doi:10.1038/nature12311

File contains supplementary methods, along with Supplementary Tables 1-4 and Supplementary Figures 1-6.

**RNAcompete pool design**

This description is partially redundant with the online methods but adds additional details.

The RNA pool design is related to our previous design[11] except that highly stable RNA stem-loop structures were replaced with larger numbers of unstructured probes. To generate this new probe set, we started with a de Bruijn sequence of order 11 (generated using Linear Feedback Shift Registers[58] with the primitive polynomial $x^{22}+x^{21}+x^{20}+x^{19}+x^{18}+x^{17}+x^{16}+x^{15}+x^{13}+x^{12}+x^{11}+x^{10}+x^{9}+x^{4}+x^{3}+x^{2}$) [Primitive polynomial was downloaded from http://fchabaud.free.fr/English/Poly], and then partitioned it with sliding windows of 35nts, while overlapping by 10 nts to prevent the loss of any 11-mers and prepending each probe with the T7 initiator (AGA or AGG) that forms a less structured probe of length 38nt. This resulted in 167,773 probes. We identified less structured probes using RNAshapes[59] with the option to enumerate all secondary structures with free energies within 70% of the minimum free energy (MFE) with the following call: `RNAshapes -s –c 70.0 –r –M 30 –t 1 –o 2.` We then summed the probabilities of the structures (output by RNAshapes) with free energies less than -2.5 kcal /mol, and used this value to quantify “structuredness”: if this value is larger than 0.5, that probe is classified as “strongly structured”. Based on this, there were 130,936 strongly structured probes and 36,837 weakly structured probes.

We applied a series of strategies to ensure that each 9-mer was represented in a weakly structured context at least 16 times. First, we split each of the strongly structured probes into two equal fragments of length 19nt. Let [i-j] represent the subsequence starting from index i and ending at index j, inclusive. We fixed the prefixes([1-19]) of the probes and tried swapping the suffixes ([20-38]) using a greedy algorithm to match prefixes and suffixes. This succeeded in forming 98,602 weakly structure probes, leaving 32,334 strongly structured. Then, we recombined the fragments [4-19] and [20-38] from two strongly structured probes, and prepended the T7 initiator sequence that results in a less structured probe. This step produced an additional 8,260 weakly structured probes. Third, we merged 16-mers that span the breakpoints of strongly structured probes (8

bases on either side). We were able to merge 107,070 16-mers that resulted in 53,535 weakly structured probes. We combined all the weakly structured probes and calculated the distribution of 9-mer occurrences. For 65,723 9-mers (including repeats) that were represented less than 16 times, we attempted to increase the number of occurrences by merging four 9-mers or three 9-mers into a single probe. For the 9-mers that did not result in a weakly structured probe when merged, we designed probes that each contain one missing 9-mer using RNAinverse (from the Vienna RNA package[60]). The final probe set contained 214,948 weakly structured probes.

Similar to the previous RNAcompete design, we sought two replicate sets for robustness and evaluation purposes. Therefore, we attempted to divide the probe set into two sets (i.e. Set A and Set B) with a balanced distribution of 9-mer occurrences. To do this, we first randomly assigned probes to Set A or Set B, and then greedily swapped individual probes between Set A and B to attempt to correct imbalances in their 9-mers distributions, and continued swapping probes until the 9-mer distributions were as balanced as possible. After this greedy swapping step, Set A had 105,527 probes and Set B had 106,558 probes. Finally, to ensure that each 9-mer appears at least 8 times in any of the sets, we added more probes (3804 for Set A and 3538 for Set B) formed by merging three 9-mers.

Our next step was to remove probes that could lead to microarray cross-hybridization or RNA-RNA interactions in the pool. We ran MegaBLAST (version 2.2.20 with command line parameters(-W 12 -D 3 -g -S 3)) in order to identify matches with at least 14 consecutive bases, or with at least 17 bases with at least 12 consecutive bases, to other sequences in either the forward or reverse orientation. Some probes can match to many other probes because the same set of 9-mers tends to get merged in the same probe when we try to combine three or four 9-mers. We removed the probes that have matches to at least four other probes. For probes with less than four matches, we attempted to disrupt the matches by modifying the two bases in the middle of matching subsequences. Among the 15 (except the original probe from 16 possible modifications) modified probes, we kept the ones that are weakly structured. We also checked for matches between the set of modified probes and the original probe set, and removed the modified probes that have matches to the original probe set. Then, we checked the distribution of 9-mers and designed probes to add missing 9-mers either by merging three 9-mers or designing a single probe for a single 9-mer (using RNAinverse) when merging was not possible. After the addition of these new probes, we re-ran MegaBLAST and repeated the procedure described above. During this iterative process, we also made sure that the Sap1 restriction sites did not appear in newly designed probes. We fixed the probe set once each

9-mer was represented at least 8 copies in each set. There were 109,642 probes in Set A and 110,348 probes in Set B. Since we had more space in the array, we duplicated some of the probes and ended up with 120,326 probes in Set A and 121,031 probes in Set B. Lastly, we added 22 control sequences which are known targets for a set of RBPs. **The final Set A and Set B each contained at least 8 copies of each 9-mer, 33 copies of each 8-mer and 155 copies of each 7-mer.** There remained 2,858 strongly structured probes (containing 9-mers that are self-structured) in the final design.

**Protein cloning**

RBP cDNA inserts were cloned into the multiple-cloning site of pDEST15 based expression vectors, pTH5325[61] and pTH6838 (a derivative of pTH5325 engineered with additional restriction enzyme sites to facilitate cloning), using standard molecular biology techniques. The vector map and sequence for pTH6838 is posted on our Supplementary Data page (http://hugheslab.ccbr.utoronto.ca/supplementary-data/RNAcompete_eukarya/). Primers were designed to amplify DNA corresponding to full-length RBPs and various RBP fragments, based on boundaries defined by Pfam (as described in supplementary section "Derivation of sequence similarity rules and construction of cisBP-RNA"). We initially investigated three types of constructs: (1) full-length proteins; (2) "core" RNA-binding regions (RBRs) which we defined to consist of a contiguous region containing all RBDs in a given RBP; (3) discrete RBDs (e.g. RBD1 and RBD2 etc. in separate constructs, for instances where an RBP contains multiple RBDs). We cloned RBRs and discrete RBDs with either an additional 90 or 150 bp (i.e. 30 or 50 amino acid residues) of respective 5'- and 3'- flanking sequence from corresponding cDNA or RNA templates, as structural studies have demonstrated that amino acids neighboring an RBD can impact RNA-binding affinity and specificity [48, 49]. Preliminary RNAcompete analysis of 62 constructs from a panel of 19 drosophila RBPs indicated that when successful, RBRs and full length RBPs yield comparable RNAcompete data, whereas the majority of discrete RBDs do not pass internal RNAcompete quality control checks. We found the success rate of RBRs in RNAcompete assays to be slightly higher (~1.25-fold) than full-length RBPs, and >4-fold higher than discrete RBDs (**Table S1**). In addition, cloning and purification of RBRs was more reliable and efficient than full-length RBPs. Thus, most of the constructs used in this study contain RBRs. Note that we also used some inserts from collaborators that did not satisfy these guidelines, and that we only included flanking sequence up to the start or end of the annotated coding region of the protein. The sequences of all inserts and their source are compiled in **Supplementary Data 2.**

## RNAcompete assay

The RNA pool generation, RNAcompete pulldown assays, microarray hybridizations, and microarray data quantification were performed as previously described[11] with the following exceptions: (i) the common 3'-end linker from the dsDNA pool was removed by digestion with BspQI instead of SapI and (ii) GST-tagged RBPs and RNA pool were typically incubated in 1 mL of Binding Buffer (20 mM Hepes pH 7.8, 80 mM KCl, 20 mM NaCl, 10% glycerol, 2 mM DTT, 0.1 μg/μL BSA) containing 20 μL glutathione sepharose 4B (GE Healthcare) beads (washed 3 times in Binding Buffer) for 30 minutes at 4°C, and subsequently washed four times for two minutes with Binding Buffer at 4°C. In some instances, alternative binding and washing conditions were used; these are listed together with individual experiments and hybridizations are listed in **Supplementary Data 2**.

## Normalization of probe intensities

This section is partially redundant with the online methods but adds additional details.

Hybridizations were batched based on whether or not they used the same initial RNA pool because arrays using the same pool tended to require similar normalization. Each batch was represented as a matrix where rows correspond to probes and columns are the pulldown intensities of each RBP profiled in that batch. Note that we treated the red and green channels of the array as separate one colour hybridizations. From this matrix, we set to NaN elements corresponding to probes that we identified by visual inspection whose intensities were affected by spatial trends or image analysis artifacts. Then, to correct for any differences in laser power and to ensure that abundance estimates in each column were in the same scale, we applied a separate global normalization to each column. Specifically, we applied an affine transformation to each column (i.e. we added a bias and rescaled the elements of the column) so that the median and inter-quartile range (IQR) of each column was equal to the median of the column medians and the median of the column IQRs, respectively. To correct for differences in the RNA oligo abundances in the initial RNA pool, we then performed a row normalization. Specifically, we subtracted the row median from each element in the row and then divided by a robust estimate of the standard deviation, which we set equal to 1.4826 times the median absolute deviation of the row. We call this row normalization a robust z-transform. We found – based on visual inspection of motifs and reproducibility of 7-mers scores for the same RBPs within and across batches – that the robust z-transform provided a better correction for differences in the abundances of RNA oligos in the initial pool than dividing by a direct measurement of the oligo abundances from a microarray

(data not shown). As a final normalization, so that we could interpret the normalized probe intensities in a column as z-scores, we performed a robust z-transform on the column.

**Testing stability of RBFOX1 target transcripts by qRT-PCR**

To generate stable cells expressing doxycycline-inducible human RBFOX1, Flp-in™-293 cells (Invitrogen) were co-transfected with the pOG44 Flp recombinase expression vector along with a modified gateway-compatible pcDNA5-FRT-FLAG vector containing human RBFOX1 cDNA (NM_018723), using Lipofectamine 2000 (Invitrogen) transfection reagent. Stable cells were selected with 200 µg/mL hygromycin B for roughly 2 weeks after which stably expressing colonies were pooled.

To test the effects of RBFOX1 on transcript stability, reporter constructs containing the CADPS (NM_003716) 3'-UTR were generated. CADPS 3'UTR sequences (mRNA nucleotide positions 4423-4773), containing either a wild-type (UGCAUG) or mutant (UGAGUC) RBFOX1 site (nucleotide position 4472), were cloned into the unique XbaI site of the pGL4.13 (Promega) mammalian luciferase expression vector.

Stable cells expressing RBFOX1 were plated in 6-well plates. To reduce the potential for RBFOX1-redundant regulators, 24 hours after plating, the cells were transfected with 30 nM of RBFOX2-targeting siRNA (SIGMA-ALDRICH: siRNA ID SASI_Hs01_00242056). After 18 hours, 1 µg/mL of doxycycline was added to half of the cells to initiate RBFOX1 production. Six hours after initiating RBFOX1 expression, cells were transfected with 1 µg of stability reporter along with 250 ng of pmCherry-C1 plasmid as transfection control. 42 hours after plasmid transfection cells were treated with 10 µM Actinomycin D for 6 hours to halt transcription prior to harvest.

Total RNA was extracted from cells using TRI reagent (SIGMA-ALDRICH) and treated with DNaseI (Roche Applied Science). For quantitative qRT-PCR, cDNA was generated using 500 ng of DNaseI-treated total RNA using SuperScriptIII Reverse Transcriptase (Invitrogen). qRT-PCR was performed in a 384-well plate using 20ng of cDNA per reaction and FastStartUniversal SYBR Green Master (Roche Applied Science). Levels of luciferase transcript were normalized to the levels of mCherry transfection control. Primer sequences used for the qRT-PCR reactions are available upon request.

