LETTER

doi:10.1038/nature11013

Systematic discovery of structural elements governing stability of mammalian messenger RNAs

Hani Goodarzi[1,2]†, Hamed S. Najafabadi[3,4]†, Panos Oikonomou[1,2]†, Todd M. Greco[2], Lisa Fish[5], Reza Salavati[3,4,6], Ileana M. Cristea[2] & Saeed Tavazoie[1,2]†

Decoding post-transcriptional regulatory programs in RNA is a critical step towards the larger goal of developing predictive dynamical models of cellular behaviour. Despite recent efforts[1–3], the vast landscape of RNA regulatory elements remains largely uncharacterized. A long-standing obstacle is the contribution of local RNA secondary structure to the definition of interaction partners in a variety of regulatory contexts, including—but not limited to—transcript stability[3], alternative splicing[4] and localization[3]. There are many documented instances where the presence of a structural regulatory element dictates alternative splicing patterns (for example, human cardiac troponin T) or affects other aspects of RNA biology[5]. Thus, a full characterization of post-transcriptional regulatory programs requires capturing information provided by both local secondary structures and the underlying sequence[3,6]. Here we present a computational framework based on context-free grammars[3,7] and mutual information[2] that systematically explores the immense space of small structural elements and reveals motifs that are significantly informative of genome-wide measurements of RNA behaviour. By applying this framework to genome-wide human mRNA stability data, we reveal eight highly significant elements with substantial structural information, for the strongest of which we show a major role in global mRNA regulation. Through biochemistry, mass spectrometry and *in vivo* binding studies, we identified human HNRPA2B1 (heterogeneous nuclear ribonucleoprotein A2/B1, also known as HNRNPA2B1) as the key regulator that binds this element and stabilizes a large number of its target genes. We created a global post-transcriptional regulatory map based on the identity of the discovered linear and structural *cis*-regulatory elements, their regulatory interactions and their target pathways. This approach could also be used to reveal the structural elements that modulate other aspects of RNA behaviour.

To isolate stability from other aspects of mRNA behaviour, we performed whole-genome mRNA stability measurements by incubating human MDA-MB-231 breast cancer cells in the presence of 4-thiouridine, which is efficiently incorporated into cellular RNA. Subsequently, 4-thiouridine-labelled transcripts were captured and quantified at different time-points after the removal of 4-thiouridine from the growth medium. We calculated a relative decay rate for each transcript based on the rate at which 4-thiouridine-labelled transcripts, in the absence of 4-thiouridine in the media, are replaced by newly synthesized unlabelled mRNAs in the population (Supplementary Fig. 1). These measurements were then used to identify the putative *cis*-regulatory elements (linear and structural) that underlie transcript stability. A number of methods have been previously introduced for discovering structural motifs mainly based on free energy minimization, local sequence alignments or a combination of both alignments and secondary structure predictions[3,6,8]. However, the extent to which these *in silico* predictions reflect stable *in vivo* molecular conformations has not been fully explored[9]. In fact, the RNA binding proteins and complexes that interact with their target transcripts may facilitate the formation of secondary structures *in vivo*. Thus, we sought to bypass the need for predicting thermodynamically stable secondary structures by efficiently enumerating a large space of potential structural motifs. We developed TEISER (Tool for Eliciting Informative Structural Elements in RNA), a framework for identifying the structural motifs that are informative of whole-genome measurements across all the transcripts. In this approach, structural motifs are defined in terms of context-free grammars[7] (CFGs) that represent hairpin structures as well as primary sequence information (see Methods and Supplementary Fig. 2). TEISER employs mutual information to measure the regulatory consequences of the presence or absence of each of roughly 100 million different seed CFGs (see Methods). Mutual information is a robust non-parametric measure that reveals general dependencies across discrete or continuous measurements[2,10]. For example, when applied to the transcript stability data, TEISER captures the dependency between the stability of each mRNA and the presence or absence of a given structural motif in its 5′ and 3′ untranslated regions (UTRs). TEISER, subsequently, uses these measurements to choose and further refine the most informative motifs, and performs a series of statistical tests—for example, randomization-based statistics and jackknifing tests—to achieve very low (<0.01) false-discovery rates (see Methods and Supplementary Fig. 2).

