Supplementary computational methods

# preparation and pre-processing for sequencing data

## Genomic reference

We used the hg19 reference genome throughout this project. Chromosome files were downloaded from the UCSC genome browser website [1].

## Preparation of the transcriptome

We prepared a modified transcriptome from RefSeq [2]. The RefSeq exon annotations were acquired from the UCSC genome browser website, as were lookup tables linking each RefSeq identifier to an official gene symbol. After associating each transcript to a gene symbol, we removed any transcript that had exons on multiple chromosomes, or on both positive and negative strands. We then constructed our reference by collapsing all isoforms for a given gene into one super-transcript. We defined the super-transcript boundaries as the 5’ end of the 5’-most exon and the 3’ end of the 3’-most.

## Preparation of regions

We defined the regions of 3’ UTR, 5’ UTR, coding sequence and introns by collapsing all such labelled regions associated with RefSeq identifiers obtained from the UCSC genome browser. In cases where there is ambiguity, we use the following order of preference: 3’ UTR, 5’ UTR, coding, and intron. We consider any region of the genome not falling on of these four categories as intergenic.

# Statistical analyses

## Corrections for multiple hypothesis testing

Unless otherwise stated, all *p*-values reported in this manuscript have been corrected for multiple hypothesis testing using the method of Benjamini and Hochberg [3].

# Read mapping and preliminary PROCESSING FOR sequencing data

## Mapping of data

We constructed two masked versions of the hg19 chromosomes. In the first, we masked all regions that are non-exonic (i.e. not 5’UTR, 3’UTR or coding) with Ns; we call this the *exonic-masked genome*. In the second, we do the same, but allow intronic regions; we call this the *transcript-masked genome*. We also construct a junction database from all of the exon-exon junctions for each super-transcript. We used RMAP [4] to map the iCLIP data to the transcript-masked genome and the junction database, while we mapped the RNA-Seq data to the exon-masked genome.

## Assignment of reads to regions, exons and genes

In the case of the RNA-Seq data, we count the number of reads mapping with their first mapped position within each exon of our super-transcript reference. The count of reads within a gene is then simply the sum of all read counts for the exons in its super-transcript.

# Identification and analysis of hnRNP H1 iCLIP sites and targets

## Peak calling in iCLIP data

We call peaks in iCLIP data using Piranha [5], using a bin size of 10nt but picking the cross-link location as the 5’ end of the peak. We consider significant peaks to be those that have a corrected p-value less than 0.05.

## Peak calling in RIPSeq data

We called peaks in RIP-Seq using Piranha [5], where a ZTNB regression model was used. In this way the responses are the counts for the experimental conditions (i.e. the IP using the hnRNP H1 antibody) and the single covariate used was the counts for the control (IP with non-specific antibody). Because RIP-Seq produces more dispersed peaks, the peaks are called in transcript level.

## 

## Target identification from iCLIP data

Target genes are defined to be any site that is supported with both iCLIP replicates.

## Analysis of motif enrichment in iCLIP data

There is a bias in iCLIP data towards cross-linking at triple-uracil sequences [6]. We observed a strong enrichment for these tri-nucleotides around our identified iCLIP sites. To ameliorate this when trying to determine enriched sequences around H1 binding sites, we computed an expected number of occurrences for each tri-nucleotide. To do this, we identified the top 1000 most enriched locations from a set of public iCLIP datasets using Piranha [5] – dataset details below – and counted the number of times each possible tri-nucleotide occurs within +/- 2nt of the cross-link location as follows:

,

where  is the normalized count for tri-nucleotide , computed over  total datasets and  is the indicator function that returns  if the tri-nucleotide starting at position  of sequence  from dataset  is equal to , and  otherwise. We compute the observed counts from our three H1 iCLIP replicates analogously:

,

where  is the number of significant iCLIP sites reported for replicate j. We use the following public iCLIP datasets to compute the expected values:

|  |  |
| --- | --- |
| RBP | Citation |
| HuR | [7] |
| TIAL | [8] |
| TIA1 | [8] |
| hnRNPC | [9] |
| TDP43 | [10] |

# Analysis of changes in mRNA levels and exon-inclusion rates

## 

## Identification of exons with changed inclusion ratio from RNA-Seq data

To identify those exons with changes in inclusion ratios between the control and the knockdown condition, we paired each control replicate with its corresponding KD replicate and for each gene we constructed a two-by-two contingency table as follows:

|  |  |  |
| --- | --- | --- |
|  | Total RNA | H1 KD |
| Within target exon | T11 | T12 |
| Within other exons (same gene) | T21 | T22 |

where:

* T11 is the number of reads mapping into the exon of interest in the total RNA sample
* T12 is the number of reads mapping into the exon of interest in the H1 KD sample
* T21 is the number of reads mapping into the gene, but outside the exon of interest in the total RNA sample
* T22 is the number of reads mapping into the gene, but outside the exon of interest in the H1 KD sample

We then performed Fisher’s exact test on this table to compute an odds-ratio and a two-tailed p-value for the significance of the change from an odds-ratio of 1. We perform this comparison for the 3 pairs of control/KD samples. We considered significant changes to be those with a corrected *p*-value less than 0.01 and an odds ratio greater than 1.5 (increase) or less than 0.66 (decrease) in three replicates.

