind to these regions, but it is unknown why RNase footprints are not detected for most expressed miRNAs.

The fact that mRNAs are associated with ribosomes makes it difficult to identify nonribosomal RNA-protein complexes that interact with protein-coding or noncanonical translated regions. In this regard, we found 95 protected RNA regions in 3′ UTRs of 69 mRNAs (**Supplementary Table 1**). For example, the protected RNA sequence in AMD1 3′ UTR also forms a stable hairpin structure (**Supplementary Fig. 3**).

Some lncRNAs interact with polycomb proteins, and it has been suggested that these interactions affect chromatin structure and transcription[15,16]. Although we detected RNase footprints for only 87 (8%) of expressed lncRNAs, this is five times as many footprints as observed for 3′ UTRs, even though the number of nucleotides in 3′ UTRs is higher than in lncRNAs. Moreover, in this subset of 87 lncRNAs, we identified 208 nonribosomal binding sites (**Supplementary Table 1**), an average of 2.4 footprints/lncRNA. For example, the telomerase component TERC contains three nonribosomal protein-binding sites (**Supplementary Fig. 4a**) that cover the H- and CAB-boxes of the small cajal body–specific (sca)RNA domain, and a 5′ single-stranded region (**Supplementary Fig. 4b**), whereas the MALAT1 lncRNA showed several RNase footprints at regions tending to form RNA hairpin structures (**Fig. 2i**). Notably, one MALAT1 region showed two distinct RNase footprints as defined by different protected fragment lengths (**Fig. 2i**) and a similar situation occurred at other lncRNAs (e.g., **Supplementary Fig. 5**). Distinct RNase footprints over the same region could reflect completely different or related RNA-protein complexes or alternative conformations of the same complex. In addition, some RNA-protein complexes were cell-type specific (**Fig. 2i** and **Supplementary Fig. 5**). Considering all RNase footprints in lncRNAs, PhastCon scores based on 44-vertebrate Multiz alignment[17] of nucleotide sequences reveals that the conservation level is about twofold higher than surrounding sequences (**Fig. 2j**; Wilcoxon rank-sum test P-value < 10^{-19}). Taken together, these observations suggest that RNase footprints in lncRNAs may represent RNA-protein complexes that carry out biological functions.

Our experimental method differs from a transcriptome-scale RNase footprinting approach described previously[5], and it is advantageous in several respects. First, by avoiding cross-linking, we are

