Transite: A computational motif-based analysis platform that identifies RNA-binding proteins modulating changes in gene expression

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

[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 02139, USA, [3]David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 500 Main Street, Cambridge, MA 02139, 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 and [8]New York Genome Center, 101 Avenue of the Americas, New York, NY 10013, USA

*To whom correspondence should be addressed. Tel: +1 617 452 2103; Fax: +1 617 452 4978; Email: myaffe@mit.edu. Correspondence may also be addressed to Ian G. Cannell. Tel: +1 646 977 7241 ; Email: icannell@nygenome.org

***Abstract*— RNA-binding proteins (RBPs) play critical roles in regulating gene expression by modulating splicing, RNA stability, and protein translation and are frequently the targets of signal transduction pathways that control RBP function through post-translational modifications such as phosphorylation. In response to various stimuli, alterations in RBP function contribute to global changes in gene expression, but identifying which specific RNA-binding protein(s) are responsible for the observed changes in gene expression patterns remains an unmet need. Here, we present *Transite* – a computational approach to systematically infer RBPs influencing gene expression changes through alterations in RNA stability and degradation. Specifically, our approach builds on pre-existing differential gene expression data and performs sequence-based enrichment analysis. By matching the enriched sequences to a compendium of RBP-binding motifs, we can identify potential RBPs responsible for the observed gene expression changes. As an example, we applied Transite to examine RBPs potentially involved in the response of human patients with non-small cell lung cancer to platinum-based chemotherapy, since RBPs have been recently identified as one of the primary classes of proteins influencing the DNA damage response. Transite implicated known RBP regulators of the DNA damage response and identified hnRNPC as a new modulator of chemotherapeutic resistance, which was subsequently validated experimentally. These data show that Transite is a generalizable framework for the identification of RBPs responsible for gene expression changes driving cell-state transitions and adds value to the vast wealth of publicly-available gene expression data. To ensure that Transite is available to a broad range of scientists for routine differential gene expression analysis workflows we have built a user-friendly web interface that is accessible at https://transite.mit.edu.**

I. Introduction

RNA-binding proteins are major modulators of gene expression at the post-transcriptional level, where they control RNA splicing, stability, localization, degradation, and translation [1,2]. For mRNAs, the role of RBPs in modulating global changes in gene expression at both the RNA and protein level becomes particularly important under conditions where new gene transcription is repressed, such as during inflammation, cell stress, and in response to genomic damage [3–5]. In addition, mutations affecting the expression of specific RBPs, or their function, have been implicated in a variety of diseases, including cancer [5–8].

RBPs appear to play an especially critical 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 [9–11]. Recent 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 [12–16]. In addition, emerging evidence from our lab and others has identified RBPs as critical targets of DDR kinases, including both upstream responder kinases such as ATM, ATR and DNA-PK, and downstream effector kinases such as Chk1 and MK2 [12,13,17–19]. 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 DNA damage response (DDR) [20,21]. For example, we found that a key target of the DNA damage-activated MK2 pathway was the RBP hnRNPA0, which was required for maintenance of the G1/S and G2/M checkpoints following cisplatin treatment [22,23]. Furthermore, this finding dictated the response of non-small cell lung cancers (NSCLCs) to chemotherapy in both mouse models and human patients, where the expression levels of two critical hnRNPA0 target RNAs, Gadd45α and p27, predicted the clinical response of mouse and human tumors to platinum therapy. Despite these types of data, and the recent surge of interest in the roles of RBPs in cancer chemosensitivity and resistance [5,9,24], methods for systematic prioritization of RBPs that influence the response to therapy in diverse clinically relevant data sets are lacking.

Motivated by our long standing interest in the DNA damage response, protein kinase signaling and the centrality of RBPs in dictating cell death decisions in response to chemotherapy, we developed Transite. Transite is a computational method that leverages the wealth of publicly available gene expression data to infer RBPs influencing mRNA

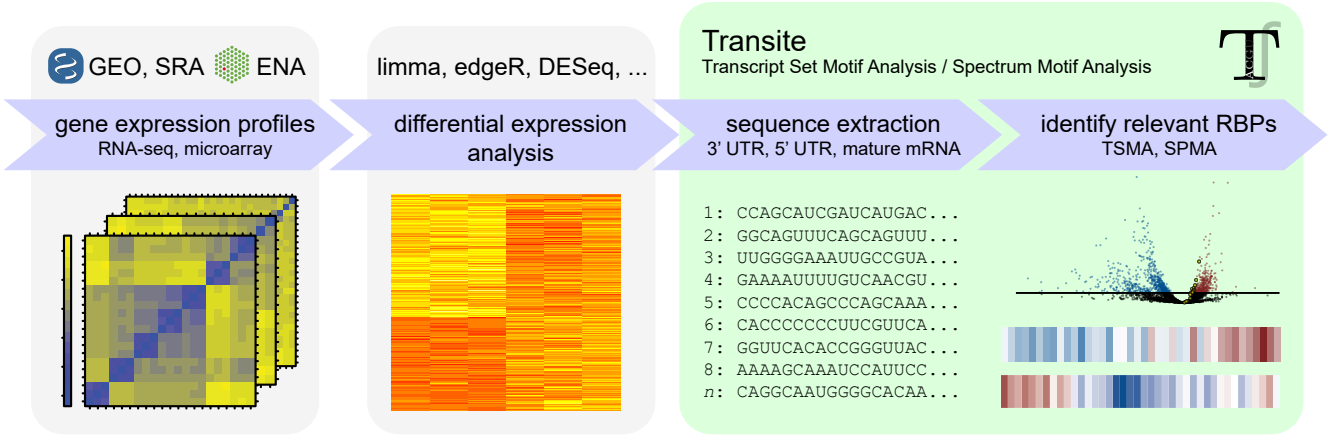


Fig. 1. Schematic of the Transite analysis pipeline: The initial steps of the canonical Transite data analysis workflow include preprocessing and differential expression analysis of gene expression profiles, which are usually obtained from NCBI and EMBL-EBI repositories such as GEO, SRA, and ENA. Differential expression analysis is used to either identify groups of upregulated and downregulated genes (for Transcript Set Motif Analysis) or establish a ranked list of genes from most upregulated to most downregulated (for Spectrum Motif Analysis). Transite then collects the sequences of all genes in the data set and identifies RBPs whose targets concordantly change their expression level.

expression changes through modulation of mRNA stability. Specifically, our method performs motif enrichment analysis on a user-specified region of the mRNA, e.g., the 3′-UTR. By identifying RBPs whose targets are overrepresented in differentially-expressed genes, Transite nominates potential RBP drivers of the observed gene expression changes, a workflow that has been effective in the study of the kinases associated with a particular biological stimulus [25]. A schematic overview of the Transite pipeline is shown in figure 1.

Application of Transite to a NSCLC patient data set of chemoresistant tumors identified hnRNPC as a top putative driver of increased mRNA levels in resistant patients. An orthogonal analysis of hnRNPC-target mRNAs derived from CLIP-Seq data confirmed upregulation of hnRNPC-target mRNAs identified by our *in silico* approach. We then experimentally validated that knock-down of hnRNPC enhanced cisplatin-induced cell death *in vitro*, while hnRNPC overexpression reduced cisplatin induced cell death. Furthermore, high levels of hnRNPC expression were associated with failure to respond to adjuvant platinum-based chemotherapy in an independent NSCLC patient cohort.

Since the data input requirements for using Transite are very general, the approach is not limited to DNA damage or cancer chemotherapy responses, but instead Transite can take advantage of the plethora of publicly available gene expression data sets for any perturbation, and can be used to investigate the effects of RBPs on gene expression using the result of any experiment measuring the expression levels of large numbers of genes simultaneously. Most prominently, these include RNA-Seq, ribosome profiling and microarray experiments. Transite is a versatile tool for inferring the RBPs modulating gene expression and as such, should be a valuable resource for the entire RNA community.

