Landes Bioscience www.landesbioscience.com

Supplemental Material to:

Oliver Rossbach, Lee-Hsueh Hung, Ekaterina Khrameeva, Silke Schreiner, Julian König, Tomaž Curk, Blaž Zupan, Jernej Ule, Mikhail S Gelfand, Albrecht Bindereif

Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L

2014; 11(2)
http://dx.doi.org/10.4161/rna.27991

www.landesbioscience.com/journals/rnabiology/article/27991/

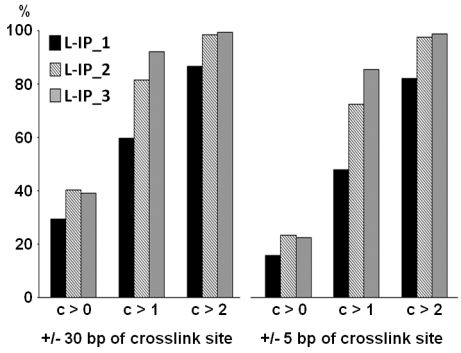
Supplementary Figures

A

	sequence reads (Solexa 50-bp)	crosslink sites in chromosomes 1-22 and X
hnRNP L-IP_1	13 327 397	590 960
hnRNP L-IP_2	1 220 636	259 045
hnRNP L-IP_3	1 210 822	303 483
FLAG-IP_1	469 975	18 857
FLAG-IP_2	8 571	1 092
FLAG-IP_3	78 225	9 279

1 109 962 hnRNP L crosslink sites
↓ *clustering*
786 192 crosslink sites clusters
↓ *filtering*
415 606 crosslink site clusters
(containing 622 798 crosslink sites)

