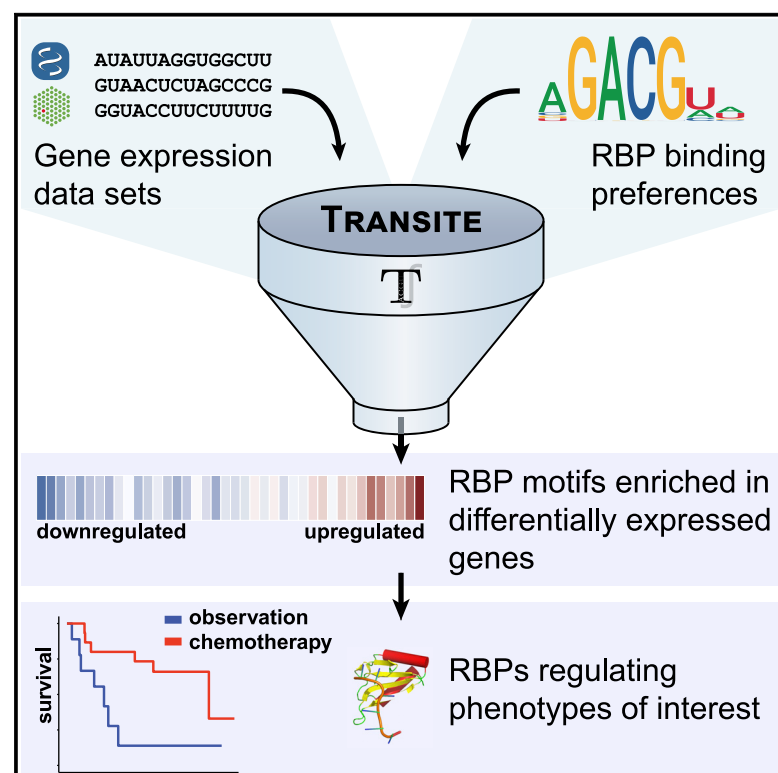**Resource**

Transite: A Computational Motif-Based Analysis Platform That Identifies RNA-Binding Proteins Modulating Changes in Gene Expression

Graphical Abstract

Authors
Konstantin Krismer, Molly A. Bird, Shohreh Varmeh, ..., Yi Wen Kong, Ian G. Cannell, Michael B. Yaffe

Correspondence
ywkong@mit.edu (Y.W.K.),
ian.cannell@cruk.cam.ac.uk (I.G.C.),
myaffe@mit.edu (M.B.Y.)

In Brief
Krismer et al. present a computational approach to identify RNA-binding proteins (RBPs) that modulate post-transcriptional control of gene expression using RNA expression data as inputs. By applying this approach to publicly available patient datasets, they identify and experimentally confirm that the RBP hnRNPC contributes to chemotherapy resistance in lung cancer.

Highlights

- Transite nominates RBPs that influence gene expression from RNA expression datasets

- Transite identifies enriched and depleted *k*-mers in differentially expressed genes

- Transite identifies hnRNPC as a modulator of cisplatin resistance in human lung cancer

- Transite is available online at https://transite.mit.edu and as an R package

CellPress

Resource

Transite: A Computational Motif-Based Analysis Platform That Identifies RNA-Binding Proteins Modulating Changes in Gene Expression

Konstantin Krismer,[1,2,3,5,7] Molly A. Bird,[2,3,5,11] Shohreh Varmeh,[2,3,11] Erika D. Handly,[2,3,5,11] Anna Gattinger,[3,7] Thomas Bernwinkler,[2,3,7] Daniel A. Anderson,[4,5] Andreas Heinzel,[7] Brian A. Joughin,[2,3,5] Yi Wen Kong,[2,3,*] Ian G. Cannell,[2,3,8,*] and Michael B. Yaffe[2,3,5,6,9,10,12,*]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA
[2]Center for Precision Cancer Medicine, Massachusetts Institute of Technology, 500 Main Street, Cambridge, MA 02142, USA
[3]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 500 Main Street, Cambridge, MA 02142, USA
[4]Synthetic Biology Center, Massachusetts Institute of Technology, 500 Technology Square, Cambridge, MA 02139, USA
[5]Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[6]Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA
[7]Department for Medical and Bioinformatics, University of Applied Sciences Upper Austria, Softwarepark 11, 4232 Hagenberg, Austria
[8]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK
[9]Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[10]Divisions of Acute Care Surgery, Trauma and Surgical Critical Care, and Surgical Oncology, Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Boston, MA 02215, USA
[11]These authors contributed equally
[12]Lead Contact
*Correspondence: ywkong@mit.edu (Y.W.K.), ian.cannell@cruk.cam.ac.uk (I.G.C.), myaffe@mit.edu (M.B.Y.)
https://doi.org/10.1016/j.celrep.2020.108064

SUMMARY

RNA-binding proteins (RBPs) play critical roles in regulating gene expression by modulating splicing, RNA stability, and protein translation. Stimulus-induced alterations in RBP function contribute to global changes in gene expression, but identifying which RBPs are responsible for the observed changes remains an unmet need. Here, we present Transite, a computational approach that systematically infers RBPs influencing gene expression through changes in RNA stability and degradation. As a proof of principle, we apply Transite to RNA expression data from human patients with non-small-cell lung cancer whose tumors were sampled at diagnosis or after recurrence following treatment with platinum-based chemotherapy. Transite implicates known RBP regulators of the DNA damage response and identifies hnRNPC as a new modulator of chemotherapeutic resistance, which we subsequently validated experimentally. Transite serves as a framework for the identification of RBPs that drive cell-state transitions and adds additional value to the vast collection of publicly available gene expression datasets.

INTRODUCTION

RNA-binding proteins (RBPs) are major modulators of gene expression at the post-transcriptional level, where they control RNA splicing, stability, localization, degradation, and translation (Gerstberger et al., 2014; Lunde et al., 2007). RBPs play critical roles in cell differentiation and tissue development, and aberrant RBP function is implicated in a wide range of diseases, including neurodegenerative disorders and neuropathies, myopathies, autoimmune paraneoplastic syndromes, and cancer (Lukong et al., 2008). For mRNAs, the role of RBPs in modulating global changes in gene expression at both the RNA and the protein level becomes particularly important under conditions in which new gene transcription is repressed, such as during inflammation, cell stress, and in response to genomic damage (Stumpo et al., 2010; Sugiura et al., 2011; Pereira et al., 2017). Under these conditions, changes in gene expression have been shown to result, in part, from alterations in RBP activity (Perron et al., 2018). Furthermore, mutations affecting the expression or function of specific RBPs have been implicated in a variety of diseases, including cancer (Cooper et al., 2009; Licatalosi and Darnell, 2010; Lukong et al., 2008; Pereira et al., 2017).

A large subset of RBPs recognize short linear sequence motifs of 6–8 nt within their target RNAs. Other RBPs recognize specific structural features in their bound RNA targets, or bind to RNA promiscuously without clear specificity. However, for those RBPs that recognize sequence-specific motifs, the presence of the motifs in the RNAs and sequence-specific binding by the RBP, as defined by cross-linking immunoprecipitation (CLIP), are highly correlated (Van Nostrand et al., 2018).

This sequence-specific recognition affords the opportunity for computational approaches to infer the activity of this class of RBPs using the sequence enrichment of differentially expressed genes (Coppin et al., 2018). The identity of these motifs has been determined for a subset of all known RBPs using various *in vitro*-based oligonucleotide selection methods such as SELEX (Tuerk and Gold, 1990), RNAcompete (Ray et al., 2009), and RNA Bind-n-Seq (Zykovich et al., 2009), and directly confirmed for a smaller set of RBPs through the experimental analysis of RBP-RNA interactions using CLIP sequencing (CLIP-seq) and various extensions thereof. The RNA targets for most RBPs, as determined by CLIP-seq, however, have not been identified due to a variety of technical challenges, including cost, limited antibody specificity, and high background binding. Furthermore, the direct experimental identification of RNA targets of RBPs likely depends on the experimental situation under which the CLIP-seq was performed. This lack of direct CLIP-seq data has limited our ability to directly map specific RBPs onto global changes in RNA levels, including those in patient-based gene expression datasets, that have been observed following various stimuli or clinical treatments. However, it is worth noting that not all RBPs recognize linear motifs; some notable examples (RNA-induced silencing complex [RISC]-loading complex subunit TARBP2) recognize structured elements, whereas others show little to no sequence specificity at all.

RBPs appear to play a particularly important role in orchestrating the DNA damage response (DDR) by regulating mRNA expression changes that control the onset and duration of cell-cycle checkpoints and drive DNA repair (Reinhardt et al., 2011; Rieger and Chu, 2004; Gasch et al., 2001). Unbiased large-scale screening efforts have converged on RBPs as one of the most enriched classes of proteins modulating the DDR, even more so than annotated DNA damage repair proteins (Matsuoka et al., 2007; Paulsen et al., 2009; Hurov et al., 2010; Floyd et al., 2013; Adamson et al., 2012). In addition, emerging evidence from a number of labs has identified RBPs as critical targets of DDR kinases, including both upstream sensor kinases such as ATM (ataxia telangiectasia mutated), ATR (ataxia telangiectasia and Rad3-related protein), and DNA-PK, and downstream effector kinases such as Chk1 (checkpoint kinase-1) and mitogen-activated protein (MAP) kinase-activated protein kinase-2 (MK2) (Matsuoka et al., 2007; Paulsen et al., 2009; Wilker et al., 2007; Fan et al., 2002; Kim et al., 2010). The discovery of RBPs as integration points of the cellular response to genomic damage has important clinical applications, since the efficacy of many commonly used chemotherapeutic drugs is dependent on the integrity (or lack thereof) of the DDR (Ciccia and Elledge, 2010; Jackson and Bartek, 2009). For example, we found that a key target of the DNA damage-activated MK2 pathway was the RBP hnRNPA0 (heterogeneous nuclear ribonucleoprotein A0), which was required for the maintenance of the G1/S and G2/M checkpoints following cisplatin treatment (Cannell et al., 2015; Reinhardt et al., 2010). Furthermore, this finding has clear clinical relevance for the response of non-small-cell lung cancers (NSCLCs) to chemotherapy in both mouse models and human patients, in which the expression levels of 2 critical hnRNPA0 target RNAs, Gadd45α (growth arrest and DNA dam-

age-inducible protein α [GADD45α]) and p27 (cyclin-dependent kinase inhibitor 1B), predicted the clinical response of mouse and human tumors to platinum-based therapy. Despite these types of data and the recent surge of interest in the roles of RBPs in cancer chemosensitivity and resistance (Hong, 2017; Pereira et al., 2017; Reinhardt et al., 2011), general methods for the systematic identification and prioritization of RBPs that influence various biological responses, including the DDR in clinically relevant patient-based gene expression datasets, are lacking.

To address this, we developed a computational approach called *Transite* that leverages preexisting gene expression data and known RBP-binding preferences to infer RBPs that may be responsible for alterations in RNA levels under a given condition or perturbation. This approach is analogous to our previous computational tool *Scansite*, which predicts the substrates of kinases and modular signaling domains based on phosphorylation and peptide-binding motifs (Obenauer et al., 2003). With *Transite*, we hope to expand the utility of RBP biology to the wider scientific community.