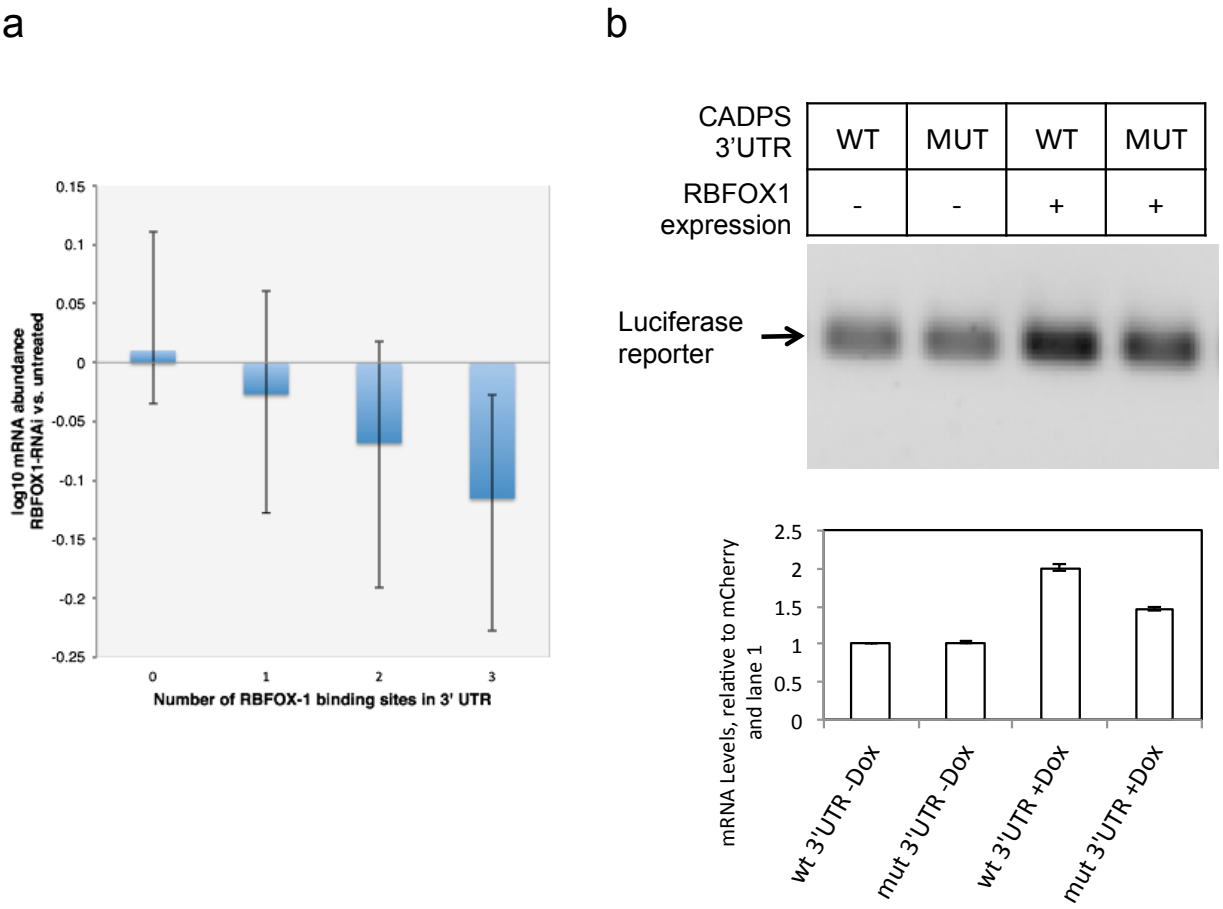



**Figure S1. Data supporting the *in vivo* relevance of individual RBFOX1 binding sites in transcripts.**

**(a)** Relative abundance of RBFOX1 predicted targets in RBFOX1 RNAi data[36]. Transcripts are binned according to the number of sites in the 3'UTR. Error bars indicate 25$^{th}$ and 75$^{th}$ percentile of the distribution. The differences between 0-1, 0-2, and 0-3 are all significant ($P < 0.01$, one-sided T-test). **(b)** Testing stability of RBFOX1 target transcripts by qRT-PCR. Cells expressing recombinant RBFOX1 under doxycycline control were sequentially transfected with RBFOX2-targeting siRNAs and pGL4.13 (Promega) mammalian luciferase expression vector encoding luciferase fused to CADPS 3'UTR containing either a wild-type or mutant RBFOX1 site, along with a constitutively-expressed mCherry transfection control plasmid. 6 hours before harvesting, transcription was shut off by treating cells with 10 μM Actinomycin D. Levels of luciferase transcript fused to either wild-type or mutant 3'UTR (wt/mut) in the presence or absence of doxycycline-induced RBFOX1 expression (-/+ Dox) was quantified using qRT-PCR. Transcript levels were normalized to mCherry control transcript. Error bars correspond to standard deviation of triplicate qRT-PCR runs performed on samples from a single transfection experiment.

## Justification for use of top 10 7-mer procedure to define motifs

We evaluated a panel of alternative approaches to motif derivation, including RNAcontext[62], Malarkey (HK and QDM, manuscript in preparation), MEME[63], MatrixREDUCE[64], BEEML-PBM[65], and the same top 10 procedure using k-mers of lengths other than 7. We tested the efficacy of the motifs in cross-validation between the A and B probe sets, reproducibility between biological replicates, similarity of motifs obtained between proteins with related amino acid sequences, similarity of motifs obtained to literature motifs, and ability to predict *in vivo* data. The 7-mer based top 10 motif derivation method was the only approach that scored consistently well across all tests. The results of this analysis will be presented elsewhere (KBC, manuscript in preparation).

## Data Availability

Data are available under NCBI GEO accession GSE41235. Data are also posted on our project website, http://hugheslab.ccbr.utoronto.ca/supplementary-data/RNAcompete_eukarya/. The cis-BP-RNA database, which is browsable and searchable, is at http://cisbp-rna.ccbr.utoronto.ca/.

## Secondary structure analyses

This section is partially redundant with the description in online methods but contains more detail.

We predicted the secondary structures of the probe sequences using an existing tool called RNAplfold[53]. RNAplfold considers the ensemble of all possible structures of an RNA sequence to calculate probabilities for each base to be in various structural contexts (e.g. hairpin loop, external loop). We modified RNAplfold so that instead of outputting the accessibility (i.e. the probability that the region of interest is single-stranded), it outputs the probabilities for the region of interest to be in four possible single-stranded contexts: hairpin loop, internal or bulge loop, external loop (i.e., ssRNA not in a loop), or multiloop (i.e., ssRNA in a loop containing 3 or more stems). These four probabilities sum up to the original accessibility. We ran this modified RNAplfold with the option –u 1 and set -W and -L arguments equal to the length of the probe. Then using the RNAplfold output, for each probe, we computed a matrix (which we call the secondary structure profile) where rows represent the accessibility and the four ssRNA structural contexts (i.e., hairpin loop, internal loop, multiloop, external loop) and columns correspond to the positions of the probe sequence. Each entry of this

matrix represents the probability of a base to appear in a particular structural context.

Our next step was to analyze these profiles to check whether an RBP displayed a specific secondary structure preference in a given RNAcompete assay. To do this, we split the probes containing one of the top 10 7-mers for each RBP into a bottom and top half according to their intensities. If a probe is selected for both bottom and top halves (because it was in the top half for one of the 10 7-mers and in the bottom half for the other), we kept the probe in both sets. Then, for each of the five structure contexts (ssRNA, and the four other contexts described over), we computed the average probability for each 7-mer in each probe and compared the distributions of these values among the probes in the top and bottom halves using Wilcoxon's rank sum test (two-sided) with multiple testing correction. We repeated this analysis separately for Set A and Set B and retained only the preferences that were found to be significant (Bonferroni-corrected $P < 0.05$) both in Set A and Set B. After performing this analysis, we found that a large number of RBPs had a preference for multiloop but this result was difficult to interpret because the probabilities for the multiloop context were very low in all cases – as such, we removed these preferences from further analysis but did not modify the Bonferroni correction.

**Supplementary Data 3** contains the results of this analysis. When an RBP had multiple RNAcompete assays associated with it, we deemed an RBP to display a secondary structure preference in RNAcompete if any of its assays demonstrated that preference.

**Success rate of multiple versus single RBD RBP constructs**

As part of our assay optimization process, we evaluated how well different RNA-binding constructs worked in RNAcompete for the same set of RBPs. We compared full-length (FL) proteins, RNA-binding regions (RBRs) as defined above, or individual RBDs. To perform this comparison, we generated 44 constructs from 12 Drosophila RBPs by cloning corresponding FL (12), RBR (12), and individual RBD (20) cDNA fragments. Successful experiments for single (e.g. RRM1, KH1, etc.) and multi-RBD (e.g. RRM x3, KH x2, etc.) containing RBPs were determined based on the presence of clear PWM motifs–represented in **Figure 2** as well as the RNAcompete website. Success rates for the various single and multi-RBD domain types are summarized in **Table S1**. Based on this analysis, we prepared RBR constructs for most of the RBPs that we assayed.

**Table S1: Comparison of RNAcompete success rates for full-length RBPs, RNA-binding regions and individual RNA-binding domains.**

| Gene name | Structure | Construct type | Success? |
|---|---|---|---|
| aret | RRM x3 | FL | Yes |
| | RRM x3 | RBR | Yes |
| | RRM1 | RBD | No |
| | RRM2 | RBD | No |
| | RRM3 | RBD | Yes |
| CG2931 | RRM x1 | FL | Yes |
| | RRM | RBR | No |
| CG3056 | RRM x2 | FL | No |
| | RRM x2 | RBR | No |
| | RRM1 | RBD | No |
| | RRM2 | RBD | No |
| CG4612 | RRM x2 | FL | No |
| | RRM x2 | RBR | No |
| | RRM1 | RBD | No |
| | RRM2 | RBD | No |
| CG7082 (PAPI) | KH x2, Tudor | FL | No |
| | KH x2, Tudor | RBR | Yes |
| | KH1 | RBD | No |
| | KH2 | RBD | No |
| Hrb27C | RRM x2 | FL | Yes |
| | RRM x2 | RBR | Yes |
| | RRM1 | RBD | No |
| | RRM2 | RBD | No |

| | | | |
|---|---|---|---|
| Hrb98DE | RRM x2 | FL | Yes |
| | RRM x2 | RBR | Yes |
| | RRM1 | RBD | No |
| | RRM2 | RBD | No |
| mub | KH x3 | FL | No |
| | KH x3 | RBR | No |
| | KH1 | RBD | No |
| | KH2 | RBD | No |
| | KH3 | RBD | Yes |
| Rsf1 | RRM x1 | FL | No |
| | RRM | RBR | Yes |
| tsu | RRM x1 | FL | No |
| | RRM | RBR | No |
| xl6 | RRM x1, zf_CCHC | FL | No |
| | RRM x1, zf_CCHC | RBR | No |
| | RRM | RBD | No |
| | zf_CCHC | RBD | No |
| yu | KH x1, Tudor | FL | No |
| | KH x1, Tudor | RBR | No |
| | KH | RBD | No |
| | Tudor | RBD | No |
| **Construct** | **# Assayed** | **# Successes** | **Success Rate (%)** |
| FL | 12 | 4 | 33.3 |
| RBR | 12 | 5 | 41.7 |
| RBD | 20 | 2 | 10.0 |

## Compilation of *in vivo* datasets

This section contains some information already provided in the online methods but describes our methodology in much greater detail.

We compiled data sets from the literature that report RNAs associated with individual proteins using genome-wide techniques. The positive and negative sets are posted on our project web site (http://hugheslab.ccbr.utoronto.ca/supplementary-data/RNAcompete_eukarya/). Note that in some cases multiple data sets were obtained for the same protein. The data sources and the procedure by which we defined "bound" and "unbound" sequences are described in **Table S2**.

Compilation of these data sets required us to extract the sequences that either correspond to the mature mRNA sequence of a gene or to the genomic locus covered by the pre-mRNA transcript of the gene. To define these sequences, we downloaded the mouse (mm9), rat (rn4) and human genome builds (hg18 and hg19) and their corresponding Refseq gene sets from the UCSC Genome Browser[66]. Fly (Drosophila melanogaster) genes were downloaded from Ensembl BioMART in August 2012 and represent the BDGP 5.4 release of gene models. When there are multiple isoforms for the same gene we used the longest isoform to define its mature mRNA sequence and the genomic locus covered by its pre-mRNA sequence.

To perform the ROC analyses for assessing how well RNAcompete motifs reproduce *in vivo* binding data, we needed to define a set of bound and unbound sequences. For most CLIP data sets, we applied a common procedure where we either used all or a defined subset of the identified peaks to be the bound sequences – often these peaks are described as "clusters of reads" in the corresponding papers. For these datasets, we also often needed to define "unbound sequences" – to do so, we selected random non-peak windows of matching length from the pre-mRNA sequence (defined as described above) from the same set of genes. Hereafter, we call this the "random windows" procedure. Note that although these windows are selected from the same set of genes as the peaks, we did not require the procedure to select at least one window from each gene and, as such, multiple non-peak windows could be selected from the same gene as long as they are at least 300 nts away from the ends of the peaks. We utilized the features of the BEDTools suite both for extracting sequences that correspond to genomic locations (covered by pre-mRNA sequences) and for selecting random regions to define unbound sequences.

RIP-based *in vivo* binding data typically only has transcript resolution and measures binding to mature mRNAs. Unless otherwise indicated below, we used the mature mRNA sequences defined as described above for the "bound" and "unbound" sequences.

Note that the actual number of sequences in bound and unbound set of the compiled data set can be lower than the selected number of sequences when the length of a cluster is too short (<12) or the cluster does not reside within a gene for CLIP data or the reported gene IDs do not have a matching Refseq mRNA sequence for RIP data.