B

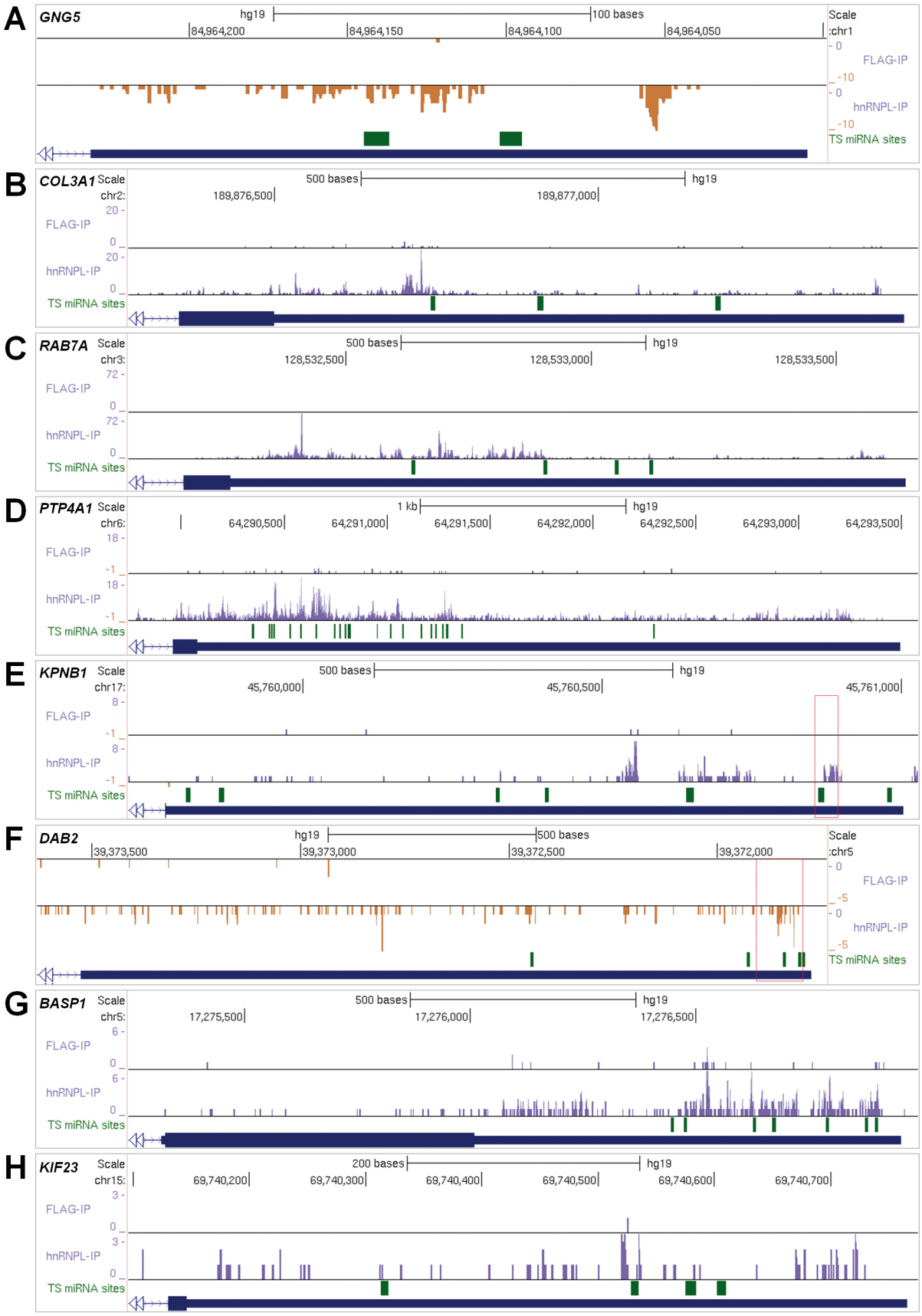


Supplementary Figure S1.

HnRNP L crosslink-site data processing and reproducibility.

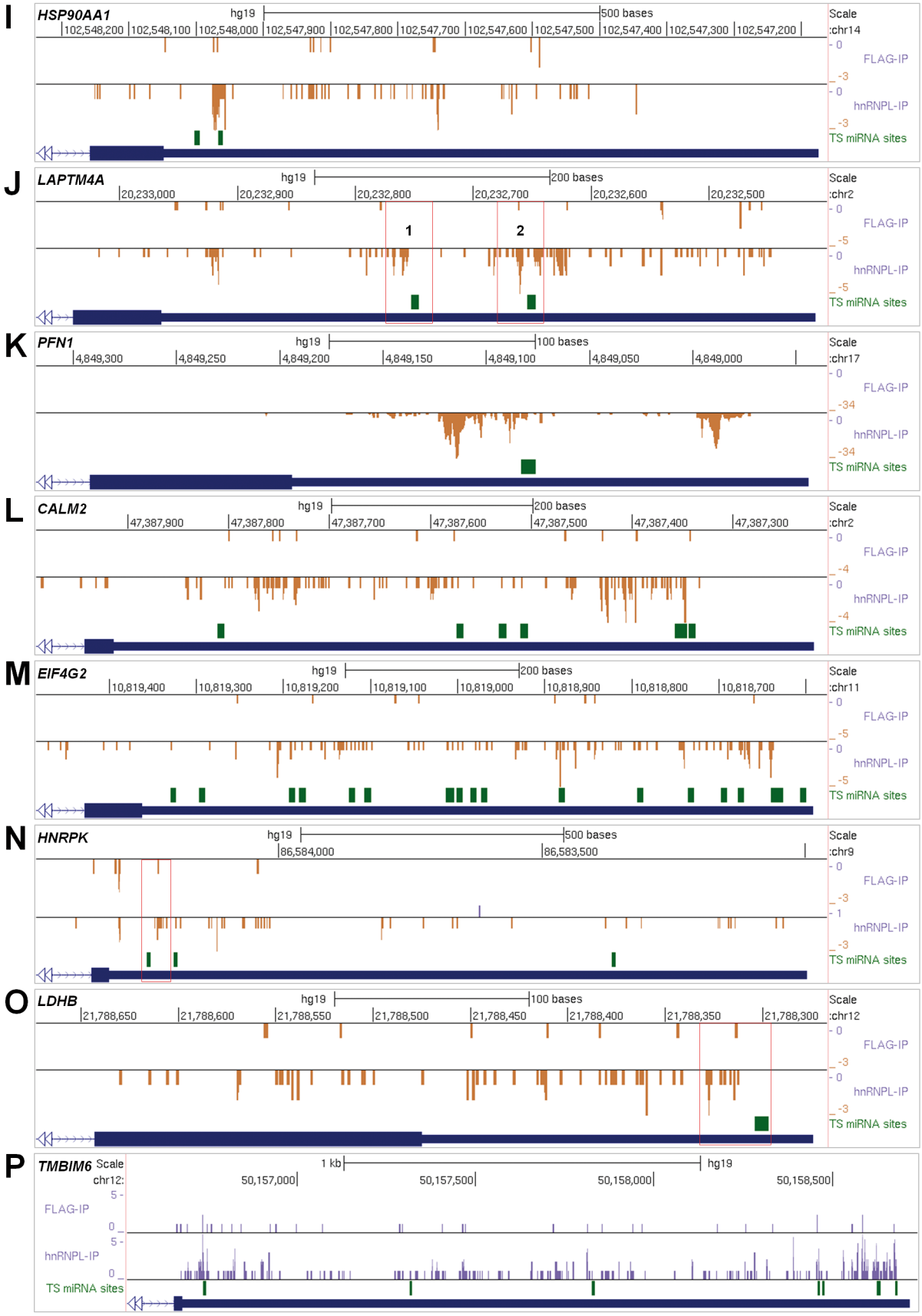
(**A**) Flow chart of data processing from Solexa sequence reads, derived from three independent hnRNP L iCLIP experiments (hnRNP L-IP_1 to _3) and three FLAG controls (FLAG-IP_1 to _3), to filtered crosslink sites, which entered functional analysis. Note that the initial sequence read number does not represent the entire data content of the experiments. Only random barcode-filtered, uniquely mapped reads have an absolute quantitative value, since the deviation between sequence reads and crosslink sites is due to multiple sequencing of the same PCR product. The twofold difference between hnRNP L-IP_1 and the other two experiments is due to experiment 1 being sequenced separately, and experiments 2 and 3 together in one Solexa flow-cell lane.