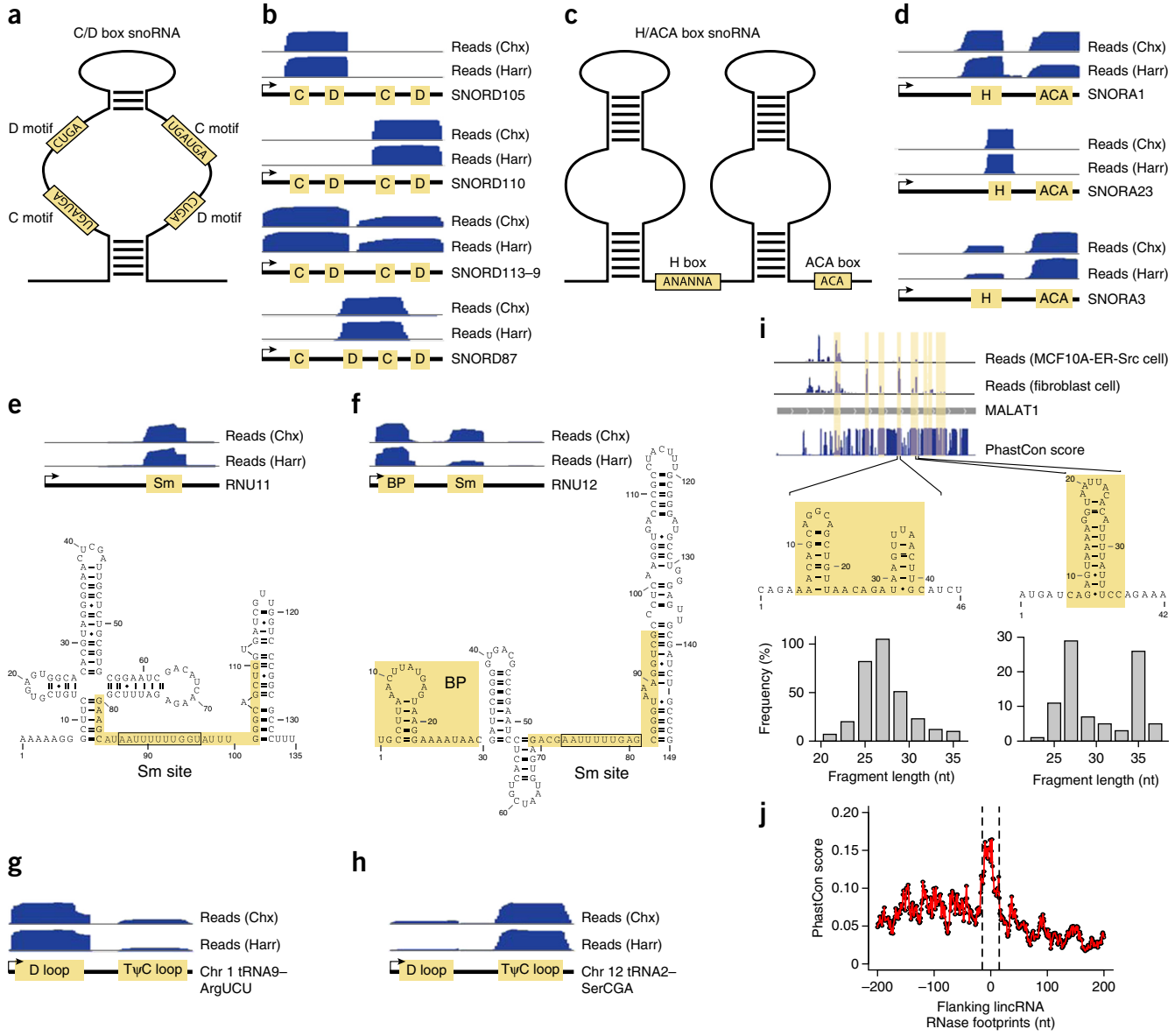

Figure 2 Footprinted regions on various classes of RNA. (**a**) Structure of C/D box snoRNAs. (**b**) Read distribution of the indicated C/D box snoRNAs with respect to the C and D motifs. (**c**) Structure of H/ACA box snoRNAs. (**d**) Read distribution of the indicated H/ACA box snoRNAs with respect to the H and ACA motifs. (**e**,**f**) Read distribution in RNU11 (**e**) and RNU12 (**f**) spliceosomal RNAs with respect to the indicated motifs and secondary structures. (**g**,**h**) Read distribution in chr1.tRNA9-ArgUCU (**g**) and chr12.tRNA2-SerCGA (**h**) tRNAs with respect to the D and TΨC loops. (**i**) Read distribution in the MALAT1 lncRNA along with protected regions and PhastCon scores based on 44-vertebrate Multiz alignment. Read distributions in the indicated cell types and fragment lengths and RNA structures in two protected regions are shown. The two fragment length peaks in the protected region on the right indicate structurally and/or conformationally distinct RNA-protein complexes. (**j**) Distribution of mean PhastCon scores around lncRNA RNase footprints.

able to identify native RNA-protein complexes. Cross-linking can cause artifacts, although it also enables the detection of less stable complexes. Second, whole-cell extracts are subject to a crude purification step that enriches for RNA-protein complexes and removes degraded RNA, thereby eliminating sequence reads corresponding to RNA not associated with proteins. In principle, distinct RNA-protein complexes could be enriched by fractionation based on molecular weight or by immunoprecipitation with an antibody against a specific protein (analogous to CLIP-seq). In addition, factors important for RNase footprints can be identified by comparing cells depleted of an individual factor with their wild-type counterparts. Third, each sequencing read corresponds to a complete protected region for an individual RNA molecule. By examining the size distribution of the protected region of individual RNase footprints, we detected distinct RNA-protein complexes for some footprints of MALAT1 and several other lncRNAs. In contrast, RNase footprints obtained with the previous approach represent averages over many molecules such that distinct RNA-protein complexes cannot be detected.