II. Materials and Methods

A. Differential gene expression analysis

The analysis was done with the R/Bioconductor package *limma* [26]. The design matrix for this analysis contains two sample groups, *untreated adenocarcinoma* (10 microarray samples) and *recurrent adenocarcinoma* (15 microarray samples). A linear model was fit to each row of the $\log_2$-transformed expression value matrix, where rows correspond to transcripts and columns correspond to samples. The coefficients of the fitted models describe the differences between the *untreated adenocarcinoma* and *recurrent adenocarcinoma* groups. An empirical Bayes method was used to obtain the significance and the strength of the log fold change between sample groups for each transcript [27], resulting in an estimate of the fold-change analysis between groups and the significance of the change. Raw p-values were adjusted using the Benjamini-Hochberg procedure [28].

B. Motif databases

Transite incorporates sequence motifs of RBP binding sites from two databases: CIS-BP, the Catalog of Inferred Sequence Binding Preferences [29], and RBPDB, a database of RNA-binding specificities [30]. Together these contribute 174 sequence motifs of varying lengths (between six and 18 nucleotides). All motifs were obtained using *in vitro* techniques for determining RNA targets. The majority of motifs was determined by either systematic evolution of ligands by exponential enrichment (SELEX) [31] or RNAcompete [32]. The RNA binding specificities of two further RBPs were obtained by electrophoretic mobility shift assays (EMSA) [33].

C. 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 PSSM is based, is too small to contain at least one occurrence of each nucleotide per position. In this case, pseudocounts are introduced [34].

D. CLIP-seq data analysis

The BED files (output from Piranha analysis) for all CLIP-Seq data sets were downloaded from CLIPdb (http://lulab.life.tsinghua.edu.cn/clipdb/). 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 data set GSE7880. To generate the non-targets list, the entire mature mRNA list was subtracted from the GSE7880 list.

E. Package and web development

R package development and documentation was streamlined with *devtools* and *roxygen2*, respectively. Core algorithms were implemented in C++. *ggplot2* [35] 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 serves as an R wrapper for the Twitter Bootstrap HTML/CSS/JavaScript components.

F. Cell culture and colony formation assays

T6a (mouse lung adenocarcinoma) cells were grown in RPMI-1640 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 [22]. Briefly, 48 hours after transfection with siRNAs or pcDNA vectors, cells were treated with either 4 or 8 μM cisplatin or vehicle for 4 hours. Cells were then 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 either SYTO 60 (Thermo Fisher Scientific) or modified Wright-stain (Sigma-Aldrich). Colonies were scanned and counted using Odyssey® CLx Imaging System (LI-COR Biosciences).

G. siRNA transfection

Silencer select siRNAs (Ambion) transfection was performed using RNAiMax following manufacturer instructions (Life Technologies) with a final concentration of 5 nM. Cells were then treated as described in the previous section.

H. Overexpression of hnRNPC

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 (Life Technologies). cDNAs were used as templates in PCR reactions using PfuUltra II HF DNA polymerase (Agilent) and the following primers: 5′-GCCCAT**AAGCTT**ATG-GACTACAAAGACGATGACGACAAGGCTAGCAAT-GTTACCAACAAGACAGATCCTCGG-3′ (forward) and 5′-GCCCAT**TCTAGA**TTATTAAGAGTCATCCTCCCCA-TTGGCGCTGTCTCTG-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 sub-cloned into pcDNA3.1 vectors. The integrity of the plasmids were confirmed by sequencing (Eton Bioscience Inc).

I. Immunoblotting

Cells were harvested 24 (siRNA transfected) or 48 (pcDNA vectors transfected) hours after cisplatin treatment and re-plating. Cells were then lysed in RIPA buffer and subjected to standard SDS/PAGE electrophoresis and transferred to nitrocellulose membranes. The membranes were immunoblotted with hnRNPC (ab10294, Abcam Inc., Cambridge, MA) and γ-tubulin (Sigma-Aldrich) following manufacturers instructions.

III. Results

RNA-binding proteins (RBPs) influence all stages of the mRNA life cycle through specific interactions involving short linear sequence motifs containing 6 - 8 nucleotides within their target RNAs [36]. The identity of these RBP-binding motifs has been determined for a subset of all known RBPs using various *in vitro* based oligonucleotide selection methods such as SELEX [31], RNAcompete [32] and Bind-n-Seq [37], and directly confirmed for a smaller set of RBPs through experimental analysis of RBP-RNA interactions using CLIP-Seq and various extensions thereof. These latter techniques are laborious and costly to perform, and the direct experimental identification of the complete set

of RNA targets for most RBPs is therefore not yet known. Furthermore, identification of these RNA targets are likely to differ depending upon the experimental situation under which CLIP-Seq was performed. This has limited our ability to understand which RBPs are critical mediators of changes in RNA levels in pre-existing gene expression datasets like those contained in the gene expression omnibus (GEO), including cancer-relevant datasets that describe treatment responses.

In order to systematically mine these existing data sets for RBPs that may influence gene expression changes, with a focus of patient response to therapy, we have developed Transite. Transite takes either a discrete set of differentially expressed genes, or a continuous list of genes, and searches for enriched short linear oligonucleotide motifs within specific regions of the transcripts they encode, and then matches these motifs to the likely RBP that binds them using a compendium of RBP motifs. By default, Transite uses 3′-UTR sequences, as our major focus is mRNA stability and motifs that determine mRNA stability are known to generally reside within the 3′-UTR. Gene sets used for Transite analysis can be defined in either a discrete or a continuous fashion. For discrete sets of genes we implement Transcript Set Motif Analysis, which takes the predefined sets of transcripts such as upregulated and downregulated transcripts and performs motif analysis based on systematic differences between these sets and the total gene expression data. For continuous collections of genes we developed Spectrum Motif Analysis, which uses a continuous quantity to establish an ordered ranking of transcripts and analyzes motif enrichment along that ordered list of transcripts, similar to the approach taken by Gene Set Enrichment Analysis [38]. For this continuous quantity, a measure of differential expression is commonly used, such as fold change or signal-to-noise ratio, thus exploiting information across the entire spectrum of changes rather than limiting analysis to the up- and down-regulated extremes.

A. Transcript Set Motif Analysis identifies RBPs with substrate sites enriched or depleted among differentially expressed genes

Transcript Set Motif Analysis (TSMA) can identify the overrepresentation or underrepresentation of putative binding sites of 174 RNA-binding proteins in a set (or sets) of transcripts, i.e., the foreground set, relative to the entire population of transcripts measured in an experiment. The latter is called background set, and is a proper superset of the foreground sets.

Foreground sets are proper subsets of the background set and their definition depends on the desired motif analysis approach. In any case, foreground and background sets define the groups of transcripts wherein the overrepresentation and underrepresentation of putative RBP binding sites is investigated.

When gene expression data is used, the two foreground sets for TSMA are usually the statistically significantly upregulated and downregulated transcripts. (Figure 2A). Various deviations of this canonical use of foreground sets are possible. Upregulated and downregulated sets can be defined in various ways, depending on the method of differential expression analysis [39]. It is not even necessary to use gene expression data to define foreground and background transcripts. For example, all human or all murine genes associated with a certain Gene Ontology (GO) [40] term could be compared to all genes of human or mouse, respectively, which are annotated with at least one GO term, to identify RBPs associated with particular ontological terms.

Two different methods are used to assign transcript targets to specific RBPs, *k*-mer-based TSMA and matrix-based TSMA.

B. k*-mer-based TSMA*

In the *k*-mer based approach of TSMA the sequence motifs recognized by each RBP are specified by lists of RBP-specific hexamers or heptamers, collated from current motif databases [29,30].

After foreground and background sets are defined and the preferred sequence region is selected (3′-UTR, 5′-UTR, or complete mature mRNA including the coding region), the sequences of both sets are broken down into overlapping hexamers (i.e., *k*-mers of length 6) (Figure 2B, left column), and for each *k*-mer its frequency in the foreground set and background set is determined. While Transite supports both hexamer- and heptamer-matching, hexamers are recommended, since run-time increases exponentially with *k* and the results for heptamers mirror those for hexamers in our experience.

1) k*-mer enrichment values:* The enrichment value of *k*-mer i, e_i, is 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 foreground and background set and n_F and n_B are the total counts of *k*-mers in the foreground and background, respectively.