(**B**) Reproducibility of three hnRNP L iCLIP experiments (for details, see *Supplementary Materials and Methods*, section 1).


A
GNG5
hg19
100 bases
84,964,200
84,964,150
84,964,100
84,964,050
Scale
chr1
FLAG-IP
hnRNPL-IP
TS miRNA sites
B
COL3A1
Scale
chr2:
500 bases
hg19
189,876,500
189,877,000
FLAG-IP
hnRNPL-IP
TS miRNA sites
C
RAB7A
Scale
chr3:
500 bases
hg19
128,532,500
128,533,000
128,533,500
FLAG-IP
hnRNPL-IP
TS miRNA sites
D
PTP4A1
Scale
chr6:
1 kb
hg19
64,290,500
64,291,000
64,291,500
64,292,000
64,292,500
64,293,000
64,293,500
FLAG-IP
hnRNPL-IP
TS miRNA sites
E
KPNB1
Scale
chr17:
500 bases
hg19
45,760,000
45,760,500
45,761,000
FLAG-IP
hnRNPL-IP
TS miRNA sites
F
DAB2
hg19
500 bases
39,373,500
39,373,000
39,372,500
39,372,000
Scale
chr5
FLAG-IP
hnRNPL-IP
TS miRNA sites
G
BASP1
Scale
chr5:
500 bases
hg19
17,275,500
17,276,000
17,276,500
FLAG-IP
hnRNPL-IP
TS miRNA sites
H
KIF23
Scale
chr15:
200 bases
hg19
69,740,200
69,740,300
69,740,400
69,740,500
69,740,600
69,740,700
FLAG-IP
hnRNPL-IP
TS miRNA sites