**Table S2: Summary of *in vivo* datasets compiled and definitions of bound and unbound sequences.**

| RBP | Method | Selection of bound and unbound sequences | Reference (# refers to reference section) | Name of in vivo dataset (# of bound/# unbound transcripts) |
|---|---|---|---|---|
| Vts1p | RIP-chip | Bound and unbound sequences were obtained from the authors of a previous study[13] that analyzed this data. | [39] | Vts1p (121 / 1449) |
| ELAVL1[67-69,70]<br>FUS[14]<br>TAF15[14]<br>IGF2BP1-3[71]<br>PUM2[71]<br>QKI[71]<br>SFRS1[72]<br>TIA1[73]<br>TIAL1[73]<br>TARDBP[74] | CLIP-seq | We defined sequences with doRINA[75] scores (please see the doRINA paper for more details on the definition of peaks and the calculation of scores associated with these peaks) in the top five percentile as bound sequences. When necessary, we reduced the percentile cutoff to include a minimum of 1,000 sequences. We used the "random windows" procedure to define the unbound sequences.<br><br>Note: The first four entries of the fifth column correspond to ELAVL1 data sets which are compiled from [67]; from [68] doRINA ids ELAVL1-MNASE PAR-CLIP; from [68] doRINA ids ELAVL1-PARCLIP; and from [69] respectively. Subsequent entries appear in the same order as the RBPs in the first column. | [75] | ELAVL1_Lebedeva (1,445 / 1,445)<br>ELAVL1_MNASE (1000 / 1000)<br>ELAVL1_Mukharjee (5,625 / 5,625)<br>ELAVL1_Hafner (1000 / 1000)<br>FUS (1,568 / 1,568)<br>TAF15 (1,000 / 1,000)<br>IGF2BP1-3 (3,799 / 3,799)<br>PUM2 (1,000/ 1,000)<br>QKI (1,000 / 1,000)<br>SFRS1 (310 / 314)<br>TIA1 (1,000 / 968)<br>TIAL1 (2,117 / |

| | | | | |
|---|---|---|---|---|
| | | | | 2,093)<br>TARDBP_iCLIP (4,755 / 4,745) |
| FOX-2 | CLIP-seq | We downloaded CLIP-derived clusters from UCSC Genome Browser under 'Regulation' track. We used all the identified clusters as bound sequences, and used the "random windows" procedure to define the unbound sequences. | [24] | FOX-2 (3,547 / 3,547) |
| Mbnl1 | CLIP-seq | We downloaded CLIP-derived clusters from the corresponding GEO submission (GSM1226-30). We used all the identified clusters as bound sequences, and defined the unbound sequences using the "random windows" procedure.<br><br>Note: The fifth column contains five entries that correspond to data sets compiled from GSM1226 (B6Brain), GSM1227 (129Brain), GSM1228 (B6Heart), GSM1229 (B6Muscle), GSM1230 (C2C12). | [27] | Mbnl1_B6Brain (3,177 / 3,177)<br>Mbnl1_B129Brain (11,580 / 11,580)<br>Mbnl1_B6Heart (645 / 645)<br>Mbnl1_B6Muscle (443 / 443)<br>Mbnl1_C2C12 (24,191 / 24,191) |
| LIN28 | CLIP-seq | Bound and unbound sequences were obtained from the authors.<br><br>Note: Two different cell lines were used in this study: H9 human ES (hES) and LIN28-V5 293. The four entries in the fifth column correspond to data sets compiled from hES clusters in 3' UTRs, hES clusters in coding regions, LIN28-V5 293 clusters in 3'UREs and LIN28-V5 293 clusters in coding regions, respectively. | [76] | LIN28_hES_3UTR (12,399 / 3,945)<br>LIN28_hES_coding_exons (6,461 / 1,647)<br>LIN28_v5_3UTR (6,525 / 1,582)<br>LIN28_v5_coding_exons (3,554 / 668) |
| RBM4 | PAR-CLIP | We downloaded the list of genes associated with the RBP from the supplementary data of the original study. We defined the mature mRNA sequences of top 1,000 genes with highest number of matching reads as the bound sequences. Unbound sequences were randomly selected mature mRNA sequences from the remaining set of human genes (hg18 build, Refseq gene models as described above). | [77] | RBM4 (824 / 1000) |
| Lark | RIP-chip | We used the list of genes identified in the original study (Supplementary Table 1) as bound sequences. We prepared two data sets; one contained the union of genes identified in two replicate experiments (Expt 1 and 2), other contained the genes identified in both of the | [78] | Lark_union (168 / 221)<br>Lark_shared (65 / 80) |

| | | | | |
|---|---|---|---|---|
| | | experiments.<br><br>Unbound sequences were randomly selected from the remaining set of fly genes (BDGP 5.4, as defined above). | | |
| CPEB4 | RIP-seq | We used the p-value cutoff used in the original study to define genes whose mature RNA sequences were used as the bound sequences (Supplementary Table 2, p-value < 0.05). We selected unbound sequences from the mature mRNA sequences associated with the 942 genes with the highest p-values. | [79] | CPEB4<br>(927 / 942) |
| TARDBP | RIP-seq | We downloaded the data from the corresponding GEO submission (GSM614808). We first filtered out the genes that have less than 10 reads mapped. We then sorted the genes based on either "exonic read density" or "intronic read density" (as defined in [80]), obtaining two lists. We then found genes that appeared in the top 1000 of both lists and used their mature mRNA sequences as the bound sequences. Similarly, we used the genes that appear in the bottom 1000 of both lists to define the unbound sequences. | [80] | TARDBP_RIP<br>(422 / 565) |
| MSI | RIP-chip | We downloaded the data from the corresponding GEO submission (GSE30904). As suggested by the authors, we used the mature mRNA sequences of the top 50 genes with highest enrichment ratios as the bound sequences. We randomly chose genes from the remaining set of human genes (hg19 build, Refseq gene model) to define the unbound set. | [81] | (MSI)<br>42 / 50 |
| hnRNPA1<br><br>hnRNPA2B1 | CLIP-seq | Bound and unbound sequences were obtained from the authors. | [19] | hnRNPA1<br>(433 / 433)<br>hnRNPA2B1<br>(1361 / 1361) |
| SHEP | RIP-seq | Unpublished RIP-seq data for Shep were obtained from the authors of the referenced study. Genes that are enriched in the immunoprecipitates (adjusted p-value < 0.05, fold change > 1.5) were defined as the bound genes. Unbound genes were selected from those that have the p-values equal to 1.<br><br>We also used a more stringent definition of enrichment where we include only the genes with average number of background counts greater than 220.<br><br>Note: The fifth column contains four entries that correspond to data sets compiled from bg3 cell lines with default constraints, bg3 cell lines with stringent constraints, kc cell lines with default constraints and kc cell lines with stringent constraints, respectively. | [82] | SHEP_bg3_normal<br>(168 / 290)<br>SHEP_bg3_stringent<br>(110 / 221)<br>SHEP_kc_normal<br>(373 / 674)<br>SHEP_kc_stringent<br>(262 / 527) |

| | | | | |
|---|---|---|---|---|
| FMR1 | CLIP-seq and RIP-seq | We compiled the CLIP data sets from Supplementary Table 2a and 2b of the original paper. We prepared two data sets from each table, where we include the top and bottom 1000 or 5000 clusters based on PARalyzer peak score.<br><br>We prepared the RIP-seq data set from Supplementary Table 6 of the original paper. We defined the bound sequences as the mature mRNA sequences associated with the genes that have the highest 1000 enrichment scores. Similarly, unbound sequences are defined as the genes with lowest 1000 enrichment scores.<br><br>Note: The first two entries of the fifth column correspond to data sets prepared from Table 2a with top (and bottom) 1000 and 5000 clusters, respectively. The third and fourth entries correspond to data sets prepared from Table 2b with top (and bottom) 1000 and 5000 clusters, respectively. The last entry corresponds to the RIP-seq data set. | 83 | FMR1_table2a_top 1K<br>(995 / 876)<br>FMR1_table2a_top 5K<br>(4,653 / 4,352)<br>FMR1_table2b_top 1K<br>(901 / 853)<br>FMR1_table2b_top 5K<br>(4,369 / 4,312)<br>FMR1_top1K<br>(1000 / 1000) |
| PTBP1 | CLIP-seq | We used the peaks compiled by the original study (GSE19323) as the bound set, and we used the "random windows" procedure to define the unbound sequences. | 34 | PTBP1<br>(2553 / 2547) |

## Learning Malarkey motif models from *in vivo* datasets

Malarkey is a motif finding method that infers both sequence and structure binding preferences of an RBP from experimental binding data (manuscript in preparation). Malarkey fits its model parameters by using multilinear regression to maximize the agreement between Malarkey-predicted affinities and experimental data for the input set of sequences.

Malarkey motif models are fit to *in vivo* data sets where bound sequences are labeled as 1 and unbound sequences are labeled as 0. In order to make a fair comparison against RNAcompete-derived motifs, we fitted Malarkey without the secondary structure model and with a fixed motif length of 7. In this mode, except for the differences described below, Malarkey's motif finding algorithm is nearly identical to MatrixREDUCE[84]. To evaluate the predictive performance of Malarkey motifs, we used a 10-fold cross validation scheme and calculated the average AUROC across the 10 held-out sets. Similarly, we scanned the same held-out sets with RNAcompete-derived PFMs and compared the average AUROCs.

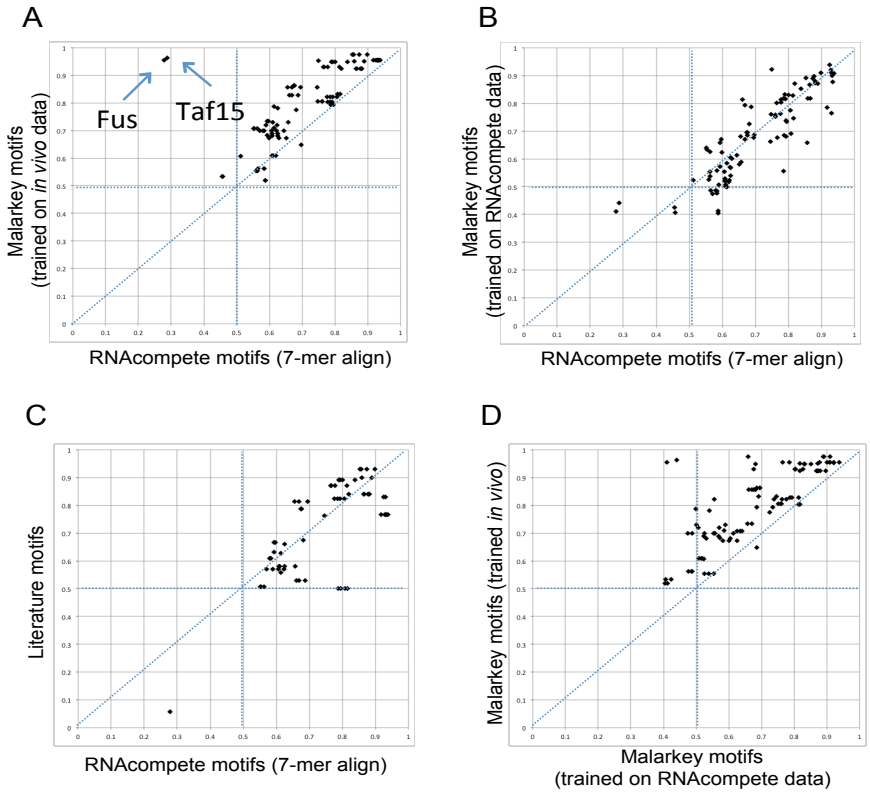



**Figure S2: Comparison of AUROCs of RNAcompete and Malarkey defined motifs on *in vivo* binding data.**

Plots in each scatterplot are AUROCs for a pair of columns in **Suppl. Data 6** (**A**) Shows that with the exception of Fus and Taf15, there is a close correspondence between the performance of RNAcompete motifs and Malarkey motifs obtained from the in vivo data; (**B**) Shows that the slight increase in AUROC obtained from Malarkey in (**A**) is not due to the Malarkey algorithm, but instead due to factors present in vivo but not in vitro. (**C**) Shows that the RNAcompete motifs generally perform comparably or better than literature motifs for the same protein. (**D**) Direct comparison of Malarkey motifs in vivo and in vitro.

## Analysis of Drosophila post-transcriptional data sets

This section contains information also presented in the online methods but provides greater detail.