2) Significance of k*-mer enrichment values:* The statistical significance of the enrichment for all possible *k*-mers is then determined. First, a contingency table C_i for *k*-mer i is defined as

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

Then, the p-value p_i for C_i is approximated with Pearson's χ^2 test. If $p_i < 5\alpha$, p_i is replaced by the p-value obtained by Fisher's exact test for C_i. This step-wise procedure reduces computation time dramatically (approximately 50-fold), because the computationally expensive Fisher's exact test is only used in cases where the approximate p-value from the computationally inexpensive χ^2 test is close to the decision boundary (α) and is avoided in cases where a precise p-value is unnecessary. Furthermore, Fisher's exact test is always used if at least one of the expected counts is less than five, because this constitutes a violation of the assumptions of the approximate test. The p-values are subsequently adjusted for

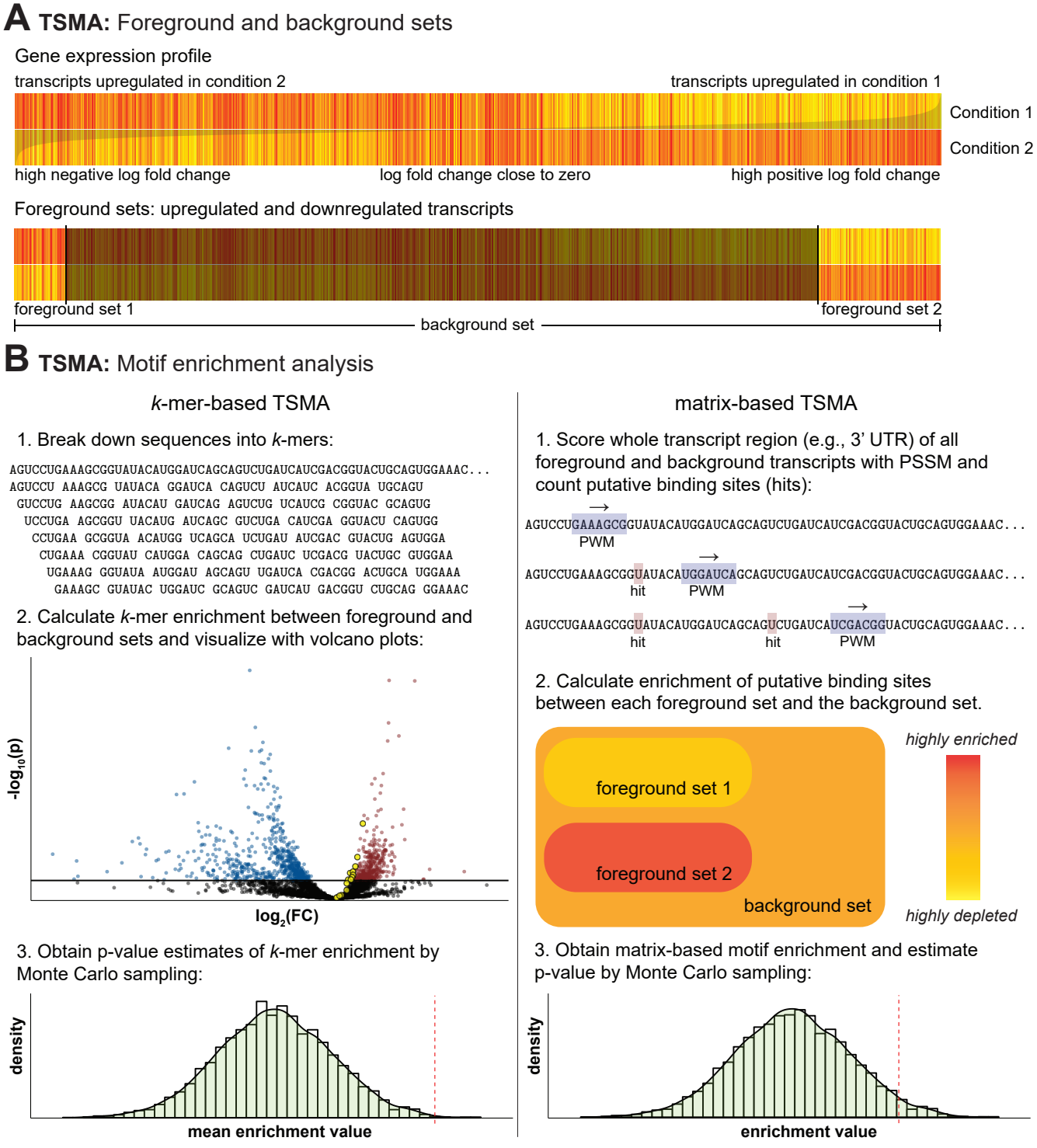


Fig. 2. Schematic figure of Transcript Set Motif Analysis. (**A**) The foreground sets in TSMA are usually defined by differential gene expression analysis of RNA-seq or microarray data. In this case, the foreground sets are most naturally defined as statistically significantly upregulated and downregulated genes, whereas the background set is all genes of the microarray platform or all measured genes in RNA-seq. In the heatmap of the gene expression profile in panel **A**, 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 then estimates the enrichment or depletion of putative binding sites between each foreground set and the background set. There are two ways to describe 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).

multiple hypothesis testing. The available p-value adjustment methods are described in section 5 of the supplement.

3) k-mer volcano plots: Volcano plots are then used to visualize the enrichment values (x-coordinate, log transformed) and associated p-values (y-coordinate, log transformed and multiplied by -1) for all 4^k *k*-mers in a Transite TSMA run (Figure 2B, left column, step 2). The black dots represent *k*-mers without significant enrichment or depletion, while blue dots denote significantly depleted and red dots significantly enriched *k*-mers. Yellow *k*-mers are part of the RBP motif.

4) RBP assignment by average of k-mer enrichment values: As a way to quantify the overrepresentation (or underrepresentation, respectively) of putative binding sites for a particular RBP, the geometric mean of the enrichment values of all *k*-mers associated with that RBP is used. Monte Carlo tests (permutation tests) are used to obtain an estimate of their significance (see section 2 in the supplement for details on Monte Carlo sampling). In step 3 of figure 2, panel B, an example histogram is shown, which depicts the empirical null distribution of mean enrichment values associated with an RBP's *k*-mers after repeated random selection of foregrounds from the background. The dashed red line denotes the mean enrichment value of motif-associated hexamers (which are the yellow dots in the volcano plot in step 2) which were actually observed in the true, unpermuted foreground.

C. Matrix-based TSMA

In the matrix-based TSMA approach, the sequence motifs of 174 RBPs are represented as PWMs. Next, all sequence positions in all transcripts in foreground and background gene sets are scored by these PWMs, as shown in step 1 of the right column of panel B, figure 2. The PWM slides along the sequence, assigns a score to each position, and scores above a certain threshold are considered putative binding sites (*hits*). These hits are tallied in both the foreground and the background set and enrichment values and associated p-values are calculated analogously to the *k*-mer-based approach. Again, all p-values are multiple testing corrected.

A 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 is the possibility of detecting clusters of putative binding sites. This can be done by counting regions with many hits using positional hit information or by simply applying a hit count threshold per sequence, e.g., only sequences with more than some number of hits are considered. Homotypic clusters of RBP binding sites may play a similar role as clusters of transcription factors [41].

D. Spectrum Motif Analysis identifies RBPs with nonrandom arrangement of substrate 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 substrate sequences are statistically significantly enriched among a foreground defined using a collection of either the most or least differentially-regulated genes. As an alternative method, we present Spectrum Motif Analysis (SPMA), which is an effort to more broadly and generally identify non-random distributions of RBP substrate sequences 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-value less than or equal to 0.05) to assign transcripts to a single foreground set, SPMA subdivides the entire list of rank-ordered transcripts into a number of foreground sets (bins) of equal width, calculates enrichment scores for *k*-mers or PWM motifs in each bin as described above, and then searches for non-random bin-wise assortment of *k*-mer or matrix hit frequencies associated with individual RBPs.

SPMA thereby helps to illuminate the relationship between RBP binding evidence (putative binding site enrichment) and the transcript sorting criterion (e.g., fold change between *treatment* and *control* samples). Figure 3A illustrates how the sorted list of transcripts is divided into bins.