Application of TEISER to the mRNA stability measurements in MDA-MB-231 cells revealed eight strong structural motif predictions that passed our statistical tests aimed at finding the most likely elements causally involved in mRNA stability (Fig. 1 and Supplementary Fig. 3). Apart from being highly informative of mRNA stability measurements, these putative regulatory elements show a variety of other characteristics that support their functionality. For example, four of the discovered motifs are also informative of transcript stability measurements in mouse[11] (Supplementary Fig. 4a). Furthermore, these motifs are highly conserved between human and mouse genomes (see Methods and Supplementary Fig. 3) and are also informative of co-expression clusters discovered across independent whole-genome data sets (Supplementary Fig. 4b).

Among the putative structural motifs discovered by TEISER, we chose sRSM1 (structural RNA stability motif 1)—the most statistically significant 3′ UTR element (z-score = 122)—for further analysis. In order to probe the functionality of sRSM1 instances across the genome, we performed *in vivo* titration experiments using synthetic oligonucleotides[10,12]. Upon transfecting MDA-MB-231 cells with decoy RNA molecules harbouring sRSM1 instances (Supplementary Fig. 5), we observed a notable reduction in the level of endogenous transcripts that carried this motif, in comparison to their level in the control cells transfected with scrambled RNA molecules (Fig. 2). This global

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08540, USA. [2]Department of Molecular Biology, Princeton University, Princeton, New Jersey 08540, USA. [3]Institute of Parasitology, McGill University, Montreal, Quebec H3G1Y6, Canada. [4]McGill Centre for Bioinformatics, McGill University, Montreal, Quebec H3G1Y6, Canada. [5]Laboratory of Systems Cancer Biology, Rockefeller University, New York, New York 10065, USA. [6]Department of Biochemistry, McGill University, Montreal, Quebec H3G1Y6, Canada. †Present addresses: Department of Biochemistry and Molecular Biophysics, and Initiative in Systems Biology, Columbia University, New York, New York 10032, USA (H.G., P.O., S.T.); The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada (H.S.N.).

