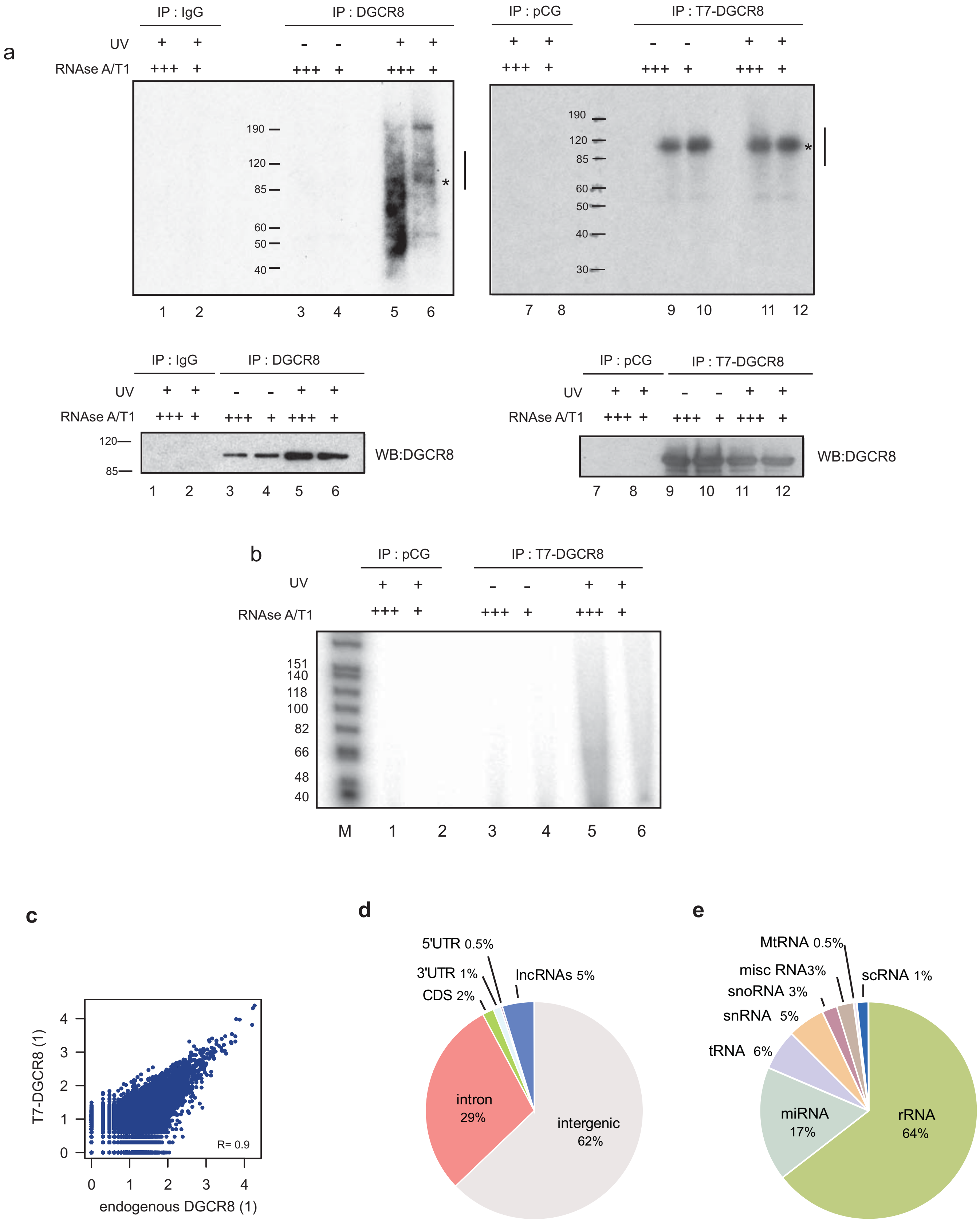
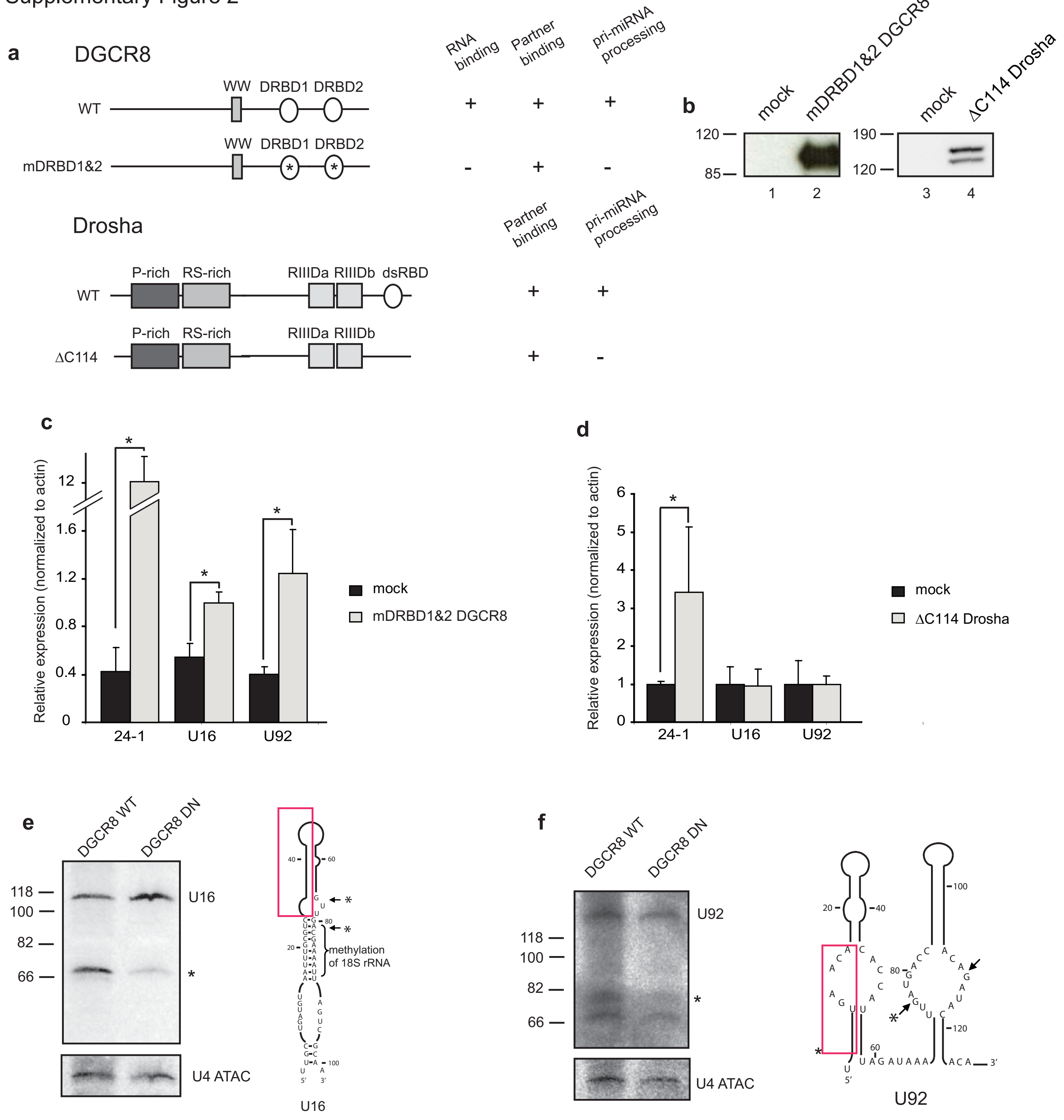
Supplementary Figure 1