Our method can analyze reported and future ribosome profiling datasets for RNase footprints on nonribosomal RNA-protein complexes. In this regard, we performed Rfoot analysis on published ribosomal profiling datasets from mouse cell lines[18,19]. In accord with our results in human cells, 14.5% of the sequencing reads corresponded to nonribosomal RNA-protein complexes, and the PME profiles of the mouse (**Supplementary Fig. 6a**) and human (**Fig. 1b**) samples were similar. Furthermore, RNA-protein complexes representing all types

of RNA species are identified in these mouse cell lines, and the relative proportion of these types of complexes are roughly comparable to what we observed in human cells (compare **Fig. 1d** with **Supplementary Fig. 6b**). Analyzing translation (ribosome footprints) and nonribosomal RNA-protein complexes in the same sample cannot be done by other methods. Lastly, we note that most of the RNA-protein complexes identified here have not been described previously. Therefore, our method represents a distinct and complementary approach to identifying RNA-protein complexes on a transcriptome scale.

METHODS

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

Accession codes. Experimental data for the identification of RNA-protein complexes is available at GEO (GSE65885), and the Rfoot package can be downloaded from http://www.broadinstitute.org/~zheji/software/Rfoot.0.1.tar.gz

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

ACKNOWLEDGMENTS

This work was supported by grants to K.S. from the National Institutes of Health (CA 107486). A.R. is a Howard Hughes Investigator.

AUTHOR CONTRIBUTIONS

Z.J., R.S., A.R. and K.S. conceived of and designed experiments, R.S. performed experiments, Z.J. and H.H. performed the data analysis, and Z.J., R.S., A.R. and K.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Zhang, C. & Darnell, R.B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* **29**, 607–614 (2011).
2. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).
3. Baltz, A.G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690 (2012).
4. Freeberg, M.A. *et al.* Pervasive and dynamic protein binding sites of the mRNA transcriptome in *Saccharomyces cerevisiae*. *Genome Biol.* **14**, R13 (2013).
5. Silverman, I.M. *et al.* RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biol.* **15**, R3 (2014).
6. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
7. Ji, Z., Song, R., Huang, H., Regev, A. & Struhl, K. Many lincRNAs are translated and some are likely to express functional proteins. *eLife* 08890, doi:10.7554/eLife.08890 (19 December 2015).
8. Hirsch, H.A. *et al.* A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases. *Cancer Cell* **17**, 348–361 (2010).
9. Matera, A.G., Terns, R.M. & Terns, M.P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* **8**, 209–220 (2007).
10. Kiss-László, Z., Henry, Y. & Kiss, T. Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO J.* **17**, 797–807 (1998).
11. Will, C.L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3**, a003707 (2011).
12. Russell, A.G., Charette, J.M., Spencer, D.F. & Gray, M.W. An early evolutionary origin for the minor spliceosome. *Nature* **443**, 863–866 (2006).
13. Hendrickson, T.L. Recognizing the D-loop of transfer RNAs. *Proc. Natl. Acad. Sci. USA* **98**, 13473–13475 (2001).
14. Peattie, D.A. & Herr, W. Chemical probing of the tRNA–ribosome complex. *Proc. Natl. Acad. Sci. USA* **78**, 2273–2277 (1981).
15. Batista, P.J. & Chang, H.Y. Long noncoding RNAs: cellular address codes in development and disease. *Cell* **152**, 1298–1307 (2013).
16. Rinn, J.L. & Chang, H.Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).
17. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
18. Eichhorn, S.W. *et al.* mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol. Cell* **56**, 104–115 (2014).
19. Diaz-Muñoz, M.D. *et al.* The RNA-binding protein HuR is essential for the B cell antibody response. *Nat. Immunol.* **16**, 415–425 (2015).