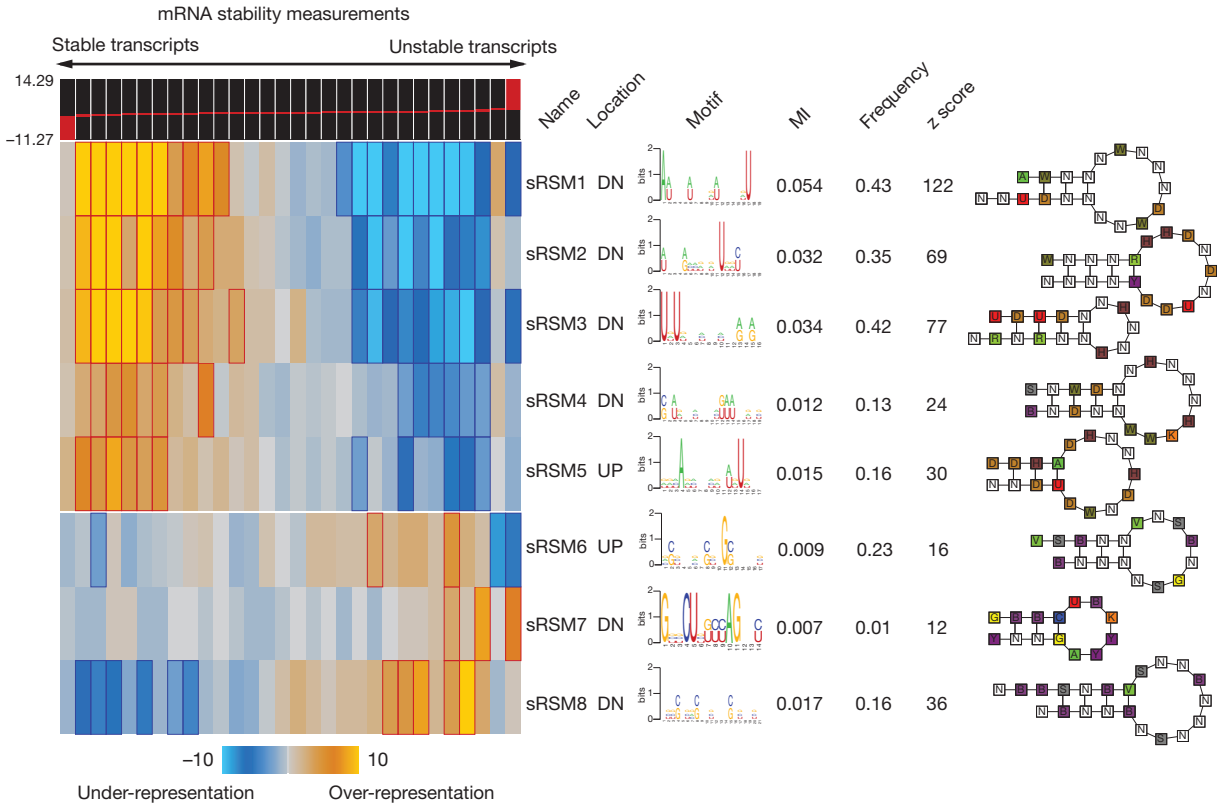

Figure 1 | Discovery of RNA structural motifs informative of genome-wide transcript stability. Each RNA structural motif is shown (far right) along with its pattern of enrichment/depletion across the range of mRNA stability measurements throughout the genome (far left). The panel labelled mRNA stability measurements shows how the transcripts are partitioned into equally populated bins based on their stability measures, going from left (highly stable) to right (unstable). In the heatmap representation, a gold entry marks the enrichment of the given motif in its corresponding stability bin (measured by log-transformed hypergeometric *P*-values), while a light-blue entry indicates motif depletion in the bin. Red and blue borders mark highly significant motif enrichments and depletions, respectively. From left to right, we show the motif names, their location (UP for 5′ UTR and DN for 3′ UTR), their sequence information ('motif', in the form of an alphanumeric plot), their associated mutual information values (MI; see below), their frequency (the fraction of transcripts that carry at least one instance of the motif), and their *z* score (see below). Each MI value is used to calculate a *z* score, which is the number of standard deviations of the actual MI relative to MIs calculated for 1.5 million randomly shuffled stability profiles. A structural illustration of each motif is also presented (far right) using the following single letter nucleotide code: Y = [UC], R = [AG], K = [UG], M = [AC], S = [GC], W = [AU], B = [GUC], D = [GAU], H = [ACU], V = [GCA] and N = any nucleotide.

downregulation points to the presence of a *trans*-acting factor that, upon interaction with sRSM1, stabilizes its target transcripts. The decoy (synthetic) sRSM1 elements compete with endogenous mRNAs for the putative *trans*-acting factor, which results in the observed reduction in the level of its target mRNAs. Furthermore, reporter constructs carrying instances of sRSM1 showed a marked decrease in transcript decay rate in comparison to scrambled controls, further suggesting a direct role for this structural element in transcript stability (Supplementary Fig. 6).

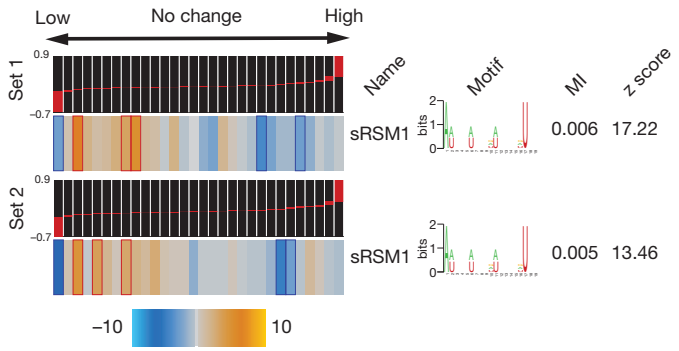

Figure 2 | The regulatory role of sRSM1. Whole-genome expression levels were measured in decoy-transfected samples relative to the controls transfected with scrambled RNA molecules (see Methods). The measurements were performed in duplicate, for two independent decoy/scrambled sets (the relative transcript levels were subsequently averaged across the two replicates in each set). Genes were sorted and quantized into equally populated bins based on the average log-ratio of their expression levels in the decoy samples relative to the scrambled controls. TEISER was used to show the enrichment/depletion patterns of transcripts harbouring sRSM1 in their 3′ UTRs. From left to right, we also show motif name, sequence, MI values and the associated *z* scores.