We used previously published Drosophila post-transcriptional regulation (PTR) datasets (i.e. the flyFISH website and supplementary data from references[40, 41, 55, 56]) to define a set of 112 categories of post-transcriptional fate and for each category defined two sets of transcripts: a "positive set" and a "negative set". The positive set consisted of those transcripts with the post-transcriptional fate described by that category and the negative set consisted of those transcripts that were expressed under the same conditions as the positives but were not annotated as having the given fate. These sets and further details of their definition will be provided in a forthcoming publication (XL, HDL, and QM, in preparation). For each compiled dataset, we performed a likelihood ratio test to assess whether any of the motifs from our collection could better distinguish the positive set from the negative set when provided to a regression algorithm that also had access to a control set of features that consisted of all the dinucleotides contained within the corresponding motif as well as the length of the target sequence; the construction of these regression models is described below. The comparisons between the motif and the control features were restricted to either the 3' UTR or the coding region of the transcripts. We scored each 3' UTR or coding region using a given motif by summing the accessibility of all the target sites, where a target site was defined as a perfect match to the IUPAC representation of the motif (see **Supplementary Data 8** for IUPAC motifs used in these analysis) and the accessibility of a target site was defined as the average single base accessibility of the bases in the site. A score of zero was assigned to those transcripts whose 3' UTRs or coding regions did not contain a motif match. The single base accessibility was assessed using RNAplfold[53] as described previously[13] and in the "**Secondary Structure Analysis**" section above. We used the parameters with $W$=80, $L$=40 and $U$=1. Although the analysis was applied in the 3' UTR or the coding region, the entire transcript was input into RNAplfold to ensure correct folding of the bases close to the start codon and stop codon. We used the glmnet.R package (version 1.8) [85] to apply Lasso penalized logistic regression to predict the particular PTR dataset using the feature sets containing the score calculated for one motif and the relevant control features. In the Lasso regression, the hyper-parameter lambda (i.e. the regularization strength) was selected through a five-fold cross-validation procedure, from the

lambda sequence computed by glmnet using the default settings of nlambda and lambda.min.ratio. The final value for lambda was the one (from the sequence) with the smallest average generalization error across the five folds. We then used this value of lambda with the 'glmnet.fit' object on the entire dataset to compute the weights for the features. The features with non-zero weights were selected as contributing most to the prediction. After the non-zero weight features were defined, we trained two standard logistic regression models: one using all non-zero weight features (including the motif) and one that contained only the non-zero weighted control features, and then assessed whether there was a significant difference in predictive power between these two nested models using a log-likelihood ratio test (as per the procedure recommended in [86]). We then used these P-values to compute a false-discovery rate using the Benjamini-Hochberg procedure. The motifs, RBPs, and categories that with FDR < 0.1 are provided in **Supplementary Data 8**.

**Assessing tissue alternative splicing levels using RNA-Seq data**

This section expands on methods presented in online methods.

Information on intron-exon structures was extracted from Ensembl annotations (release 65) for the human (hg19) genome. This information was used to generate a Bowtie library of non-redundant exon-exon junction (EEJ) sequences by combining every possible (forward combination) splicing donor and acceptor within each gene. For each EEJ sequence, we determined the effective number of unique mappable positions for a given read length (k). We extracted the L-k+1 (L being the EEJ length) k-mers from each EEJ sequence and then aligned the full set of k-mers against the EEJ library plus the respective genome using Bowtie[87], allowing for a maximum of two mismatches along the entire length of the read. The number of k-mers with one unique alignment was counted; this corresponds to the junction's effective number of unique mappable positions for a given set of RNA-Seq k-mers.

RNA-Seq reads from the different samples were then mapped to the EEJ libraries using Bowtie with –m 1 –v 2 parameters. Reads were trimmed to 50 nucleotides, if longer, and reads that had full-length mappings to the genome were discarded because EEJs should not exist as contiguous sequences in the genome. A minimum of eight mapped nucleotides was required for each of the two exons forming a given EEJ. Next, the outputs were parsed to identify cassette exons – exons that are either included or fully excluded from the transcripts – by identifying exons that have associated reads mapping to (i) both EEJs supporting the inclusion of the exon (constitutive upstream exon (C1)-cassette exon (A) and A-constitutive downstream exon (C2), or C1A and AC2) and (ii) the EEJ for the exclusion of the exon (i.e. C1C2).

The inclusion level of an exon was defined as the percentage of gene transcripts in which a given exon is spliced in (PSI). This was estimated using read counts mapping to EEJs. The initial read counts for each EEJ $k$ ($EEJ_{k,count}$) were corrected for mappability (i.e. the uniqueness of the EEJ among the transcriptome) as follows ($EEJ_{k,corrected} = EEJ_{k,count} / MAP_k * MAPMAX$) where $MAP_k$ is the mappability for the EEJ for read length $k$ as described above, and MAPMAX is the maximum mappability for a EEJ for a given read length (e.g., $MAPMAX = 35$ for $k = 50$nt). After correction, we renamed each corrected EEJ count according the position of the EEJ relative to the alternative exon under consideration, and computed the PSI as follows:

$$PSI = 100\% * EEJ\_Reads\_Supporting\_A / EEJ\_Reads\_Mapping\_to\_A\_or\_Adjacent\_Exons,$$

where $EEJ\_Reads\_Supporting\_A = [\Sigma_i C_iA] + [\Sigma_i AC_j]$ and

$$EEJ\_Reads\_Mapping\_to\_A\_or\_Adjacent\_Exons = [\Sigma_i C_iA] + [\Sigma_i AC_j] + [\Sigma_i C_iC2] + [\Sigma_i C1C_j]$$

where $C_i$ is any possible splicing donor upstream of the alternative exon (including C1); $C_j$ is any possible splicing acceptor downstream of the alternative exon (including C2) and $C_iA$, $AC_j$, $C_iC2$, and $C1C_j$ represent the corrected read count mapping to the indicated EEJ ($EEJ_{k,corrected}$ as defined above). Alternative exons were only included when a minimal transcript coverage requirement was met of (i) ≥15 corrected reads mapping to the exclusion EEJs, or (ii) ≥15 corrected reads mapping to one of the sets of inclusion EEJs ($C_iA$ or $AC_j$), and ≥10 to the other set of inclusion EEJs. For alternative exons with multiple acceptor/donor splice sites, we used the splice site combination with the highest read support. When several putative C1 and/or C2 exons could be defined, we used the one with the highest read support as reference.

**Associating motifs with alternative splicing regulation**

This section repeats and expands on methods presented in the online methods.

We processed a collection of 34 RNA-seq experiments from diverse human tissues and cell lines (listed in **Table S3**) to measure the expression level of genes as well as abundance of splicing events in each sample. In particular, we downloaded the raw read data from GEO and reprocessed the data using an in-house pipeline described in detail in the previous section. This pipeline computed percent-spliced-in (PSI) of alternatively spliced cassette exons for a previously defined set of alternatively spliced cassette exons across the 34 experiments, as well as corrected RPKM (cRPKM) profiles (reads per kilobase per million mapped reads corrected for mappability as described in the previous section) for each gene across the 34 experiments. The PSI value is an estimate of the proportion of transcripts that include the alternative exon in a particular tissue or

cell line, and cRPKM is a measure of the abundance of transcripts from a given gene in a tissue or cell line. We hypothesized that if RBP *x* is involved in regulating splicing, the cRPKM profile of its gene should be either correlated with the PSI profiles of its target exons (indicating a role of RBP *x* in promoting exon inclusion), or anti-correlated (indicating a role in promoting exon exclusion), where its target exons were identified based on matches to one or more motifs associated with that RBP *x* within a defined splicing regulatory region associated with the target exon.

We associated each target exon with 32 different possible regulatory regions; these regions were defined based on their positions relative to splice boundaries of the target exon or its neighboring exons. In the following definitions, the target exon is called "exon *A*" (because it is Alternative), its upstream exon (i.e. 5' to exon *A*) is called "exon $C_1$", its upstream intron (i.e. lying between $C_1$ and *A*) is called "intron $I_1$", its downstream exon is called "exon $C_2$", and its downstream intron is called "intron $I_2$". We removed from consideration any cassette exon event for which any of $C_1$, *A*, or $C_2$ were less than 100nt in length or either $I_1$ or $I_2$ were less than 300nt in length. We then defined eight regulatory areas (i)-(viii) as follows: (i) the 100-nucleotide exonic region upstream of the 3' end of the exon $C_1$, (ii) the 300-nucleotide intronic region downstream of the 5' end of the intron $I_1$, (iii) the 300-nucleotide intronic region upstream of the 3' end of the intron $I_1$, (iv) the 100-nucleotide exonic region downstream of the 5' end of exon *A*, (v) the 100-nucleotide exonic region upstream of the 3' end of exon *A*, (vi) the 300-nucleotide intronic region downstream of the 5' end of the intron $I_2$, (vii) the 300-nucleotide intronic region upstream of the 3' end of the intron $I_2$, and (viii) the 100-nucleotide exonic region downstream of the 5' end of the exon $C_2$. Each of the eight regulatory areas was divided into 50-nucleotide-long bins, resulting in a total of 32 regulatory regions. We analyzed each of these region types separately as described in the following paragraph. The sequences for regulatory areas (i)-(viii) were retrieved from the hg19 assembly of the human genome based on Ensembl annotations (release 69).

To identify whether an RBP *x* may promote inclusion or exclusion of its target exons by binding in regulatory region *r*, we first sorted all alternatively spliced exons by the descending order of the Pearson correlation of their PSI profiles with the cRPKM profile of RBP *x*, resulting in the sorted list $L_x$. We then determined whether exons with significant matches to one or more motifs associated with RBP *x* in region *r* were significantly enriched at the top of list $L_x$ (indicating that binding of RBP *x* in *r* promotes inclusion) or at the bottom of list $L_x$ (indicating that binding of *x* to *r* promotes exclusion). We used a two-tailed Mann-Whitney U test of ranks to measure enrichment of exons with binding sites at the top or bottom of list $L_x$. The test produces a normalized splicing z-score that

follows a standard normal distribution, based on which a *p*-value can be calculated. Benjamini-corrected *p*-values were used to identify significant associations at a false discovery rate (FDR) <0.1.

To determine target exons that contained a significant match in region *r* to a motif associated with RBP *x*, we first identified all motifs associated with RBP *x* by collecting all motifs (either RNAcompete-derived or literature-derived) from our cisbp-rna database that had at least 70% sequence identity and matching RBD domain patterns to this RBP. We then transformed the position-specific frequency matrices provided by cisbp-rna to position-specific affinity matrices (PSAMs) by dividing each column by its maximum element. To determine whether a particular regulatory region *r* in a particular exon was significantly enriched for matches to a motif, we calculated the "regulatory region affinity value" of that motif to region *r* using the PSAM as described previously[88] – in brief, we summed the PSAM scores of each *k*-mer in the regulatory region, where *k* is the width of the PSAM. We then transformed these affinity values to z-scores by subtracting the mean of these values in region *r* of all cassette exons in our dataset and divided by the standard deviation of this distribution. Empirically, the distribution of these z-scores was similar to a standard normal distribution, so we associated *p*-values to z-scores using a one-tailed Z-test, and deemed that a region *r* in a particular target exon had a significant match to the binding site of RBP *x* if the Benjamini-corrected false discovery rate of its affinity z-score was less than 10% (where the multiple test correction was applied based on all *p*-values calculated for region *r* for a given motif).

**Table S3: List of 34 tissues and cell lines used in human post-transcriptional regulation analysis**

| Sample Type | Sample Name | Platform | GEO Series | Notes |
|---|---|---|---|---|
| ESC | H1 (a) | Illumina | GSE23316 | GEO: GSM591680 |
| | H1 (b) | Illumina | GSE16256 | PMID: 20944595 |
| | H9 (a) | Illumina | GSE30992 | PMID: 21924763 |
| | H9 (b) | Illumina | GSE22666 | PMID: 21324177 |
| | hESC2 | SOLiD | GSE25842 | PMID: 22042643 |
| iPS | iPS (a) | Illumina | GSE32625 | PMID: 21915259 |
| | iPS (b) | SOLiD | GSE16256 | GEO: GSM706050 |

| | | | | |
|---|---|---|---|---|
| Cell line | Fibroblast | Illumina | GSE30554 | PMID: 21890647 |
| | HNEK | Illumina | GSE30567 | GEO: GSM765401 |
| | HUVEK | Illumina | GSE30567 | GEO: GSM758563 |
| | MCF7 | Illumina | GSE30567 | GEO: GSM765388 |
| | GM12878 | Illumina | GSE23316 | GEO: GSM591664 |
| Tissue | Whole Brain | Illumina | GSE30611 | Human Body Map |
| | Cortex | Illumina | GSE30352 | PMID: 22012392 |
| | Cerebellum | Illumina | GSE30352 | PMID: 22012392 |
| | Liver (a) | Illumina | GSE30611 | Human Body Map |
| | Liver (b) | Illumina | GSE30352 | PMID: 22012392 |
| | Kidney (a) | Illumina | GSE30611 | Human Body Map |
| | Kidney (b) | Illumina | GSE30352 | PMID: 22012392 |
| | Heart (a) | Illumina | GSE30611 | Human Body Map |
| | Heart (b) | Illumina | GSE30352 | PMID: 22012392 |
| | Muscle | Illumina | GSE30611 | Human Body Map |
| | Testis (a) | Illumina | GSE30611 | Human Body Map |
| | Testis (b) | Illumina | GSE30352 | PMID: 22012392 |
| | Adipose | Illumina | GSE30611 | Human Body Map |
| | Adrenal | Illumina | GSE30611 | Human Body Map |
| | Breast | Illumina | GSE30611 | Human Body |

| | | | | |
|---|---|---|---|---|
| | | | | Map |
| | Colon | Illumina | GSE30611 | Human Body Map |
| | Lung | Illumina | GSE30611 | Human Body Map |
| | Lymph node | Illumina | GSE30611 | Human Body Map |
| | Ovary | Illumina | GSE30611 | Human Body Map |
| | Prostate | Illumina | GSE30611 | Human Body Map |
| | Thyroid | Illumina | GSE30611 | Human Body Map |
| | WBC | Illumina | GSE30611 | Human Body Map |

**Defining the exons that are regulated by each splicing-related RBPs using leading-edge analysis**

This section repeats and expands on methods presented in the online methods.