1

2


I
HSP90AA1
hg19
500 bases
102,548,200 102,548,100 102,548,000 102,547,900 102,547,800 102,547,700 102,547,600 102,547,500 102,547,400 102,547,300 102,547,200
Scale
chr14
FLAG-IP
hnRNPL-IP
TS miRNA sites
J
LAPTM4A
hg19
200 bases
20,233,000 20,232,900 20,232,800 20,232,700 20,232,600 20,232,500
1
2
Scale
chr2
FLAG-IP
hnRNPL-IP
TS miRNA sites
K
PFN1
hg19
100 bases
4,849,300 4,849,250 4,849,200 4,849,150 4,849,100 4,849,050 4,849,000
Scale
chr17
FLAG-IP
hnRNPL-IP
TS miRNA sites
L
CALM2
hg19
200 bases
47,387,900 47,387,800 47,387,700 47,387,600 47,387,500 47,387,400 47,387,300
Scale
chr2
FLAG-IP
hnRNPL-IP
TS miRNA sites
M
EIF4G2
hg19
200 bases
10,819,400 10,819,300 10,819,200 10,819,100 10,819,000 10,818,900 10,818,800 10,818,700
Scale
chr11
FLAG-IP
hnRNPL-IP
TS miRNA sites
N
HNRPK
hg19
500 bases
86,584,000 86,583,500
Scale
chr9
FLAG-IP
hnRNPL-IP
TS miRNA sites
O
LDHB
hg19
100 bases
21,788,650 21,788,600 21,788,550 21,788,500 21,788,450 21,788,400 21,788,350 21,788,300
Scale
chr12
FLAG-IP
hnRNPL-IP
TS miRNA sites
P
TMBIM6
Scale
chr12:
1 kb
hg19
50,157,000 50,157,500 50,158,000 50,158,500
FLAG-IP
hnRNPL-IP
TS miRNA sites

1

2

3

Supplementary Figure S2.

HnRNP L binds preferentially in close proximity to predicted miRNA binding sites: searching for co-regulated targets.

Labeling in each panel from top: scale with genome annotation, chromosome position, crosslink sites obtained from the control (FLAG-IP) and the hnRNP L iCLIP experiment (hnRNPL-IP). Crosslink sites are represented either in blue (plus strand) or in orange (minus strand). Below (in green) the TargetScan-predicted miRNA binding sites (TS miRNA sites), and on the bottom, the gene structure in the 3' UTR region are represented.

(**A**)-(**P**) UCSC browser views of candidate 3' UTR targets predicted to be coregulated by hnRNP L and a specific miRNA. Targets shown in panels A-G were derived from scanning for dense binding clusters (based on hnRNP L iCLIP data) that overlap with TargetScan-predicted miRNA binding sites of miRNAs expressed in HeLa cells.[44] Targets shown in panels H-P were derived from scanning the top Ago2 binding clusters (based on PAR-CLIP data[45]) that overlap with dense hnRNP L binding clusters and TargetScan-predicted miRNA binding sites.

The target 3' UTR regions that were cloned in the firefly luciferase reporter (**Fig. 6E**) are indicated by red boxes. Note that in the case of *LAPTM4A* (**Fig. S2J**), two different 3' UTR regions were investigated (labeled by digits 1 and 2; *LAPTM4A*.1 and *LAPTM4A*.2 in **Fig. 6E**).

Supplementary Materials and Methods

1. Data processing of hnRNP L crosslink sites from three independent experiments.

All crosslink sites in chromosomes 1-22 and X derived from three independent iCLIP experiments of hnRNP L were combined, yielding a total number of 1,109,962 sites (see **Fig. S1A** for detailed numbers from each dataset). To assess the reproducibility between the three experiments, we took each crosslink site (*position_A*) from one experiment and checked whether the crosslink site could be found in at least one of the two experiments within 30 bp (or 5 bp) up- and downstream of position_A (defined here as *positive*). The crosslink sites were further grouped according to their cDNA count (uniquely mapped sequenced reads, random barcode filtered; see ref. 25 for details) as following:

a) all sites taken (c > 0),

b) minimal cDNA count of 2 (c > 1),

c) minimal cDNA count of 3 (c > 2).

Fig. S1B shows the percentage of crosslink sites, which are *positive*, if a 30-bp (left panel) or a 5-bp (right panel) distance was used in each group. The values in experiment L-IP_1 were lower than in experiments L-IP_2 and L-IP_3, due to the approximately tenfold higher crosslink site count in experiment L-IP_1. The very high percentage (87% in 30-bp; 82% in 5-bp) of *positive* crosslink sites (in group c>2) confirms the high reproducibility of these three experiment.

