Online Methods

TEISER: detailed description of the algorithm

Genome profile

A *genome profile* is defined across the genes in the genome, where each gene is associated with a unique measurement. Whole-genome measurements, discrete or continuous, can be obtained from a variety of experimental or computational sources (*e.g.* Supplementary Fig. 1).

Structural motif definition

Each structural motif is defined as a series of context-free statements that define the structure and sequence of the motif (Supplementary Fig. 2). A context-free grammar is a set of production rules that describes how phrases are made from their building blocks. Considering a structured RNA molecule as a phrase, its potential building blocks are the different base pairs and bulges. Loops can be considered as bulges that happen at the beginning of phrases. Also, internal loops can be considered as combination of left and right bulges in the middle of phrases. The context-free grammar that we have used contains the following production rules: $S \rightarrow S$[AUCGN], $S \rightarrow$[AUCGN]S, $S \rightarrow$[AUCGN]S[AUCGN]; wherein the first production rule depicts a right bulge, the second production rule results in a left bulge, and the third production rule creates a base-pairing. For example, consider the stem loop AAACGCUUU (the stem region is underlined). Let the symbol *S* be a non-terminal symbol that stands for this stem loop; The production rule $S \rightarrow S$G adds a G to the 3' end of the molecule, creating a new *S*, AAACGCUUUG, which has an unpaired 3'-end G. Next, using the production rule $S \rightarrow$GSC, we can add a G to the 5' end and a C to the 3' end of the molecule and make them pair with each

other, again creating a new *S*, GAAACGCUUUGC, which can be further expanded in this way. Note that the G that we added in the previous step has now become a right bulge.

Motif profile

For every given motif, we create a binary vector across all the genes, in which '1' denotes the presence and '0' denotes the absence of that motif. This vector is called a *motif profile*.

Creating seed CFGs

We used, as the seed motifs, an exhaustive set of context-free statements that represented all possible stem-loop structures that satisfied the following criteria: stem length of at least 4 bp and at most 7 bp; loop length of at least 4 nt and at most 9 nt; at least 4 and at most 6 production rules representing non-degenerate bases (i.e. production rules that are not $S \rightarrow SN$, $S \rightarrow NS$, or $S \rightarrow NSN$); and information content of at least 14 bits and at most 20 bits. The information content of the motif *M*, which is represented by *n* production rules, was defined as $-\log_2(p_M)$, wherein p_M is the probability that a random sequence of length *l* matches the *n* production rules of motif *M*, with *l* being equal to $2 \times n_1 + n_2$ in which n_1 is the number of production rules that represent base pairings and n_2 is the number of production rules that represent bulges ($n_1+n_2=n$).

Quantizing continuous genome profiles

Mutual information is defined for both continuous and discrete random variables; however, in practice, continuous data are discretized before calculating the mutual information (MI) values. Our quantization procedure involves using equally populated "bins". Thus, the discretization step only requires a single parameter, *i.e.*, the number of genes in each bin. In TEISER, we have set the default number of bins to 30 (N_e=30). It should be noted that the results are not sensitive to

variations in the value of N_e as long as N_e is >10 and each bin has more than ~100 associated transcripts.

Removing recently duplicated genes

Recently duplicated members of gene families or transposons often share a significant amount of sequence identity in their UTRs. They also tend to cross-hybridize on the arrays and show a high artificial correlation. This would in turn bias our search towards conserved elements in the UTRs of these genes. In TEISER, similar to FIRE[2], we remove the duplicates that have similar values (*e.g.* fall in the same bin after quantization of the input genome profile). A MegaBlast *E*-value cutoff of 1×10^{-15} was used to identify duplicates.

Calculating the mutual information values

We performed mutual information (MI) calculations between the *genome profile* and the *motif profiles* using algorithms introduced and described elsewhere[2,10]. These algorithms take the necessary steps to ensure reliable MI calculations (*e.g.* minimum sample sizes for reliable estimation of joint distributions).

Randomization-based statistical testing

To assess the statistical significance of the calculated MI values, TEISER uses a non-parametric randomization-based statistical test. In this test, the *genome profile* is shuffled 1,500,000 times and the corresponding MI values are calculated. A motif is deemed significant only if the real MI value is greater than all of the randomly generated ones. In TEISER, in order to minimize the required number of tests, structural motifs are first sorted based on the MI values (from high to low) and the statistical test is applied in order. When 20 contiguous motifs in the sorted list do not pass the test, the procedure is terminated.

Optimization of the identified seeds into more informative motifs

Our initial collection of structural motifs, despite being large, is a coarse-grained sampling of the entire space. Mainly, it provides us with a set of informative seeds that should be later optimized into closer representations of their actual form[2]. Accordingly, all the structural motifs that pass the previous stage are further optimized and elongated. The process involves:

1. Optimization: randomly choose one of the context-free statements (production rules) from the motif and convert its sequence information to all possible combinations of nucleotides. Evaluate all the resulting structural motifs and accept the one that results in the highest MI value.