1) Spectrum plots: The results of SPMA are displayed as spectrum plots, compact graphical representations of the distribution of putative binding sites for a single RBP, across a range of transcripts (which are sorted in a meaningful way). A spectrum plot visualizes putative binding site enrichment or depletion, and associated p-values, for an RBP motif across the spectrum of transcripts. Spectrum plots are one-dimensional heatmaps, where red-blue coloring encodes the putative binding site enrichment values and the columns are the individual bins of transcripts. Significance levels are indicated by one, two, or three asterisks (p-value less than or equal to 0.05, 0.01, and 0.001, respectively). Examples of spectrum plots are shown in panels B and C of figure 3.

2) Spectrum plot classification: SPMA generates one spectrum plot for each RBP motif in the motif database. With 174 motifs currently available, it is imperative to provide a means to aid in the identification of biologically meaningful spectrum plots that exhibit non-random patterns. A typical non-random pattern is shown in the first spectrum plot in figure 3C, where the enrichment values are observed to positively correlate with the sorting criterion. This type of positive linear relationship between RBP motif enrichment values and sorting criterion might arise in a situation where transcripts are sorted according to their fold change between treatment and control groups and the target transcripts of the RBP are collectively upregulated in the treatment group, perhaps via stabilization by the RBP itself. Transite further aids the user in the process of identifying spectrum plots with a meaningful pattern—one that might be indicative of an underlying biological process—by separating them from spectrum plots with more random distributions of motif enrichment, which are more likely to occur by chance. Each spectrum plot is automatically labeled either *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. For (1), polynomial regression models

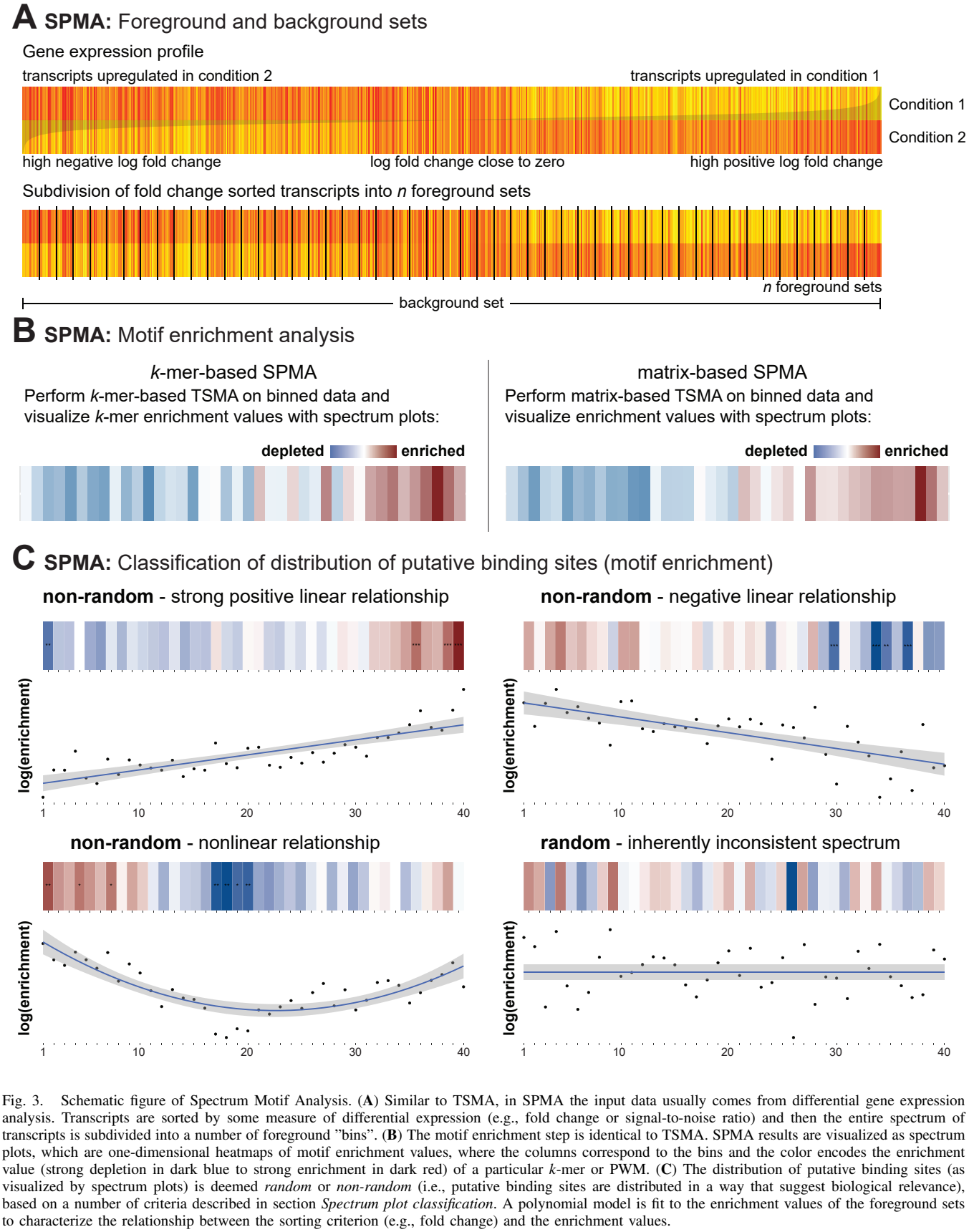


Fig. 3. Schematic figure of Spectrum Motif Analysis. (**A**) Similar to TSMA, in SPMA the input data usually comes from differential gene expression analysis. Transcripts are sorted by some measure of differential expression (e.g., fold change or signal-to-noise ratio) and then the entire spectrum of transcripts is subdivided into a number of foreground "bins". (**B**) The motif enrichment step is identical to TSMA. SPMA results are visualized as spectrum plots, which are one-dimensional heatmaps of motif enrichment values, where 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 a number of criteria described in section *Spectrum plot classification*. A polynomial model is fit to the enrichment values of the foreground sets to characterize the relationship between the sorting criterion (e.g., fold change) and the enrichment values.