RESULTS

The overall approach used by *Transite* to map RBPs to sets of differentially expressed genes is illustrated in Figure 1. *Transite* starts with a list of differentially expressed genes between two conditions (i.e., treated versus untreated samples), identifies short linear oligonucleotide motifs or *k*-mers that are enriched or depleted within specific regions of the transcripts they encode (i.e., 5′ UTR, CDS, or 3′ UTR), and then matches these motifs to likely RBPs that bind them using a compendium of known RBP motifs (see STAR Methods). *Transite*'s default setting is to analyze 3′ UTR sequences, since motifs that regulate mRNA stability typically reside within the 3′ UTR, but also allow the same analysis to be performed on the CDS or the 5′ UTR. Two different approaches are used, depending on whether the set of differentially expressed genes is separated into distinct foreground and background sets or instead is analyzed as a continuous list of genes ordered by change in expression level. For the former approach, in which foreground sets are predetermined by differentially expressed genes, we developed transcript set motif analysis (TSMA), which looks for enriched or depleted oligonucleotide motifs based on systematic differences between the foreground sets and the total gene expression data (i.e., the background). For the latter approach (i.e., a list of ranked genes) we developed spectrum motif analysis (SPMA), which analyzes motif enrichment along that ordered list of transcripts, similar to the approach taken by gene set enrichment analysis (GSEA) (Subramanian et al., 2005). This approach exploits information across the entire spectrum of changes rather than limiting analysis to the up- and downregulated extremes, and allows motif enrichment or depletion to be visually displayed as a color spectrum. Both TSMA and SPMA then use two distinct methods, a *k*-mer-based and a matrix-based method, to score for and infer candidate RBP in the differentially expressed genes. The *k*-mer-based and matrix-based implementations of TSMA and SPMA are explained in more detail below.
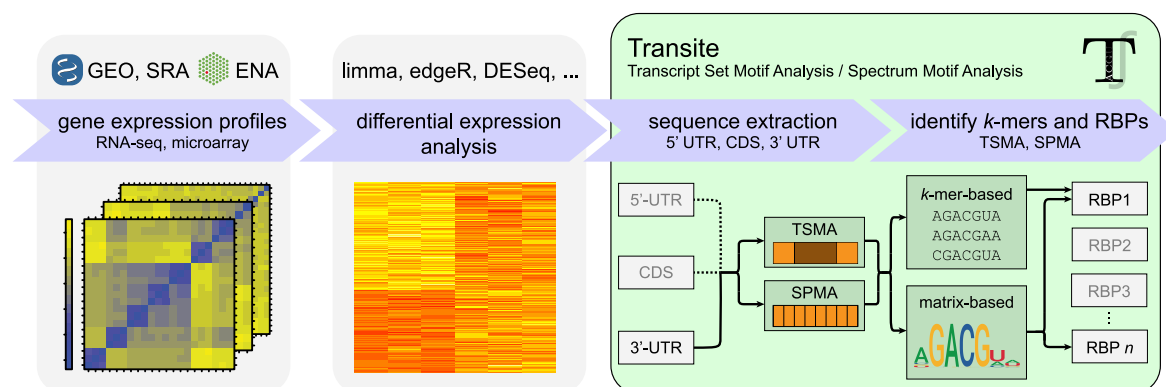
Figure 1. Schematic Figure of the Transite Analysis Pipeline
The initial steps of the Transite data analysis workflow include preprocessing and differential expression analysis of gene expression profiles, which could be collected in-house or obtained from NCBI and EMBL-EBI repositories such as GEO, SRA, and ENA. Differential expression analysis is used to either identify groups of upregulated and downregulated genes (for transcript set motif analysis [TSMA]) or to establish a ranked list of genes from most upregulated to most downregulated (for spectrum motif analysis [SPMA]). Transite then analyzes regions within these genes to identify k-mers and RBPs whose motifs are enriched or depleted in the differentially expressed genes.

TSMA Identifies Enriched and Depleted k-mers within Assigned Sets of Upregulated and Downregulated Genes and Maps Them onto RBPs

TSMA identifies the overrepresentation or underrepresentation of all of the possible hexamers or heptamers, as well as binding motifs for 174 well-characterized RBPs in a set (or sets) of transcripts (i.e., a foreground set), relative to the background of the entire population of transcripts measured in an experiment (Figure 2A).

Two different methods are used to assign transcript targets to specific RBPs. One of the methods, k-mer-based TSMA, also identifies statistically significantly overrepresented and underrepresented hexamers or heptamers within the foreground set, irrespective of whether they can be associated with a known RBP motif. Matrix-based TSMA leverages the full position weight matrix (PWM) representations (see STAR Methods for details) of known RBP motifs to nominate RBPs whose motifs are overrepresented or underrepresented in the foreground set.

In the k-mer-based approach, after foreground and background sets are defined (Figure 2A) and the preferred sequence region is selected (5′ UTR, CDS, or 3′ UTR), the sequences of both sets are broken down into overlapping hexamers or heptamers (i.e., k-mers of length 6 or 7, respectively) (Figure 2B, left column, step 1), and for each k-mer, its frequencies in the foreground and background set are determined. While *Transite* supports both hexamer and heptamer matching, hexamers are recommended, since computer runtime increases exponentially with k, and the results for heptamers mirror those for hexamers in our experience.

The enrichment value of a particular k-mer i, e_i, is then calculated as follows:

$$e_i = \frac{f_i/n_F}{b_i/n_B},$$

where f_i and b_i are the absolute counts of k-mer i in the foreground and background set and n_F and n_B are the total counts of all k-mers in the foreground and background, respectively.

The statistical significance of the enrichment for each k-mer is then determined. First, a contingency table C_i for k-mer i is defined as

$$C_i = \begin{pmatrix} f_i & (n_F - f_i) \\ b_i & (n_B - b_i) \end{pmatrix}.$$

Second, the p value p_i for C_i is approximated with Pearson's χ^2 test. If $p_i < 5\alpha$, where α is the decision boundary, p_i is replaced by the p value obtained by Fisher's exact test for C_i. This stepwise procedure reduces the computation time dramatically (~50-fold), because the computationally expensive Fisher's exact test is used only in cases in which the approximate p value from the computationally inexpensive χ^2 test is close to the decision boundary and is avoided in cases in which a precise p value is unnecessary. Furthermore, Fisher's exact test is always used if at least one of the expected counts is <5, because this constitutes a violation of the assumptions of the approximate test. The p values are subsequently adjusted for multiple hypothesis testing. Available p value adjustment methods in Transite include Holm's method (Holm, 1979), Hochberg's method (Hochberg, 1988), Bonferroni's method (Dunn and Dunn, 1961), Benjamini and Hochberg's method (Benjamini and Hochberg, 1995), and Benjamini and Yekutieli's method (Benjamini and Yekutieli, 2001).

The list of all of the k-mers with their associated enrichment values and statistical significance in the foreground sets is then reported. This is particularly important because it provides an unbiased way to identify overrepresented and underrepresented sequences and novel motifs regardless of whether they conform to known RBP-binding motifs. Although we think that the most likely explanation for RBP motif enrichment in 3′ UTRs of differentially expressed mRNAs is related to changes in mRNA stability, when analyzing preexisting datasets, it is important to acknowledge the possibility that these enrichments could also reflect the transcriptional effects of DNA-binding proteins' binding to these motifs in the corresponding DNA. The results are

A **Transcript Set Motif Analysis: Foreground and background sets**
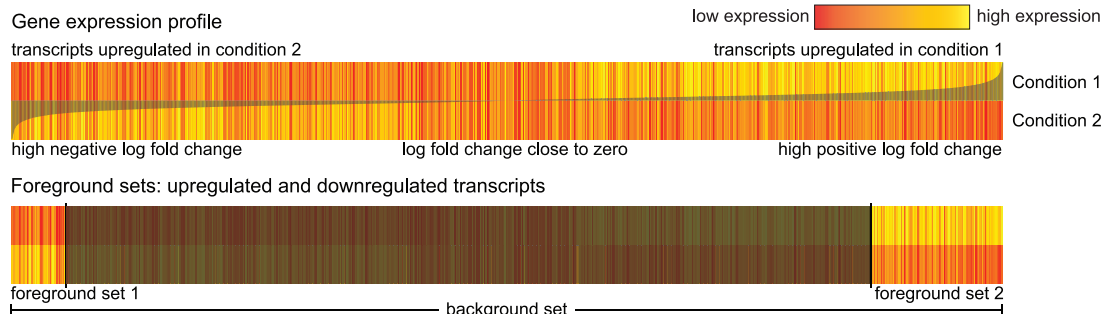
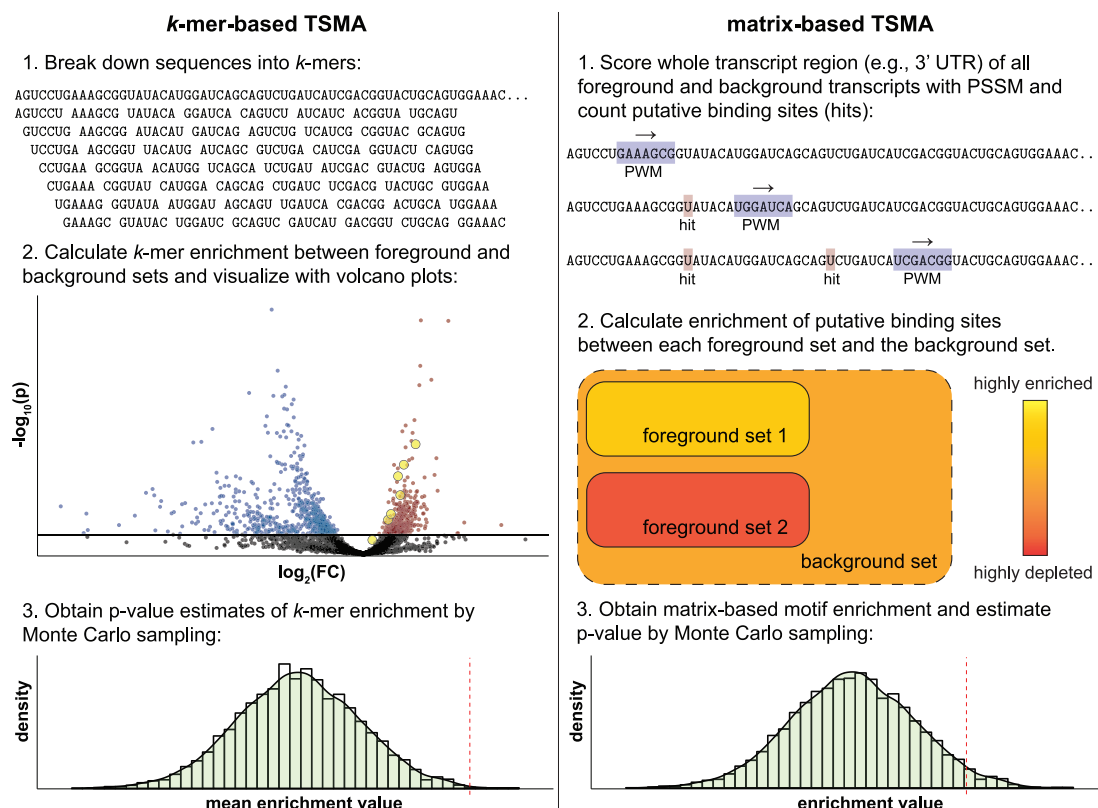**B** **TSMA: Motif enrichment analysis**

Figure 2. TSMA
(A) Foreground sets in TSMA are defined by differential gene expression analysis of RNA-seq or microarray datasets, usually by selecting statistically significantly upregulated and downregulated genes. The background set is all of the genes in the microarray platform or all of the measured genes in RNA-seq. In the heatmap of the gene expression profile, the two rows (condition 1, condition 2) are the mean gene expression values of the replicates of the respective groups (e.g., condition 1 could be treated with drug A and condition 2 untreated). The columns of the heatmap correspond to the genes, and the superimposed gray curve is the log fold change between condition 1 and condition 2.
(B) TSMA estimates the statistical significance of putative RBP-binding site enrichment between each foreground set and the background set. There are 2 ways to describe the putative binding sites of RNA-binding proteins (i.e., the motif). The column on the left depicts *k*-mer-based TSMA, which uses a list of *k*-mers to describe putative binding sites. The column on the right is matrix-based TSMA, which instead uses position weight matrices (PWMs). See text for details.

visualized using volcano plots that show the enrichment values on the x coordinate (log transformed) and the associated p values on the y coordinate (log transformed and multiplied by −1) for all *k*-mers. An example is shown in Figure 2B, in which the black dots represent *k*-mers without significant enrichment or depletion, while blue dots denote significantly depleted

k-mers and red dots significantly enriched *k*-mers. The *k*-mers corresponding to the motif of one particular RBP are indicated by yellow circles.

Over- and underrepresented *k*-mers are then mapped onto specific RBPs. A set of *k*-mers associated with each RBP is generated from the known RBP motif PWMs, as described in

STAR Methods. These RBP-specific k-mers are then assigned the enrichment values calculated from the data, as shown by the yellow dots in the volcano plot in Figure 2B. The geometric mean of the enrichment values of all k-mers that are associated with that particular RBP is then calculated and analyzed for its statistical significance using Monte Carlo sampling. A null distribution of mean enrichment values associated with the k-mers of an RBP is generated by repeated random selection of foreground sets from the background. The null distribution is used to obtain an estimate of the significance of the true mean enrichment value of the RBP-associated k-mers observed in the experimental data, which is shown as a red dashed line in the histogram in Figure 2B, step 3. A ranked list of RBPs and their associated p values, corrected for multiple hypothesis testing, is then provided.