Crosslink sites were grouped into one cluster, if the distance between two sites was less than 31 nts. Sequence motif analysis was applied with these 785,192 cluster sequences (including 30 bp up- and downstream regions of each cluster). Based on our SELEX consensus for hnRNP L, the frequency of each 4-mer sequence motif found in all SELEX-derived binding sequences was calulated.[1] We defined the *score-index* of each cluster sequence as the sum of the frequencies for all 4-mer motifs divided by the length of the sequence. Sequences in the clusters with very low *score-indices* are T- or A- rich and contain T- or A-stretches at a high frequency. We interpret this to mean that the T-rich

sequences are due to the chemical bias and uridine preference of UV-crosslinking. Since hnRNP L binds frequently in 3' UTR regions (see **Fig. 2B**) and the crosslink site density increases in the proximity of the poly(A) sites (see **Fig. 4**), the sequence reads containing A-stretches were most likely mapped incorrectly, contributing to the enrichment of A-rich sequence motifs. Therefore data filtering to exclude cross-link sites within these low-score clusters would improve our functional analysis. As a result, we selected 415,606 clusters (~53% of total) with a *score-index* higher than 0.06. All crosslink sites (n = 622,798) in the 415,606 selected clusters were used for downstream data analysis.

2. Pentamer enrichment analysis.

To investigate the hnRNP L binding motif in the iCLIP dataset, we evaluated the occurrence of all pentamers within a -30 to -10 and +10 to +30 window around the crosslink sites, excluding the central 20 positions because of the strong chemical uridine bias of the UV-crosslink.[47] The enrichment was determined by comparing the occurrence of each pentamer around the true crosslink site with the occurrence at randomized crosslink positions. The positions of crosslink sites were randomized within the same regions of the gene (i.e., within the same intron, CDS, or UTR). Pentamers were then ranked by their enrichment, and the top 10 pentamers are shown in **Fig. 2A**.

3. Distribution of hnRNP L crosslink sites around splice sites of protein-coding genes.

We used the ENCODE GENCODE V4 Gene Annotations (level 1+2+3; May 2010) for functional analysis of the hnRNP L crosslink sites.

First, all internal exons of all annotated transcripts from protein-coding genes were selected, and their 3' and 5' splice sites checked separately. All 3' (5') splice sites were grouped as **alternative** 3' (5') splice sites, if there was evidence of their alternative usage, otherwise, as **constitutive**.

The number of crosslink sites (frequency) found around the alternative 3' splice sites (positions -200 to +75, total of 60,959 splice sites; **Fig. 3A**, upper panel left) and 5' splice

sites (positions -75 to +200, total of 57,898 splice sites; **Fig. 3A**, upper panel right) were plotted with the position index on the X-axis and the frequency on the Y-axis. In analogy, the frequency around constitutive 3' splice sites (positions -200 to +75, total of 143,296 splice sites; **Fig. 3A**, middle panel left) and 5' splice sites (positions -75 to +200, total of 146,357 splice sites; **Fig. 3A**, middle panel right) were plotted.

For direct comparison, the frequency of constitutive versus alternative splice sites was normalized by the ratio of the total number of alternative to constitutive splice sites (60,959 / 143,296 = 0.396 for 3' splice sites; 57,898 / 146,357 = 0.425 for 5' splice sites) and plotted in smooth lines (**Fig. 3A**, lower panel in green: alternative, in black: constitutive).

4. Alternative splicing targets detected by exon microarray analysis after hnRNP L knockdown and cycloheximide treatment in HeLa cells.

Total RNA after hnRNP L knockdown (luciferase GL2 as control), and cycloheximide treatment was isolated from HeLa cells.[24] All 6 samples (three biological replicates each of luciferase control and hnRNP L RNAi knockdown) were processed according to Affymetrix's standard protocol (GeneChip Whole Transcript (WT) Sense Target Labeling Assay Manual). The GeneChip Human Exon 1.0 Array (http://www.affymetrix.com) was applied to monitor differences of exon expression signals between the two sample groups. Normalized signal intensities (in $\log_2$ values) of *present* (expressed) probesets were selected for further downstream analysis.[10]

First, *present* probesets were re-assigned (based on their position annotation) to the exons of the protein-coding genes from ENCODE GENCODE V4 Gene Annotations. Second, 8,376 genes (defined here as *expressed multi-exon genes*) with more than three exons and with probesets in at least 50% of their constitutive exons (both constitutive 3' and 5' splice sites as described in section 2) were selected.