## Calculation of base-pair probability

To calculate the base-pair probability, we selected a subset of the significant iCLIP sites such that each was at least 200 nucleotides from the closest other site. We then computed the base-pair probabilities for a window of 100 nucleotides around each selected iCLIP site using a modified version of the RNA Vienna package [11]. Then we took the average base pair probability for each location in the window over all sequences. Since the folding algorithm favours the ends of the sequences to be single stranded to obtain more stable structures, the base pairing probabilities of the ends of sequences are biased towards zero. In order to fix this problem, although we folded a window of 100 nucleotides around the peaks, we only took the base pairing probability of the middle 80 nucleotides into account, and dropped 10 nucleotides from each end of all sequences.

## Identification of differentially expressed genes from RNA-Seq data

We used EdgeR to identify differentially expressed gene [12]. Two replicates of control total RNA and two replicates of hnRNP H1 knockdown RNA-Seq were used to construct a 4xN matrix of gene-read counts (where N is the number of genes in our reference) that was provided to EdgeR. We considered those genes reported as having a corrected *p*-value less than 0.05 as showing significant changes in expression. We divided this set into two lists: those genes that were up-regulated on H1 KD (i.e. had a greater normalized read-count in the H1 KD condition) and those genes that were down-regulated on H1 KD (i.e. had a lesser normalised read-count in the H1 KD condition)

# The effects of hnrnp h1 on alternative polyadenylation

## Changes in poly-A

Locations of alternative poly-A sites are taken from [13,14].To identify changes, we split 3' UTRs into segments bounded by polyA sites and tested for a change in ratio of reads within a segment relative to the whole 3' UTR using Fisher's exact test (similar to the test for changes in exon usage; see above). We classified 3' UTRs as lengthened if they contained a contiguous run of segments showing decreased usage upon hnRNP H1 knockdown followed by a contiguous run of segments showing increased usage upon hnRNP H1 knockdown. Conversely, we classified 3' UTRs as shortened if we observed a contiguous run of segments showing increased usage on hnRPH H1 knockdown followed by a contiguous run of segments showing decreased usage on hnRNP H1 knockdown.

# intron retention

## Intron retention

Changes in intron usage were called similarly to changes in exon usage (see above).

# References

1. Dreszer, T.R., et al., *The UCSC Genome Browser database: extensions and updates 2011.* Nucleic Acids Research, 2012. **40**(D1): p. D918-D923.

2. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Research, 2012. **40**(D1): p. D130-D135.

3. Hochberg, Y.B.a.Y., *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.

4. Smith, A.D., et al., *Updates to the RMAP short-read mapping software.* Bioinformatics, 2009. **25**(21): p. 2841-2842.

5. Uren, P.J., et al., *Site identification in high-throughput RNA–protein interaction data.* Bioinformatics, 2012. **28**(23): p. 3013-3020.

6. Sugimoto, Y., et al., *Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions.* Genome Biology, 2012. **13**(8): p. R67.

7. Uren, P.J., et al., *Genomic Analyses of the RNA-binding Protein Hu Antigen R (HuR) Identify a Complex Network of Target Genes and Novel Characteristics of Its Binding Sites.* Journal of Biological Chemistry, 2011. **286**(43): p. 37063-37066.

8. Wang, Z., et al., *iCLIP Predicts the Dual Splicing Effects of TIA-RNA Interactions.* PLoS Biol, 2010. **8**(10): p. e1000530.

9. Konig, J., et al., *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution.* Nat Struct Mol Biol, 2010. **17**(7): p. 909-915.

10. Tollervey, J.R., et al., *Characterizing the RNA targets and position-dependent splicing regulation by TDP-43.* Nat Neurosci, 2011. **14**(4): p. 452-458.

11. Hofacker, I.L., *Vienna RNA secondary structure server.* Nucleic Acids Research, 2003. **31**(13): p. 3429-3431.

12. Robinson, M., McCarthy, D., Chen, Y., Smyth, G.K.: edger: differential expression analysis of digital gene expression data user’s guide (2011)

13. Cheng Y, Miura RM, Tian B. Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics. 2006 Oct 1;**22**(19):2320-5.

14. Zhang H, Hu J, Recce M, Tian B. PolyA\_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res. 2005 Jan 1;**33**(Database issue):D116-120.