An alternative to k-mer-based TSMA is a matrix-based approach, in which the sequence motifs of 174 RBPs are maintained as PWMs. All of the sequence positions in the transcripts within the foreground and background gene sets are then scored, as shown in step 1 of the right column of Figure 2B. The PWM slides along the sequence and assigns a score to each position; scores above a certain threshold are considered putative binding sites (hits) (see STAR Methods). These hits are tallied in both the foreground and the background set, and enrichment values and associated p values calculated analogously to the k-mer-based approach. Again, all of the p values are multiple testing corrected.

One disadvantage of the matrix-based TSMA method relative to the k-mer-based approach is that a PWM assumes independence among positions, making it impossible to construct a PWM that assigns high scores to AAAAAA and CCCCCC but a low score to ACACAC. An advantage of our matrix-based approach, however, is that it retains positional hit information within the sequence and therefore facilitates the detection of closely spaced clusters of putative binding sites. Homotypic clusters of binding sites on DNA, for example, have been shown to be important for transcription factor binding (Gotea et al., 2010) and have been postulated to be involved in RNA regulation (Plass et al., 2017; Mukherjee et al., 2011), but a clear experimental demonstration of their general importance for RBP binding to RNA has not been unambiguously shown.

SPMA Identifies RBPs with Non-random Arrangement of Putative Binding Sites in a Ranked List of Transcripts

A limitation of the TSMA method described above is that it will only capture those RBPs for which putative binding sites are statistically significantly enriched among a pre-defined foreground set of differentially expressed genes relevant to a background set. As an alternative method, we developed SPMA, an approach that more broadly and generally identifies non-random distributions of RBP target sites in an ordered list of genes without having to pre-define a specific foreground set (compare Figures 2A and 3A).

Instead of using an arbitrary threshold (e.g., p $\leq$ 0.05) to assign transcripts to a single foreground set, SPMA subdivides the entire list of rank-ordered transcripts into a number of bins of equal width. Each bin is considered its own foreground set, and enrichment scores for k-mers or PWM motifs are then calcu-
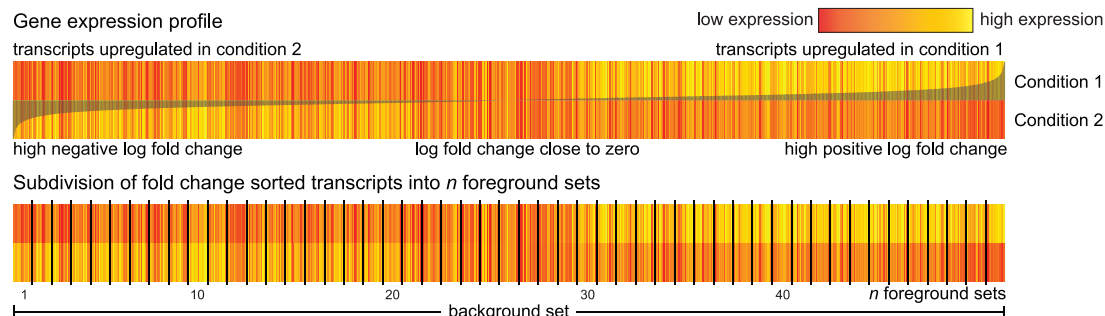
lated as described above. The enrichment scores for each RBP across the bins are then visualized as one-dimensional heatmaps, where red-blue coloring encodes the putative binding site enrichment values, as shown in Figure 3B, to generate spectrum plots. RBPs that are involved in regulating differential gene expression should show non-random red-blue color patterns in the spectrum plot, indicating progressive RBP-binding motif enrichment in the upregulated genes, the downregulated genes, or both. As shown in the upper left plot of Figure 3C, genes that are upregulated in condition 1 show a progressive overrepresentation of putative binding sites for a particular RBP, consistent with that RBP enhancing mRNA stability. In contrast, as shown in the upper right plot of the same panel, genes that are downregulated in condition 1 show a progressive overrepresentation of binding sites for a different RBP, consistent with this RBP destabilizing its mRNA targets.

SPMA generates one spectrum plot for each RBP motif in the motif database. With 174 motifs available, it is imperative to provide an analytical means to aid in the identification of biologically meaningful spectrum plots that exhibit non-random patterns. Each spectrum plot is therefore examined for whether the distribution of enrichment values among the bins is non-random or random, based on three criteria: (1) the adjusted R^2 of a polynomial model fit, (2) the local consistency score, and (3) the number of bins with a significant enrichment or depletion of putative binding sites. The significance of the enrichment values is calculated in an identical fashion to the significance calculation in TSMA. For the first approach, polynomial regression models of degrees ranging from 0 through 5 are fitted to the spectrum of enrichment values, and the model that best reflects the true nature of the data is selected by means of the F test (see STAR Methods for details on the polynomial model approach). Examining the coefficient of the linear term in the polynomial depicts the general increase or decrease in RBP enrichment along the bins, as illustrated in the first two examples of Figure 3C, respectively. If there is strong evidence for a non-linear relationship, then this can also be captured by the model, as seen in the third example shown in the lower left panel of Figure 3C. With the second approach, a local consistency score quantifies the local noise of the spectrum by calculating the deviance between the linear interpolation of the scores of two bins separated by exactly one other, and the observed score of the center bin, for each position in the spectrum. The lower the score, the more consistent the trend in the spectrum plot (see STAR Methods for a formal definition of the local consistency score). Spectrum plots are classified as non-random if (1) the adjusted R^2 of the polynomial fit is ≥ 0.4 and (2) the p value associated with the local consistency score is $\leq 5 \times 10^{-6}$, and (3) at least 10% of the bins have significant ($\alpha = 0.05$) enrichment or depletion of putative binding sites.
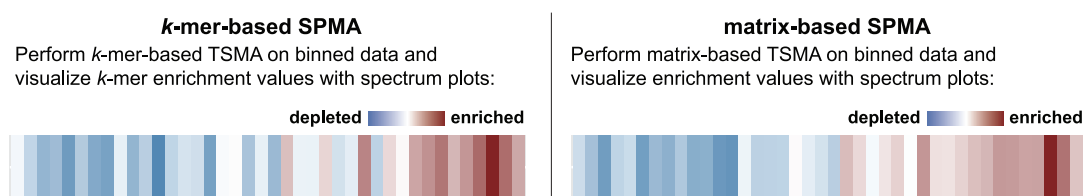
Website and R Package for Transite Available for Customizable Use

To make the RBP analysis of gene expression datasets widely available to the scientific community, the *Transite* analysis platform is hosted at https://transite.mit.edu. Both the TSMA and SPMA methods are web accessible, and familiarity with the R programming language is not required (Figure 4). The full

A   Spectrum Motif Analysis: Ranked list of genes defines foreground and backgrounds sets

B   SPMA: Motif enrichment analysis

***k*-mer-based SPMA**
Perform *k*-mer-based TSMA on binned data and
visualize *k*-mer enrichment values with spectrum plots:

matrix-based SPMA
Perform matrix-based TSMA on binned data and
visualize enrichment values with spectrum plots:

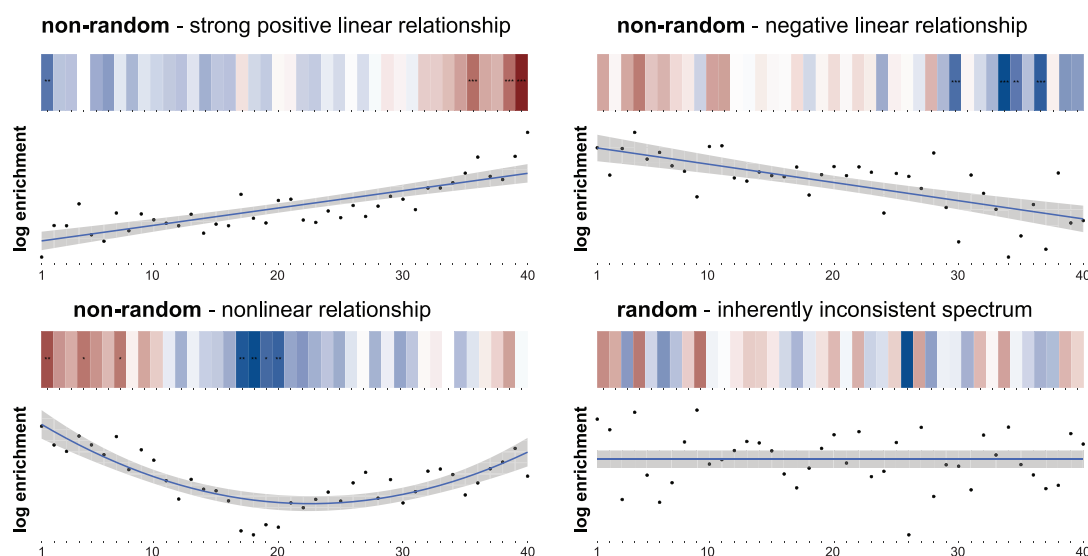**C   SPMA: Classification of distribution of putative binding sites**

Figure 3. SPMA
(A) Transcripts are sorted by some measure of differential expression (e.g., fold change or signal-to-noise ratio) and the entire spectrum of transcripts is then subdivided into a number of equally sized foreground bins.
(B) The motif enrichment step is identical to TSMA. SPMA results are visualized as spectrum plots, which are 1-dimensional heatmaps of motif enrichment values, in which the columns correspond to the bins and the color encodes the enrichment value (strong depletion in dark blue to strong enrichment in dark red) of a particular *k*-mer or PWM.
(C) The distribution of putative binding sites (as visualized by spectrum plots) is deemed random or non-random (i.e., putative binding sites are distributed in a way that suggest biological relevance), based on multiple criteria described in the text. Shown beneath each strip in the heatmap are the log enrichment values for the RBP motif being analyzed (black dots) and the best first, second, or zero order polynomial fit (blue line), along with 95% confidence intervals (shaded gray).

functionality of Transite is also provided as an R/Bioconductor package (https://doi.org/10.18129/B9.bioc.transite) to facilitate a seamless integration of these algorithms into existing bioinformatics workflows. The source code of the *Transite* package is hosted on GitHub (https://github.com/kkrismer/transite). Both

the website and the R package also allow motif enrichment analysis with user-defined motifs, in addition to the 174 motifs provided by the *Transite* motif database, enabling users to search for enrichment of any RBP motif in a discrete set of genes or a rank-ordered list.

Figure 4. *Transite* Web Interface
Datasets are analyzed using TSMA or SPMA in 4 simple steps, some of which are illustrated in (A)–(D). These involve the selection of *k*-mer or matrix-based analysis (A), the specification of foreground and background sets for TSMA, the number of bins for SPMA (B), the region of the RNA to be analyzed and the threshold for statistical significance (C), and the source of RNA-binding motifs to be used for the analysis (D).

Transite Correctly Maps Observed Changes in RNA Abundance following ZFP36 Overexpression or ELAVL1 Knockdown onto Their Respective RBPs

To test the ability of the *Transite* algorithms to correctly map changes in RNA expression onto specific RBPs, we used a publicly available dataset in which RNA expression levels were measured following overexpression of the RBP mRNA decay activator protein ZFP36 (also known as TTP). ZFP36 is known to destabilize its target RNA transcripts by binding to sequence elements in the 3′ UTR (Lai et al., 2003). Mukherjee et al., 2014 reported microarray measurements of differential RNA expression in HEK293 cells following the inducible overexpression of an EGFP-ZFP36 fusion protein (GEO: GSE53185). The RNA expression fold change and associated p values per gene between the induced and un-induced groups, as reported by the authors, were used as input for *Transite*. Genes that were statistically significantly downregulated and upregulated following ZFP36 overexpression (i.e., p < 0.05 after multiple testing corrections) were chosen as foreground sets for TSMA. Volcano plots showing *k*-mer enrichment and depletion in these gene sets are shown in Figure 5A, and the top 10 empirically identified *k*-mers are listed in Tables S1 and S2. The left panel in Figure 5A shows that *k*-mers corresponding to the ZFP36-binding motif, shown in yellow, are among the most highly enriched *k*-mers in transcripts that were found to be downregulated, while the right panel shows conversely that ZFP36-associated *k*-mers were highly depleted in the genes that were upregulated after ZFP36 overexpression. This was even more apparent in the spectrum

plot following SPMA of this dataset (Figure 5B), which revealed a highly consistent nearly monotonic increase in ZFP36-binding sites when the genes were ranked from those most upregulated to those most downregulated after ZFP36 overexpression. On this basis, ZFP36 emerged as the single best RBP out of all 174 RBPs in the database whose motif could rationalize the observed gene expression changes.