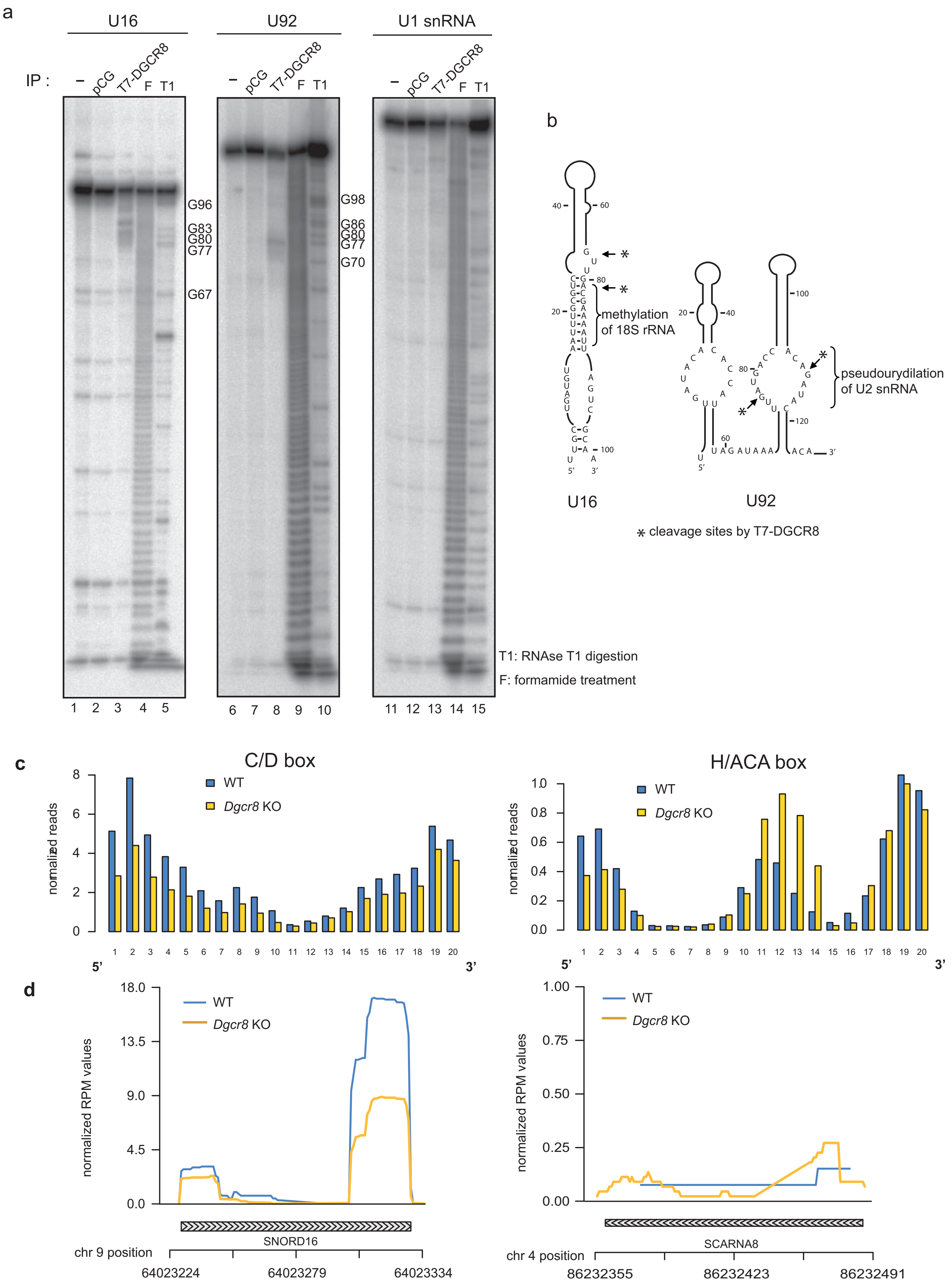
Supplementary Figure 1 HITS-CLIP analysis of endogenous and T7-tagged DGCR8 protein (**a**) Protein extracts were prepared from UV and non-UV cross-linked HEK293T cells and RNA was partially digested using high (+++) or low (+) RNase concentrations. Endogenous and overexpressed DGCR8-RNA complexes were immunopurified from cells extracts using an antibody against DGCR8 or T7 epitope, respectively. Following immunoprecipitation, RNA associated to DGCR8 was 5'end labelled using T4 PNK (polynucleotide kinase). Complexes were size-separated using denaturing gel electrophoresis and transferred to a nitrocellulose membrane. The upper left panel shows the autoradiogram for endogenous DGCR8 CLIP. No radioactive signal was detected when cells were not cross-linked (lanes 3 and 4) or matching isotype IgG was used as a control (lanes 1 and 2). The upper right panel shows the autoradiogram for overexpressed T7-DGCR8 CLIP. No radioactive signal was detected when performing immunoprecipitations with anti-T7 epitope antibody in non-transfected extracts (lanes 7 and 8). Immunoprecipitated T7-DGCR8 protein was shown to be phosphorylated in the non-UV irradiated condition (lanes 9 and 10), but no RNA was shown to be associated in this condition (compare lanes 3 and 4 with 5 and 6 in panel b). The bottom panels show the presence of both endogenous DGCR8 (left) and overexpressed T7-DGCR8 protein (right) in the immunoprepitated CLIP material (**b**) RNAs associated to overexpressed DGCR8 protein range from 40 to 100 nt long when cross-linked extracts were used for immunoprecipitation (lanes 5 and 6), instead, no RNA was found to be associated to T7-DGCR8 when non-crosslinked extracts were used (lanes 3 and 4), indicating that the radioactive signal observed in Supplementary Fig. 1a, right, corresponds to radiolabeled protein, as previously described for the HITS-CLIP of the SR protein, SRSF11 (**c**) Results for the first CLIP replica for endogenous and overexpressed DGCR8 in HEK293T cells. In this experiment 13 million reads were obtained and 47% of those were mapped to the genome. After removing read duplicates, uniquely mapped reads from each of the datasets were clustered according to their genomic positions and the reproducibility of all DGCR8 binding sites, when comparing endogenous and overexpressed DGCR8 CLIP experiments is high (Pearson correlation co-efficient R =0.9). The axis shows the amount of reads in each of the multisample clusters in log10 scale (**d**) Distribution of reproducible DGCR8 significant clusters from first CLIP experiment (FDR<0.01) at the genomic level. More than 9,000 significant clusters at the genomic level were found and distributed as follows: 62% mapped to intergenic regions, 32.5% to protein coding genes and 5% to long non-coding RNAs (lncRNAs). When analyzing protein coding genes, the majority of the clusters were located in introns (29%), followed by coding sequences (2%) and 3' and 5'UTRs (1 and 0,5%, respectively) (**e**) Location of significant clusters from first CLIP experiment (FDR<0.01) in non-coding RNAs is distributed as follows: 64% to rRNA, 17% to miRNAs, and tRNAs, snRNAs ans snoRNAs were also found (6%, 5% and 3%, respectively).