We used streptomycin-binding RNA aptamer immobilization coupled with mass spectrometry[13] to discover candidates that bind, *in vitro*, to the decoy instances of sRSM1, but not to the scrambled versions (Supplementary Fig. 7). After isolation under stringent conditions and in-solution digestion of RNA-bound proteins followed by nanoliquid chromatography-tandem mass spectrometry, we identified HNRPA2B1 as a promising candidate (Supplementary Table 1). This RNA-binding protein is a member of the A/B subfamily of heterogeneous nuclear ribonucleoproteins (hnRNPs)[14] and carries two repeats of quasi-RNA-recognition motif (qRRM) RNA binding domains (Supplementary Fig. 8). Moreover, the established roles of other members of this family, namely HNRNPD and HNRNA1, in regulating RNA stability[15] and binding terminal stem-loops[16] further suggest HNRPA2B1 as a functional regulator. Also, more than 4,000 transcripts carry potentially functional instances of sRSM1 (see Methods), implicating this motif as a major global regulator of mRNA stability. The HNRPA2B1 transcript, at the same time, is highly abundant in the cell (one standard deviation higher than average[17]), thus making it a promising candidate for global modulation of mRNA stability through sRSM1.

In order to directly assess the regulatory consequences of modulating HNRPA2B1, we performed knock-down experiments followed by gene expression profiling. Consistent with our prior observations, HNRPA2B1 knock-down caused a significant decrease in the expression level of transcripts carrying sRSM1 (Fig. 3a). Stability measurements in the knock-down cells confirmed that the observed downregulation of these transcripts was in fact due to changes in stability (see Methods), with the transcripts carrying sRSM1 elements showing a marked increase in their corresponding relative decay rates (Fig. 3b).

In principle, our observations are consistent with a possible indirect role for HNRPA2B1—brought about, for instance, by a common partner that binds both HNRPA2B1 and sRSM1 sites. The direct interaction between HNRPA2B1 and its potential target genes can be tested through cross-linking and immunoprecipitation of HNRPA2B1, which, through local ultraviolet photoreactivity of bases and amino acids, can detect direct physical interactions[18]. We expressed a tagged clone of HNRPA2B1 in MDA-MB-231 cells, and after ultraviolet-crosslinking, immunoprecipitated this protein and the target mRNA molecules that were bound to it. We then labelled the isolated RNA population and hybridized it to microarrays with the input total RNA as control (a method called RIP-chip[19]). We observed a highly significant enrichment of sRSM1 in the immunoprecipitated population (Fig. 3c). In order to reduce the background and better pinpoint the HNRPA2B1 binding sites, we treated the samples with nuclease before immunoprecipitation under denaturing conditions and sequenced the HNRPA2B1-bound RNA population (HITS-CLIP[20]). We observed that sRSM1 elements were significantly enriched in the identified putative binding sites, in comparison with randomly selected sequences[21] (Fig. 3d). These observations demonstrate that HNRPA2B1 directly