2. Elongation: production rules are added to the end of the context-free phrase that represents the motif, thus extending its effective length in the form of a base pair or a bulge. The increase in length is similarly accepted only if it results in a higher MI value.

Removing redundantly informative structural motifs

Motifs that redundantly represent the same potential *cis*-regulatory elements are identified and removed using the concept of conditional information as described before[2,10].

Finding robust motifs

TEISER also performs jack-knife resampling to find robust motifs that are not over-sensitive to the composition of the input data. For each predicted motif, we perform 10 jackknifing trials where, in each trial, one third of the genes are randomly removed and the mutual information value and its statistical significance is evaluated. The *robustness* score is then defined as the number of trials in which the motif remains significant (scores better in the original genome profile than in all the randomly shuffled genome profiles) after resampling, ranging from 0/10 to

10/10. By default, TEISER requires the motif to be significant in more than half of the trials (a *robustness* score equal to or greater than 6/10). While this parameter can be changed at the user's discretion, our experience with both TEISER and FIRE[2] suggests that this threshold results in very low false discovery rates across a variety of datasets (discrete and continuous).

Patterns of motif enrichment and depletion

For a given motif, a high mutual information value results from the non-random distribution of its targets across the input range. This results in significant patterns of enrichment and depletions across the *genome profile*, which can be quantified by calculating enrichment/depletion scores. These scores result from the log transformation of *p*-values calculated based on the hypergeometric distribution, as described previously[2].

Final statistical tests

In case the *genome profile* is continuous, one can require TEISER to return motifs that are enriched at one end of the data range or the other (*e.g.* structural motifs in Fig. 1). TEISER accomplishes this through calculating the Spearman correlation between the enrichment scores and the average data value across all the bins. For the structural motifs in Fig. 1, the *p*-value threshold for these Spearman correlations was set to 0.001 (for, Supplementary Fig. 3, this value is 0.01 which puts the FDR at 10%). It should be noted, however, that other statistical tests could be used in this step at the discretion of the user. The goal, ultimately, is to identify the motifs that show significant enrichments at either end of the data range.

Inter-species conservation

For each motif, we also calculate a conservation score based on its network-level conservation with respect to a related genome[2]. For this, orthologous transcripts in both genomes are scanned

for the presence/absence of the motif. The overlap of positive sequences between the orthologous sequences is used to calculate a hypergeometric *p*-value[2]. The conservation score is then defined as 1-*p*, which ranges between 0 and 1 (1 being highly conserved between the two genomes). In this study, we have used the human and mouse genomes to calculate the conservation scores associated with each structural motif.

Finding potentially active instances of each motif

As described previously[2], we defined the target genes of a predicted motif as all transcripts whose 3' or 5' UTRs contain the motif and are associated with a category/bin where the motif is enriched. In other words, these are the transcripts whose UTRs contain potentially "active" motif occurrences. Upon identifying these likely targets for each structural motif, a weight-matrix can be generated from these potentially functional instances as a post-processing step (Supplementary Table 2).

False-discovery rate

In order to assess the false discovery rate, we ran 30 trials with shuffled 5' and 3' UTR sequences. In all the trials, not a single motif passed all the statistical tests. Thus, in case of the stability dataset, the number of false positives in each trial, on average, is smaller than $1/30 \approx 0.34$, which corresponds to an FDR of <0.01.

Predicting functional interactions

Given two motifs, structural or linear, one can assess their putative functional interaction through measuring how informative the presence of one would be about the presence or absence of the other. For revealing these interactions, we again use mutual information values calculated for pairwise motif profiles of structural and linear motifs. Randomization-based statistical tests are

then used to find the significant interactions. For this, one of the motif profiles is shuffled 10,000 times and the interaction is deemed significant only if the real mutual information value is higher than all the 10,000 random ones.

Predicting the target pathways

iPAGE[10], with default settings, was used to identify the likely pathways that are regulated by the discovered structural and linear motifs.

Availability

TEISER is available online for download at https://tavazoielab.c2b2.columbia.edu/TEISER.

Measuring mRNA stability

RNA stability measurements were performed based on a previously published protocol[1]. In short, MDA-MB-231 cells at 70% confluency were incubated in the presence of 25μM 4-thiouridine (sigma) for 4hr. Then the cells were washed with fresh media (DMEM+10% FBS) and incubated for 0, 1, 2 and 4h. At each time point, cells were washed with cold PBS and RNA extraction was performed using total RNA purification kit (Norgen Biotek). The 4sU thiol groups were then biotinylated using EZ-Link Biotin-HPDP (Pierce). We subsequently used μMacs magnetic columns (Miltenyi Biotec) to capture the labelled RNAs. The resulting samples were then processed for one-color hybridization using one-color low-input quick-amp labelling kit (Agilent) and hybridized according to the manufacture's instructions. One-color RNA spike-in kit (Agilent) was used as endogenous control to normalize values between arrays.