Supplementary Figure 2



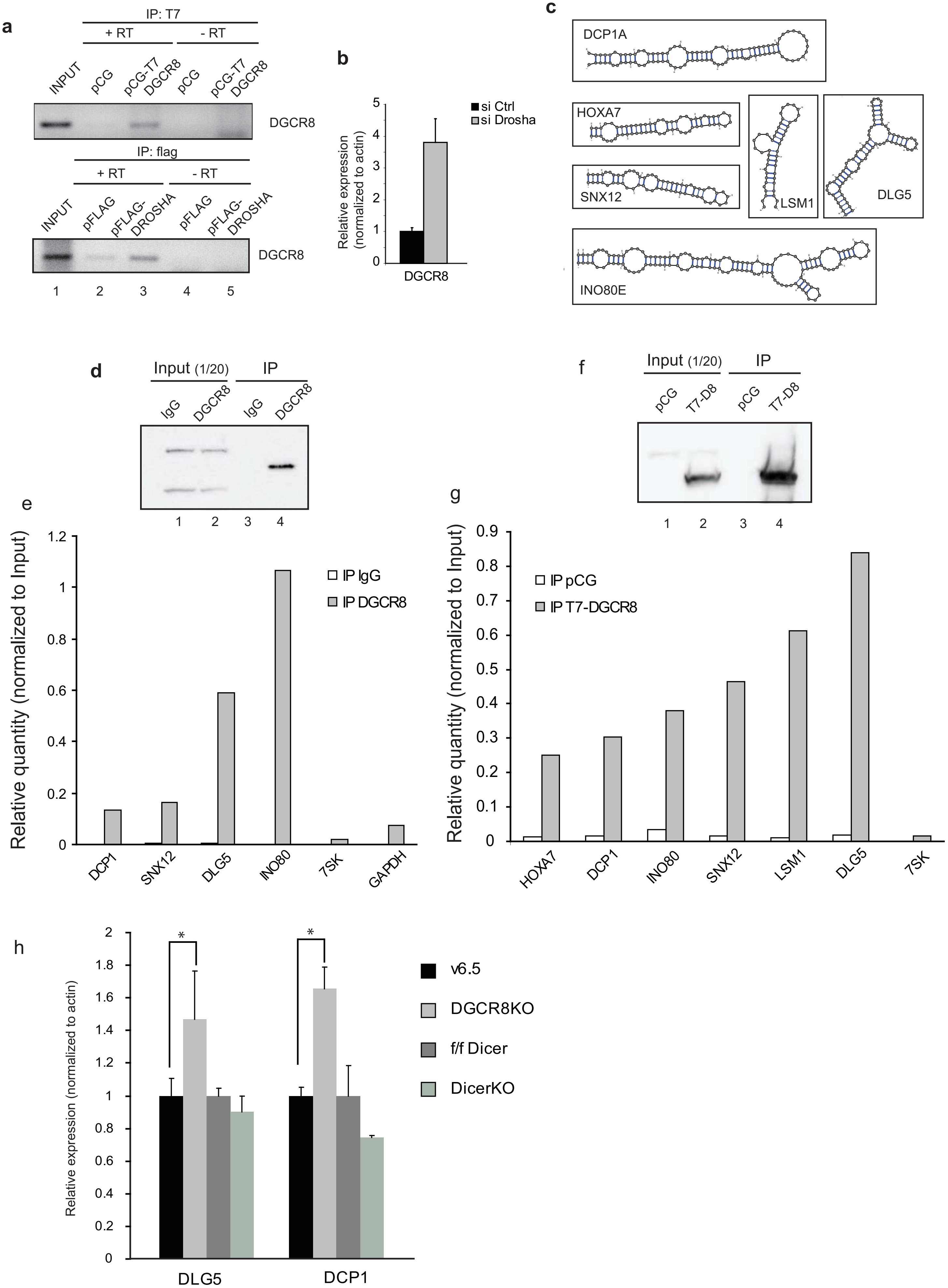
Supplementary Figure 2 Overexpression of a dominant negative form of DGCR8 significantly affects mature snoRNA levels (**a**) Cartoon depicting Dominant negative versions of DGCR8 and Drosha (**b**) Western blot analysis of transiently expressed dominant negative forms of DGCR8 (mDRBD1&2 DGCR8) and Drosha (ΔC114 Drosha) in HEK293T cells lysates, as revealed with DGCR8 and Drosha antibodies, respectively (**c**) Overexpression of dominant negative DGCR8 leads to the accumulation of snoRNAs (U16 and U92) and unprocessed pri-miR-24-1 (24-1) as shown by qRT-PCR (**d**) Overexpression of dominant negative Drosha does not significantly alter mature snoRNA levels, but leads to the accumulation of unprocessed pri-miR-24-1 **(e)** Overexpression of a dominant negative form of DGCR8 leads to the reduction of a 70 nt long fragment (marked by asterisk) detected with a probe mapping to the 5'end of U16 (as depicted on the right and marked by a red box) (**f**) A similar behavior was observed with U92 (also marked by asterisk)

Supplementary Figure 3



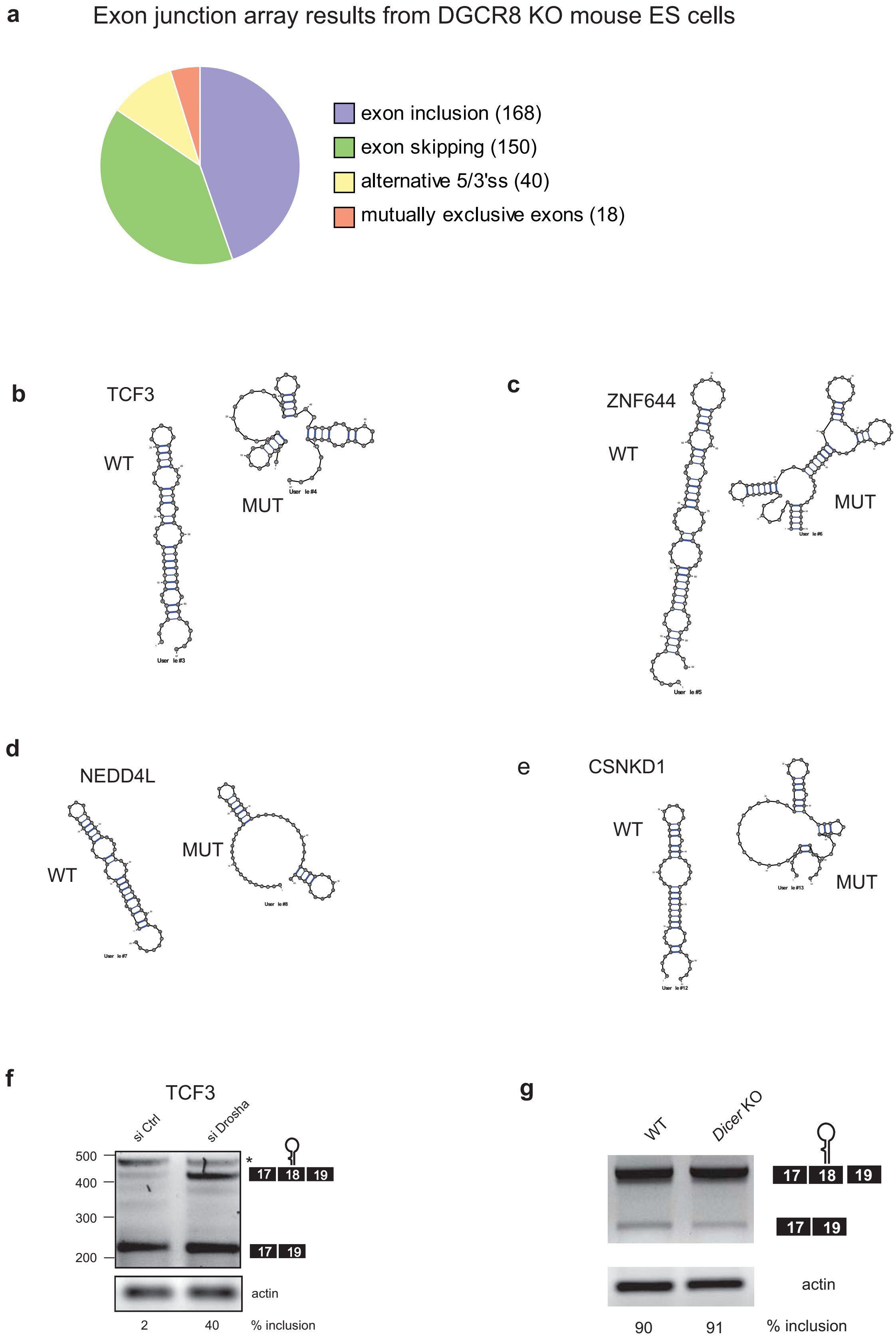
Supplementary Figure 3 DGCR8 directs cleavage of C/D and H/ACA box snoRNAs to generate small RNAs (**a**) Mature U16, U92 snoRNAs and U1 snRNA (control) were 5'end labeled. In vitro processing reactions with immunopurified DGCR8 (lanes 3, 8 and 13) and control immunoprecipitation (lanes 2, 7 and 12) were run in a polyacrylamide-urea gel. Position of the in vitro cleavage sites in snoRNAs was obtained by comparison of the migration of the cleaved products with markers generated by formamide treatment (F, single nucleotide) and RNase T1 (T1, G-specific cleavage) of the U16, U92 and U1 probes labeled at their 5'end. (**b**) In vitro cleavage analyses revealed position G77 in U16 and G80 in U92 to be cleaved by T7-DGCR8 immunoprecipitation, as depicted by asterisks. (**c**) Global analysis of the effect of DGCR8 depletion on snoRNA-derived small RNAs. Small RNA libraries from WT and Dgcr8 KO cells were mapped to snoRNA sequences, which were divided in 20 bins. The amount of reads mapped to each bin was represented as an average for C/D box snoRNAs (left) and H/ACA box (right). The distribution shows that for both snoRNAs, C/D box (left) and H/ACA box, there are less small RNAs originating from both ends when DGCR8 is absent, indicating a major stability of the mature form. In addition the central region of the H/ACA box snoRNAs (corresponding to the H box) generates more small RNAs when DGCR8 is absent (**d**) Small RNAs distribution for U16 snoRNA (SNORD16) (left). In the absence of DGCR8, less small RNAs originate from the 3'end, coinciding with the cleavage site observed in vitro (**Fig. 3e** and **Supplementary Fig. 3a**). The same analysis was performed with U92 (SCARNA8) but not enough small RNA reads were obtained, probably because the RNA fragments generated (around 40 nt) are bigger than the maximum size of the small RNA library (18-32nt) (right)