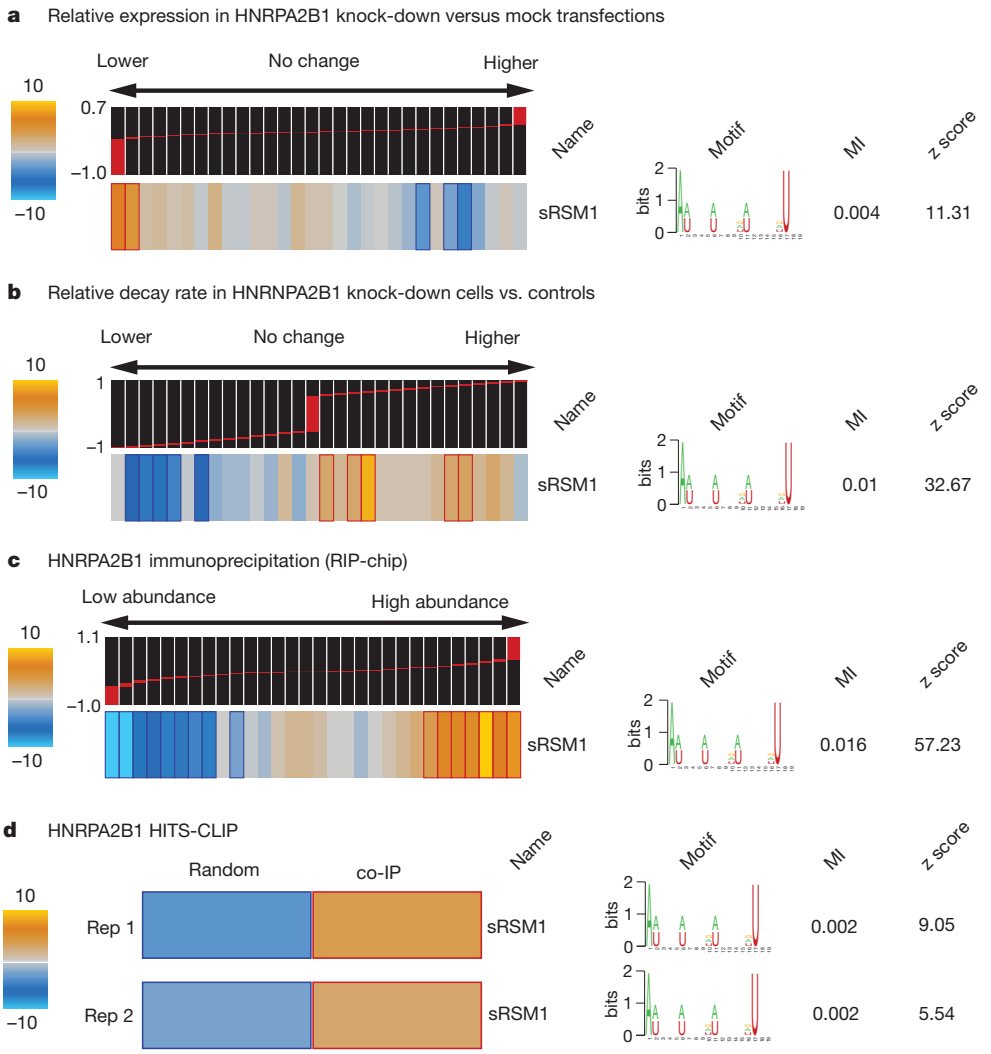

Figure 3 | HNRPA2B1 stabilizes transcripts through direct *in vivo* binding to sRSM1 structural motifs. **a**, Genome-wide expression levels were measured in HNRPA2B1 siRNA-transfected samples relative to mock-transfected controls. TEISER was used to capture the enrichment/depletion pattern of transcripts carrying sRSM1 across the relative expression values. Experiments were performed in triplicate, each with an independent siRNA targeting HNRPA2B1 and the resulting log ratios were averaged for each transcript. **b**, Transcript decay rates were compared in HNRPA2B1 knock-downs versus mock-transfected controls. These measurements were then analysed by TEISER to visualize the extent to which the decay rates of transcripts carrying sRSM1 elements were increased following HNRPA2B1 knock-down. **c**, Using ultraviolet-crosslinking followed by immunoprecipitation, mRNAs that bind HNRPA2B1 were extracted and compared against the input mRNA population (RIP-chip). The log ratio calculated for each mRNA denotes its abundance in the immunoprecipitated sample relative to the input control. Bins to the right contain the mRNAs that were captured as interacting partners with HNRPA2B1. Similar to the prior examples, TEISER was used to show the enrichment/depletion pattern of transcripts carrying sRSM1 in their 3′ UTRs. The values associated with each transcript were calculated as the average of log ratios from biological replicates. **d**, HNRPA2B1 binding sites were identified using immunoprecipitation followed by high-throughput sequencing (HITS-CLIP). Instances of the sRSM1 element are significantly enriched in these sites relative to a population of random sequences from 3′ UTRs that are not represented in the sequenced population.

interacts with sRSM1 *in vivo* and acts to stabilize its target transcripts through this regulatory element. These transcripts, in turn, modulate a variety of cellular processes and pathways. For example, we observed a significant positive correlation between sRSM1 target transcripts and doubling-time in NCI-60 breast cancer cell lines (Fig. 4a). Indeed, knocking-down HNRPA2B1 resulted in a slight but significant increase in growth rate (by 10%, *P*-value $<10^{-8}$), further highlighting the regulatory role of this global modulator in a key cellular process (Fig. 4b).