Here, we sought to connect RBPs to the exons that they regulate. Some RBPs were associated with more than one recognition motif (e.g. from multiple experiments, or by inferring multiple motifs through similarity of RBDs). In the previous section, we analyzed each recognition motif separately. After grouping motifs by RBP, we found that in general different recognition motifs of each RBP resulted in similar conclusions regarding the role of the RBP in regulating splicing as well as the regulatory region that the RBP binds (**Figure S3**). Therefore, for each RBP, we combined the set of exons that had at least one significant match – in the inferred relevant regulatory region(s) – to one of the motifs with significant splicing z-scores. This resulted in a list of exons $E_x$ for each RBP $x$. Re-analysis of PSI profiles of exon set $E_x$ using Mann-Whitney U test of ranks as in the previous section showed that this combined set invariably obtains higher scores than exon sets defined based on any of the individual motifs of RBP $x$. We further refined the exon set $E_x$ by analyzing the list $L_x$ as described before[89] whereby, in brief, we identified a new, stringent correlation or anti-correlation

threshold by finding the threshold that maximized the modified KS-test *p*-value described in [89]. This refinement resulted in a high-confidence "leading-edge" list of exons that (i) have a binding site for RBP *x* in the relevant regulatory region based on at least one of the significant splicing-associated motifs of *x*, and (ii) have PSI profiles whose correlation or anti-correlation with the expression profile of *x* is above or below the stringent threshold depending on the inferred role of *x* in promoting inclusion or exclusion, respectively. The splicing network that this procedure produced is provided in **Supplementary Data 7**.

**Defining human RBP motifs involved in regulating mRNA stability**

This section repeats and expands on methods presented in the online methods.

Using the same set of 34 tissues and cell lines as described above, we identified RBPs that are involved in regulating mRNA stability. We employed similar methods as described above, with the main difference that we used log-transformed cRPKM profiles instead of PSI profiles. In other words, we examined whether the binding sites of RBP *x* are enriched in 3' UTRs of genes whose log-transformed cRPKM profiles are correlated or anti-correlated with the log-transformed cRPKM profile of RBP *x*, suggesting a role of *x* in stabilizing or destabilizing its target genes, respectively. We used log-transformed cRPKM values because the logarithm of mRNA abundance is presumed to have an inverse linear relationship with the logarithm of mRNA decay rate at steady-state conditions[90]. We used a Mann-Whitney U test of ranks to identify significant motif-stability associations, similar to the motif-splicing association analysis described above. RBP binding sites were examined in the 300-nucleotide region immediately downstream of the stop codon of the longest isoform of each gene. Only genes whose 3' UTR consisted of a single exon were considered for this analysis, in order to rule out the possibility of erroneous identification of splicing factors as stability factors. Note that this rule should exclude exons with annotated 3' UTR alternative splicing sites. The sequences of all of the transcripts associated with each gene were downloaded from the UCSC genome browser based on the hg19 annotation of the human genome.

Unlike alternative splicing, we found that mRNA abundance/stability is greatly influenced by the GC content of the 3' UTR. To filter out RBPs whose inferred role in regulating stability was confounded by differences in dinucleotide bias among 3' UTRs, we randomly shuffled the 3' UTR sequences 100 times, each time calculating the Mann-Whitney U z-scores of all RBP motifs for association with stability. This procedure created a null distribution of z-scores for each motif, to which we compared the original z-score of the motif (i.e. the score that was obtained using real 3' UTR sequences). Specifically, we used the random scores to calculate the mean and standard deviation of the null distribution for each

motif, which was used to transform the original z-score to "z-of-z". Similar to the z-score, we observed that the z-of-z score appears to follow a standard normal distribution, so we used a two-tailed Z-test to compute a new *p*-value for z-of-z score. A motif is deemed significantly associated with stability if (i) the *p*-value associated with its original z-score is significant (Benjamini correction, FDR <0.1), (ii) its z-score has the same sign as its z-of-z score, and (iii) the *p*-value associated with its z-of-z score is significant (Benjamini correction, FDR <0.1).

Similar to the procedure described for splicing, we combined the set of genes that had binding sites based on different significant motifs of each RBP, creating the union set $G_x$ for each RBP $x$. The set $G_x$ for each RBP was further refined using leading-edge analysis as described in the previous section, resulting in a high-confidence stability network that is provided in **Supplementary Data 7**.

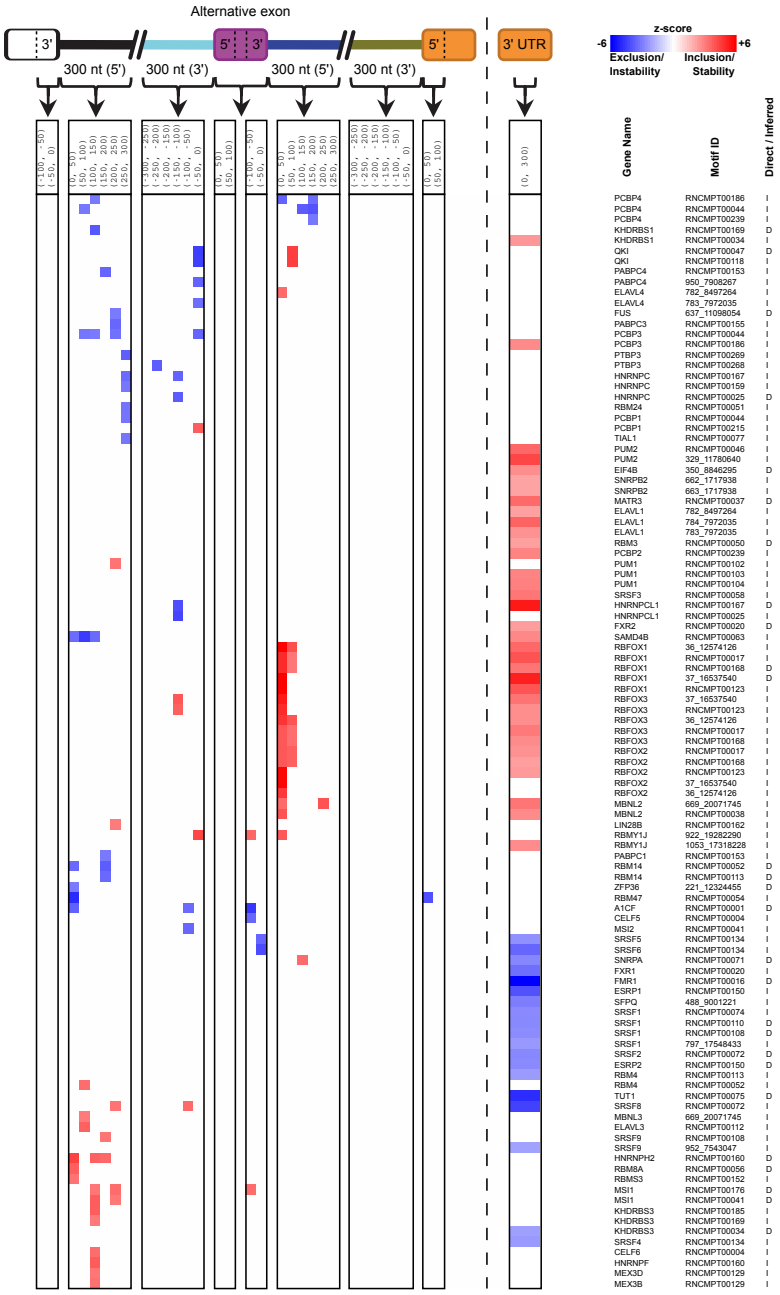



**Figure S3: The binding profile of RBPs that are involved in regulating splicing and/or stability.**

Red indicates that binding of the RBP to the corresponding region promotes inclusion of the alternative exon or, in the case of binding to 3' UTR, stability of the mRNA. Blue indicates promoting exclusion/instability. The z-scores are based on Mann-Whitney U test of enrichment. For 3' UTRs, z-of-z as defined above is indicated. Motif IDs without RNCMPT prefixes are motif IDs from RBPDB (http://rbpdb.ccbr.utoronto.ca/).


**Figure S4: Gene Ontology (GO) enrichment analysis of human RBP motifs in 3' UTRs**

For each RBP with an inferred role in regulating mRNA stability, we examined the enrichment and depletion of GO terms among genes in their region target sets. In this figure, each column is an RBP, and each row is a GO term. Red indicates significant enrichment of the GO term among target genes of the corresponding RBP, and blue means significant depletion (Fisher's exact test, Benjamini correction, FDR < 0.1). The color gradient shows the logarithm of p-value of enrichment or depletion.

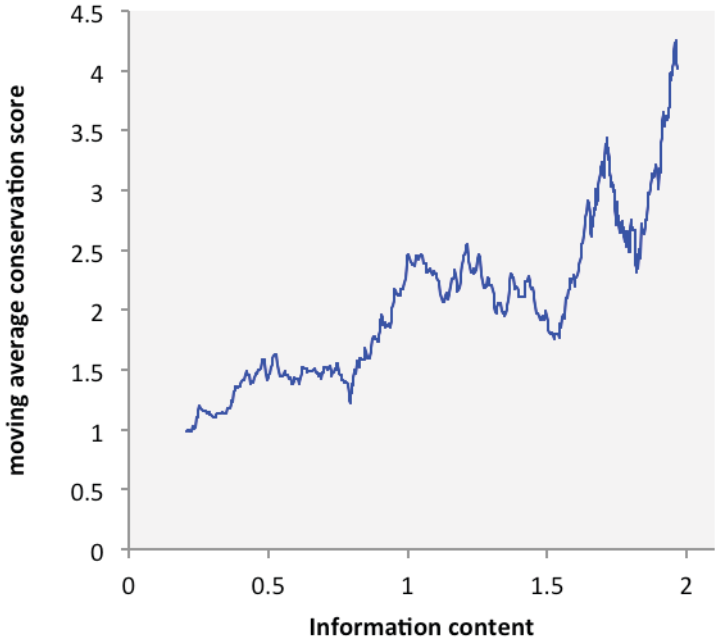



**Figure S5: Information content of motifs versus conservation of bases in motif matches**

Bases at degenerate positions of motifs are less conserved than bases at positions with high information content. In this figure, the relationship between conservation and information content is shown for the non-redundant motifs that are represented in **Figure 4**. The information content (2 – entropy of the column measured in bits) and aggregated conservation score (-log10(P-value)) of each column of each motif were calculated. The resulting pairs of values were then sorted by ascending order of entropy, and running average of conservation was calculated for every 100 instances.