Supplementary Figure 4



Supplementary Figure 4 The Microprocessor regulates endogenous mRNA levels (**a**) IP-RT-PCR experiments showed that both overexpressed DGCR8 and Drosha bind to the DGCR8 mRNA in HEK293T cells (**b**) Depletion of Drosha from HeLa cells results in an up-regulation of DGCR8 mRNA (**c**) Predicted RNA secondary structures of mRNAs that were stabilized upon Drosha and DGCR8 knock-down (as shown on Figs. 4 d, e) (**d**) Endogenous DGCR8 was specifically immunoprecipitated from HEK293T cells, as detected by Western blot (lane 4) (**e**) RNAs associated to endogenous DGCR8 protein were analyzed by qRT-PCR with specific primers, 7SK and GAPDH were used as negative controls. The graphs show the relative amount of immunoprecipitated RNA relative to the amount present in the Input fraction (**f**) Overexpressed T7-DGCR8 was specifically immunoprecipitated with anti-T7 antibody as detected by Western blot with antibodies against DGCR8 (lane 4) (**g**) RNAs associated to overexpressed T7-DGCR8 protein were analyzed by qRT-PCR with specific primers, 7SK was used as negative control. The graphs show the relative amount of immunoprecipitated RNA relative to the amount present in the Input fraction (**h**) mRNAs shown to be up-regulated in Drosha-depleted human cells, were also up-regulated in DGCR8KO mES cells (compared to the parental cell line, v6.5), but not in Dicer KO (compared to the parental strain, termed f/f Dicer). This strongly suggest a direct effect of the Microprocessor in the stability of these mRNAs All the values shown are averages of at least three independent experiments, and $P<0.05$ (t test) were deemed significant and marked with asterisks. Errors bars represent standard deviation.

Supplementary Figure 5



Supplementary Figure 5 The Microprocessor regulates inclusion of DGCR8-bound cassette exons through recognition of pri-miRNA-like RNA secondary structures (**a**) A mouse exon junction array containing probes to detect 6,465 cassette exons, 4,204 3'/5'alternative splice sites and 366 mutually exclusive exons was hybridized RNA from wild-type and DGCR8KO cells. The analysis revealed 168 changes in exon inclusion events, 150 skipping events, 40 changes in alternative 5'/3' splice site (ss) usage and 18 changes in mutually exclusive exons abundance in the absence of DGCR8 protein (**b**) Predicted RNA secondary structure overlapping the DGCR8 binding site on TCF3 cassette exon 18 (depicted as WT) and disruption of the secondary structure by introduced mutations (MUT) (**c**) The same for exon 3 in ZNF644 (**d**) Exon 18 in NEDD4L and (**e**) Exon 9 in CSNK1D (**f**) The TCF3 isoform including exon 18 is upregulated in Drosha-depleted HeLa cells, as it was observed in DGCR8KO cells (**Fig. 5b**) (the asterisk denotes a non-specific band) (**g**) Dicer KO cells do not display any change in the levels of inclusion of mouse TCF3 exon 18.

Supplementary Table 1 Summary of HITS-CLIP mapping statistics

	D8.2	T7.2	D8.1	T7.1
Raw read	37,065,975	36,158,041	7,546,307	5,755,741
Filtered reads	37,065,534	36,157,679	7,474,974	5,706,006
Mapped reads	34,637,699	35,480,053	3,902,514	2,273,603
Uniquely mapped length> 20 nt (genome)	15,438,321	10,789,845	2,540,736	1,611,114
Uniquely mapped length> 20 nt (genome) no duplicates	1,525,963	954,160	1,629,713	925,009
Clusters (groups of 1 or more overlapping reads)	436,359	423,026	776,418	391,412
Significant Clusters (FDR < 0.01)	126,682	39,429	59,149	33,620
Significant Clusters overlapping a HPPR	124,363	38,584	58,711	33,484
Significant Clusters overlapping optimal secondary structure	76,550	24,544	35,577	20,301
Overlap significant clusters and small RNA reads data (Karginov et al, Mol Cell 2010)	10.55%	20.81%	1.98%	2.24%