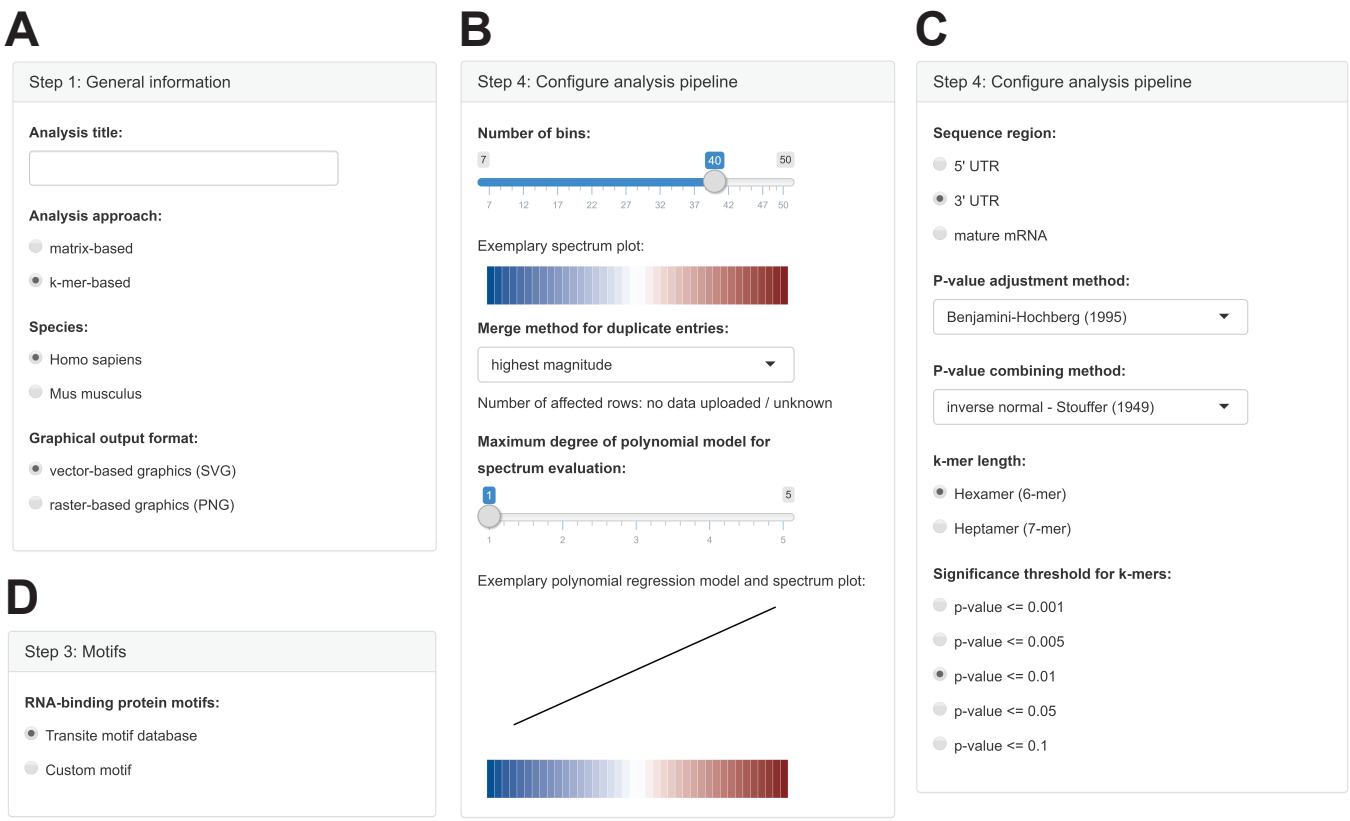


Fig. 4. The Transite website analysis submission forms for TSMA and SPMA make the Transite functionality accessible to scientists outside the R community. (**A**) Users of the website can submit TSMA and SPMA jobs in five simple steps, which include the specification of foreground and background sets and additional optional parameters. (**B**) For SPMA, users are asked to upload a file that contains two columns, an identifier column that holds RefSeq identifiers or gene symbols, and a value column, which is used to sort the transcripts (e.g., by fold change). Shown here is the configuration panel for SPMA. (**C**) Part of the *k*-mer-based TSMA submission form, where sequence region, *k*-mer length, and other parameters can be specified. (**D**) The website supports analysis runs with the Transite motif database as well as with user-defined motifs, where both PWMs and lists of hexamers and heptamers are supported.

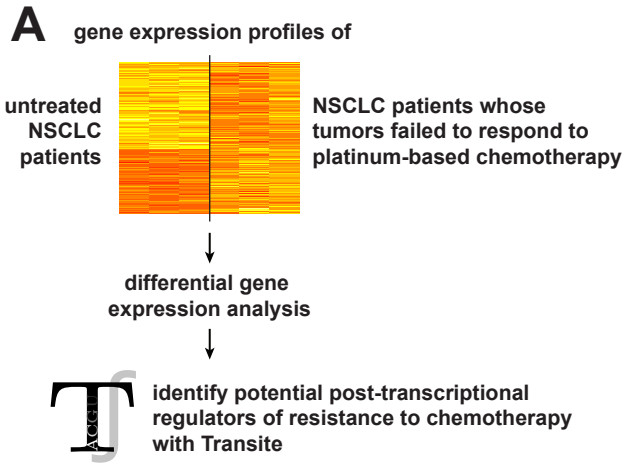
of various degrees are fitted to the spectrum of enrichment values, and the model that best reflects the true nature of the data is selected by means of the F-test. Models with positive and negative coefficients of the linear term (depicting increasing and decreasing linear relationships) are illustrated in the first two examples of figure 3C, respectively (see section 6.2 of the supplement for details on the polynomial model approach). With approach (2), 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 middle bin, for each position in the spectrum. The lower the score, the more consistent the trend in the spectrum plot (see section 6.1 of the supplement for a formal definition of the local consistency score and section 2 for details on the Monte Carlo sampling procedure of the null distribution of the score). Spectrum plots are classified as non-random if (1) the adjusted R^2 of the polynomial fit is greater than or equal to 0.4, and (2) the p-value associated with the local consistency score is less than or equal to $5 * 10^{-6}$, and (3) at least 10% of the bins have significant ($\alpha = 0.05$) enrichment or depletion of putative binding sites.

E. Transite R package and website

To make gene expression dataset analysis for assigning putative RBPs widely available to the scientific community, the Transite analysis platform is hosted at https://transite.mit.edu. Both Transcript Set Motif Analysis and Spectrum Motif Analysis are available with customizable user-friendly forms and familiarity with the R programming language is not required (Figure 4). The full functionality of Transite is also provided as an R/Bioconductor package to ensure a seamless integration into existing bioinformatics workflows. The source code of the Transite package is hosted on GitHub. Both website and R package support 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 motif in a discrete set of genes or a rank ordered list.

F. Transite identifies RBPs known to be involved in the DNA damage response

As an application of Transite-based RBP scoring, we analyzed a data set of non-small cell lung cancer (NSCLC) patients at diagnosis or at recurrence after cisplatin-based



B

RBP	R^2	slope	p-value
ELAVL1, ELAVL3	0.841	0.994	0.000002
TIA1	0.804	0.868	0.000002
RALY	0.804	0.868	0.000002
HNRNPC	0.801	0.800	0.000003
HNRNPCL1	0.801	0.800	0.000002
TIA1	0.798	0.764	0.000002
CPEB3, CPEB2	0.787	0.644	0.000003
U2AF2	0.783	0.613	0.000002
PCBP2-4	0.729	-0.695	0.000241
CPEB3, CPEB4	0.727	0.877	0.000002