**Table S4: Motifs used to represent human RBP families in Figure 4.** Non-RNCMPT motif IDs are RBPDB motif IDs (http://rbpdb.ccbr.utoronto.ca/).

| Text in Fig. 4 | Protein(s) | Motif ID |
|---|---|---|
| **EIF2S1** | EIF2S1 | RNCMPT00273 |
| **MEX3B/C/D** | MEX3B, MEX3C, MEX3D | RNCMPT00129 |
| **RBM24/38** | RBM24, RBM38 | RNCMPT00184 |
| **ACO1** | ACO1 | 1213_8021254 |
| **RBM8A** | RBM8A | RNCMPT00056 |
| **FXR1/2** | FXR1, FXR2 | RNCMPT00020 |
| **RBM5** | RBM5 | RNCMPT00154 |
| **SRSF4/5/6** | SRSF4, SRSF5, SRSF6 | RNCMPT00134 |
| **RBM45** | RBM45 | RNCMPT00241 |
| **PABPC5** | PABPC5 | RNCMPT00171 |
| **SART3** | SART3 | RNCMPT00064 |
| **HNRNPC/CL1, RALY** | HNRNPC, HNRNPCL1, RALY | RNCMPT00025 |
| **TARDBP** | TARDBP | RNCMPT00076 |
| **PABPN1/1L** | PABPN1, PABPN1L | RNCMPT00157 |
| **EIF4B** | EIF4B | 350_8846295 |
| **RBM6** | RBM6 | RNCMPT00170 |
| **CPEB2/3/4** | CPEB2, CPEB3, CPEB4 | RNCMPT00126 |
| **ANKHD1, ANKRD17** | ANKHD1, ANKRD17 | RNCMPT00002 |
| **QKI** | QKI | 149_16041388 |
| **PTBP1/2/3** | PTBP1, PTBP2, PTBP3 | RNCMPT00268 |
| **PABPC1/1L/3/4** | PABPC1, PABPC1L, PABPC3, PABPC4 | RNCMPT00153 |
| **HNRNPF/H1/H2** | HNRNPF, HNRNPH1, HNRNPH2 | RNCMPT00160 |
| **SF3B4** | SF3B4 | RNCMPT00224 |
| **ENOX1/2** | ENOX1, ENOX2 | RNCMPT00149 |

| | | |
|---|---|---|
| **SRSF2/8** | SRSF2, SRSF8 | 953_7543047 |
| **KHDRBS1/2/3** | KHDRBS1, KHDRBS2, KHDRBS3 | RNCMPT00169 |
| **PCBP1/2/3/4** | PCBP1, PCBP2, PCBP3, PCBP4 | RNCMPT00044 |
| **ZC3H10** | ZC3H10 | RNCMPT00085 |
| **CNOT4** | CNOT4 | RNCMPT00156 |
| **HNRNPK** | HNRNPK | RNCMPT00026 |
| **MBNL1/2/3** | MBNL1, MBNL2, MBNL3 | RNCMPT00038 |
| **HNRNPA1/1L2/1P7/2B1/3** | HNRNPA1, HNRNPA1L2, HNRNPA1P7, HNRNPA2B1, HNRNPA3, RP13-923O23.5 | RNCMPT00022 |
| **SRSF1/9** | SRSF1, SRSF9 | RNCMPT00110 |
| **FMR1** | FMR1 | RNCMPT00016 |
| **HuR, ELAVL2/3/4** | HuR, ELAVL2, ELAVL3, ELAVL4 | 784_7972035 |
| **RBFOX1/2/3** | RBFOX1, RBFOX2, RBFOX3 | 37_16537540 |
| **ESRP1/2** | ESRP1, ESRP2 | RNCMPT00150 |
| **NONO, SFPQ** | NONO, SFPQ | 488_9001221 |
| **SAMD4A/B** | SAMD4A, SAMD4B | RNCMPT00063 |
| **LIN28A/B** | LIN28A, LIN28B | RNCMPT00036 |
| **RBM4/4B/14** | RBM14, RBM4, RBM4B | RNCMPT00113 |
| **MATR3** | MATR3 | RNCMPT00037 |
| **HNRNPL** | HNRNPL | RNCMPT00027 |
| **CSDA, YBX1/2** | CSDA, YBX1, YBX2 | 114_7499328 |
| **CELF6** | CELF6 | RNCMPT00122 |
| **RBM28** | RBM28 | RNCMPT00049 |
| **SNRPA/B2** | SNRPA, SNRPB2 | RNCMPT00145 |
| **ZFP36/36L1/36L2** | ZFP36, ZFP36L1, ZFP36L2 | 951_12324455 |
| **PUM1/2** | PUM1, PUM2 | RNCMPT00104 |

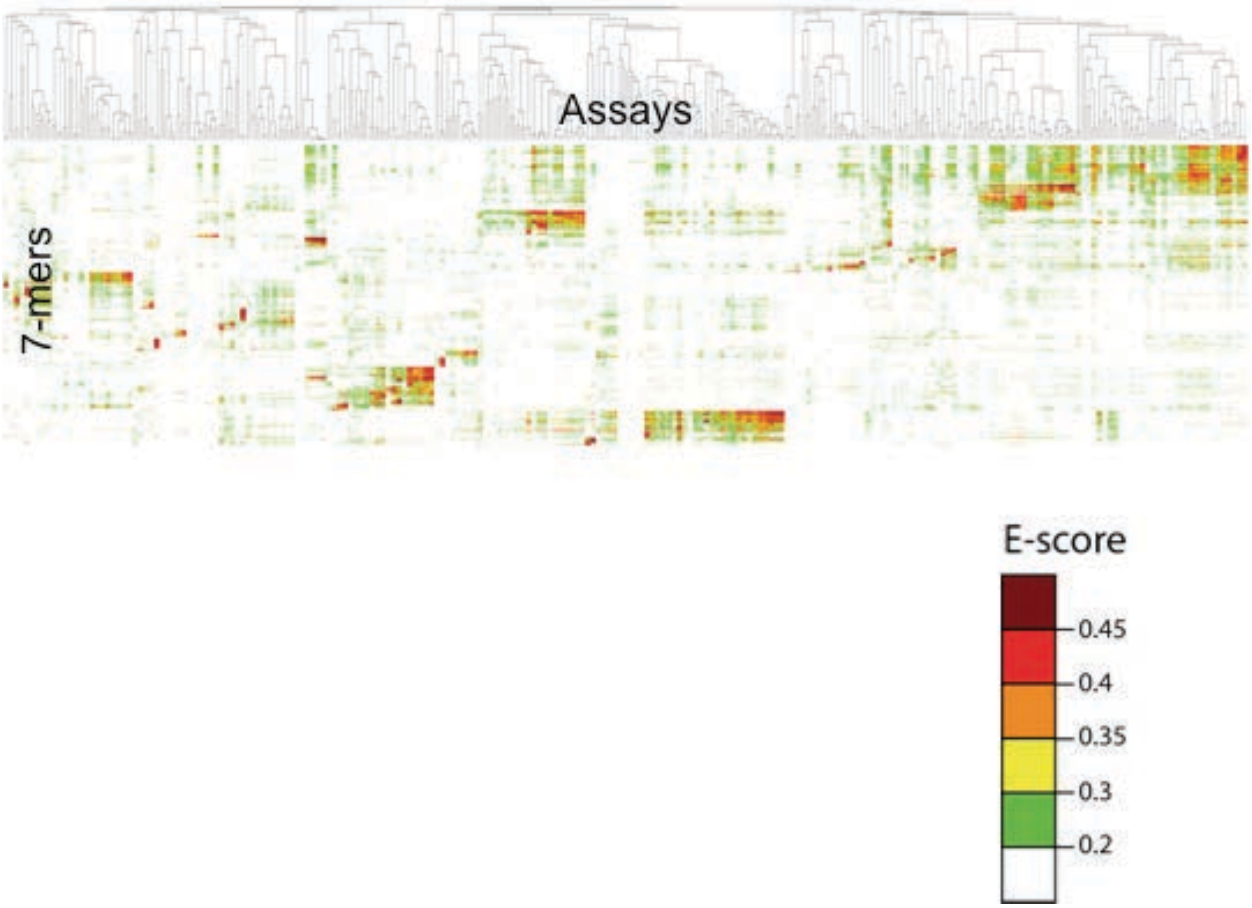
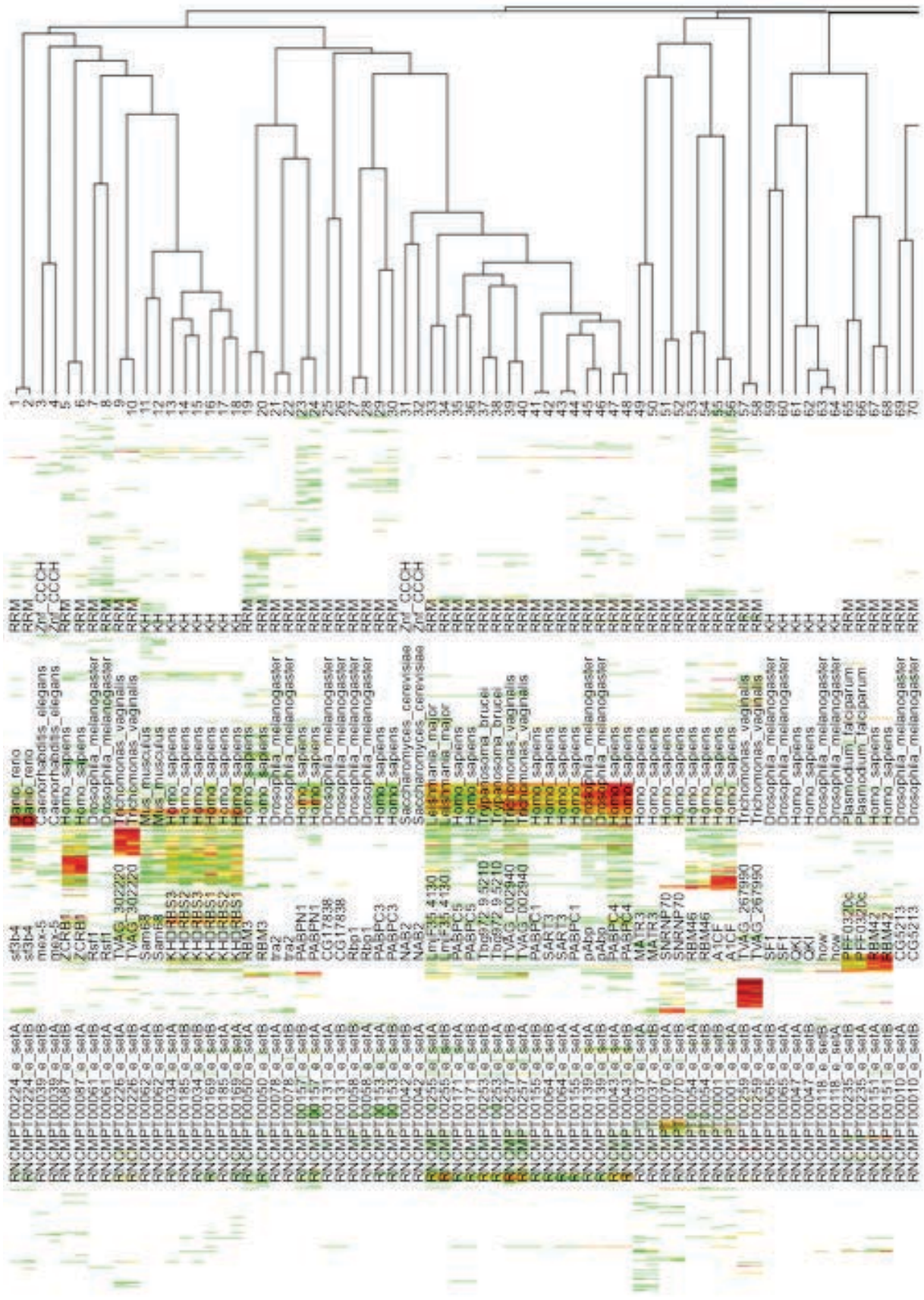
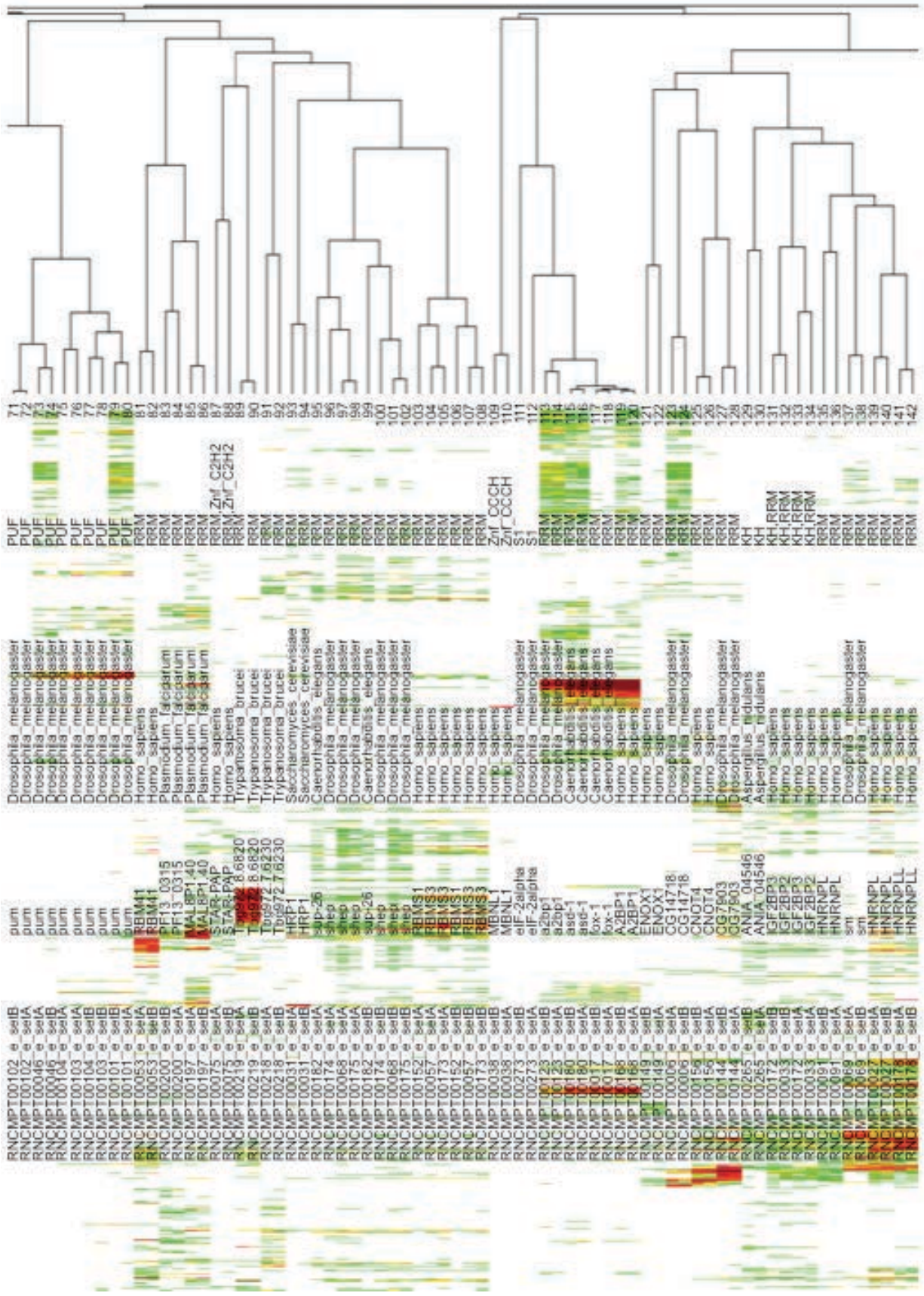
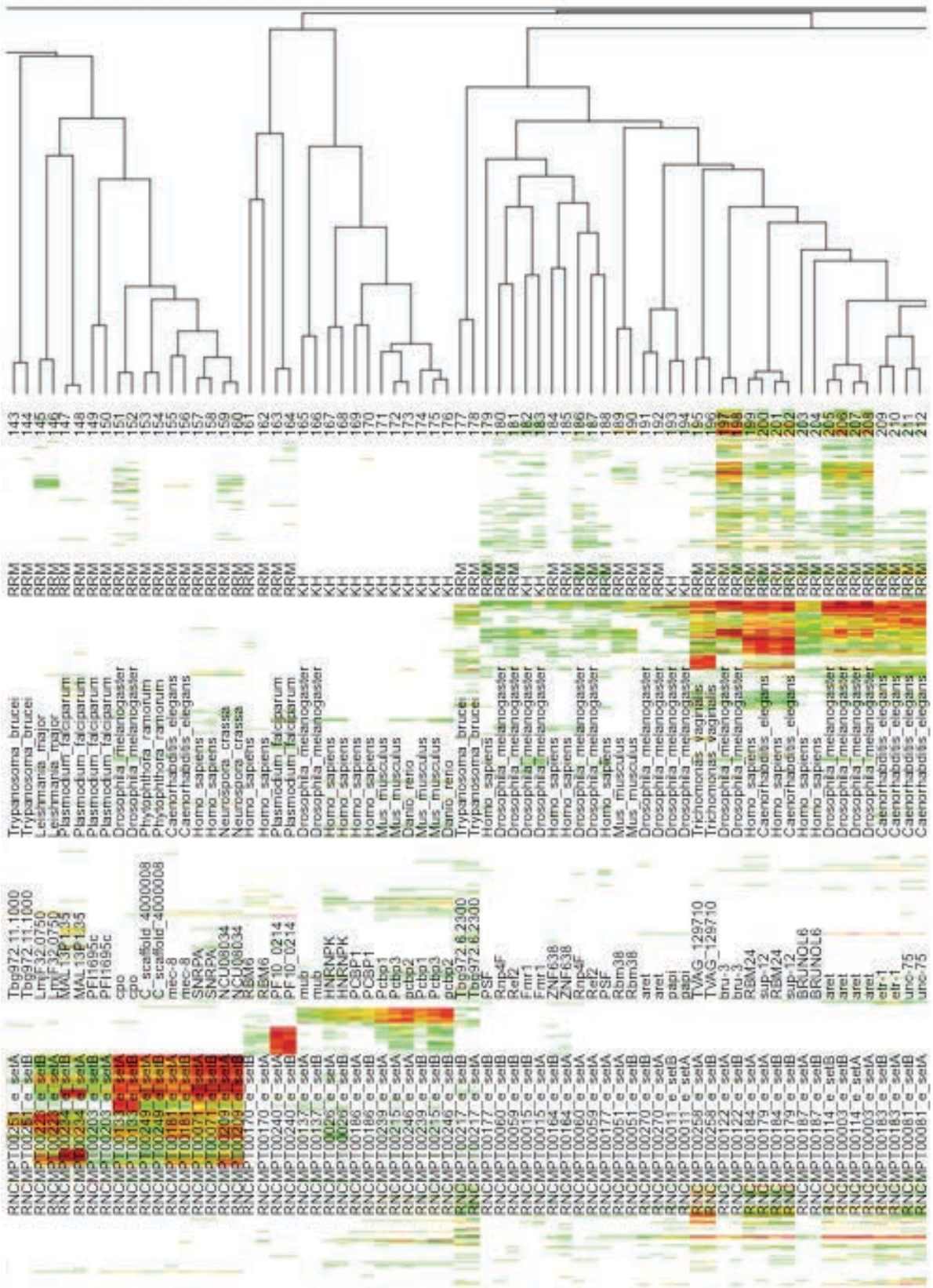