FDR: false discovery rate

HPPR: high pair probability rate

Supplementary Table 2. List of oligonucleotides used in this study

CATGCCCGAACCTACACTG	FORWARD in exon 29 RNASEN, to chek mRNA levels. HUMAN
GGTCCTTTCCCACAGCCTAT	REVERSE in exon 29 RNASEN, to chek mRNA levels. HUMAN.
GCCATCCCATGCTAGAACCT	FORWARD in exon 29 RNASEN, to chek mRNA levels. MOUSE
GAAGGAGCGCTAACGATTTG	FORWARD human MALAT1 for IP-RTPCR experiments
TCTCCAGGACTTGGCAGTCT	REVERSE human MALAT1 for IP-RTPCR experiments
GACGGAGGTTGAGATGAAGC	FORWARD human qRTPCR oligo for MALAT1 RNA
ATTCGGGGCTCTGTAGTCCT	REVERSE human qRTPCR oligo for MALAT1 RNA
CGTTTGAAGGCATGAGTTGG	FORWARD mouse qRTPCR oligo for MALAT1 RNA
TGCCTCCCAAGTGCTAGGAT	REVERSE mouse qRTPCR oligo for MALAT1 RNA
GCCTCCTACGACCAAAACAT	FORWARD HOXA7, to check mRNA abundance. HUMAN
AGGTAGCGGTTGAAGTGGAA	REVERSE HOXA7, to check mRNA abundance. HUMAN
ACCTACACGCGCTACCAGAC	FORWARD HOXA7, to check mRNA abundance. MOUSE
GTCAGCAGCTGTGGAAACG	REVERSE HOXA7, to check mRNA abundance. MOUSE
TGATTCCAGCTTCCTCAGTACA	FORWARD in terminal exon DCP1A, to check mRNA levels. HUMAN
TTCTTGGTCAGAACCTGCAA	REVERSE in terminal exon DCP1A, to check mRNA levels. HUMAN
CAGCTTCCTCAGTACGCTTCA	FORWARD in terminal exon DCP1A, to check mRNA levels. MOUSE
GGCGGTCGGTCGGTGAGGCTTTC	FORWARD qPCR for DGCR8 mRNA
GGGGCTCTCATCTGTCTCCAT	REVERSE qPCR for DGCR8 mRNA
GGCCAAGAAATCCTGTGATG	FORWARD exon DLG5 HUMAN
GAGTAAGGTGCCCACTCCTG	REVERSE exon DLG5 HUMAN
CTCCCAGAAAGTCGATGAGC	FORWARD exon DLG5 MOUSE
TGGGGTAGAGTAGGGGGTTG	REVERSE exon DLG5 MOUSE
GAATGAACGCTGCCTACACA	FORWARD in SNX12 to check levels of the terminal exon HUMAN
TTCCTGTCAATTGCCTCCTC	REVERSE in SNX12 to check levels of the terminal exon HUMAN.
GATGCCACTGCATCATCAGA	FORWARD in INO80E HUMAN
CTTGGGCCTCAGGGGACT	REVERSE in INO80E HUMAN
CAAGTGGTTCTGTGTTTTTATTG	FORWARD amplifying approx -100 from start of pre-miR-30c-1
GTACTTAGCCACAGAAGCGCA	REVERSE amplifying approx +100 from end of pre-miR-30c-1
taatacgactcactatagggCCTAGAATCAATCCCTCCTTCTC	FORWARD human HOXA7 containg T7 promoter for in vitro transcription
GCATCTCCACAGTCCTGCTAAGC	REVERSE human HOXA7
taatacgactcactatagggATTCGGTGAGCCTGGCCTATCAG	FORWARD human DLG5 containg T7 promoter for in vitro transcription
CAAGGGGACATCTGCAGAACTT	REVERSE human DLG5
taatacgactcactatagggAATTGCTGGGCACCCACTGGCTC	FORWARD human SNX12 containg T7 promoter for in vitro transcription
GGGCTCCTACTGGCGCACCTTCC	REVERSE human SNX12
CCCGGGAGGTCACTCTCCCCGGGCTCTGTCCcctgtctc	5'end of U17a snoRNA fused to T7p for miRvana northern (mouse)
CTCAGCGACAGTTGCCTGCTGTCAGcctgtctc	U16 stem-loop sequence fused to T7 promoter for MiRvana northern
TTGGGCTGAAATACTGCTCTACTTGcctgtctc	U92 5' end stem-loop sequence fused to T7 promoter for MiRvana northern
TGGGGTTGCGCTACTGTCCAcctgtctc	FORWARD 5'arm of mouse RNU4ATAC for miRvana northern
taatacgactcactatagggGGGTTTTCCGACCGAAGTC	FORWARD antisense probe for mouse U7 snRNA (loading control northerns)
AAGTGTTACAGCTCTTTTAGAATTTGTC	REVERSE antisense probe for mouse U7 snRNA (loading control northerns)

tggctcgaattccaagagtt	FORWARD to check pre-mRNA levels of the host gene of MOUSE U16
cagttggtcagttgccaaga	REVERSE to check pre-mRNA levels of the host gene of MOUSE U16
gagatgtttggctgggaact	FORWARD MOUSE region upstream of U16, in the intron (pre-snoRNA)
TTCGTCAACCTTCTGAACCA	REVERSE at the end of mouse U16 sequence (pre and m snoRNA)
CTCTGTTCACAGCGACAGTTG	FORWARD to check mature levels of mouse U16
ccaagtgctgggattaaagg	FORWARD to check pre-mRNA levels of the host gene of MOUSE U92
tgtcctcagcaccctaacaa	REVERSE to check pre-mRNA levels of the host gene of MOUSE U92
tcttcgggagagtgataCGC	FORWARD MOUSE region upstream of U92, in the intron (pre-snoRNA)
AATTGTCTGCCCCGTATCTG	REVERSE at the end of mouse U92 sequence (pre and m snoRNA)
CACTGGACCTCCCCAGAGTA	FORWARD to check mature levels of mouse U92 SM269
TGCCTGCTGTCAGTAAGCTG	FORWARD to check mature levels of human SNORD16 (U16)
TGCTCAGTAAGAATTTTCGTCAA	REVERSE to check mature levels of human SNORD16 (U16)
GTCACCATGCCTCCCTAGAA	FORWARD to check mature levels of SCARNA8 (U92)
ATCTGTCTGCCCCGTATCTG	REVERSE to check mature levels of SCARNA8 (U92)
AGCTGAGGCGCTGCTTCT	FORWARD amplifying the beginning the stem of pre-miR-24-1
CCTCGGGCACTTACAGACA	REVERSE amplifying the end the stem of pre-miR-24-1
taatacgactcactatagggTTGCAATGATGTCGTAATTTG	FORWARD U16 Cdbox adding a T7 promoter
TTGCTCAGTAAGAATTTTCGTC	REVERSE U16 Cdbox (at the end of the mature transcript)
taatacgactcactatagggTGGGAGGCTGATACACAAATTG	FORWARD U92 scaRNA adding a T7 promoter
ATCTGTCTGCCCCGTATCTG	REVERSE U92 scaRNA mRNA
taatacgactcactatagggATACTTACCTGGCAGGGGAG	FORWARD T7 promoter fused to the beginning of U1 snRNA
CAGGGGAAAGCGCGAACGCAG	REVERSE at the 3'end of U1 snRNA
TCACCTCCTCTGGCCTTG	FORWARD amplifying -100 from pre-miR-24-2, cloning pGEMt
CATCTCTGCTCCAAGCATCA	REVERSE amplifying +100 from pre-miR-24-2, cloning pGEMt
CGGAGGAGGAGAAGAAGGAG	FORWARD in exon 18 of TCF3 HUMAN
CGGAGGCATACCTTTCACAT	REVERSE in exon 20 of TCF3 HUMAN
AGATCAAGCGGGAGGAGAAA	FORWARD in exon 18 of TCF3 MOUSE
ACCACGCCAGACACCTTCT	REVERSE in exon 20 of TCF3 MOUSE
CTTGCAGCAAGATGTTAATAAGAC	FORWARD in exon 2 of ZNF644 HUMAN
GCCTCTAACATGATTTGATAATCCA	REVERSE in exon 4 of ZNF644 HUMAN
CAGTGTTACCGACACGGTTG	FORWARD exon 12 mouse NEDD4L for splicing validation
TCCAAGTTGTGGTTCGATTG	REVERSE exon 14 mouse NEDD4L for splicing validation
GGAACGAGAACGGAAAGTGA	FORWARD exon 8 mouse CSNKD1 for splicing validation
GGGGGCGTGTCACTAGTAAAG	REVERSE exon 10 mouse CSNKD1 for splicing validation
CGTCACAACCTTTCTCTCCA	FORWARD exon 5 mouse FOXM1 for splicing validation
CCAGTGGGATTTCAGTTTTGA	REVERSE exon 7 mouse FOXM1 for splicing validation
CCCACTGAAGCCATGTTTCT	FORWARD exon 3 mouse SCAMP5 for splicing validation
ACAGCCTGGATGATGCTGAT	REVERSE exon 5 mouse SCAMP5 for splicing validation
tttaagcttcttcatggagtaaaaaatg	FORWARD human MALAT1 adding Hind III to clone in pGL3 basic
tttaagctttaccttctaacttctg	REVERSE human MALAT1 adding Hind III to clone in pGL3 basic
GACATCTGTCACCCCATTGA	FORWARD human 7SK RNA IP RTPCR
GCCTCATTTGGATGTGTCTG	REVERSE human 7SK RNA IP RTPCR