Revealing the detailed post-transcriptional regulatory code relies on the discovery of all the *cis*-regulatory elements that contribute to changes in transcript abundance. In addition to the sRSMs identified through TEISER, we also discovered a large diverse set of lRSMs (linear RNA stability motifs), including six known microRNA recognition sites, that are informative of transcript stability measurements (Supplementary Fig. 9). These motifs were identified by FIRE[2] (Finding Informative Regulatory Elements), a framework for discovering informative linear motifs. Combining these two approaches provided us with an extensive set of putative regulatory elements that cover both structural and primary sequence components. The next step in deciphering the post-transcriptional regulatory program involves the identification of target pathways that are potentially modulated by each element. Using iPAGE[10] (Pathway Analysis of Gene Expression), we showed that our discovered elements probably target a diverse array of cellular processes and pathways (Supplementary Fig. 10). For example, the sRSM1 structural element is significantly enriched in the 3′ UTRs of the genes involved in 'Notch signalling', while avoiding the UTRs of other pathways such as 'nucleosome assembly' (Supplementary Fig. 11). These results demonstrate that while post-transcriptional regulatory mechanisms are poorly characterized, they have potentially far-reaching impact on specific cellular processes.

Regulatory programs often employ combinatorial interactions between various *cis*-regulatory elements to modulate gene expression[2,22]. We used mutual information to reveal such potential interactions in the post-transcriptional regulatory programs governing mRNA stability (Supplementary Figs 12 and 13). For example, sRSM1 showed significant interactions with a number of structural and linear motifs, including sRSM8 and sRSM3 (Supplementary Fig. 11). These observed interactions might reflect cross-talk, or insulation, between the underlying regulatory processes that act upstream of these elements. The full map of such interactions (Supplementary Figs 14 and 15) reveals a complex network of motif-pathway relationships that set the stage for molecular dissection and predictive modelling of post-transcriptional regulation from sequence.

Whereas we have studied mRNA stability under normal and static conditions in a single cell line, the full regulatory program that governs mRNA stability is likely to involve a much richer repertoire of *cis*-regulatory elements operating within a more complex regulatory network. Also, although we have focused on transcript stability, our framework is general in concept and can be employed to study regulatory programs governing other aspects of RNA biology. For example, the established role of local secondary structures in shaping the splicing code[4,23] suggests alternative splicing as a prominent area for analysis using this framework. The large repertoire of publicly available whole-genome expression data sets similarly offers a rich resource for identifying the post-transcriptional regulatory modules that underlie steady-state measurements.

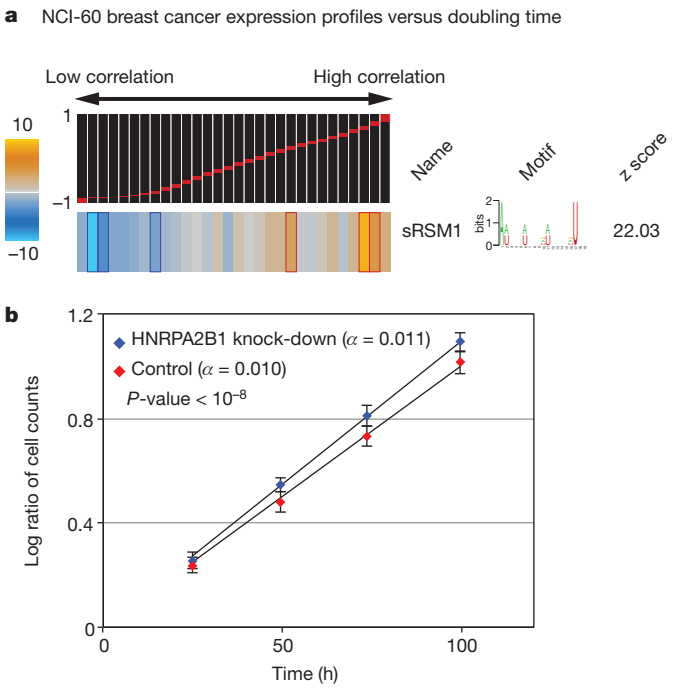

Figure 4 | HNRPA2B1 regulates growth rate. **a**, Whole genome expression levels across five breast cancer cell lines (MCF7, MDA-MB-231, HS578T, BT-549 and T47D) were correlated against their doubling times[17]. The resulting values, ranging from −1 to 1, were analysed by TEISER to probe the enrichment/depletion pattern of transcripts carrying sRSM1. **b**, The growth of HNRPA2B1 siRNA-transfected samples was compared to those of mock-transfected controls. For each time-point, the number of cells in four independent samples was counted in duplicates ($n = 8$), yielding an estimated growth-rate (α). Shown are the average log-ratios, their standard deviation at each time-point, and the statistical significance of the observed difference in growth-rate.