To further validate the utility of *Transite* to infer RBPs that modulate gene expression changes, we used a second publicly available dataset (GEO: GSE29778), in which gene expression changes were measured following small interfering RNA (siRNA) knockdown of ELAV-like protein 1 (ELAVL1, also known as HuR) to 20% of its endogenous levels (Mukherjee et al., 2011). ELAVL1 stabilizes its target RNA transcripts and likely facilitates their pre-mRNA processing; hence, its knockdown should result in the reduced expression of its target RNAs. In addition to RNA stability, ELAVL1 is known to play a role in the regulation of alternative splicing of some of its targets (Chang et al., 2014; Akaike et al., 2014). As shown in Figure 5C, analysis of this dataset using SPMA resulted in spectrum plots in which the enrichment values for the ELAVL1 motifs closely varied in direct proportion to the extent of RNA downregulation that was observed (Figures 5C, S1, and S2). Figure 5D shows the top 5 RBP motifs that were enriched in the upregulated and downregulated genes, revealing that genes downregulated after ELAVL1 knockdown were enriched in U-rich RBP motifs, including those that correspond to the ELAVL1 motifs in the *Transite* motif database. In contrast, genes that were upregulated after ELAVL1 knockdown were
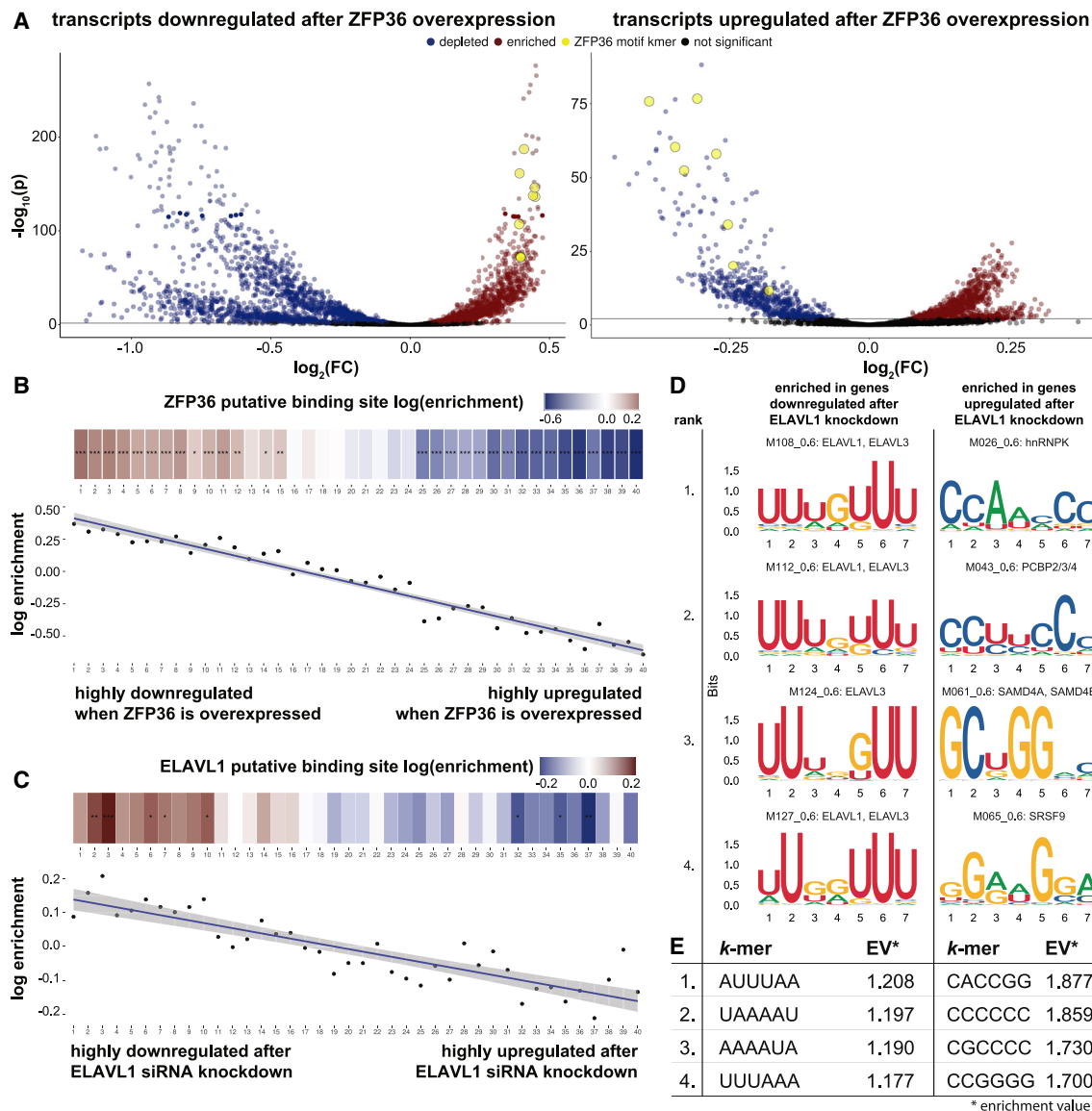
Figure 5. Unbiased Identification of Drivers of Differential Expression after the Overexpression of ZFP36 or the Knockdown of ELAVL1

(A) TSMA volcano plot showing enriched and depleted *k*-mers in downregulated transcripts after ZFP36 overexpression (left panel). *k*-mers associated with ZFP36 (shown in yellow) are highly enriched. A TSMA volcano plot of *k*-mer enrichment values in upregulated transcripts after ZFP36 overexpression shows strong depletion of ZFP36 associated *k*-mers (right panel).

(B) SPMA spectrum plot depicts the relationship between ZFP36 overexpression and downregulation of ZFP36 targets.

(C) SPMA spectrum plot of 1 ELAVL1 motif depicting the global downregulation of ELAVL1 target transcripts after ELAVL1 siRNA knockdown. When comparing spectrum plots, please note that the color scales are adjusted for each plot individually and as a result are different between plots with different enrichment score ranges. Volcano plots of ELAVL1 *k*-mers before and after knockdown are shown in Figure S1.

(D) Sequence logos of motifs highly enriched in transcripts upregulated (left column) and downregulated (right column) after ELAVL1 knockdown. U-rich ELAVL1 motifs are highly enriched in the 3′ UTRs of downregulated transcripts (GEO: GSE29778).

(E) The 4 most highly enriched hexamers in transcripts upregulated (left column) and downregulated (right column) after ELAVL1 knockdown, as identified by *k*-mer-based TSMA.

enriched in alternative RBP motifs that lacked U-rich regions and corresponded to the binding motifs of other RBPs. Furthermore, the single most highly enriched *k*-mer in the set of downregulated genes, AUUUAA, which was empirically identified by *k*-mer-based TSMA (Figure 5E; Tables S3 and S4), perfectly matches the motif of ELAVL1 that was experimentally determined using photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) and RNP immunoprecipitation-microarray (RIP-ChIP) (Mukherjee et al., 2011). These data indicate that *Transite* can capture the specific RBPs responsible for the gene expression changes caused by the manipulation of RBP levels, thus validating our approach and providing confidence that predictions derived
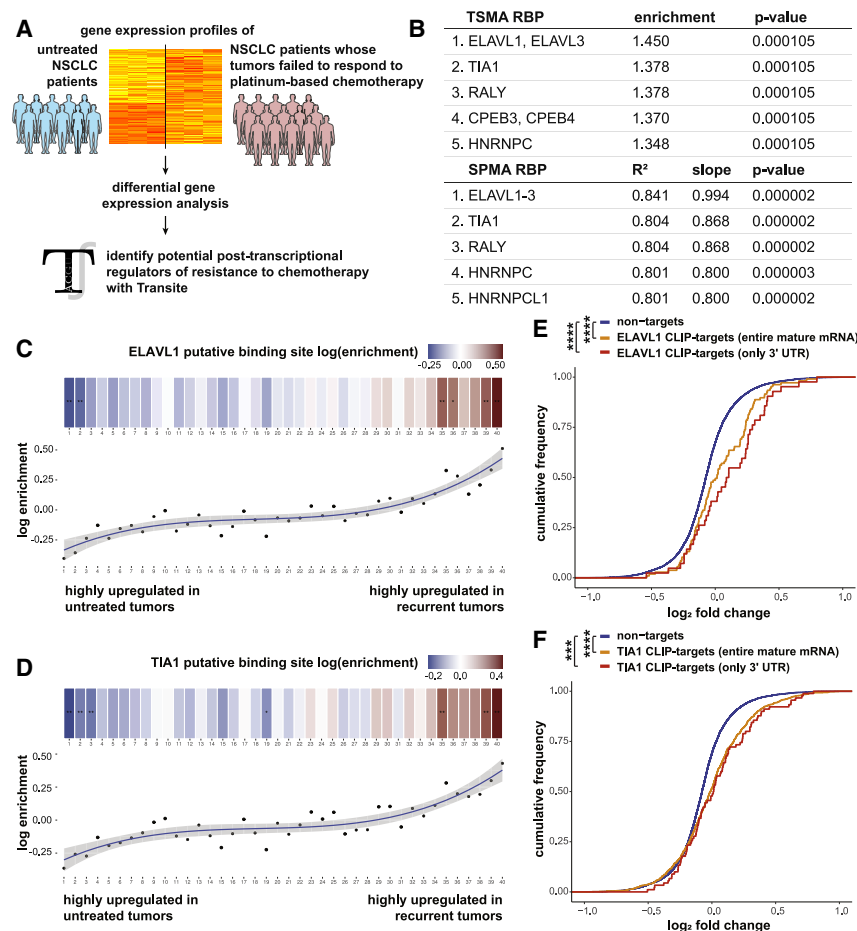
A

gene expression profiles of

untreated NSCLC patients

NSCLC patients whose tumors failed to respond to platinum-based chemotherapy

↓

differential gene expression analysis

↓

T identify potential post-transcriptional regulators of resistance to chemotherapy with Transite

B

TSMA RBP	enrichment		p-value
1. ELAVL1, ELAVL3	1.450		0.000105
2. TIA1	1.378		0.000105
3. RALY	1.378		0.000105
4. CPEB3, CPEB4	1.370		0.000105
5. HNRNPC	1.348		0.000105
SPMA RBP	R²	slope	p-value
1. ELAVL1-3	0.841	0.994	0.000002
2. TIA1	0.804	0.868	0.000002
3. RALY	0.804	0.868	0.000002
4. HNRNPC	0.801	0.800	0.000003
5. HNRNPCL1	0.801	0.800	0.000002

C

ELAVL1 putative binding site log(enrichment)
-0.25 0.00 0.50

highly upregulated in untreated tumors

highly upregulated in recurrent tumors

D

TIA1 putative binding site log(enrichment)
-0.2 0.0 0.4

highly upregulated in untreated tumors

highly upregulated in recurrent tumors

E

- non-targets
- ELAVL1 CLIP-targets (entire mature mRNA)
- ELAVL1 CLIP-targets (only 3' UTR)

F

- non-targets
- TIA1 CLIP-targets (entire mature mRNA)
- TIA1 CLIP-targets (only 3' UTR)

Figure 6. SPMA Identifies ELAVL1 and TIA1 Motifs as Highly Enriched in Recurrent NSCLC Patients

(A) Differential gene expression analysis was performed on samples from patients with untreated NSCLC tumors and patients with recurrent tumors.

(B) Transite was used to identify RBPs whose targets were overrepresented among upregulated genes in the samples of recurrent tumors. Shown are 2 tables of k-mer-based TSMA and SPMA displaying RBPs with highly enriched motifs for TSMA and highly non-random motif enrichment patterns for SPMA. Among the top hits are ELAVL1, TIA1, and hnRNPC. P values were obtained by Monte Carlo sampling and corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure.

(C) Spectrum plot from SPMA depicting the distribution of putative ELAVL1-binding sites across all of the transcripts. The transcripts are sorted by ascending signal-to-noise ratios. The transcripts downregulated in resistant samples relative to untreated samples are on the left, and those upregulated are on the right of the spectrum. The putative binding sites of ELAVL1 are highly enriched in transcripts upregulated in resistant cells (shown in red) and highly depleted in transcripts downregulated in resistant cells (shown in blue).

(D) Spectrum plot of putative TIA1 binding sites using the same transcript order as in (C).

(E) Enrichment of ELAVL1 targets in resistant NSCLC cells is recapitulated in an independent high-throughput sequencing of RNA isolated by CLIP (HITS-CLIP) experiment (publicly available data). The distribution of fold changes of transcripts that have ELAVL1-binding sites is shifted in the positive direction, even more so when the binding sites are in the 3′ UTR. The p values were calculated with the 1-sided Kolmogorov-Smirnov test.