ONLINE METHODS

RNase-footprinting. BJ fibroblast cell lines (EH, EL and ELR) were cultured on Knockout DMEM (Invitrogen) with 10% FBS, medium 199, glutamine and penicillin-streptomycin[20]. The breast epithelial cell line (MCF-10A-ER-Src) was grown in DMEM/F12 with 5% charcoal-stripped FBS (Invitrogen) and supplements[21]. Cells were seeded at 1×10^6 cells per 10-cm culture dish and cultured overnight. Cells were tested negative for mycoplasma. MCF-10A-ER-Src cells were treated by 1 μM 4-hydroxy-tamoxifen for various time points (1, 4 and 24 h) to induce transformation. The cells and tamoxifen induction were validated by marker gene expression levels based on RNA-seq. Cells were pretreated with cycloheximide (100 μg/ml; Sigma-Aldrich) for 90 s or harringtonine (2 μg/ml; Santa Cruz) for 5 min, and detergent lysis was then performed with flash-freezing in liquid nitrogen. DNase I–treated lysates were then treated with RNase I; RNA-protein complexes were enriched by sedimentation through 34% sucrose as previously described for ribosome profiling[22]. Based on the centrifugation parameters, we estimate that complexes >7–10S will be significantly enriched by this procedure. The protected RNA fragments were prepared for Illumina TruSeq library construction[22], and RNase-footprinting libraries were sequenced with Illumina HiSeq 2500.

Sequence read mapping and transcription annotations. After removing sequence reads corresponding to human rRNA sequences (5S, 5.8S, 18S and 28S), we aligned reads to human reference transcriptome and genome sequence (hg19) using Tophat[23] with default parameters. We obtained 65 million sequenced reads mapping to a unique location in the genome, and these were used for subsequent analyses. Protein coding genes were defined by the RefSeq database. Short noncoding RNAs were defined by RefSeq and GENCODE databases[24]. lincRNAs were defined by a union set of RefSeq, GENCODE lncRNAs[24] and Human Body Map lincRNAs[25]. We required a lncRNA to have introns, or otherwise, to have length greater than 500 nt, and not overlap with a protein coding gene in the same strand.

Rfoot to identify nonribosomal RNase footprints. We used the middle position of a sequence read to represent the genomic location of the protected RNA. As nonribosomal protein-associated fragments are highly localized, we developed a method named Percentage of Maximum Entropy (PME) to measure the uniformity of read distribution in a defined region (see below). Low values indicate highly localized distribution in the region, whereas high PME values indicate that reads are evenly distributed. For each transcript, we used a scanning window of 60 nt, and calculated the PME value in each window. The 60-nt window size was chosen to optimize separation between nonribosomal protein-associated regions and translated ORFs (**Fig. 1a**,**b**). We excluded the regions with high PME values (PME > 0.6) and translated regions (based on RibORF[7]) with predicted probability > 0.5 to minimize the false-negative rate. For the remaining genomic regions, we clustered reads located within 5 nt. We considered a nonribosomal protein-associated site to have >10 reads with >3 reads in the peak site. To further remove reads corresponding to translation, we excluded regions showing 3-nt periodicity considering locations supported by most reads, even if these reads are not in an intact candidate ORF.

PME to measure uniformity of read distribution. For a candidate window, suppose total read number is N, the region length is L nt. We divided the window into smaller regions based on N and L in the following way. If $N > L/3$, we define a region length as 3 nt. Otherwise, a region length is defined as floor $((L/3)/N)$. For each region i, we calculated the fraction of reads in the region: $P(X_i) = N_i/N$, where N_i represents number of reads in region i. The entropy value is defined as:

$$H(X) = \sum_{i=1}^{n} (P(X_i) * \log_2 P(X_i))$$

We then calculate the PME value as PME = $H(X)/\max(H)$. $\max(H)$ represents the entropy value assuming the reads are perfectly evenly distributed across the window.

Identify nonribosomal RNA-protein complexes in mouse cells from published ribosome profiling data. We downloaded two published mouse ribosome profiling data sets (GSE60426 and GSE62134) from the GEO database[18,19], and used the Rfoot algorithm to identify nonribosomal RNA-protein complexes (**Supplementary Fig. 6**). To compare the mouse and human data sets, we used 65 million uniquely mappable reads from the published mouse data sets. We note that the apparent small number of spliceosomal RNA-protein complexes in the mouse data sets likely reflects poor annotation, not a meaningful biological difference.

20. Hahn, W.C. *et al.* Creation of human tumour cells with defined genetic elements. *Nature* **400**, 464–468 (1999).
21. Iliopoulos, D., Hirsch, H.A. & Struhl, K. An epigenetic switch involving NF-kappaB, Lin28, Let-7 MicroRNA, and IL6 links inflammation to cell transformation. *Cell* **139**, 693–706 (2009).
22. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M. & Weissman, J.S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550 (2012).
23. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
24. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
25. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).

doi:10.1038/nbt.3441