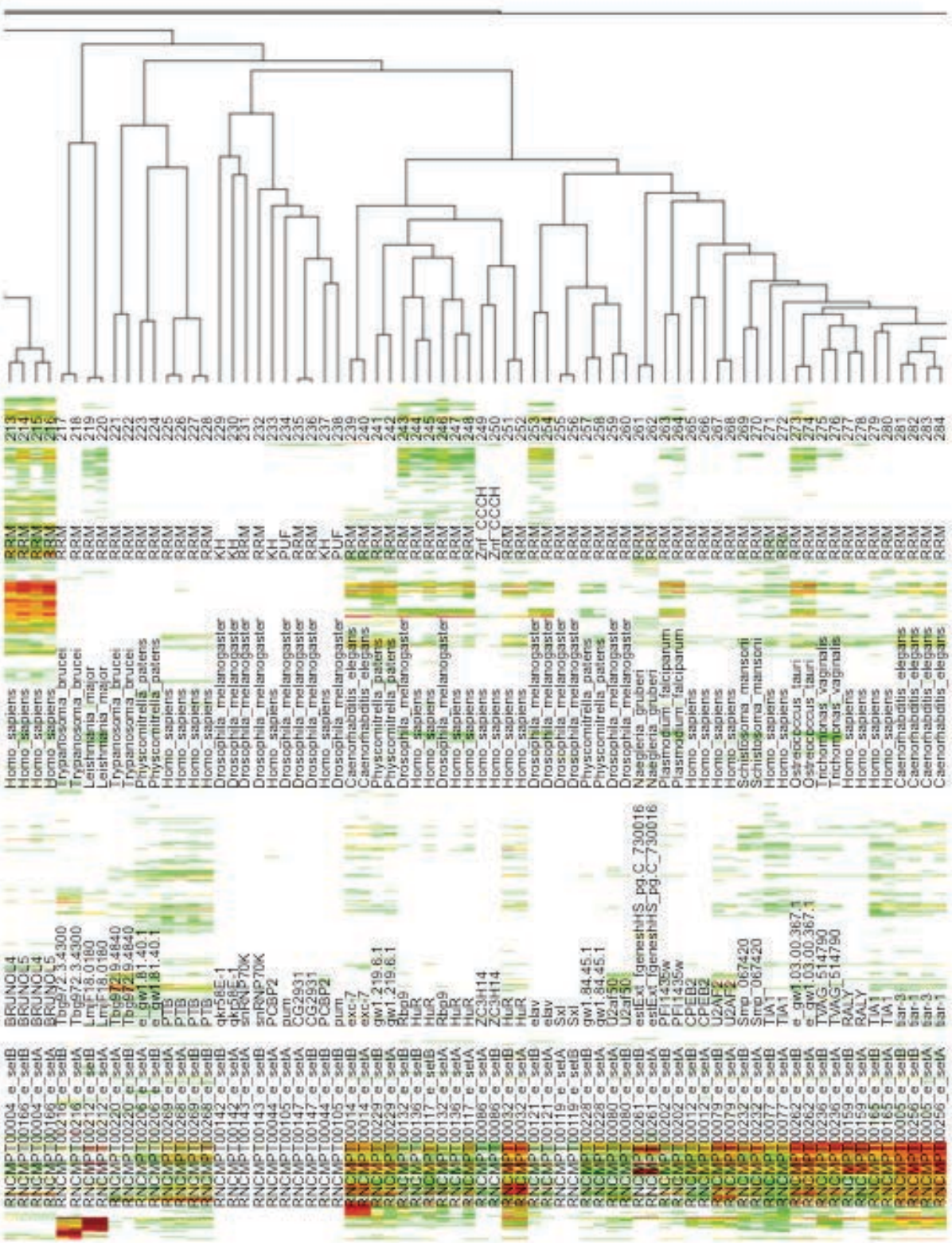

**Figure S6: 2-D hierarchical clustering analysis (Pearson correlation, average linkage) of E-scores for all experimental data, with the two halves of the array kept as separate columns.**

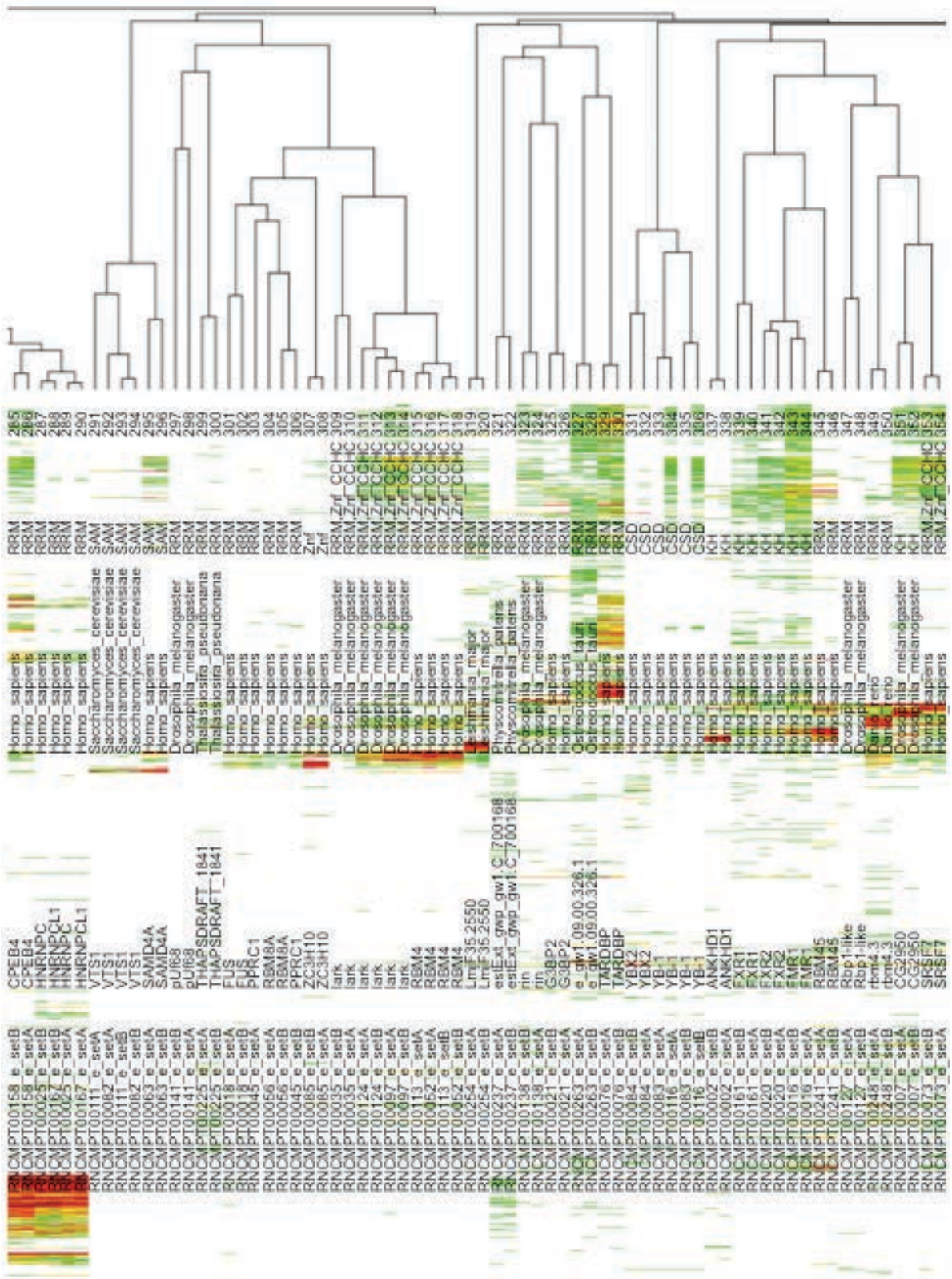
The 3,954 7-mers with E>0.4 in at least one experiment are included. To emphasize higher E-scores, the data were transformed to E' = $10^{10*E-3}$ prior to clustering. This figure is identical to that in **Figure 1C**, with the axes transposed for display. The following pages show segments of the heatmap and dendrogram of experiments, from left to right, with individual experiments labeled. Note that a smaller version of the figure is shown above and a multi-page blow-up of the figure follows this legend and the clustered E-scores are available in **Supplementary Data 5**.