(F) As in (E), transcripts with TIA1-binding sites are upregulated in resistant cells according to an iCLIP experiment, confirming results from SPMA.

from more complex perturbations are more likely to reflect real changes in RBP binding or activity.

RBPs Involved in the DDR Are Identified by *Transite* Using Cancer Patient RNA Expression Data

As an application of *Transite*-based RBP scoring, we next analyzed a gene expression dataset from patients with NSCLC who were either treatment naive or their cancer had recurred after platinum-based chemotherapy treatment (GEO: GSE7880). Differences in RNA transcript abundance were ranked between the set of tumors that were sampled pre-treatment and the separate set of tumors that were sampled after recurrence following treatment, and the ranked transcripts were then analyzed by *Transite* to identify potential RBPs that may influence the response to Pt treatment. Changes in transcript abundance were ranked based on the signal-to-noise ratio, in which transcripts upregulated in recurrent patients had positive values and those upregulated in treatment-naive patients had negative values (see Figure 6A for schematic). k-mer-based TSMA, focusing on the 3′ UTRs of the differentially regulated genes, revealed a set of enriched k-mers in the patients whose tumors

failed Pt treatment that were largely U-rich (Table S5). These k-mers mapped to the motifs of ELAVL1 and TIA1 as the top 2 hits (Figure 6B, top). SPMA revealed these same top 2 RBPs, as shown in the bottom part of Figure 6B. Individual spectrum plots for ELAVL1 (Figure 6C) and TIA1 (Figure 6D) demonstrated consistent behavior of these motifs across the gene expression continuum, being enriched in 3′ UTRs of genes that were upregulated in patients with recurrent tumors after Pt treatment and depleted in 3′ UTRs of genes that were upregulated in naive patients. The upregulation of ELAVL1 and TIA1-target mRNAs was further validated by analyzing the distribution of previously known CLIP-seq-identified targets (Kishore et al., 2011; Wang et al., 2010) for these 2 RBPs (Figure 6E and 6F), suggesting that our motif-based approach can identify bona fide target genes of a given RBP for which CLIP-seq data are available. Moreover, both ELAVL1 and TIA1 are known to be involved in the DDR (Masuda et al., 2011; Lal et al., 2006; Mehta et al., 2016; Díaz-Muñoz et al., 2017). The fact that two well-known players in the DDR were among the top hits of the motif analysis provides confidence that the predictions of *Transite* are likely to reflect the regulators of the DDR and drivers of chemoresistance.

Motif Analysis of Recurrent NSCLCs after Cisplatin Treatment Identifies hnRNPC as a Potential Modulator of Drug Resistance

We were particularly interested in using *Transite* as a tool to identify new RBPs potentially involved in chemosensitivity or resistance to DNA-damaging chemotherapy agents using data from human clinical trials. We therefore chose to focus on hnRNPC, one of the highest-scoring RBPs that emerged from both TSMA and SPMA analyses of chemoresistant NSCLC patients, and one that, to our knowledge, has not been strongly implicated in the response to chemotherapy-induced DNA damage (Shkreta and Chabot, 2015). As shown in Figure 7A, the spectrum plot of the distribution of putative hnRNPC binding sites shows a strong enrichment of mRNAs with hnRNPC motifs in their 3′ UTRs in patients whose tumors recurred after Pt therapy. This *Transite* prediction was independently confirmed by the analysis of individual-nucleotide resolution CLIP (iCLIP)-defined target mRNAs for hnRNPC (König et al., 2010) (which also showed an overrepresentation of hnRNPC targets in upregulated transcripts in recurrent patients; Figure 7B), with those with binding in the 3′ UTR showing the strongest enrichment. Since hnRNPC has not been studied closely in the context of mRNA stability regulation, we wanted to ascertain whether hnRNPC does indeed regulate mRNA levels. To do this, we used publicly available data from the ENCODE Project Consortium (2012), in which hnRNPC was knocked down in HepG2 cells by small hairpin RNA (shRNA) followed by RNA-seq to measure gene expression. Transite analysis of the gene expression changes upon hnRNPC knockdown (Figure 7C) demonstrated clear enrichment of hnRNPC motifs in the 3′ UTRs of downregulated transcripts, suggesting a general role for hnRNPC in maintaining the expression of its 3′ UTR target mRNAs. These observations were further confirmed by examining CLIP-defined target mRNAs of hnRNPC (Figure 7D). Further analysis of these knockdown data by GSEA (Figure 7E) revealed that downregulated genes upon hnRNPC knockdown were enriched for numerous gene sets relating to resistance to chemotherapy and regulation of the DDR, highlighting a potential functional role for hnRNPC in chemoresistance.

To experimentally test these *Transite* and GSEA predictions, we examined the effect of knockdown or overexpression of hnRNPC in T6a murine lung carcinoma cells on their sensitivity and resistance to cisplatin treatment. As shown in Figure 7H, colony-formation assays in T6a cells demonstrated that hnRNPC overexpression promoted resistance to cisplatin, as shown by a 1.6-fold increase in the number of surviving colonies (Figure 7H, red bar). Conversely, siRNA downregulation of hnRNPC significantly enhanced T6a cell sensitivity to cisplatin, as shown by a 5-fold decrease in the number of colonies formed by cells treated with hnRNPC siRNA compared to those of control siRNA-treated cells after cisplatin treatment (Figure 7H, blue bar). We further tested these predictions by performing colony-formation assays in H2009 human lung carcinoma cells. As shown in Figure 7I, hnRNPC overexpression in H2009 cells promoted resistance to cisplatin, as shown by a 1.6-fold increase in the number of surviving colonies (Figure 7I, red bar). Conversely, siRNA knockdown of hnRNPC enhanced H2009 cell sensitivity to cisplatin as demonstrated by a 1.75-fold decrease in the number of

colonies formed (Figure 7I, blue bar). Western blots of hnRNPC in T6a and H2009 cells are shown in Figures 7F and 7G, respectively. These data indicate that hnRNPC mediates the resistance of NSCLC cells to cisplatin chemotherapy, which is consistent with what was seen in the patient data, and demonstrates that our computational approach can identify new RBPs influencing the DDR.

To independently validate the importance of hnRNPC in mediating chemotherapy response in patients, we took advantage of data from a unique adjuvant chemotherapy trial, JBR.10 (Figure 7J; Winton et al., 2005). In this trial, early-stage NSCLC patients had their tumors surgically resected and subjected to gene expression profiling (GEO: GSE14814). Patients were then randomized to receive cisplatin/vinorelbine combination chemotherapy or observation and palliative care, allowing us to specifically query the role of hnRNPC in the response to chemotherapy. We focused our analysis on stage 2 patients, since the benefit from adjuvant chemotherapy is most pronounced in this population. The separation of patients based on hnRNPC expression level revealed that patients whose tumors displayed a low expression of hnRNPC benefited significantly from chemotherapy in terms of survival (Figure 7J, right panel, p = 0.028), while patients whose tumors had high levels of hnRNPC expression did not show significant benefit (Figure 7J, left panel, p = 0.68). We did not find a similar relationship between chemoresistance and expression of RBPs not implicated by *Transite* (Figure S3). The data in Figure 7 identify hnRNPC as a new RBP involved in the response to Pt drug treatment in NSCLC and suggest that Transite is an effective tool for identifying novel RBPs that contribute to chemoresistance in human cancer patient RNA expression datasets.

DISCUSSION

Despite their crucial role in the post-transcriptional regulation of gene expression, the majority of RBPs have unknown functions. To help understand the influence of RBPs on their target transcripts, we developed *Transite*, a computational method for the analysis of the regulatory role of RBPs in various cellular processes for which differential gene expression data or other relevant gene sets are available. Our analysis is based on the fact that most RBPs recognize short linear oligonucleotide sequences whose overrepresentation can be computed from gene expression data, and that a large collection of preexisting motif data for RBPs has been compiled in publicly available databases (Ray et al., 2013; Cook et al., 2011).

It is important to note that *Transite*, in its current form, has significant limitations. Not all RBPs have strong motif preferences that are amenable to this type of motif-based analysis. Furthermore, there may be considerable redundancy in motif recognition by different RBPs, making prediction of a single RBP challenging. Moreover, the *in vitro*-derived motifs for RBPs may not always reflect *in vivo* binding preferences. These caveats have raised questions about the ability of consensus motifs and PWMs to uniquely predict individual RBP mRNA targets *a priori* on a genome-wide scale and have led to the development of more sophisticated approaches for predicting specific RBP RNA targets (Perron et al., 2018; Weyn-Vanhentenryck and
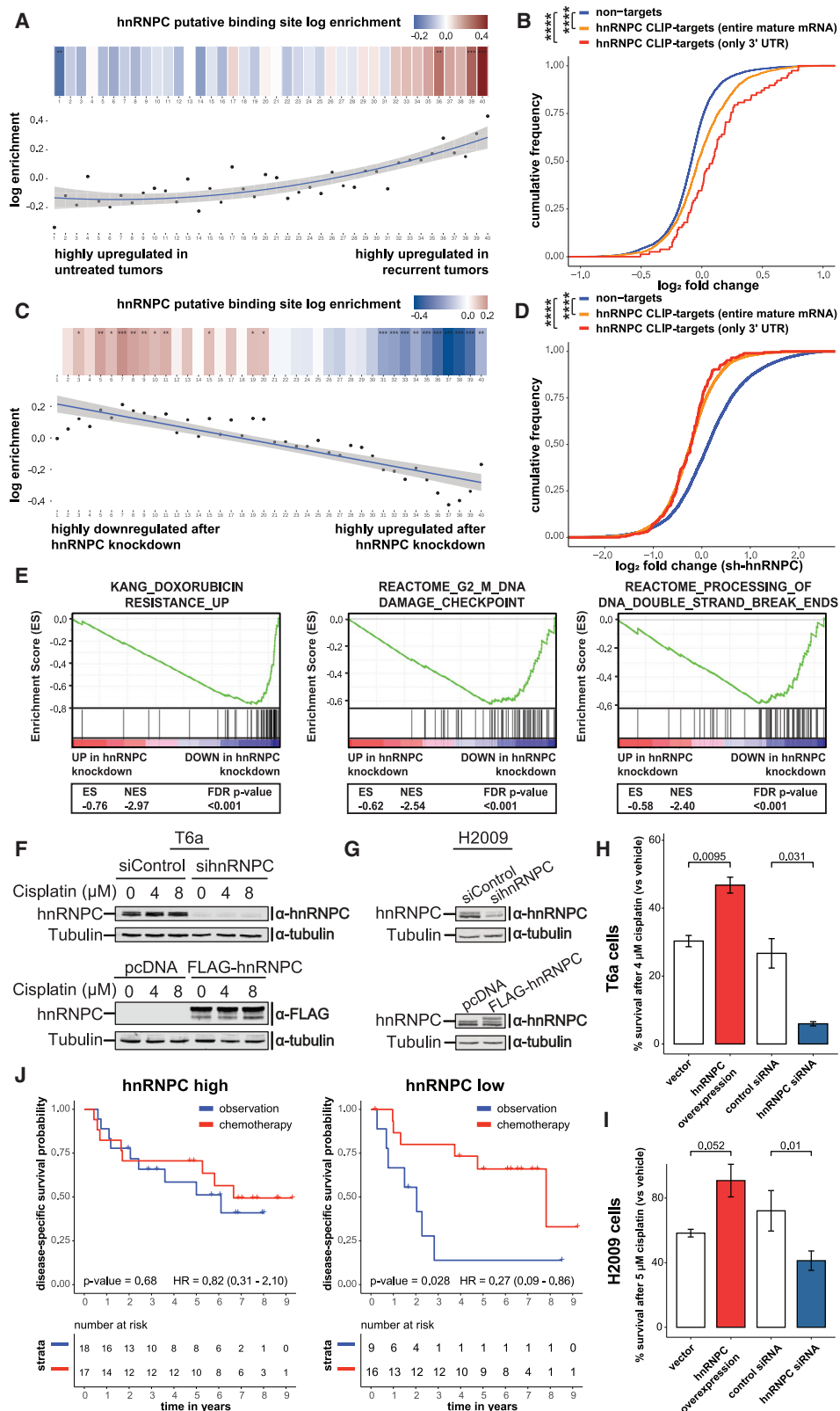
Zhang, 2016). In contrast to those approaches, *Transite* does not attempt to predict specific mRNAs bound by a particular RBP. Instead, *Transite* simply looks at the statistical distribution of RBP motif representation in sets of expressed genes to infer putative roles for specific RBPs in some biological processes, which can then be directly tested experimentally. It is worth noting that *Transite* is best suited to picking up signals from cytoplasmic 3′ UTR binding proteins that affect mRNA stability or translation, if measured by ribosome profiling.