METHODS SUMMARY

TEISER relies on calculating mutual information (MI) values between whole-genome measurements and millions of predefined structural motifs. The statistically significant motifs are then optimized and elongated through a heuristic search algorithm. The mRNA stability measurements were performed using a previously published method[1]. The decoy/scrambled experiments and siRNA knock-downs were performed using lipofectamin 2000 reagent (Invitrogen). For hybridizations, we used human 4 × 44k whole-genome human arrays (Agilent). Isolation and identification of RNA-binding proteins were based on previously published protocols[13,24]. HNRPA2B1 target transcripts were isolated based on the CLIP protocol[18].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 August 2011; accepted 2 March 2012.
Published online 8 April 2012.

1. Dölken, L. *et al.* High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14,** 1959–1972 (2008).
2. Elemento, O., Slonim, N. & Tavazoie, S. A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell* **28,** 337–350 (2007).
3. Rabani, M., Kertesz, M. & Segal, E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl Acad. Sci. USA* **105,** 14885–14890 (2008).
4. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (2010).
5. Wan, Y., Kertesz, M., Spitale, R. C., Segal, E. & Chang, H. Y. Understanding the transcriptome through RNA structure. *Nature Rev. Genet.* **12,** 641–655 (2011).
6. Pavesi, G., Mauri, G., Stefani, M. & Pesole, G. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucleic Acids Res.* **32,** 3258–3269 (2004).
7. Searls, D. B. The language of genes. *Nature* **420,** 211–217 (2002).
8. Hofacker, I. L., Fekete, M. & Stadler, P. F. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319,** 1059–1066 (2002).
9. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).
10. Goodarzi, H., Elemento, O. & Tavazoie, S. Revealing global regulatory perturbations across human cancers. *Mol. Cell* **36,** 900–911 (2009).
11. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
12. Cutroneo, K. R. & Ehrlich, H. Silencing or knocking out eukaryotic gene expression by oligodeoxynucleotide decoys. *Crit. Rev. Eukaryot. Gene Expr.* **16,** 23–30 (2006).

13. Windbichler, N. & Schroeder, R. Isolation of specific RNA-binding proteins using the streptomycin-binding RNA aptamer. *Nature Protocols* **1,** 637–640 (2006).
14. Biamonti, G., Ruggiu, M., Saccone, S., Della Valle, G. & Riva, S. Two homologous genes, originated by duplication, encode the human hnRNP proteins A2 and A1. *Nucleic Acids Res.* **22,** 1996–2002 (1994).
15. Wilusz, C. J., Wormington, M. & Peltz, S. W. The cap-to-tail guide to mRNA turnover. *Nature Rev. Mol. Cell Biol.* **2,** 237–246 (2001).
16. Michlewski, G. & Caceres, J. F. Antagonistic role of hnRNP A1 and KSRP in the regulation of let-7a biogenesis. *Nature Struct. Mol. Biol.* **17,** 1011–1018 (2010).
17. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24,** 227–235 (2000).
18. Jensen, K. B. & Darnell, R. B. CLIP: crosslinking and immunoprecipitation of *in vivo* RNA targets of RNA-binding proteins. *Methods Mol. Biol.* **488,** 85–98 (2008).
19. Keene, J. D., Komisarow, J. M. & Friedersdorf, M. B. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nature Protocols* **1,** 302–307 (2006).
20. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456,** 464–469 (2008).
21. Giannopoulou, E. G. & Elemento, O. An integrated ChIP-seq analysis platform with customizable workflows. *BMC Bioinformatics* **12,** 277–294 (2011).
22. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117,** 185–198 (2004).
23. Yang, Y. *et al.* RNA secondary structure in mutually exclusive splicing. *Nature Struct. Mol. Biol.* **18,** 159–168 (2011).
24. Greco, T. M., Yu, F., Guise, A. J. & Cristea, I. M. Nuclear import of histone deacetylase 5 by requisite nuclear localization signal phosphorylation. *Mol. Cell Proteomics* **10,** M110.004317 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the members of the Tavazoie laboratory for comments on the project and manuscript. We are also grateful to N. Pencheva, B. Tsui, S. Tavazoie and L. Dölken for their intellectual and technical contributions. L.F. was supported by a Ruth L. Kirschstein National Research Service Award (T32-GM066699). S.T. was supported by grants from NHGRI (2R01HG003219) and the NIH Director's Pioneer Award.