CATCCCCGATAGAGGAGGAC	FORWARD human 7SK RNA for q RTPCR
GCGCAGCTACTCGTATACCC	REVERSE human 7SK RNA for qRTPCR
gcttcgaattctgcATGGAGACATATGAGAGTCCCTC	FORWARD mouse DGCR8 to clone in IRESRED containing EcoRI site
gcagtcgacggtTCATACATCGACTGTGCACAAGGGC	REVERSE mouse DGCR8 to clone in IRESRED containing Sal I site
GCTGCAGGAGTAAGGACAGG	FORWARD mouse DGCR8 for qRTPCR
TCGAGCACTGCATACTCCAC	REVERSE mouse DGCR8 for qRTPCR
taatacgactcactatagggCTGGAGACTAAGAAAATAGAG	FORWARD T7 promoter fused to the beginning of ACA45
TGCTGTTGGTAGATAAGTAGG	REVERSE at the 3'end of ACA45
CAACGGATTTGGTCGTATTG	FORWARD human GAPDH
GGAAGATGGTGATGGGATTT	REVERSE human GAPDH
CTGTCTGCCTGCCATCCT	FORWARD amplifying the beginning the stem of pre-miR-24-2
CTCTGCTCCAAGCATCAGC	REVERSE amplifying the end the stem of pre-miR-24-2
TGGCTCAGTTCAGCAGGAACAGcctgtctc	miR-24 target sequence fused to T7 promoter for MiRvana northern
CAGACCAAGCTGCTCATCCTGCAGCAGGCC	
GTGCAGGTCATCCTGGGGCTGGAGCAGCAGGTGCGAG	FORWARD exon 18 TCF3 human WT for annealing
CTCGCACCTGCTGCTCCAGCCCCAGGATGAC	
CTGCACGGCCTGCTGCAGGATGAGCAGCTTGGTCTG	REVERSE exon 18 TCF3 human WT for annealing
CAGTTTAATAAAAACATAATAAATTAGGCCGTGCAGG	
TCATCCTGGGGCTGGAGCAGCAGGTGCGAG	FORWARD exon 18 TCF3 human MUT for annealing
CTCGCACCTGCTGCTCCAGCCCCAGGATGACC	
TGCACGGCCTAATTTATTATGTTTTTATTAAACTG	REVERSE exon 18 TCF3 human MUT for annealing
TATTCCCTTTATTTAAGAGCAAAG	FORWARD exon 3 ZNF644 human WT to PCR
TATTCAAAATAAAAGCCATCGTT	REVERSE exon 3 ZNF644 human WT to PCR
TATTCCCTTTTTTCCTTTCTTTGTGGAAGGTCAGGAG	FORWARD exon 3 ZNF644 human MUT
CCCAGCACCCCAACTAGGCGAGCCCGTT	
CGTCAACTGTCACGGGTGGTGAGGAATCCACG	FORWARD exon 18 NEDD4L mouse WT
CGTGGATTCCTCACCACCCGTGACAGTTGACGA	
ACGGGCTCGCCTAGTTGGGGTGCTGGG	REVERSE exon 18 NEDD4L mouse WT
CCTTTTTTTTCAACTAGGCGAGCCCGTTCG	
TCAACTGTCACGGGTGGTGAGGAATCCACG	FORWARD exon 18 NEDD4L mouse MUT
CGTGGATTCCTCACCACCCGTGACAGTTGACGA	
ACGGGCTCGCCTAGTTGAAAAAAAAGG	REVERSE exon 18 NEDD4L mouse MUT
AATAGCATTCCTTTCGAACACCACGGCAAGTA	
GCTGCTCGTCTCCATCGGAAGGCAGCACTGG	FORWARD exon 9 CSNKD1 mouse WT
CCAGTGCTGCCTTCCGATGGAGACGAGCAGCTA	
CTTGCCGTGGTGTTCGAAAGGAATGCTATT	REVERSE exon 9 CSNKD1 mouse WT
AATAGCATTTTTTTTTTACACCACGGCAAGT	
AGCTGCTCGTCTCCATCGGAATTTAGCACTGG	FORWARD exon 9 CSNKD1 mouse MUT
CCAGTGCTAAATTCCGATGGAGACGAGCAGCT	
ACTTGCCGTGGTGTAAAAAAAAAATGCTATT	REVERSE exon 9 CSNKD1 mouse MUT

Supplementary Note for

DGCR8 HITS-CLIP reveals novel functions for the Microprocessor

Sara Macias, Mireya Plass, Agata Stajuda, Gracjan Michlewski, Eduardo Eyras and Javier F. Cáceres

HITS-CLIP protocol

HITS-CLIP for DGCR8 was based on a described protocol with some modifications[1]. In summary, HEK293T cells, either non-transfected, mock-transfected (pCG) or overexpressing T7-DGCR8 were grown in 15 cm plates. Cells were washed with PBS and 6 ml of ice-cold PBS was added before UV irradiation, which was performed in a Stratalinker 1800 at 4,000µJ/cm2. After irradiation, cells were scrapped and suspensions were centrifuged at 2,500 rpm for 3 min, washed again with fresh PBS and after spinning were stored at -80°C. For each immunoprecipitation 50µl of protein A beads (GE Healthcare) were used. Beads were first washed 3 times with lysis buffer (50mM Tris-HCl pH 7.4, 100mM NaCl, 1mM $MgCl_2$, 0.1mM $CaCl_2$, 1% NP40, 0.5% sodium deoxycholate, 0.1% SDS). Beads were resuspended in 200µl of lysis buffer and antibodies were added for binding to beads at 4°C for at least one hour. For immunoprecipitation of endogenous DGCR8 protein, anti-DGCR8 antibodies ab36865 and ab90579 (Abcam) were used (1µg) (first and second CLIP experiment, respectively), whereas an anti-rabbit IgG was used, as a negative control. For overexpressed T7-DGCR8 protein, anti-T7 antibody 69522-3 (Novagen) (1µg) was used. As a negative control, the same antibody was used in mock-transfected extracts. Extracts from UV-irradiated cells were prepared by resuspending first the pellets in 50µl of cold PBS and after getting a cell suspension, 1 ml of lysis buffer supplemented with EDTA-free cocktail protease inhibitors and RNAse inhibitor (Invitrogen, 10µl) was added. Cells lysates were sonicated on ice using Bioruptor (3 times 30 sec on and 30 sec off). After sonication 5µl of Turbo DNAse and 10µl of RNAse A/T1 dilution were added and incubated at 37°C for 10 min with shaking intervals. Two RNAse A/T1 dilutions were used, a high dilution that corresponds to 1µl of the enzyme mix in 50µl of lysis buffer, and a low dilution that corresponds to 1µl in 500µl of lysis buffer. After incubation of the lysates with RNAse and DNAse, extracts were centrifuged for 3 min at maximum speed. Before adding the extracts to the antibody-conjugated beads, a pre-clear step was performed. Extracts were incubated with ready washed protein A beads for 1hour at 4°C. Before adding antibody-conjugated beads to extracts, beads were washed 3 times with lysis buffer to remove any traces of unbound antibody. After pre-clearing, extracts were added to antibody-conjugated beads and rotated overnight at 4°C. Beads were washed 4 times with high-salt wash buffer (50mM Tris-HCl pH7.4, 1M NaCl, 1mM EDTA, 1% NP-40, 0.5% sodium deoxycholate and 0.1% SDS), the microtube was changed in every wash to reduce the background. This was followed by two washes with 1x PNK buffer (20mM Tris-HCl pH 7.4, 10mM $MgCl_2$ and 0.2% Tween-20). Beads where then resuspended in 50µl of PNK buffer and 1µl of T4 PNK enzyme (M0201, NEB) was diluted in 10µl of 1x PNK buffer. For each tube, a mix containing PNK buffer, 1µl of ^{32}P-γATP, and 1µl of diluted PNK enzyme was added to beads. This reaction was incubated for 3 min at 37°C, mixing at 1,000 rpm (Thermomixer), the reaction was stopped by adding a cold ATP mix (18µl PNK, 2µl 10mM ATP) and incubated for additional 3 min at 37°C. After labeling, the beads were washed 4-6 times with 1 ml of high-salt wash buffer and two final washes with PNK buffer. Beads were then resuspended in 20µl