By using 2 approaches to identify non-random distributions of RBP-binding motifs, followed by back-mapping of those motifs onto those of 174 known RBPs, *Transite* identified 3 RBPs involved in the human DDR, which we could further validate based on independent CLIP-seq data of their known mRNA targets in cells, rather than using motifs derived from *in vitro* sequence libraries. These findings suggest that, although there are limitations to using *in vitro*-derived motifs, *Transite* serves as a discovery tool for new biology. Moreover, since users can define their own motifs in addition to those from the database, users are able to upload motifs from CLIP-seq data of their favorite RBP and use that as a means to analyze enrichment in preexisting datasets. As more RBP motifs become available, they will be incorporated in future versions of the *Transite* analysis platform.

To further demonstrate the utility of *Transite*, we performed an analysis of human NSCLC patient data and were able to recover previously known RBP biology and also identify novel sources of RBP-mediated chemoresistance. Well-known players in the DDR such as ELAVL1 and TIA1 were among the top hits in the tumor resistance gene expression dataset, showing that our approach is consistent with previous DDR literature. *Transite* was also able to identify hnRNPC as a new potential modulator of cisplatin sensitivity in NSCLC patients. Experimental validation of the *in silico* prediction further provides independent support for a critical role for hnRNPC in mediating the resistance of NSCLC cells to chemotherapy, which was independently correlated with clinical responses in an additional NSCLC patient dataset.

Transite is a versatile tool that can be used with any type of gene expression data, the only requirements being a list of gene identifiers and some means to separate foreground and background sets or rank the gene list. Examples of the other types of data that are compatible with a *Transite* style of analysis include (1) searching for RBP motif enrichment in 5′ or 3′ UTRs of genes whose translational efficiency changes in response to some stimulus as measured by ribosome or polysome profiling, (2) searching for enrichment of RBP motifs in mRNAs that are localized to specific subcellular compartments, and (3) *de novo* motif analysis in the entire mRNA of gene expression changes upon knockdown of a nuclease of unknown function. The *Transite* website (https://transite.mit.edu) makes this tool accessible to a broad group of scientists, provides a means by which the large body of preexisting gene expression data from microarray and RNA-seq experiments, for example, can be further leveraged to identify changes in mRNA expression associated with specific RBPs, and reveals potential insights into how RBPs may contribute to the concerted regulation and function of specific cellular processes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell culture and colony formation assays

Figure 7. hnRNPC Modulates Sensitivity to Cisplatin

(A) Spectrum plot from *k*-mer-based SPMA depicting the distribution of putative hnRNPC binding sites across all transcripts in samples from patients with untreated NSCLC tumors and patients with recurrent tumors, as in Figure 6. The transcripts are sorted by ascending signal-to-noise ratio from lowest to highest abundance in resistance relative to untreated samples. Putative hnRNPC binding sites are highly enriched in the upregulated fraction of transcripts (GEO: GSE7880).

(B) Enrichment of hnRNPC binding sites in upregulated transcripts is independently confirmed by CLIP experiments. The p values were calculated with the 1-sided Kolmogorov-Smirnov test.

(C) Spectrum plot from *k*-mer-based SPMA depicting the distribution of putative hnRNPC binding sites across all transcripts, sorted by fold change after hnRNPC knockdown by shRNA in HepG2 cells (GEO: GSE87993).

(D) Same as (B), now with transcript fold change from RNA-seq experiments before and after knockdown of hnRNPC (GEO: GSE87993).

(E) GSEA plots of select gene sets from analysis of hnRNPC knockdown RNA-seq data.

(F) Western blots of hnRNPC levels in T6a cells transfected with hnRNPC-specific siRNA (top) or pcDNA3.1 vector expressing FLAG-hnRNPC (bottom) 24 or 48 h, respectively, after being treated with the indicated doses of cisplatin. The blot is representative of 3 independent experiments.

(G) Western blots of hnRNPC levels in H2009 cells transfected with hnRNPC-specific siRNA (left, representative blot of 3 independent experiments) or pcDNA3.1 vector expressing FLAG-hnRNPC (right) 48 h after transfection.

In (F) and (G), the blotting antibodies are indicated next to the images. Tubulin was used as a loading control.

(H) siRNA-mediated reduction in hnRNPC levels significantly impairs long-term survival of T6a cells in response to cisplatin (blue bar). The overexpression of hnRNPC (red bar) protects against cisplatin-induced cell death in T6a cells in colony-formation assays. The bar graphs represent the percent number of colonies formed, normalized to untreated control cells. The white bars represent cells transfected with control vehicles (control siRNA or empty pcDNA3.1 vector). The error bars indicate the standard error of the mean of 3 replicates.

(I) Same as (H), except performed in H2009 cells.

(J) High expression of hnRNPC is associated with decreased efficacy of Pt-based chemotherapy in patients with stage 2 disease from the JBR.10 lung cancer adjuvant chemotherapy trial (GEO: GSE14814). The p value was calculated with the log-rank test (HR is hazard ratio, with the confidence interval in brackets). The hnRNPC low group is patients with hnRNPC expression Z scores of ≤ -0.2, and the hnRNPC high group is patients with hnRNPC expression Z scores ≥ 0.2.

- ○ siRNA transfection
- ○ Overexpression of hnRNPC in T6a cells
- ○ Overexpression of hnRNPC in H2009 cells
- ○ Immunoblotting
- ● METHOD DETAILS
  - ○ Differential gene expression analysis
  - ○ Motif databases
  - ○ Motif representations
  - ○ Combining enrichment of motif-associated k-mers
  - ○ Local consistency score
  - ○ Polynomial regression
  - ○ CLIP-seq data analysis
  - ○ Analysis of hnRNPC knockdown RNA-seq data
  - ○ Package and web development
- ● QUANTIFICATION AND STATISTICAL ANALYSIS

AUTHOR CONTRIBUTIONS

Conceptualization, I.G.C. and M.B.Y.; Methodology, K.K., A.G., I.G.C., and M.B.Y.; Software, K.K., A.G., and T.B.; Validation, M.A.B., S.V., E.D.H., and Y.W.K.; Formal Analysis, K.K., A.G., D.A.A., and M.B.Y.; Investigation, K.K., M.A.B., S.V., and E.D.H.; Resources, M.B.Y.; Data Curation, K.K.; Writing – Original Draft, K.K. and M.B.Y.; Writing – Review & Editing, K.K., M.A.B., E.D.H., T.B., D.A.A., A.H., B.A.J., Y.W.K., I.G.C., and M.B.Y.; Visualization, K.K.; Supervision, A.H., B.A.J., Y.W.K., I.G.C., and M.B.Y.; Funding Acquisition, K.K., A.G., T.B., I.G.C., and M.B.Y.

REFERENCES

Adamson, B., Smogorzewska, A., Sigoillot, F.D., King, R.W., and Elledge, S.J. (2012). A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. Nat. Cell Biol. 14, 318–328.

Akaike, Y., Masuda, K., Kuwano, Y., Nishida, K., Kajita, K., Kurokawa, K., Satake, Y., Shoda, K., Imoto, I., and Rokutan, K. (2014). HuR regulates alternative splicing of the TRA2β gene in human colon cancer cells under oxidative stress. Mol. Cell. Biol. 34, 2857–2873.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188.

Cannell, I.G., Merrick, K.A., Morandell, S., Zhu, C.Q., Braun, C.J., Grant, R.A., Cameron, E.R., Tsao, M.S., Hemann, M.T., and Yaffe, M.B. (2015). A Pleiotropic RNA-Binding Protein Controls Distinct Cell Cycle Checkpoints to Drive Resistance of p53-Defective Tumors to Chemotherapy. Cancer Cell 28, 623–637.

Chambers, J., Hastie, T., and Pregibon, D. (1990). Statistical models in S. In Compstat: Proceedings in Computational Statistics, 9th Symposium held at Dubrovnik, Yugoslavia, 1990, K. Momirović and V. Mildner, eds. (Springer), pp. 317–321.

Chang, S.H., Elemento, O., Zhang, J., Zhuang, Z.W., Simons, M., and Hla, T. (2014). ELAVL1 regulates alternative splicing of eIF4E transporter to promote postnatal angiogenesis. Proc. Natl. Acad. Sci. USA 111, 18309–18314.

Ciccia, A., and Elledge, S.J. (2010). The DNA damage response: making it safe to play with knives. Mol. Cell 40, 179–204.

Cook, K.B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T.R. (2011). RBPDB: a database of RNA-binding specificities. Nucleic Acids Res. 39, D301–D308.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell 136, 777–793.

Coppin, L., Leclerc, J., Vincent, A., Porchet, N., and Pigny, P. (2018). Messenger RNA Life-Cycle in Cancer Cells: Emerging Role of Conventional and Non-Conventional RNA-Binding Proteins? Int. J. Mol. Sci. 19, E650.

Díaz-Muñoz, M.D., Kiselev, V.Y., Le Novère, N., Curk, T., Ule, J., and Turner, M. (2017). Tia1 dependent regulation of mRNA subcellular location and translation controls p53 expression in B cells. Nat. Commun. 8, 530.

Dimitrova, N., Gocheva, V., Bhutkar, A., Resnick, R., Jong, R.M., Miller, K.M., Bendor, J., and Jacks, T. (2016). Stromal Expression of miR-143/145 Promotes Neoangiogenesis in Lung Cancer Development. Cancer Discov. 6, 188–201.

Dunn, J., and Dunn, O.J. (1961). Multiple Comparisons among Means (American Statistical Association), pp. 52–64.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Fan, J., Yang, X., Wang, W., Wood, W.H., 3rd, Becker, K.G., and Gorospe, M. (2002). Global analysis of stress-regulated mRNA turnover by using cDNA arrays. Proc. Natl. Acad. Sci. USA 99, 10611–10616.

Floyd, S.R., Pacold, M.E., Huang, Q., Clarke, S.M., Lam, F.C., Cannell, I.G., Bryson, B.D., Rameseder, J., Lee, M.J., Blake, E.J., et al. (2013). The bromodomain protein Brd4 insulates chromatin from DNA damage signalling. Nature 498, 246–250.

Garner, M.M., and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res. 9, 3047–3060.

Gasch, A.P., Huang, M., Metzner, S., Botstein, D., Elledge, S.J., and Brown, P.O. (2001). Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. Mol. Biol. Cell 12, 2987–3003.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. 15, 829–845.

Gotea, V., Visel, A., Westlund, J.M., Nobrega, M.A., Pennacchio, L.A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. 20, 565–577.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Hong, S. (2017). RNA Binding Protein as an Emerging Therapeutic Target for Cancer Prevention and Treatment. J. Cancer Prev. 22, 203–210.

Hurov, K.E., Cotta-Ramusino, C., and Elledge, S.J. (2010). A genetic screen identifies the Triple T complex required for DNA damage signaling and ATM and ATR stability. Genes Dev. 24, 1939–1950.

Jackson, S.P., and Bartek, J. (2009). The DNA-damage response in human biology and disease. Nature 461, 1071–1078.

Kim, H.H., Abdelmohsen, K., and Gorospe, M. (2010). Regulation of HuR by DNA Damage Response Kinases. J. Nucleic Acids 2010, 981487.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat. Methods 8, 559–564.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. 17, 909–915.

Lai, W.S., Kennington, E.A., and Blackshear, P.J. (2003). Tristetraprolin and its family members can promote the cell-free deadenylation of AU-rich element-containing mRNAs by poly(A) ribonuclease. Mol. Cell. Biol. 23, 3798–3812.

Lal, A., Abdelmohsen, K., Pullmann, R., Kawai, T., Galban, S., Yang, X., Brewer, G., and Gorospe, M. (2006). Posttranscriptional derepression of GAD-D45alpha by genotoxic stress. Mol. Cell 22, 117–128.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics 12, 323.

Licatalosi, D.D., and Darnell, R.B. (2010). RNA processing and its regulation: global insights into biological networks. Nat. Rev. Genet. 11, 75–87.

Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet. 24, 416–425.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. 8, 479–490.

Masuda, K., Abdelmohsen, K., Kim, M.M., Srikantan, S., Lee, E.K., Tominaga, K., Selimyan, R., Martindale, J.L., Yang, X., Lehrmann, E., et al. (2011). Global dissociation of HuR-mRNA complexes promotes cell survival after ionizing radiation. EMBO J. 30, 1040–1053.

Matsuoka, S., Ballif, B.A., Smogorzewska, A., McDonald, E.R., 3rd, Hurov, K.E., Luo, J., Bakalarski, C.E., Zhao, Z., Solimini, N., Lerenthal, Y., et al. (2007). ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science 316, 1160–1166.