For each transcript, the drop in signal as a function of time was used as a measure of mRNA stability (Supplementary Fig. 1): $r = -\ln\left(\frac{S_t}{S_0}\right)\Big/ t$, where S_t denotes signal at time t. Linear regression was used to calculate r for each transcript based on the hybridization signals from the four time points. It should be noted that TEISER is a non-parametric approach, thus it is the ranking rather than the actual stability values that underlies our motif discovery.

Transfection of decoy and scrambled oligonucleotides

We chose real instances of the sRSM1 structural motifs from NM_014363, which contains four instances of sRSM1, to create two decoy sets of sequences, each containing two of these instances (underlined) along with part of the real sequences as context.

Set1:

AAAACTATTTTGAAGATGGTGGTGA**GCTGCAAAATAGCTGGATGG**<u>**ATTTGAATGATTGGGATGA**</u>**TA**
CATCATTGAACTGCACTTTATATAACCAAAGCTTAGCAGTT**TGTTAGATAAGAGTCTATGT**<u>**ATGTC**
TCTGGTTAGGATGA</u>**AGTTAATTTTATGTTTTTAA**CATGGTATTTTTGAAGGAGCTAATGAAACACTG
G

Set2:

ATTGTTTC**TGGAAACTGCTTGCCAAGAC**<u>**AACATTTATTAACTGTTAG**</u>**AACACTTGCTTTATGTTTG**
TGTGTACATATTTTCCACAAATGTTATAATTTATATAGTGTGGTTGAACAGGATGCAATCTTTTGTTG**T**
CTAAAGGTGCTGCAGTTAA<u>**AAAAAAAACAACCTTTTCT**</u>**TTCAATATGGCATGTAGTGG**AGTTTTT

For the scrambled controls, we used the shuffled version of the putative binding sites (the bold sequences in these sets; also see Supplementary Fig. 5). These two decoy/scrambled sets were then chemically synthesized (IDT). An upstream T7 promoter was used to transcribe the constructs *in vitro* using Megascript T7 kit (Ambion). In order to reduce cytotoxicity, RNA

molecules were capped and poly-A tailed using Cap Analog (Ambion) and poly-A polymerase (NEB). MDA-MB-231 cells at 80% confluency were transfected with the resulting RNA oligos using Lipofectamin 2000 reagent (Invitrogen) according to manufacturer's recommendations. Experiments were performed in duplicates for each set. Forty-eight hours post-transfection, we extracted RNA and differentially labeled the samples with Cy3 or Cy5 dyes. The samples were then hybridized on Agilent human gene expression arrays (4×44k). The Cy3/Cy5 ratios from the two biological replicates were then averaged into a single dataset as log of ratios, which was then analyzed by TEISER.

Reporter system for testing the functionality of sRSM1 instances

The plasmid pcDNA5/FRT/TOPO (Invitogen) was used to clone a GFP-coding sequence along with a gateway cloning site downstream of GFP (in its 3' UTR). Decoy and scrambled sequences (Set1 in the previous section) were subsequently cloned into the resulting construct using the gateway site. The resulting plasmids were transfected into the Flp-In 293 cell line (Invitrogen), and the cells were grown in Hygromycin for selecting stably transfected cells. The resulting cell lines, named Flp-In 293 GFP-Decoy and Flp-In 293 GFP-Shuffled, were subjected to FACS measurements to quantify GFP expression. For the decay rate measurements, cells were incubated in media with 5μg/mL of α-Amanitine (Sigma). Time points were taken at 0, 1.5, 3 and 6 hours in duplicates for Flp-In 293 GFP-Decoy and Flp-In 293 GFP-Shuffled cells. Quantitative PCR (Fast SYBR Greem Master Mix, Ambion) was then used to determine the relative quantity of GFP transcript in each cell line at different time-points using 18S rRNA as endogenous control.

Identifying binding candidates of sRSM1

We used a published protocol[13] to isolate potential RNA-binding proteins that bind sRSM1. In short, the StreptoTag aptamer was added downstream of the Set1 decoy and scrambled sequences. The resulting RNAs were then immobilized on a dihydrostreptomycin Sepharose column (GE Healthcare) and were used to immunoprecipitate potential partners. Total protein was extracted from MDA-MB-231 cells (Total Protein Extraction Kit, Millipore), 1,000 μg of which was used as input to each column. Samples were then washed, eluted in 10μM streptomycin and subjected to in-solution digestion[25]. Tryptic peptides were then analyzed by nanoliquid chromatography-tandem mass spectrometry using an Ultimate 3000 nRSLC (Dionex) coupled online to an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific), as previously described[24].