NuPAGE loading buffer (Invitrogen) heated for 10 min at 70°C with shaking and loaded on a 4-20% Tris-Glycine gel. After electrophoresis, the gel was transferred to a nitrocellulose membrane, which was exposed to a film at -80°C for a few hours to overnight. The bands corresponding to DGCR8 molecular size were excised from the membrane and RNA was extracted from them. For RNA extraction, 200µl of Proteinase K solution (2mg/ml in PK buffer: 100mM Tris-HCl pH 7.4, 50mM NaCl and 10mM EDTA) was added to each tube and incubated at 37°C for 20 min, shaking in Thermomixer at 1,000 rpm. This solution was transferred to a new tube, and the membrane pieces were extensively washed again with PK buffer containing 7M urea. The two solutions were mixed and incubated for 20 min at 60°C. Reactions were cooled down to 37°C prior to addition of 500µl phenol/chloroform (Ambion, acidic pH) and incubated for 5 min at 37°C at 1,000 rpm shaking. Samples were centrifuged at maximum speed for 3 min, and the aqueous phase was recovered in a new tube where 1µl of glycoblue (Ambion), 50µl of sodium acetate 3M and 1 ml of ethanol: isopropanol (1:1) were added. RNA precipitation was performed overnight. Ligation of RNA linkers and RT-PCR amplification steps were preformed at the Ultrasequencing unit from CRG (Center for Genomic Regulation, Barcelona), using Illumina Small RNA Kit sequencing (v1.5) for the first CLIP experiment, and at the Beijing Genomics Institute for the second CLIP experiment. Briefly, first, samples were decapped with tobacco acid pirophosphatase (TAP) for 2 hr at 37°C. Next, RNA was dephosphorylated using CIP (calf intestinal phosphatase) for 30 min at 37°C. RNA was precipitated and purified to ligate first the 3'adapter for 6hr at 20°C. RNA was rephosphorylated and an additional gel was run to re-purify RNA and ligate the 5'adapter for 6hr at 20°C. The products were run in another gel and RNA was purified to perform the retrotranscriptase step and PCR. For size determination of the associated RNAs to DGCR8, precipitated RNA (after elution from nitrocellulose membrane) was separated in a 10% TBE- urea polyacrylamide gel. The RNA-protein complexes were visualized by autoradiography.

Read Processing and Mapping

Reads obtained from the sequencing protocol were processed with MIRO (Kofler, unpublished, 2009. Web: http://seq.crg.es/main/bin/view/Home/MiroPipeline) to remove all those reads that contained more than 1 N in their sequence and with a complexity below 0.5. The complexity c of a read is calculated as

$$c = 1 - f(A)^2 - f(T)^2 - f(G)^2 - f(C)^2,$$

where $f(A)$, $f(T)$, $f(G)$ and $f(C)$ are the frequencies of the nucleotides in the sequence.

To maximize the amount of mapped reads, a strategy was used that combines the mapping of the reads simultaneously to the genome and the transcriptome, and includes a sequential trimming of the 3' end of reads. A mapping index file was created using bowtie[2], containing the hg18 genome[3], all Ensembl 54 transcripts and the sequence of the lncRNAs from Ensembl 59[4], which were mapped back to hg18. The coordinates of lncRNAs from Ensembl 59 were converted to hg18 using the liftover tool from UCSC[3]. In this step, all those lncRNAs that could not be fully mapped back to hg18 were discarded. The sequence corresponding to the lncRNAs that were fully mapped in the hg18 genome was extracted and included in the index file. The reads were mapped using this index file with bowtie[2], allowing up to two mismatches in the entire read, and keeping all possible locations in the genome. The mapping protocol consists of several steps. Initially, all reads are mapped to the index file. After the mapping, all those reads that could not be mapped are trimmed 1 base at the 3' end of the read and mapped again, as described. This procedure is repeated until the reads have a minimum length of 21nt, which is the length at which the highest proportion of uniquely mapped reads in the genome is found. For further analyses, all those reads mapping in the genome at most 25 times were used. On the other hand, all the reads mapping in the transcriptome were used (**Supplementary Table 1**).

Dataset correlation

We analyzed the reproducibility of the HITS-CLIP experiment by measuring the correlation between biological replicate experiments done twice with two different antibodies (endogenous DGCR8 and pCG T7-DGCR8). First, for each pair of biological replicates multisample clusters containing at least one read from each of the samples were defined. This resulted in 86,538 overlapping clusters between the first pair of HITS-CLIP experiments and 84,199 overlapping clusters between the second pair of HITS-CLIP experiments. For each pair of biological replicates the correlation in the amount of reads belonging to each cluster was measured. Pearson correlation analysis shows a significant correlation for both pairs of biological replicates (Pearson correlation coefficient R = 0.90 and R = 0.82 respectively). The correlation between the second pair of biological replicates is shown in (**Fig.** 1a).

Identification of significant clusters

Reads mapped to the same strand were clustered according to their position in the reference sequence to generate clusters. To identify significant clusters we applied a modified false discovery rate (mFDR) similarly to what was been done previously[6] using Pyicos[5]. This modification of the FDR takes into account the amount of reads in a given region to calculate the probability of a cluster. These regions were defined according to the location of the cluster in the genome or the transcriptome. For clusters of reads mapped to the transcriptome, the whole transcript was considered as the region where to calculate significance. Clusters mapped to the genome were classified according to their location as genic, promoter associated or intergenic. Promoters were defined as 1,000 nt upstream of the annotated gene start. For those clusters inside genes or promoters, the whole gene or promoter was considered as the region to calculate significance. For clusters in intergenic regions, a region of 2,000 nt around the center of the cluster was considered. To avoid possible biases, we removed all regions overlapping centromeres, gaps and satellites. The positions of genomic and centromeric gaps were obtained from hg18 Gap Track from UCSC. Satellite regions were extracted from hg18 Repeat Masker Track from UCSC[6]. The overlap of satellites with clusters was calculated using fjoin[7]. In all cases, only those clusters with an mFDR p-value <0.01 were considered for further analyses. These clusters can be visualized at UCSC genome browser as custom tracks using the links provided in http://regulatorygenomics.upf.edu/Data/DGCR8/.

Overlap between small RNA reads and DGCR8 clusters

We downloaded the global 5' dependent RACE libraries from HEK293 cells from[8] from GEO database (accession number GSE21975). Reads were mapped to the hg18 human genome assembly[4] using bowtie[3], and kept all reads that were longer than 20nt and that matched at the most 500 times on the genome allowing up to two mutations in the entire read. The overlap between HITS-CLIP clusters and the small RNA reads was calculated using fjoin[8].