Mehta, M., Basalingappa, K., Griffith, J.N., Andrade, D., Babu, A., Amreddy, N., Muralidharan, R., Gorospe, M., Herman, T., Ding, W.Q., et al. (2016). HuR silencing elicits oxidative stress and DNA damage and sensitizes human triple-negative breast cancer cells to radiotherapy. Oncotarget 7, 64820–64835.

Mukherjee, N., Corcoran, D.L., Nusbaum, J.D., Reid, D.W., Georgiev, S., Hafner, M., Ascano, M., Jr., Tuschl, T., Ohler, U., and Keene, J.D. (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. Mol. Cell 43, 327–339.

Mukherjee, N., Jacobs, N.C., Hafner, M., Kennington, E.A., Nusbaum, J.D., Tuschl, T., Blackshear, P.J., and Ohler, U. (2014). Global target mRNA specification and regulation by the RNA-binding protein ZFP36. Genome Biol. 15, R12.

Nishida, K., Frith, M.C., and Nakai, K. (2009). Pseudocounts for transcription factor binding sites. Nucleic Acids Res. 37, 939–944.

Obenauer, J.C., Cantley, L.C., and Yaffe, M.B. (2003). Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res. 31, 3635–3641.

Paulsen, R.D., Soni, D.V., Wollman, R., Hahn, A.T., Yee, M.C., Guan, A., Hesley, J.A., Miller, S.C., Cromwell, E.F., Solow-Cordero, D.E., et al. (2009). A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. Mol. Cell 35, 228–239.

Pereira, B., Billaud, M., and Almeida, R. (2017). RNA-Binding Proteins in Cancer: Old Players and New Actors. Trends Cancer 3, 506–528.

Perron, G., Jandaghi, P., Solanki, S., Safisamghabadi, M., Storoz, C., Karimzadeh, M., Papadakis, A.I., Arseneault, M., Scelo, G., Banks, R.E., et al. (2018). A General Framework for Interrogation of mRNA Stability Programs Identifies RNA-Binding Proteins that Govern Cancer Transcriptomes. Cell Rep. 23, 1639–1650.

Plass, M., Rasmussen, S.H., and Krogh, A. (2017). Highly accessible AU-rich regions in 3′ untranslated regions are hotspots for binding of regulatory factors. PLOS Comput. Biol. 13, e1005460.

Ray, D., Kazan, H., Chan, E.T., Peña Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat. Biotechnol. 27, 667–670.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature 499, 172–177.

Reinhardt, H.C., Hasskamp, P., Schmedding, I., Morandell, S., van Vugt, M.A., Wang, X., Linding, R., Ong, S.E., Weaver, D., Carr, S.A., and Yaffe, M.B. (2010). DNA damage activates a spatially distinct late cytoplasmic cell-cycle checkpoint network controlled by MK2-mediated RNA stabilization. Mol. Cell 40, 34–49.

Reinhardt, H.C., Cannell, I.G., Morandell, S., and Yaffe, M.B. (2011). Is posttranscriptional stabilization, splicing and translation of selective mRNAs a key to the DNA damage response? Cell Cycle 10, 23–27.

Rieger, K.E., and Chu, G. (2004). Portrait of transcriptional responses to ultraviolet and ionizing radiation in human cells. Nucleic Acids Res. 32, 4786–4803.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Shkreta, L., and Chabot, B. (2015). The RNA Splicing Response to DNA Damage. Biomolecules 5, 2935–2977.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3, Article3.

Stumpo, D.J., Lai, W.S., and Blackshear, P.J. (2010). Inflammation: cytokines and RNA-based regulation. Wiley Interdiscip. Rev. RNA 1, 60–80.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550.

Sugiura, R., Satoh, R., Ishiwata, S., Umeda, N., and Kita, A. (2011). Role of RNA-Binding Proteins in MAPK Signal Transduction Pathway. J. Signal Transduct. 2011, 109746.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249, 505–510.

Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A., Dominguez, D., et al. (2018). A large-scale binding and functional map of human RNA binding proteins. bioRxiv. https://doi.org/10.1101/179648.

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N.M., Rot, G., Zupan, B., Curk, T., and Ule, J. (2010). iCLIP predicts the dual splicing effects of TIA-RNA interactions. PLOS Biol. 8, e1000530.

Weyn-Vanhentenryck, S.M., and Zhang, C. (2016). mCarts: Genome-Wide Prediction of Clustered Sequence Motifs as Binding Sites for RNA-Binding Proteins. Methods Mol. Biol. *1421*, 215–226.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. http://ggplot2.org.

Wilker, E.W., van Vugt, M.A., Artim, S.A., Huang, P.H., Petersen, C.P., Reinhardt, H.C., Feng, Y., Sharp, P.A., Sonenberg, N., White, F.M., and Yaffe, M.B. (2007). 14-3-3sigma controls mitotic translation to facilitate cytokinesis. Nature *446*, 329–332.

Winton, T., Livingston, R., Johnson, D., Rigas, J., Johnston, M., Butts, C., Cormier, Y., Goss, G., Inculet, R., Vallieres, E., et al.; National Cancer Institute of Canada Clinical Trials Group; National Cancer Institute of the United States Intergroup JBR.10 Trial Investigators (2005). Vinorelbine plus cisplatin vs. observation in resected non-small-cell lung cancer. N. Engl. J. Med. *352*, 2589–2597.

Yang, Y.C., Di, C., Hu, B., Zhou, M., Liu, Y., Song, N., Li, Y., Umetsu, J., and Lu, Z.J. (2015). CLIPdb: a CLIP-seq database for protein-RNA interactions. BMC Genomics *16*, 51.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res. *37*, e151.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse monoclonal anti-hnRNPC1/C2	Santa Cruz Biotechnology	Cat#sc-32308; RRID:AB_627731
Mouse monoclonal anti-hnRNPC1/C2	Abcam Inc	Cat#ab10294; RRID:AB_297034
Rabbit polyclonal anti-y-Tubulin	Sigma Aldrich	Cat#T5192; RRID:AB_261690
Chemicals, Peptides, and Recombinant Proteins		
Attractene Transfection Reagent	QIAGEN	301005
X-tremeGene 9	Sigma Aldrich	XTG9-RO
Lipofectamine RNAiMAX	Thermo Fisher	13778150
SYTO 60	Invitrogen	S11342
cis-Diammineplatinum(II) dichloride (cisplatin)	Sigma	CAS# 15663-27-1
N,N-Dimethylformamide (DMF)	Burdick and Jackson	Cat# AS076
G418 Sulfate	Corning	30-234
Deposited Data		
Human reference genome NCBI build 38, GRCh38	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Experimental Models: Cell Lines		
NCI-H2009 (human lung adenocarcinoma)	ATCC	CRL-5911; RRID:CVCL_1514
LG1233/T6a line (mouse lung adenocarcinoma)	Laboratory of Tyler Jacks	LG1233; Dimitrova et al., 2016
Oligonucleotides		
Silencer Select siRNA targeting mouse hnRNPC	Ambion	Cat# s67639
Silencer Select siRNA targeting human hnRNPC	Ambion	Cat# s6720
Silencer Select Negative Control No. 1 siRNA	Ambion	Cat# 4390843
FLAG-tagged mouse hnRNPC Forward primer (5′ → 3′)– GCCCATAAGCTTATGGACTACAAAGACGATGACGACAAGGCTAGCAATGTTACCAACAAGACAGATCCTCGG	This paper	N/A
FLAG-tagged mouse hnRNPC Reverse primer (5′ → 3′)-GCCCATTCTAGATTATTAAGAGTCATCCTCCCCATTGGCGCTGTCTCTG	This paper	N/A
FLAG-tagged human hnRNPC Forward primer (5′ → 3′)- CCATAAGCTTATGGACTACAAAGACGATGACGACAAGTCAGGCGGGATCCGCCAGCAACGTTACCAACAAGACAGATCC	This paper	N/A
FLAG-tagged human hnRNPC Reverse primer (5′ → 3′)- TCAGGAATTCTTAAGAGTCATCCTCGCCATTGGC	This paper	N/A
Recombinant DNA		
pcDNA3.1 Mammalian Expression Vector	Invitrogen	Cat# V79020
pcDNA3.1 FLAG-tagged mouse hnRNPC	This paper	N/A
pcDNA3.1 FLAG-tagged human hnRNPC	This paper	N/A
Software and Algorithms		
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
FastQC v0.11.7	Andrews, 2010	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
RSEM v1.3.0	Li and Dewey, 2011	https://deweylab.github.io/RSEM/
R 3.4.4 / 3.6.0 / 4.0.0	R Core Team	https://www.R-project.org/

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R Bioconductor package - edgeR 3.30.0	Robinson, McCarthy, and Smyth, 2010	https://bioconductor.org/packages/edgeR
R Bioconductor package - BSgenome.Hsapiens. UCSC.hg38 1.4.3	Bioconductor Core Team	https://bioconductor.org/packages/BSgenome. Hsapiens.UCSC.hg38
R Bioconductor package - TxDb.Hsapiens.UCSC. hg38.knownGene 3.10.0	Bioconductor Core Team	https://bioconductor.org/packages/TxDb. Hsapiens.UCSC.hg38.knownGene
R Bioconductor package - org.Hs.eg.db 3.11.1	Bioconductor Core Team	https://bioconductor.org/packages/org. Hs.eg.db
R Bioconductor package - limma 3.44.1	Ritchie, Phipson, Wu, Hu, Law, Shi, Smyth, 2015	https://bioconductor.org/packages/limma
R Bioconductor package - transite 1.6.1	This paper	https://bioconductor.org/packages/transite
R package - survminer 0.4.6	Alboukadel Kassambara, Marcin Kosinski, Przemyslaw Biecek, Scheipl Fabian	https://cran.r-project.org/web/packages/ survminer/index.html
R package - survival 3.1-12	Terry Therneau	https://cran.r-project.org/web/packages/ survival/index.html

RESOURCE AVAILABILITY

Lead Contact
Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Michael B. Yaffe (myaffe@mit.edu).

Materials Availability
This study did not generate new unique reagents.

Data and Code Availability
The code generated during this study is hosted on GitHub (https://github.com/kkrismer/transite) and licensed under the MIT free software license. The Transite method can be accessed online at https://transite.mit.edu. For workflow integration and advanced analysis, the Transite functionality is also offered as an R/Bioconductor package at https://doi.org/10.18129/B9.bioc.transite.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture and colony formation assays
LG1233/T6a cells (mouse lung adenocarcinoma, in the following referred to as T6a) (Dimitrova et al., 2016) were grown in RPMI-1640 medium supplemented with 10% fetal bovine serum at 37 °C in a humidified incubator supplied with 5% CO_2. NCI-H2009 cells (human lung adenocarcinoma, ATCC CRL-5911, in the following referred to as H2009) were grown in DMEM medium supplemented with 10% fetal bovine serum at 37 °C in a humidified incubator supplied with 5% CO_2. Colony formation assays were performed as previously described (Cannell et al., 2015). Briefly, 48 hours after transfection with siRNAs or pcDNA vectors, T6a and H2009 cells were treated with 4 and 5 μM cisplatin or vehicle (DMF), respectively, for 4 hours. Cells were then washed in PBS and re-plated in 6-well plates using 1000 mock-treated or 10,000 cisplatin-treated cells per well. In overexpression assays, 500 μg/ml G418 was added to the media to select for cells transfected with pcDNA vectors. After 10 to 14 days, cells were fixed with 4% formaldehyde and stained with SYTO 60 (Thermo Fisher Scientific). Colonies were scanned and quantified using Odyssey ® CLx Imaging System (LI-COR Biosciences).

siRNA transfection
Silencer Select siRNA (Ambion) transfection was performed using Lipofectamine RNAiMAX following manufacturer instructions (Thermo Fisher) with a final siRNA concentration of 5 nM for T6a cells (mouse siRNA s67639) and 20 nM for H2009 cells (human siRNA s6720). 24 hours after transfection, cells were given fresh media. Cells were then treated as described in the previous section.