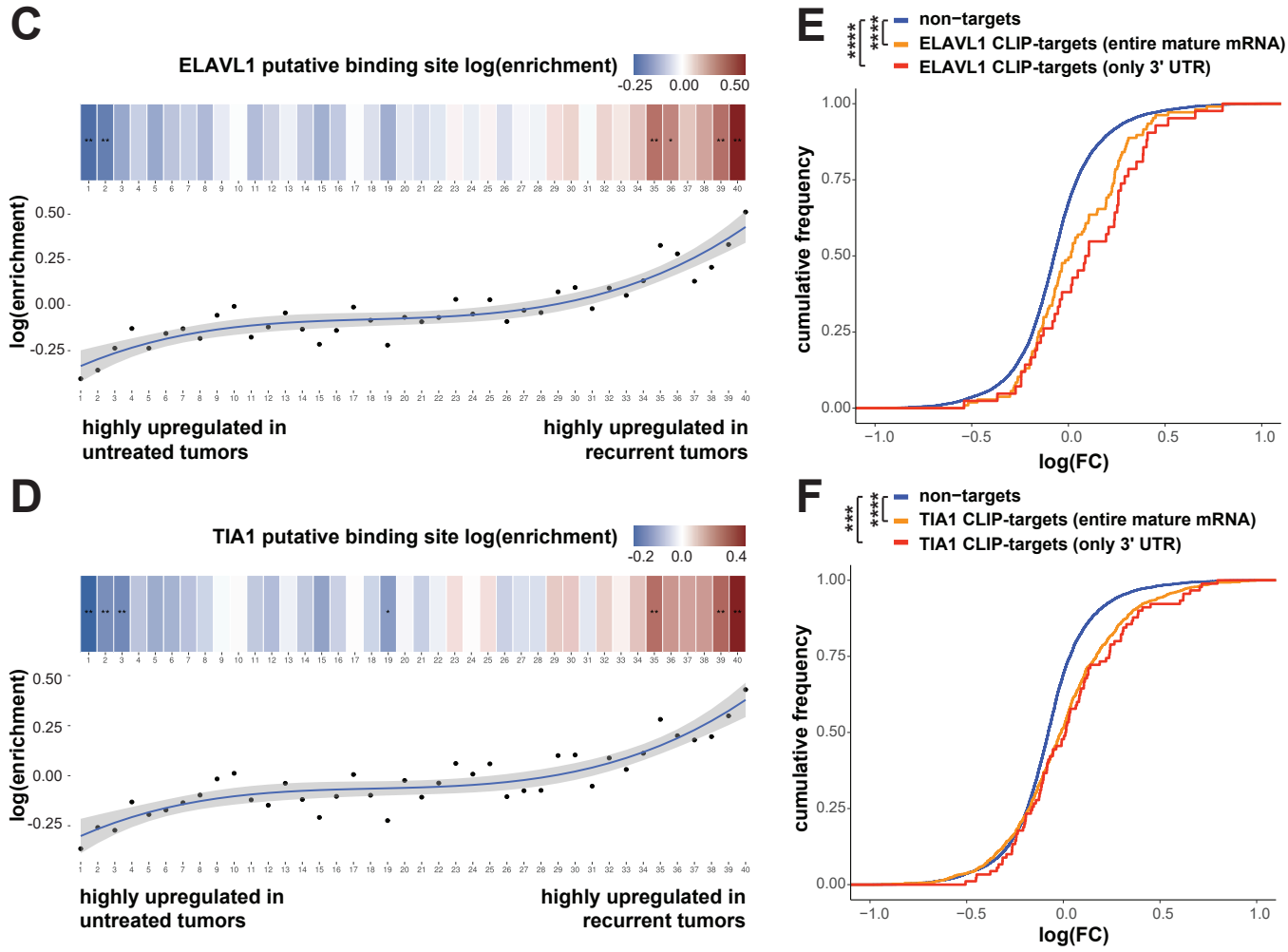


Fig. 5. SPMA identifies ELAVL1 and TIA1 motifs as highly enriched in recurrent NSCLC patients. (**A**) Differential gene expression analysis was performed on samples of patients with untreated NSCLC tumors and patients with recurrent tumors. (**B**) Transite was used to identify RBPs whose targets were overrepresented among upregulated genes in samples of recurrent tumors. Shown is a table of *k*-mer-based SPMA showing RBPs with highly non-random motif enrichment pattern. Among the top hits are ELAVL1, TIA1, and hnRNPC. (**C**) Spectrum plot from SPMA depicting the distribution of putative ELAVL1 binding sites across all transcripts. The transcripts are sorted by ascending signal-to-noise ratio. Transcripts downregulated in resistant samples relative to untreated samples are on the left, and those upregulated are on the right of the spectrum. 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 same transcript order as in panel C. (**E**) Enrichment of ELAVL1 targets in resistant NSCLC cells is recapitulated in an independent 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 one-sided Kolmogorov-Smirnov test. (**F**) As in panel E, transcripts with TIA1 binding sites are upregulated in resistant cells according to an iCLIP experimenmt, confirming results from SPMA.

chemotherapy (GEO Series accession GSE7880) in order to prioritize RBPs that may influence the response of these patients to chemotherapy. We ranked the gene expression changes between pre-treatment and recurrent patients based on signal-to-noise ratio where transcripts present in higher quantities in recurrent patients had positive values and those upregulated in naive patients had negative values (see Figure 5A for schematic). The results of *k*-mer-based and matrix-based TSMA and SPMA (including all 174 motifs) are available as sample output on the Transite website, *About* page, *Example output* section. Ranking the Transite output of both SPMA runs by adjusted R squared (ARS) revealed ELAVL1 and TIA1 to be among the top RBPs predicted to be associated with the 3′-UTRs of transcripts upregulated in recurrent patients relative to untreated patients (Figure 5B shows top 10 RBPs in *k*-mer-based SPMA). Individual spectrum plots for ELAVL1 (Figure 5C) or TIA1 (Figure 5D) demonstrated consistent behavior of these motifs across the gene expression continuum, being enriched in 3′-UTRs of genes that are up in recurrent patients and depleted in 3′-UTRs of genes that are up in naive patients. Importantly, upregulation of ELAVL1 and TIA1-target mRNAs was further validated by analyzing the distribution of known CLIP-Seq identified targets [42,43] for these two RBPs (Figure 5E and 5F). Moreover, both ELAVL1 [44] and TIA1 [45] are known to be involved in the DNA damage response. The fact that two well-known players in the DNA damage response are among the top hits of the motif analysis provides confidence that Transite's predictions are likely to reflect bona fide regulators of the DNA damage response and drivers of chemoresistance. Although CLIP-Seq-defined target mRNAs remain the gold standard for known RBP targets, there are very few RBPs that have been subjected to extensive CLIP-Seq analysis. In the absence of this data, Transite currently utilizes information about putative binding sites from 174 motifs, covering 142 distinct RBPs. Therefore, Transite presents the possibility of identifying RBPs whose true targets as identified by CLIP-Seq and related methods are currently unknown, thus nominating novel RBPs as putative modulators of chemoresistance or other biological processes, in order to prioritize those RBPs for further analysis by CLIP.

G. Motif analysis of the non-small cell lung cancer response to cisplatin treatment identfies hnRNPC as a potential modulator of resistance to chemotherapy

The data in Figure 5 shows that Transite analysis can identify known RBPs involved in the DNA damage response. We were particularly interested in using Transite as a tool to discover new biology related to the DDR in data from human clinical trials. We therefore focused on hnRNPC, one of the highest-scoring RBPs that emerged from our analysis of chemoresistant NSCLC patients, and has not to our knowledge been strongly implicated in the response to chemotherapy-induced DNA damage [46]. As shown in Figure 6A, 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 with tumor recurrence after platinum therapy. This Transite prediction was independently confirmed by analysis of iCLIP-defined target mRNAs [47], which also showed an overrepresentation of hnRNPC targets in upregulated transcripts in recurrent patients (Figure 6B), with those with binding in the 3′-UTR showing the strongest enrichment.

H. hnRNPC modulates sensitivity to cisplatin

To experimentally validate these Transite results, we examined the effect of knockdown or over-expression of hnRNPC on sensitivity to cisplatin treatment in T6a murine lung carcinoma cells. Colony formation assays in T6a cells demonstrated that hnRNPC over-expression promoted resistance to cisplatin as evidenced by a 1.6 fold increase in the number of surviving colonies (Figure 6C, red bar). Conversely, siRNA-downregulation of hnRNPC significantly enhanced T6a cell sensitivity to cisplatin as evidenced 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 6C, blue bar). These data indicate that hnRNPC is a key player in mediating resistance of NSCLC cells to chemotherapy, and demonstrate 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 a unique adjuvant chemotherapy trial, JBR.10 (Figure 6D). In this trial, early stage NSCLC patients had their tumors surgically resected and subjected to gene expression profiling. Patients were then randomized to receive cisplatin / vinorelbine doublet chemotherapy or observation and palliative care [48](GSE14814), allowing us to specifically assess the role of hnRNPC in the response to chemotherapy. We focused our analysis on stage 2 patients, as their benefit from adjuvant chemotherapy is most pronounced. Separation of patients based on hnRNPC expression revealed that patients whose tumors display low expression of hnRNPC benefited significantly from chemotherapy in terms of survival (Figure 6D, right panel, $p = 0.019$) while patients whose tumors have high hnRNPC expression did not benefit (Figure 6D, left panel, $p = 0.68$). Together these data define hnRNPC as an important new RBP involved in the chemotherapeutic response in NSCLC and suggest that Transite is a highly effective tool to pinpoint novel RBPs that drive chemoresistance in human cancer patients.