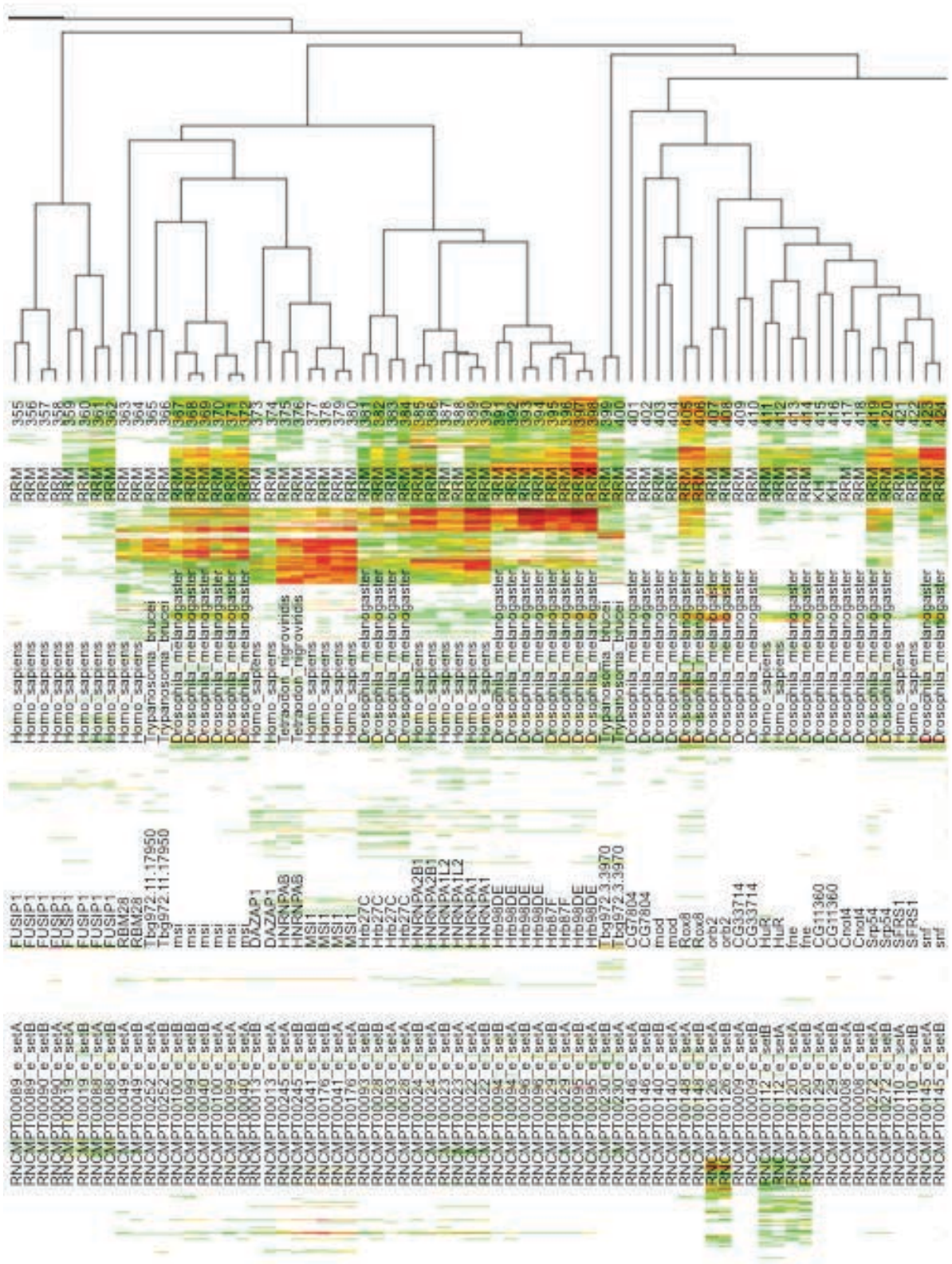


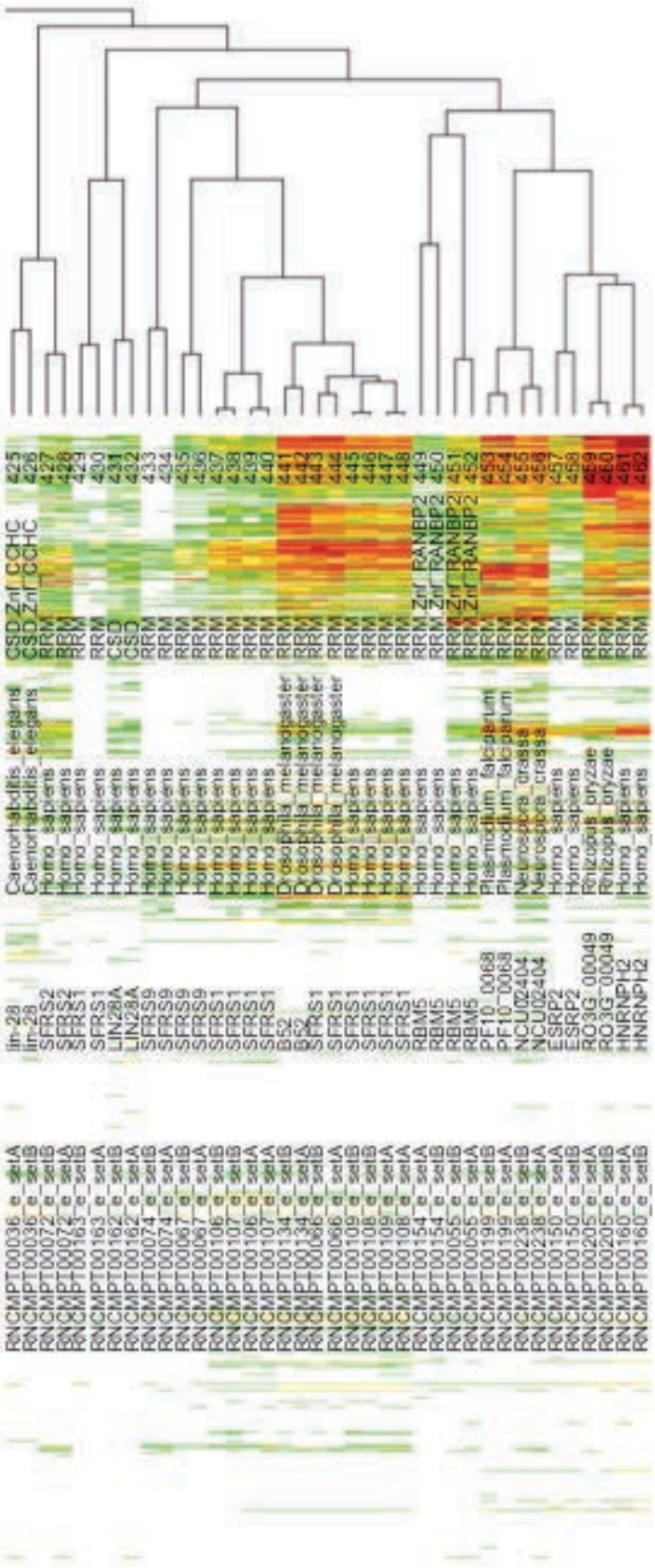











CSD Znf_CCHC
RRM
Znf_RANBP2
Caenorhabditis_elegans
Homo_sapiens
Drosophila_melanogaster
Plasmodium_falciparum
Neurospora_crassa
Rhizopus_oryzae
lin-28
SRSF1
SRSF2
SRSF7
LIN28A
SRSF9
B52
RBM5
PF10_0068
NCU02404
ESRP2
RO3G_00049
HNRNPH2
RNCMPT00036_e_setA
RNCMPT00036_e_setB
RNCMPT00072_e_setA
RNCMPT00072_e_setB
RNCMPT00163_e_setA
RNCMPT00163_e_setB
RNCMPT00162_e_setA
RNCMPT00162_e_setB
RNCMPT00074_e_setA
RNCMPT00074_e_setB
RNCMPT00067_e_setA
RNCMPT00067_e_setB
RNCMPT00106_e_setA
RNCMPT00106_e_setB
RNCMPT00107_e_setA
RNCMPT00107_e_setB
RNCMPT00134_e_setA
RNCMPT00134_e_setB
RNCMPT00066_e_setA
RNCMPT00066_e_setB
RNCMPT00109_e_setA
RNCMPT00109_e_setB
RNCMPT00108_e_setA
RNCMPT00108_e_setB
RNCMPT00154_e_setA
RNCMPT00154_e_setB
RNCMPT00055_e_setA
RNCMPT00055_e_setB
RNCMPT00199_e_setA
RNCMPT00199_e_setB
RNCMPT00238_e_setA
RNCMPT00238_e_setB
RNCMPT00150_e_setA
RNCMPT00150_e_setB
RNCMPT00205_e_setA
RNCMPT00205_e_setB
RNCMPT00160_e_setA
RNCMPT00160_e_setB

## References


58. Philippakis, A.A., Qureshi, A.M., Berger, M.F. & Bulyk, M.L. Design of compact, universal DNA microarrays for protein binding microarray experiments. *J Comput Biol* **15**, 655-665 (2008).
59. Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J. & Giegerich, R. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22**, 500-503 (2006).
60. Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res* **31**, 3429-3431 (2003).
61. Lam, K.N., van Bakel, H., Cote, A.G., van der Ven, A. & Hughes, T.R. Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Res* **39**, 4680-4690 (2011).
62. Kazan, H., Ray, D., Chan, E.T., Hughes, T.R. & Morris, Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol* **6**, e1000832 (2010).
63. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **2**, 28-36 (1994).
64. Foat, B.C., Houshmandi, S.S., Olivas, W.M. & Bussemaker, H.J. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci U S A* **102**, 17675-17680 (2005).
65. Zhao, Y. & Stormo, G.D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol* **29**, 480-483 (2011).
66. Karolchik, D., Hinrichs, A.S. & Kent, W.J. The UCSC Genome Browser. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 1**, Unit1 4 (2012).
67. Lebedeva, S. et al. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol Cell* **43**, 340-352 (2011).
68. Kishore, S. et al. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods* **8**, 559-564 (2011).
69. Mukherjee, N. et al. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol Cell* **43**, 327-339 (2011).
70. Hafner, M. et al. PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp* (2010).
71. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-141 (2010).
72. Sanford, J.R. et al. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res* **19**, 381-394 (2009).
73. Wang, Z. et al. iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol* **8**, e1000530 (2010).
74. Tollervey, J.R. et al. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat Neurosci* **14**, 452-458 (2011).
75. Anders, G. et al. doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* **40**, D180-186 (2012).

76. Wilbert, M.L. et al. LIN28 Binds Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance. *Mol Cell* **48**, 195-206 (2012).
77. Uniacke, J. et al. An oxygen-regulated switch in the protein synthesis machinery. *Nature* **486**, 126-129 (2012).
78. Huang, Y., Genova, G., Roberts, M. & Jackson, F.R. The LARK RNA-binding protein selectively regulates the circadian eclosion rhythm by controlling E74 protein expression. *PLoS ONE* **2**, e1107 (2007).
79. Ortiz-Zapater, E. et al. Key contribution of CPEB4-mediated translational control to cancer progression. *Nat Med* **18**, 83-90 (2012).
80. Sephton, C.F. et al. Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes. *J Biol Chem* **286**, 1204-1215 (2011).
81. Vo, D.T. et al. The RNA-Binding Protein Musashi1 Affects Medulloblastoma Growth via a Network of Cancer-Related Genes and Is an Indicator of Poor Prognosis. *Am J Pathol* **181**, 1762-1772 (2012).
82. Matzat, L.H., Dale, R.K., Moshkovich, N. & Lei, E.P. Tissue-specific regulation of chromatin insulator function. *PLoS Genet* **8**, e1003069 (2012).
83. Ascano, M., Jr. et al. FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492**, 382-386 (2012).
84. Foat, B.C., Morozov, A.V. & Bussemaker, H.J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141-149 (2006).
85. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33**, 1-22 (2010).
86. Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E. & Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714-721 (2009).
87. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
88. Lee, E. & Bussemaker, H.J. Identifying the genetic determinants of transcription factor activity. *Mol Syst Biol* **6**, 412 (2010).
89. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
90. Huang, J.C. et al. Using expression profiling data to identify human microRNA targets. *Nat Methods* **4**, 1045-1049 (2007).