Secondary Structure Prediction

DGCR8 binds to structured regions similar to that of pri-miRNAs. Therefore, to identify a set of candidates in which to validate DGCR8 binding, a computational method was designed in order to identify regions with CLIP-Seq reads that could contain secondary structures. This method is based on the assumptions that (1) all the regions contain a secondary structure and (2) this structure can be identified in the sequence as regions with a high pair probability that can be separated from each other by regions with a low pair probability. The algorithm works as follows:

I. The sequence of the read cluster plus 100 nt upstream and downstream is retrieved. If there is not enough sequence (i.e. the cluster is at the end of a transcript), the available sequence is recovered.

II. For each of the nucleotides, the average pair probability (pp) is calculated using RNAfold[9].

III. High pair probability regions (HPRs) are defined as stretches of nucleotides with a mean pair probability higher that the mean pair probability of the total sequence plus the standard deviation. All those HPRs separated by less than 5nt are joined together.

IV. If there is more than one HPR, for each pair of HPRs, e.g. HPR_A and HPR_B, an interaction score (IS) is defined as the sum of the pair probabilities of the nucleotides in HPR_A that interact with the nucleotides in HPR_B.

V. If there is more than one HPR, the HPR with the highest area, defined as the sum of pp values of all its nucleotides, is joined with the HPR whose area difference is the lowest; hence, a new HPR, HPR_2, is defined.

VI. For each of the remaining HPRs, e.g. HPR_A, starting from the one downstream of HPR2, the HPR, e.g. HPR_B, with which HPR_A has the higher interaction score (IS) is searched for. If HPR_B is equal to HPR_2 or is located in the other side of HPR_2 (relative to HPR_A), the boundaries of HPR_2 are extended. The same operation is also applied for the HPRs upstream of HPR_2.

VII. Once an HPR is identified, all clusters that are not overlapping with their corresponding HPR are discarded.

VIII. An optimal secondary structure is predicted then in the maximum region defined by the overlap between the read cluster and the HPR using RNAfold [9].

IX. For this sequence, all stems are identified and ranked according to the amount of overlap with the read cluster in the region defined. As sometimes the prediction of an optimal secondary structure is dependent on the amount of sequence given, each of the stems is extended at most 30 nt on each side, and the one that overlaps the most with the read cluster is kept as final prediction.

To validate the method this algorithm was applied to the known miRNAs annotated in Ensembl 54 that overlap with clusters in our samples (a total of 261 different pri-miRNAs). For each of the clusters overlapping pri-miRNAs, the HPR calculated overlapped the position of the pri-miRNA, suggesting that the HPR calculation is a good way to define the region where to predict the secondary structure.

Mouse exon junction arrays

Labeled RNA from three biological replicas for each cell line (WT and *Dgcr8* KO) was hybridized to a non-commercial exon junction array from Affymetrix. This array contains probe sets for approximately 30,000 genes in the mouse genome, consisting of a total of 319,769 exon-probe sets and 237,871 junction-probe sets. In this way, gene expression and alternative splicing analyses can be done. This array determines the relative abundance of 6,465 cassette exons, 4,204 5'/3' alternative splice sites and 366 mutually exclusive exons. Alternative splicing events were identified using a splicing index approach (as described in *MADS+: discovery of differential splicing events from Affymetrix exon junction array data*)[10]. For cassette exons, only those exhibiting reciprocal behaviour for inclusion and skipping probes were selected. For detailed results of the exon junction array data please see **Supplementary Excel file**.

Identification of homologous human genes from mouse exon junction arrays

The annotation of cassette exons, alternative 5' and 3' exons and regulated genes identified with the mouse array were mapped from mm7 to mm9 using the liftover tool from UCSC[4]. For each of the genes, we checked the overlap with Ensembl genes annotated from Ensembl54[5] and kept all those genes from the microarray whose overlap with annotated genes was at least 85%. We used one-to-one homolog gene pairs from Ensembl to identify a total of 2211 human homologs. In the case of exons, we kept only those exons whose from the array whit a 100% overlap with the annotated Exons from Ensembl54. From these, we identified 106 regulated cassette exons conserved in human, 53 upregulated and 53 downregulated.

SnoRNA analysis

The human snoRNA annotation, including H/ACA, C/D box and small Cajal body RNAs (scaRNAs) was downloaded from UCSC[3]. For each of the snoRNAs (402), the genomic sequence was retrieved and aligned to the mouse genome using exonerate[11]. To select homologous mouse snoRNAs, the percentage identity and the coverage of the best mouse alignment was checked for each human snoRNA, and the first quartiles of the percentage identity (81.48) and coverage (98.40) were taken as thresholds. Using these parameters, 413 snoRNAs homologous to human snoRNAs were found. Small reads mapped to the genome from wt and *dgcr8* Δ/Δ datasets overlapping each snoRNA were identified using fjoin[7]. To check the effect of DGCR8 in snoRNA processing each snoRNA was divided into 20 equal-size bins and the amount of normalized reads mapped in each of them was calculated for wt and *dgcr8* Δ/Δ datasets. As it can be observed in **Supplementary Fig. 3c**, the amount of small RNA reads produced from different regions of the snoRNAs is in general lower in the absence of DGCR8.

Small RNA reads analyses for snoRNAs

The sequences of the small RNA reads for wt ES cells and *dgcr8*Δ/Δ, which correspond to the libraries GSM314552 and GSM314557[12], respectively, were retrieved from deepBase[13]. These reads were mapped to the mouse mm9 genome using bowtie[2] allowing 0 mismatches. Only those reads that mapped at most 500 times in the genome were further used. The mapped reads from *dgcr8*Δ/Δ were normalized by the fold change of the reads mapped to tRNAs, srpRNAs and snRNA, as described in the original paper[12].

References

1. Wang,Z., Tollervey,J., Briese,M., Turner,D., & Ule,J. CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. *Methods* **48**, 287-293 (2009).

2. 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).

3. Fujita,P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876-D882 (2011).

4. Flicek,P. *et al.* Ensembl's 10th year. *Nucleic Acids Res.* **38**, D557-D562 (2010).

5. Althammer,S., Gonzalez-Vallinas,J., Ballare,C., Beato,M., & Eyras,E. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics.* **27**, 3333-3340 (2011).

6. Karolchik,D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-D496 (2004).

7. Richardson,J.E. fjoin: simple and efficient computation of feature overlaps. *J. Comput. Biol.* **13**, 1457-1464 (2006).

8. Karginov,F.V. *et al.* Diverse endonucleolytic cleavage sites in the mammalian transcriptome depend upon microRNAs, Drosha, and additional nucleases. *Mol. Cell* **38**, 781-788 (2010).

9. Hofacker,I.L. RNA secondary structure analysis using the Vienna RNA package. *Curr. Protoc. Bioinformatics.* **Chapter 12**, Unit12 (2009).

10. Shen,S., Warzecha,C.C., Carstens,R.P., & Xing,Y. MADS+: discovery of differential splicing events from Affymetrix exon junction array data. *Bioinformatics.* **26**, 268-269 (2010).

11. Slater,G.S. & Birney,E. Automated generation of heuristics for biological sequence comparison. *BMC. Bioinformatics.* **6**, 31 (2005).

12. Babiarz,J.E., Ruby,J.G., Wang,Y., Bartel,D.P., & Blelloch,R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev.* **22**, 2773-2785 (2008).

13. Yang,J.H., Shao,P., Zhou,H., Chen,Y.Q., & Qu,L.H. deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res.* **38**, D123-D130 (2010).