IV. Discussion

Despite their crucial role in post-transcriptional regulation of gene expression, the majority of RNA-binding proteins (RBPs) have unknown functions. To help understand the influence of RBPs on their target transcripts, we developed Transite, a novel 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

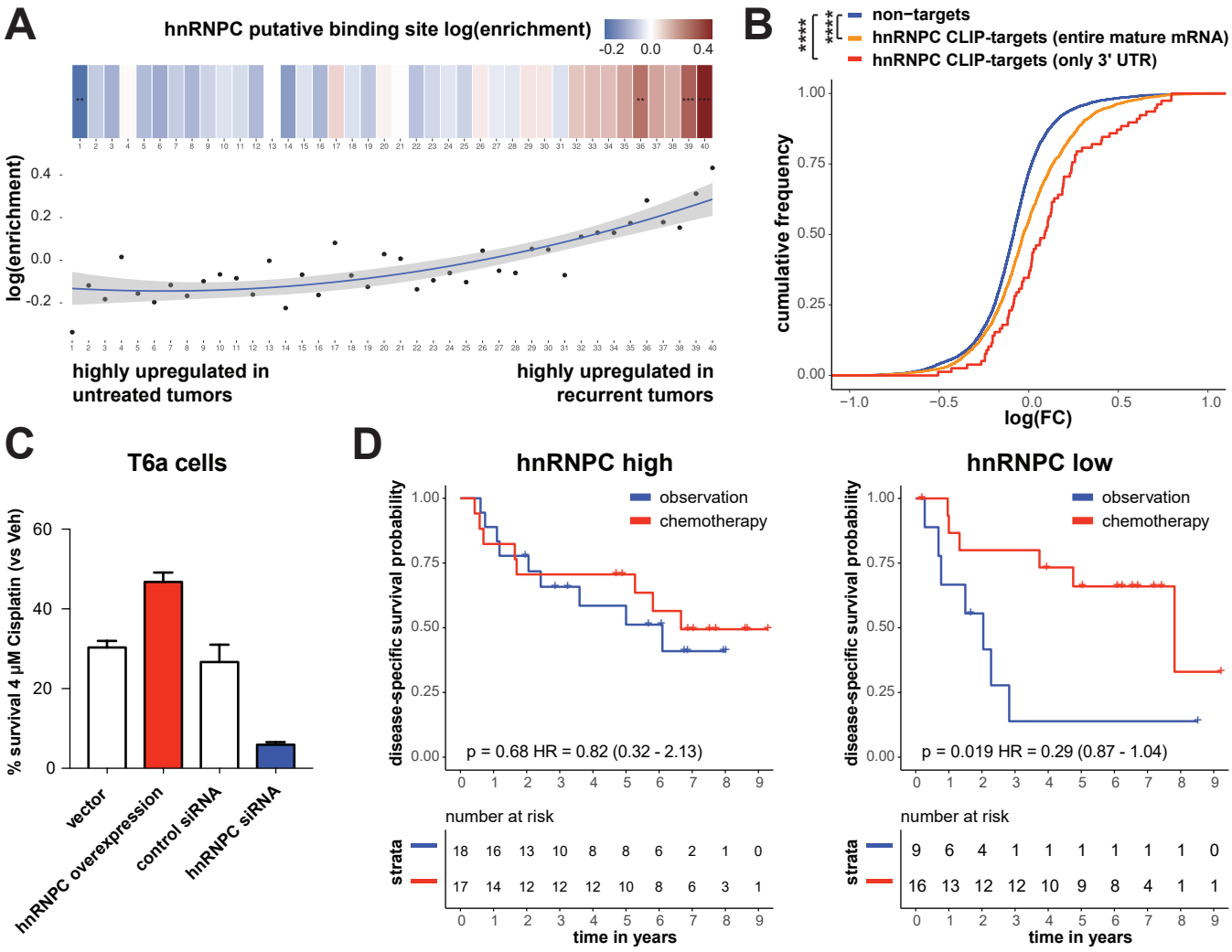


Fig. 6. hnRNPC modulates sensitivity to cisplatin. (**A**) Spectrum plot from *k*-mer-based SPMA depicting the distribution of putative hnRNPC binding sites across all transcripts. The transcripts are sorted by ascending signal-to-noise ratio from lowest to highest abundance in resistant relative to untreated samples. Putative hnRNPC binding sites are highly enriched in the upregulated fraction of transcripts. (**B**) Enrichment of hnRNPC binding sites in upregulated transcripts is independently confirmed by CLIP experiments. The p-values were calculated with the one-sided Kolmogorov-Smirnov test. (**C**) siRNA-mediated reduction in hnRNPC levels significantly impairs long-term survival of T6a cells in response to cisplatin (blue bar). Overexpression of hnRNPC (red bar) protects against cisplatin-induced cell death of T6a cells in colony formation assays. Bar graphs represent percent number of colonies formed, normalized to untreated control cells. White bars represent control cells transfected with control vehicles (control siRNA or empty pcDNA). Error bars represent standard deviation among 3 replicates. (**D**) High expression of hnRNPC impedes the efficacy of platinum-based chemotherapy in patients with stage 2 disease from the JBR.10 lung cancer adjuvant chemotherapy trial (GSE14814). The p-value was calculated with the log-rank test (HR is Hazard Ratio). hnRNPC low group = patients with hnRNPC expression Z-scores of less than or equal to -0.2, and hnRNPC high group = patients with hnRNPC expression Z-scores greater than or equal to 0.2.

sequences whose over-representation can be computed from gene expression data, and that a large collection pre-existing motif data for RBPs has been compiled in publicly available databases [29,30].

It is important to note that not all RBPs have strong motif preferences, and that there may be considerable redundancy in motif recognition by multiple RBPs. Furthermore, their *in vitro*-derived motifs may not always reflect motifs derived from *in vivo* binding analysis. These caveats have raised questions about the ability of consensus motifs and PWMs to accurately predict RBP targets a priori on a genome-wide scale, and have led to the development of more sophisticated approaches for predicting specific RBP RNA targets [49,50]. In contrast, Transite does not make any specific RBP RNA target predictions, and instead 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 process.

By using two 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 mo-

tifs derived from *in vitro* sequence libraries. These findings suggest that, although there are limitations to utilizing *in vitro*-derived motifs, Transite serves as an excellent 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 data sets.

To further demonstrate the utility of Transite, we performed an analysis of human NSCLC patient data we were able to recover previously-known biology and also identify novel sources of chemoresistance. Well-known players in the DNA damage response such as ELAVL1 and TIA1 were among the top hits in the tumor resistance gene expression data set, showing that our approach is consistent with previous DNA damage response 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 resistance of NSCLC cells to chemotherapy, which was independently validated in an additional NSCLC patients data set.

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 types of data users may utilize Transite to analyze are: (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 sub-cellular compartments. (3) *de novo* motif analysis in the entire mRNA of gene expression changes upon knockdown of a nuclease of unknown function. These are just a few examples of the versatility of Transite. The Transite website (https://transite.mit.edu) makes this tool accessible to a broad group of scientists and provides insight as to how key post-transcriptional regulators contribute to the concerted regulation and function of specific cellular processes. With Transite, the large body of gene expression data from microarray and RNA sequencing experiments can be further leveraged to identify changes in mRNA expression associated with specific RBPs. In this way, hypotheses can be generated regarding which RBPs interact preferentially with mRNAs that are specific to a particular condition.

Availability

The Transite website is available at https://transite.mit.edu. For workflow integration and advanced analysis, the Transite functionality is also offered as an R/Bioconductor package at https://www.bioconductor.org. The Transite source code is hosted on GitHub (https://github.com/kkrismer/transite).

Acknowledgements

We wish to thank all members of the Yaffe and Hemann labs for helpful advice and discussions. Additionally, we thank Anne E. van Vlimmeren for feedback on the manuscript.

Funding

This work was supported by scholarships of the Marshall Plan Foundation and the Austrian Federal Ministry for Education (to K.K., A.G., and T.B.), National Institutes of Health (NIH) grants R01-ES015339, R35-ES028374, U54-CA112967, the Charles and Marjorie Holloway Foundation, the and a Starr Cancer Consortium Award I9-A9-077 (to M.B.Y. and I.G.C.).

Conflict of interest statement: None declared.

References

[1] Gerstberger, S., Hafner, M., and Tuschl, T. (12, 2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.,* **15**(12), 829–845.

[2] Lunde, B. M., Moore, C., and Varani, G. (Jun, 2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.,* **8**(6), 479–490.

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

[4] 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.

[5] Pereira, B., Billaud, M., and Almeida, R. (07, 2017) RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends Cancer,* **3**(7), 506–528.

[6] Cooper, T. A., Wan, L., and Dreyfuss, G. (Feb, 2009) RNA and disease. *Cell,* **136**(4), 777–793.

[7] Licatalosi, D. D. and Darnell, R. B. (Jan, 2010) RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.,* **11**(1), 75–87.

[8] Lukong, K. E., Chang, K. W., Khandjian, E. W., and Richard, S. (Aug, 2008) RNA-binding proteins in human genetic disease. *Trends Genet.,* **24**(8), 416–425.