HNRPA2B1 knock-down

ON-Target*plus* (Dharcomon) set of siRNAs for HNRPA2B1 (target sequences: GAGGAGGAUCUGAUGGAUA, GGAGAGUAGUUGAGCCAAA, and GCUGUUUGUUGGCGGAAUU) were used to transfect MDA-MB-231 cells (grown in D10F medium) using Lipofectamine 2000 (Invitrogen). Three of the four tested siRNAs resulted in a substantial knock-down in HNRPA2B1 (more than 2-fold reduction in expression, log ratio>0.4 and p<1e-7) and their corresponding samples were used for hybridization. Forty-eight hours post-transfection, we extracted total RNA from each sample along with mock-transfected controls. We then differentially labeled the RNA samples with Cy3 and Cy5 dyes and hybridized them to Agilent human gene expression arrays (4×44k). The log of signal ratios was used as a measure of differential expression between the samples and controls. These values were

averaged across the three samples and were subsequently analyzed by TEISER to assess the enrichment/depletion pattern of sRSM1 across the distribution.

For the decay rate measurements, forty-eight hours post-transfection, cells were incubated in media with 5μg/mL of α-Amanitine (Sigma). Time points were taken at 0, 1, 2 and 4 hours in duplicates for the siRNA-transfected samples and mock-transfected controls. Each sample was then Cy3-labeled and hybridized to expression arrays (Agilent 4×44k) in duplicates and the reported signals were used to calculate decay rates. Following this procedure, for each transcript, four decay rates (two biological replicates, each having two technical replicates) were calculated from the siRNA-transfected samples and four decay rates from the controls. For each transcript, we then calculated a value according to $s.(1-p)$, where p is the t-test p-value between the two sets and s denotes whether the decay rates are higher in the siRNA samples (+1) or the mock controls (-1). After this transformation, the data range is between -1 and 1 with the background genes (the transcripts that show little change between the two samples) around 0. TEISER was then used to visualize the enrichment pattern of sRSM1 across this data range.

Identifying transcripts that interact with HNRPA2B1 (RIP-chip)

A myc-tagged ORF clone of HNRPA2B1 (variant A2, OriGene) was transfected into MDA-MB-231 cells (grown in D10F medium) using Lipofectamine LTX and Plus reagent (Invitrogen). Seventy-two hours post-transfection, the cells were washed with cold PBS and UV-irradiated at 4000 mJ/cm^2. The cells were then collected and lysed with 1mL M-PER Reagent (Pierce) and 10 uL RNasin (NEB). The samples were subjected to DNAse treatment (baseline ZERO DNAse) for 15min at 37°C. Samples were then centrifuged at 16,000×g at 4°C for 20 minutes to pellet the cell debris. Immunoprecipitation of tagged HNRPA2B1 protein was performed using

Mammalian c-Myc Tag IP/Co-IP Kit (Pierce) per manufacturer's instructions. Upon elution, samples were subjected to proteinase K digestion and polyadenylation. The RNA molecules in each sample were extracted using RNeasy MinElute Cleanup Kit (Qiagen) and Cy3-labeled using low-input quick-amp labeling kit (Agilent). As control, we used Cy5-labeled RNA samples extracted prior to HNRPA2B1 immunoprecipitation. The samples were hybridized to Agilent human gene expression arrays (4×44k) and the log of signal ratios was used as a measure of transcript affinity to HNRPA2B1. For each transcript, affinity values were averaged across two biological replicates and TEISER was used to assess the enrichment/depletion pattern of sRSM1.

Identifying 3'UTR binding sites of HNRPA2B1 (HITS-CLIP)

A strategy similar to that of target transcript identification was used to discover the HNRPA2B1 binding sites. Upon UV-irradiation of mycHNRPA2B1-transfected cells, the samples were subjected to the HITS-CLIP protocol previously described elsewhere[26]. ChIPSeeqer[21], an integrated ChIP-seq analysis platform, was used to identify binding sites and extract real and random sequences (default parameters) for analysis with TEISER.

Measuring growth-rates in HNRPA2B1 knock-down cells

HNRPA2B1 siRNAs (Dharcomon) were used to knock-down the expression of this regulator. Seventy-two hours post-transfection, four independent samples were harvested and counted in duplicates as the baseline number of cells at time zero. Similarly, samples were counted at 25, 49.5, 73.5 and 99.5 hour time-points. The same experiment was performed for mock-transfected cells. Using an exponential growth model, the log-ratio of the counted cells at each time-point was used to estimate a growth rate for siRNA-transfected and mock-transfected samples.

ANCOVA was used to determine the *p*-value associated with the observed differences between the two growth rates.