Second, the expression level of each gene was calculated with both median (*gene_mean*) and mean (*gene_mrdian*) of the signal intensity of all exons, and the mean values of the three replicates in each sample group were assigned. The differential

expression of each gene between the two sample groups (hnRNP L knockdown versus luciferase) was calculated with both *gene_mean* (*diff_gene_mean*) and *gene_median* (*diff_gene_median*) values. A total of 5,828 genes without significant gene expression changes (both *diff_gene_mea*n and *diff_gene_median* value within the range of -0.25 to 0.25) were selected for alternative splicing target analysis.

Third, only internal exons in the 5,828 genes selected were further analyzed. The differential expression of each exon (*diff_exon*; hnRNP L knockdown – luciferase) between the two sample groups (hnRNP L knockdown - luciferase) was calculated with the mean values of the exon sinal from the three replicates. Exon inclusion index (EI) of each exon was then calculated by three methods:

1) *EI_mean: diff_exon - diff_gene_mean*,

2) *EI_median: diff_exon - diff_gene_median*,

3) *EI_Zscore: z-score of each diff_exon from all exons of the gene*.

Two types of alternative splicing targets of hnRNP L were detected:

1) hnRNP L-activated: Exon inclusion index lower than 0 (exon inclusion rate decreased upon hnRNP L knockdown). 890 exons meet the criteria below and were selected as targets where hnRNP L acts as splicing activator:

EI_mean ≤ -0.75 AND *EI_median* ≤ -0.75 AND *EI_Zscore* ≤ -2.

2) hnRNP L-repressed: Exon inclusion index higher than 0 (exon inclusion rate increased upon hnRNP L knockdown). 574 exons meet the criteria below and were selected as targets where hnRNP L acts as splicing repressor:

EI_mean ≥ 0.75 AND *EI_median* ≥ 0.75 AND *EI_Zscore* ≥ 2.

5. Distribution of hnRNP L crosslink sites around alternative splicing target exons.

Three groups of exons were defined (as described above in section 3): L-activated, L-repressed, and background (all internal exons from the 5,828 genes expressed and without

significant expression changes as described in section 3). The number of crosslink sites (frequency) found around the 3' (positions -300 to +75) and 5' splice sites (positions -75 to +300) of exon grouped as activated (n = 890, **Fig. 3B** upper panel, red) and repressed (n = 574, **Fig. 3B** middle panel, blue) were plotted with position index on the X-axis and frequency on the Y-axis. The frequency of background exons were plotted in black. For comparison, the frequency of activated and repressed target were plotted in smooth lines (**Fig. 3B** lower panel).

6. Distribution of hnRNP L crosslink sites around polyadenylation sites.

All polyadenylation sites from protein-coding genes were selected, and grouped into 1) *internal*: all poly(A) sites upstream of the annotated most downstream poly(A) site, 2) *terminal*: most downstream poly(A) site of each gene.

The number of crosslink sites (frequency) found around the poly(A) sites (positions -300 to +300) of the *internal* group (total of 17,576 sites; **Fig. 4**, upper panel) and *terminal* group (total of 19,687 sites; **Fig. 4**, middle panel) were plotted with the position index on the X-axis and the frequency on the Y-axis. For comparison, the frequency of *internal* versus *terminal* poly(A) sites – after normalization by the ratio of *internal*-to-*terminal* sites (17,576 / 19,687 = 0.893) were plotted in smooth lines (**Fig. 4**, lower panel).

7. Distribution of hnRNP L crosslink sites in 3'UTR and miRNA target sites.

For miRNA target sites we extracted the conserved mammalian miRNA regulatory target sites in the 3' UTR regions of RefSeq genes as predicted by TargetScanHuman 5.1 (http://www.targetscan.org). The crosslink site counts (frequency) around the miRNA target sites (position -80 to +80 of target site, on X-axis) were plotted on the Y-axis (**Fig. 5C**)

From the 3' UTRs of the protein-coding genes from ENCODE annotation, 5,062 3'UTRs containing both crosslink and miRNA target sites were selected for further analysis. We compared the crosslink site density (crosslink site count / sequence length) around miRNA target sites (**MS**, from position -20 to +20 of target site) with the density outside the

miRNA target sites in 3' UTR (**NS**). For each of the 5 062 3' UTRs, the density of **MS** was plotted on the X-axis and **NS** on the Y-axis (**Fig. 5D**). The red line shows the slope (0.59) using the Linear Model. The density in miRNA target sites was significantly higher than non-target region in 3' UTR (p value = 2.2×10^{-16}).