[9] Reinhardt, H. C., Cannell, I. G., Morandell, S., and Yaffe, M. B. (Jan, 2011) Is post-transcriptional stabilization, splicing and translation of selective mRNAs a key to the DNA damage response?. *Cell Cycle,* **10**(1), 23–27.

[10] Rieger, K. E. and Chu, G. (2004) Portrait of transcriptional responses to ultraviolet and ionizing radiation in human cells. *Nucleic Acids Res.,* **32**(16), 4786–4803.

[11] Gasch, A. P., Huang, M., Metzner, S., Botstein, D., Elledge, S. J., and Brown, P. O. (Oct, 2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell,* **12**(10), 2987–3003.

[12] Matsuoka, S., Ballif, B. A., Smogorzewska, A., McDonald, E. R., Hurov, K. E., Luo, J., Bakalarski, C. E., Zhao, Z., Solimini, N., Lerenthal, Y., Shiloh, Y., Gygi, S. P., and Elledge, S. J. (May, 2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science,* **316**(5828), 1160–1166.

[13] 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., Meyer, T., and Cimprich, K. A. (Jul, 2009) A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. *Mol. Cell,* **35**(2), 228–239.

[14] Hurov, K. E., Cotta-Ramusino, C., and Elledge, S. J. (Sep, 2010) A genetic screen identifies the Triple T complex required for DNA damage signaling and ATM and ATR stability. *Genes Dev.,* **24**(17), 1939–1950.

[15] 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., Fydrych, A., Ho, R., Greenberger, B. A., Chen, G. C., Maffa, A., Del Rosario, A. M., Root, D. E., Carpenter, A. E., Hahn, W. C., Sabatini, D. M., Chen, C. C., White, F. M., Bradner, J. E., and Yaffe, M. B. (Jun, 2013) The bromodomain protein Brd4 insulates chromatin from DNA damage signalling. *Nature,* **498**(7453), 246–250.

[16] Adamson, B., Smogorzewska, A., Sigoillot, F. D., King, R. W., and Elledge, S. J. (Feb, 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**(3), 318–328.

[17] 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. (Mar, 2007) 14-3-3sigma controls mitotic translation to facilitate cytokinesis. *Nature,* **446**(7133), 329–332.

[18] Fan, J., Yang, X., Wang, W., Wood, W. H., Becker, K. G., and Gorospe, M. (Aug, 2002) Global analysis of stress-regulated mRNA turnover by using cDNA arrays. *Proc. Natl. Acad. Sci. U.S.A.,* **99**(16), 10611–10616.

[19] Kim, H. H., Abdelmohsen, K., and Gorospe, M. (Jul, 2010) Regulation of HuR by DNA Damage Response Kinases. *J Nucleic Acids,* **2010**.

[20] Ciccia, A. and Elledge, S. J. (Oct, 2010) The DNA damage response: making it safe to play with knives. *Mol. Cell,* **40**(2), 179–204.

[21] Jackson, S. P. and Bartek, J. (Oct, 2009) The DNA-damage response in human biology and disease. *Nature,* **461**(7267), 1071–1078.

[22] 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. (Nov, 2015) A Pleiotropic RNA-Binding Protein Controls Distinct Cell Cycle Checkpoints to Drive Resistance of p53-Defective Tumors to Chemotherapy. *Cancer Cell,* **28**(5), 623–637.

[23] 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. (Oct, 2010) DNA damage activates a spatially distinct late cytoplasmic cell-cycle checkpoint network controlled by MK2-mediated RNA stabilization. *Mol. Cell,* **40**(1), 34–49.

[24] Hong, S. (Dec, 2017) RNA Binding Protein as an Emerging Therapeutic Target for Cancer Prevention and Treatment. *J Cancer Prev,* **22**(4), 203–210.

[25] Joughin, B. A., Naegle, K. M., Huang, P. H., Yaffe, M. B., Lauffenburger, D. A., and White, F. M. (Jan, 2009) An integrated comparative phosphoproteomic and bioinformatic approach reveals a novel class of MPM-2 motifs upregulated in EGFRvIII-expressing glioblastoma cells. *Mol Biosyst,* **5**(1), 59–67.

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

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

[28] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological),* **57**(1), 289–300.

[29] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J. C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J., Bouget, F. Y., Ratsch, G., Larrondo, L. F., Ecker, J. R., and Hughes, T. R. (Sep, 2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell,* **158**(6), 1431–1443.

[30] Cook, K. B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T. R. (Jan, 2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.,* **39**(Database issue), D301–308.

[31] Tuerk, C. and Gold, L. (Aug, 1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science,* **249**(4968), 505–510.

[32] Ray, D., Kazan, H., Chan, E. T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B. J., Morris, Q., and Hughes, T. R. (Jul, 2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.,* **27**(7), 667–670.

[33] Garner, M. M. and Revzin, A. (Jul, 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**(13), 3047–3060.

[34] Nishida, K., Frith, M. C., and Nakai, K. (Feb, 2009) Pseudocounts for transcription factor binding sites. *Nucleic Acids Res.,* **37**(3), 939–944.

[35] Wickham, H. (2009) ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, .

[36] Coppin, L., Leclerc, J., Vincent, A., Porchet, N., and Pigny, P. (Feb, 2018) Messenger RNA Life-Cycle in Cancer Cells: Emerging Role of Conventional and Non-Conventional RNA-Binding Proteins?. *Int J Mol Sci,* **19**(3).

[37] Zykovich, A., Korf, I., and Segal, D. J. (Dec, 2009) Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res.,* **37**(22), e151.

[38] 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. (Oct, 2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.,* **102**(43), 15545–15550.

[39] Costa-Silva, J., Domingues, D., and Lopes, F. M. (2017) RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE,* **12**(12), e0190152.

[40] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (May, 2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.,* **25**(1), 25–29.

[41] Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A., and Ovcharenko, I. (May, 2010) Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.,* **20**(5), 565–577.

[42] Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (May, 2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods,* **8**(7), 559–564.

[43] Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N. M., Rot, G., Zupan, B., Curk, T., and Ule, J. (Oct, 2010) iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol.,* **8**(10), e1000530.

[44] Masuda, K., Abdelmohsen, K., Kim, M. M., Srikantan, S., Lee, E. K., Tominaga, K., Selimyan, R., Martindale, J. L., Yang, X., Lehrmann, E., Zhang, Y., Becker, K. G., Wang, J. Y., Kim, H. H., and Gorospe, M. (Mar, 2011) Global dissociation of HuR-mRNA complexes promotes cell survival after ionizing radiation. *EMBO J.,* **30**(6), 1040–1053.

[45] Lal, A., Abdelmohsen, K., Pullmann, R., Kawai, T., Galban, S., Yang, X., Brewer, G., and Gorospe, M. (Apr, 2006) Posttranscriptional derepression of GADD45alpha by genotoxic stress. *Mol. Cell,* **22**(1), 117–128.

[46] Shkreta, L. and Chabot, B. (Oct, 2015) The RNA Splicing Response to DNA Damage. *Biomolecules,* **5**(4), 2935–2977.

[47] Knig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (Jul, 2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.,* **17**(7), 909–915.

[48] Winton, T., Livingston, R., Johnson, D., Rigas, J., Johnston, M., Butts, C., Cormier, Y., Goss, G., Inculet, R., Vallieres, E., Fry, W., Bethune, D., Ayoub, J., Ding, K., Seymour, L., Graham, B., Tsao, M. S., Gandara, D., Kesler, K., Demmy, T., and Shepherd, F. (Jun, 2005) Vinorelbine plus cisplatin vs. observation in resected non-small-cell lung cancer. *N. Engl. J. Med.,* **352**(25), 2589–2597.

[49] Perron, G., Jandaghi, P., Solanki, S., Safisamghabadi, M., Storoz, C., Karimzadeh, M., Papadakis, A. I., Arseneault, M., Scelo, G., Banks, R. E., Tost, J., Lathrop, M., Tanguay, S., Brazma, A., Huang, S., Brimo, F., Najafabadi, H. S., and Riazalhosseini, Y. (May, 2018) A General Framework for Interrogation of mRNA Stability Programs Identifies RNA-Binding Proteins that Govern Cancer Transcriptomes. *Cell Rep,* **23**(6), 1639–1650.

[50] 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.