Author Contributions H.G., H.S.N. and S.T. conceived and designed the study. H.G. and H.S.N. developed TEISER. R.S. contributed to the execution of the study. H.G., H.S.N., T.M.G., P.O., I.M.C. and S.T. designed the experiments. H.G., P.O., L.F. and T.M.G. performed the experiments. H.G., H.S.N. and T.M.G. analysed the results. H.G., H.S.N. and S.T. wrote the paper.

Author Information The microarray and high-throughput sequencing data are deposited at GEO under the umbrella accession number GSE35800. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.T. (st2744@columbia.edu).

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 (for example, 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 $S \rightarrow SG$ 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 (that is, 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, that is, 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 $\sim$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 (for example, 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 (for example, 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 data sets (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 (for example, 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 data set, 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 25 μM 4-thiouridine (Sigma) for 4 h. Then the cells were washed with fresh media (DMEM + 10% FBS) and incubated for 0, 1, 2 and 4 h. At each time point, cells were washed with cold PBS and RNA extraction was performed using a total RNA purification kit (Norgen Biotek). The 4-thiouridine 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-colour hybridization using a one-colour low-input quick-amp labelling kit (Agilent) and hybridized according to the manufacture's instructions. A one-colour 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. Set 1: AAAACTATTTTGAAGATGGTGGTGAGCTGCAAAATAGCTGGATGGATTTGAATGATTGGGATGATACATCATTGAACTGCACTTTATATAACCAAAGCTTAGCAGTTTGTTAGATAAGAGTCTATGTATGTCTCTGGTTAGGATGAAGTTAATTTTATGTTTTTAACATGGTATTTTTGAAGGAGCTAATGAAACACTGG. Set2: ATTGTTTCTGGAAACTGCTTGCCAAGACAACATTTATTAACTGTTAGAACACTTGCTTTATGTTTGTGTGTACATATTTTCCACAAATGTTATAATTTATATAGTGTGGTTGAACAGGATGCAATCTTTTGTTGTCTAAAGGTGCTGCAGTTAAAAAAAAAACAACCTTTTCTTTCAATATGGCATGTAGTGGAGTTTTT. For the scrambled controls, we used the shuffled version of the putative binding sites (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 Analogue (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 labelled 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 data set as log of ratios, which was then analysed 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 (Set 1 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 5 μg ml^{-1} of α-amanitine (Sigma). Time points were taken at 0, 1.5, 3 and 6 h in duplicates for Flp-In 293 GFP-Decoy and Flp-In 293 GFP-Shuffled cells. Quantitative PCR (Fast SYBR Green 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 Set 1 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 10 μM streptomycin and subjected to in-solution digestion[24,25]. Tryptic peptides were then analysed 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. The ON-Target*plus* (Dharmacon) 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 twofold reduction in expression, log ratio >0.4 and $P < 10^{-7}$) 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 labelled 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 analysed 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 5 μg ml^{-1} of α-amanitine (Sigma). Time points were taken at 0, 1, 2 and 4 h in duplicates for the siRNA-transfected samples and mock-transfected controls. Each sample was then Cy3-labelled 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 ultraviolet-irradiated at 4,000 mJ cm^{-2}. The cells were then collected and lysed with 1 ml M-PER Reagent (Pierce) and 10 μl RNasin (NEB). The samples were subjected to DNase treatment (baseline ZERO DNase) for 15 min at 37 °C. Samples were then centrifuged at 16,000*g* at 4 °C for 20 min 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-labelled using low-input quick-amp labelling kit (Agilent). As control, we used Cy5-labelled RNA samples extracted before 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 ultraviolet-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 (Dharmacon) 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 h 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.

25. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nature Methods* **6,** 359–362 (2009).
26. Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460,** 479–486 (2009).