8. Oligonucleotides.

RT-PCR primers (Sigma):

GNG5 fwd	AAGTGGTTCAACAGCTCCGG
GNG5 rev	CTGGGGTCTGAAGGGATTTGT
COL3A1 fwd	TCGAGGCAGTGATGGTCAAC
COL3A1 rev	CGGGACCCATTTCGCCTTTA
RAB7A-1 fwd	GGAGGTGATGGTGGATGACA
RAB7A-1 rev	GCACCTCTGTAGAAGGCCAC
RAB7A-2 fwd	AAACGGAGGTGGAGCTGTAC
RAB7A-2 rev	TGTGCTCAACTCTCACTGCC
PTP4A1 fwd	GCTCCAGTACTTGTTGCCCT
PTP4A1 rev	TGGAATCTTTGAAACGCAGCC
KPNB1 fwd	ACCACATTGCTGGAGATGAGG
KPNB1 rev	CGATCTCCGCCCTTCAGTTA
DAB2 fwd	TCTGTCCAGTCCTCACCACA
DAB2 rev	CTGAGACGGGAGGAGCAAAG
BASP1 fwd	GCTAACTCAGGGGCTGCATA
BASP1 rev	CTCCTTGGCTTTCTCGTCGT
KIF23 fwd	GGACCAGAGCAGAAGGGAAC
KIF23 rev	ACACACGATCATCCGCACTT
HSP90AA1 fwd	GTACGCTTGGGAGTCCTCAG
HSP90AA1 rev	TCTTCAGCCTCATCATCGCTT
LAPTM4A fwd	TCCATGCCAGCTGTCAACAT

LAPTM4A rev	GGAATCAGCCAACCCACTTG	
PFN1 fwd	CGCCTACATCGACAACCTCA	
PFN1 rev	AAATTCCCCATCCTGCAGCA	
CALM2 fwd	GGCAGAATCCCACAGAAGCA	
CALM2 rev	CACATGGCGAAGTTCTGCAG	
EIF4G2 fwd	GTGGAAATGCAAATGAGGCTGT	
EIF4G2 rev	ACCAGCTCTGAAATGATGGCA	
HNRNPK fwd	TGCGAGTTGAGGCTGTTGAT	
HNRNPK rev	TAAGGCTGTGCACGTCCTTT	
LDHB fwd	CAGCAAGAAGGGGAGAGTCG	
LDHB rev	CACGCGGTGTTTGGGTAATC	
TMBIM6 fwd	GTCATGTGTGGCTTCGTCCT	
TMBIM6 rev	TTGGGAAAGGCTGGATGGTC	
ACTB fwd	TGGACTTCGAGCAAGAGATG	
ACTB rev	GTGATCTCCTTCTGCATCCTG	
GAPDH fwd	GAGTCAACGGATTTGGTCGT	
GAPDH rev	GATCTCGCTCCTGGAAGATG	
U1 snRNA fwd	GGGGAGATACCATGATCACG	
U1 snRNA rev	GTCGAGTTTCCCACATTTGG	

siRNAs:

hnRNP L 3'-UTR	GACAUUUCUCUUUCCUUUAdTdT	(Sigma)
hnRNP L H1 (exon 4)	GAAUGGAGUUCAGGCGAUGdTdT	(MWG Biotech)
luciferase GL2	CGUACGCGGAAUACUUCGAdTdT	(MWG Biotech)

iCLIP RNA 3'-linker (Dharmacon):

P-UGAGAUCGGAAGAGCGGUUCAG-Puromycin

iCLIP RT-primers (Eurogentech):

iCLIP-RT1	P-NNAACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC
iCLIP-RT2	P-NNACAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC
iCLIP-RT3	P-NNCTAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC
iCLIP-RT4	P-NNCATTNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC
iCLIP-RT5	P-NNGCCANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC

iCLIP BamHI-linearization oligonucleotide (MWG Biotech):

Cut_oligo	GTTCAGGATCCACGACGCTCTTCAAAA

Supplementary Material References

47. Sugimoto Y, König J, Hussain S, Zupan B, Curk T, Frye M, Ule J. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol 2012; 13:R67.