Overexpression of hnRNPC in T6a cells
pcDNA3.1 vectors expressing FLAG-tagged mouse hnRNPC were generated as follows. First, total RNA was prepared from KP7B (mouse lung carcinoma) cells using RNeasy purification kit (QIAGEN) and was used to synthesize cDNAs using SuperScript cDNA

Synthesis System (Thermo Fisher). cDNAs were used as templates in PCR reactions using PfuUltra II HF DNA polymerase (Agilent) and the following primers: 5′-GCCCAT**AAGCTT**ATG<u>GACTACAAAGACGATGACGACAAG</u>GCTAGCAATGTTACCAACAAGACA GATCCTCGG-3′ (forward) and 5′-GCCCAT**TCTAGA**TT<u>ATT</u>AAGAGTCATCCTCCCCATTGGCGCTGTCTCTG-3′ (reverse). Restriction sites for HindIII (in forward primer) and XbaI (in reverse primer) are in bold. Sequences encoding FLAG are underlined. The PCR products were cleaved with the indicated restriction enzymes (New England BioLabs Inc), purified (QIAquick PCR Purification Kit, QIAGEN) and cloned into pcDNA3.1 vectors. The integrity of the plasmids were confirmed by sequencing (Eton Bioscience, Inc.).

Attractene (QIAGEN) transfection was performed following manufacturer instructions using 4 g of DNA (empty pcDNA3.1 vector or vector expressing FLAG-tagged mouse hnRNPC) and 15 μL of Attractene Reagent per 10cm plate. 24 hours after transfection, cells were given fresh media. Cells were then treated with cisplatin as described in a previous section.

Overexpression of hnRNPC in H2009 cells

pcDNA3.1 vectors expressing FLAG-tagged human hnRNPC were generated as follows. First, total RNA was prepared from BT20 (human breast carcinoma) cells using RNeasy purification kit (QIAGEN) and was used to synthesize cDNAs using SuperScript cDNA Synthesis System (Thermo Fisher). cDNAs were used as templates in PCR reactions using PfuUltra II HF DNA polymerase (Agilent) and the following primers: 5′-CCAT**AAGCTT**ATG<u>GACTACAAAGACGATGACGACAAG</u>TCAGGCGGATCCGCCAGCAACGTTAC CAACAAGACAGATCC-3′ (forward) and 5′-TCAG**GAATTC**TTAAGAGTCATCCTCGCCATTGGC-3′ (reverse). Restriction sites for Hin-dIII (in forward primer) and EcoR1 (in reverse primer) are in bold. Sequences encoding FLAG are underlined. The PCR products were cleaved with the indicated restriction enzymes (New England BioLabs Inc), purified (QIAquick PCR Purification Kit, QIAGEN) and cloned into pcDNA3.1 vectors. The integrity of the plasmids were confirmed by sequencing (Eton Bioscience, Inc.).

X-tremeGENE 9 (Sigma Aldrich) transfection was performed following manufacturer instructions with a 3:1 ratio of transfection reagent to DNA. 5 μg of DNA (empty pcDNA3.1 vector or vector expressing FLAG-tagged human hnRNPC) was transfected per 10cm plate. 24 hours after transfection, cells were given fresh media. Cells were then treated with cisplatin as described in a previous section.

Immunoblotting

T6a cells were harvested 24 (siRNA-transfected) or 48 (pcDNA vectors-transfected) hours after cisplatin treatment and re-plating. H2009 cells were harvested 48 hours after siRNA of pcDNA transfection. Cells were then lysed in RIPA buffer and subjected to standard SDS/PAGE electrophoresis and transferred to nitrocellulose membranes. The membranes were immunoblotted with antibodies against hnRNPC (T6a - ab10294, Abcam Inc.; H2009 - sc-32308, Santa Cruz) and γ-tubulin (T5192, Sigma-Aldrich) following manufacturer's instructions.

METHOD DETAILS

Differential gene expression analysis

Differential gene expression analysis for datasets used in this manuscript was performed with the R/Bioconductor package *limma* (Ritchie et al., 2015). A linear model was fit to each row of the log_2-transformed expression value matrices, where rows correspond to transcripts and columns correspond to samples. An empirical Bayes method was used to obtain the magnitude and significance of the log fold change between sample groups for each transcript (Smyth, 2004). Raw p values were adjusted using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Motif databases

Transite incorporates sequence motifs of RBP binding sites from two databases: CISBP-RNA, a catalog of inferred sequence preferences of RNA binding proteins (Ray et al., 2013), and RBPDB, a database of RNA-binding specificities (Cook et al., 2011). Together these contribute 174 sequence motifs of varying lengths (between 6 and 18 nucleotides). All motifs were obtained using *in vitro* techniques for determining RNA targets. The majority of motifs were determined by either systematic evolution of ligands by exponential enrichment (SELEX) (Tuerk and Gold 1990) or RNAcompete (Ray et al., 2009). The RNA binding specificities of two further RBPs were obtained by electrophoretic mobility shift assays (EMSA) (Garner and Revzin 1981).

Motif representations

Motif descriptions provided from the databases described above were converted from count matrices to position weight matrices (PWMs), obtained by normalizing each nucleotide's probability at each position by the mean probability of each nucleotide, 25%.

For *k*-mer-based analyses, PWMs were converted to hexamers and heptamers by generating all *k*-mers for which each position has a probability higher than a certain threshold. In the work presented here, we used a threshold probability of 0.215, which is a stringency level that works well empirically with the motifs from the motif databases.

Laplace smoothing (also known as additive smoothing) is applied to avoid zeros in count matrices before conversion to PWMs. Zeros might occur if the number of sequences on which the position-specific scoring matrix (PSSM) is based, is too small to contain at least one occurrence of each nucleotide per position. In this case, pseudocounts are introduced (Nishida et al., 2009).

Combining enrichment of motif-associated k-mers

The overall enrichment of a motif for k-mer TSMA is calculated as the geometric mean of the enrichments of associated k-mers:

$$\bar{e} = exp\left(\frac{1}{n}\sum_{i=1}^{n} log(e_i)\right),$$

where $\boldsymbol{e}$ is the vector of enrichment values of motif-associated k-mers. The sum of logarithms is used instead of the product to avoid arithmetic underflow.

Local consistency score

The local consistency score quantifies the local noise of the gradient in the spectrum by calculating the deviance between the linear interpolation of the scores of two bins separated by one other, and the score of the middle bin, for each position in the spectrum. The lower the score, the more consistent the trend in the spectrum plot. Formally, the local consistency score x_c of scores vector $\boldsymbol{s}$ is defined as

$$x_c(\boldsymbol{s}) = \frac{1}{n}\sum_{i=1}^{n-2}\left|\frac{s_i + s_{i+2}}{2} - s_{i+1}\right|.$$

In order to obtain an estimate of the significance of a particular score x_c', Monte Carlo sampling is performed by randomly permuting the coordinates of the scores vector $\boldsymbol{s}$ and recomputing x_c. The probability estimate $\hat{p}$ is given by the lower tail version of the cumulative distribution function

$$\widehat{Pr_L}(T(x)) = \frac{\sum_{i=1}^{n} 1(T(y_i) \leq T(x)) + 1}{n + 1},$$

where T equals x_c.

Polynomial regression

An alternative approach to assess the consistency of a spectrum plot is via polynomial regression. In a first step, polynomial regression models of various degrees are used to fit $\boldsymbol{s}$, the vector of scores, as a function of $\boldsymbol{b}$, the vector of bin numbers. Then the model that reflects best the true nature of the data is selected by means of the F-test. Finally, the adjusted R^2 are calculated to indicate how well the model fits the data. These statistics are used as scores to rank the spectrum plots.

In general, the polynomial regression equation is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \epsilon_i,$$

where m is the degree of the polynomial (usually $m \leq 2$), and ϵ_i is the error term. The dependent variable $\boldsymbol{y}$ is the vector of scores $\boldsymbol{s}$ and $\boldsymbol{x}$ to $\boldsymbol{x}^m$ are the orthogonal polynomials of the vector of bin numbers $\boldsymbol{b}$.

Orthogonal polynomials are used in order to reduce the correlation between the different powers of $\boldsymbol{b}$ and therefore avoid multicollinearity in the model. This is important, because correlated predictors lead to unstable coefficients, i.e., the coefficients of a polynomial regression model of degree m can be greatly different from a model of degree $m + 1$.

The orthogonal polynomials of vector $\boldsymbol{b}$ are obtained by centering (subtracting the mean), QR decomposition, and subsequent normalization (Chambers et al., 1990).

Given the dependent variable $\boldsymbol{y}$ and the orthogonal polynomials of $\boldsymbol{b}$ $\boldsymbol{x}$ to $\boldsymbol{x}^m$, the model coefficients β are chosen in a way to minimize the deviance between the actual and the predicted values. Ordinary least-squares is used as the estimation method for the model coefficients. After polynomial models of various degrees have been fitted to the data, the F-test is used to select the model that best fits the data. After a model has been selected, the adjusted R^2 is calculated as an additional way to evaluate the goodness of fit.

CLIP-seq data analysis

The BED files (output from Piranha analysis) for all CLIP-Seq datasets were retrieved from CLIPdb (Yang et al., 2015). Read counts were mapped to RefSeq identifiers using a UCSC table with either just 3′-UTR sequences or the entire mature mRNA of all human mRNAs in Hg19 coordinates. RefSeq identifiers were then summarized to gene symbols. For gene symbols with multiple RefSeq identifiers, the one with the maximum counts was taken, as it was assumed this indicated the most highly expressed transcript. This analysis created two gene lists, one where there was binding in the 3′-UTR (3′-UTR targets) or where there was binding in any region of the mRNA (entire mature mRNA targets). These gene lists were then merged with fold change lists from GEO gene expression dataset GSE7880. To generate the non-targets list, the entire mature mRNA list was subtracted from the GSE7880 list.

Analysis of hnRNPC knockdown RNA-seq data

RNA-seq data for HepG2 cells expressing shRNA targeting hnRNPC were retrieved from GEO. Briefly, HepG2 cells were infected with either lenti shRNA for hnRNPC (GSE87993) or a non-target shRNA (GSE88174) for 24 hours in duplicate. RNA was collected 6 days later. 100 nucleotide paired-end reads were obtained from Illumina HiSeq2000 (ENCODE Project Consortium, 2012).

The quality of the RNA-seq data was assessed using the FastQC tool v0.11.7 (Andrews, 2010) before and after removing reads that mapped to ribosomal RNAs. Paired-end RNA-seq data were mapped to the human genome (GRCh38 build from Gencode) using RSEM (Li and Dewey, 2011) with Bowtie 2 (Langmead and Salzberg, 2012) as the aligner. Expected counts, expected counts rounded, TPM, and FPKM files were generated.

EdgeR (Robinson et al., 2010) was used to perform differential gene expression analysis using the expected counts rounded values file from the previous section. Low count reads were filtered out and TMM normalization (Trimmed Mean of M-values) was used to account for differences in total reads between samples. Exact test was performed for shRNA hnRNPC versus non-targeting shRNA.

The log fold change-sorted list of genes was used for the Transite analysis. k-mer-based SPMA was performed using the default settings (Transite motif database, 40 bins, merge method of highest magnitude, 1 degree polynomial, 3′-UTR, Benjamini-Hochberg p value adjustment, and 5 maximum binding sites per mRNA).

The same list of genes was used for gene set enrichment analysis (GSEA) (Subramanian et al., 2005), which we queried against the *C2_all* gene sets from the molecular signatures database using the preranked feature of GSEA.

Package and web development

R package development and documentation was streamlined with *devtools* and *roxygen2*, respectively. Core algorithms were implemented in C++. *ggplot2* (Wickham, 2009) was used for data visualization.

The website was developed in R with the reactive web application framework *shiny* from RStudio. The components of the graphical user interface were provided by *shiny* and *shinyBS*, which serve as an R wrapper for the components of the Bootstrap front-end web development framework.

Human 3′-UTR sequence annotations were obtained from the Bioconductor UCSC hg38 annotation packages.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were carried out in R. Detailed information about R packages and versions is part of the Transite analysis reports, which are available online at https://transite.mit.edu/p/transite-reports.pdf.

All p values reported by Transite are multiple hypothesis corrected using the Benjamini-Hochberg procedure. Significance levels are denoted by asterisks as follows: **** (p values $\leq$ 0.0001), *** (p values $\leq$ 0.001), ** (p values $\leq$ 0.01), and * (p values $\leq$ 0.05). Shaded areas around log enrichment curves in spectrum plots are 95% confidence intervals.

Significance levels of difference between cumulative distribution functions were determined by one-sided Kolmogorov-Smirnov tests (Figures 6E, 6F, 7B, and 7D).

In Figures 7H and 7I, significance of difference in means was calculated with Welch's two-sample t tests (unpaired, two-sided) with $n = 3$. Error bars in Figures 7H and 7I are standard error of the mean.

In Figures 7J and S3, significance of difference between Kaplan-Meier survival curves was calculated with log-rank tests. The range values for the hazard ratios are 95% confidence intervals.