Supplementary Materials

Andreas J. Gruber[1], Ralf Schmidt[1], Andreas R. Gruber[1], Georges Martin[1], Souvik Ghosh[1], Manuel Belmadani[1,2], Walter Keller[1] & Mihaela Zavolan[1*]
1 Biozentrum, University of Basel, Klingelberstrasse 50-70, CH-4056 Basel, Switzerland
2 Current address: University of British Columbia, 177 Michael Smith Laboratories, 2185 East Mall Vancouver BC V6T 1Z4
∗ To whom correspondence should be addressed (mihaela.zavolan@unibas.ch)

1 3' end sequencing protocols

1.1 2P-Seq

In the 2P-Seq protocol, reverse transcription is accomplished by an anchored oligo(dT) primer. The products of reverse transcription and PCR amplification are expected to have 20 As preceding the 3' adapter. Libraries are sequenced in anti-sense direction with a custom primer. Reads should be reverse complemented [1, 2].

1.2 3'-Seq

In the 3'-Seq protocol of Mayr and colleagues, reverse transcription is accomplished by an anchored oligo(dT) primer. The products of reverse transcription and PCR amplification are expected to have 17 As preceding the 3' adapter. Libraries are sequenced in sense direction requiring removal of the 3' adapter sequence and preceding As to pinpoint the 3' end [3].

1.3 3P-Seq

In the 3P-seq protocol, a biotinylated adapter is ligated to the end of the poly(A) tail via splint-ligation. After partial digestion, poly(A) regions are captured with streptavidin and reverse transcription is carried out only with dTTP. Most of the poly(A) tail is then removed through RNase H digestion. Adapter ligation, reverse transcription and PCR amplification follow before the library is sequenced in anti-sense direction. Consequently, pinpointing the 3' end requires the reads to be reverse complemented [4, 5].

1.4 3'READS

3' region extraction and deep sequencing (3'READS) is a protocol that utilizes a special primer (45 thymidines followed by 5 uridines) to capture poly(A) containing RNA fragments. RNase H digestion releases transcripts 3' ends from the most of the poly(A) tail. Subsequently, the fragments are subjected to adapter ligation, reverse transcription, and PCR amplification before they are sequenced in anti-sense direction. The cleavage site is inferred as the first non-A of the 3' end of the read's reverse complement [6, 7].

1.5 A-seq

In the A-seq protocol, reverse transcription is accomplished by an anchored oligo-dT primer. The products of reverse transcription and PCR amplification are expected to have six As preceding the 3' adapter. Libraries are sequenced in sense direction requiring removal of the 3' adapter sequence and preceding As to pinpoint the 3' end [8].

1.6 A-seq (version 2)

The second version of the A-seq protocol has the following changes: (1) The steps of the protocol are conducted such that the generation of adapter dimers is minimized. (2) Libraries are sequenced in anti-sense direction and the mRNA cleavage site is inferred as the first nucleotide after a stretch of 4 random nucleotides and 3 Ts [9].

1.7 DRS

In the direct RNA sequencing (DRS) protocol, 3' ends of transcripts are hybridized to poly(dT)-coated flow cell surfaces where antisense strand synthesis is initiated. This has the advantage that no prior reverse transcription or cDNA amplication is needed [10–13].

1.8 PAS-seq

In the PAS-Seq protocol, reverse transcription is accomplished with an anchored oligo-dT primer. The products of reverse transcription and PCR amplification are expected to have 20 As preceding the 3' adapter. Libraries are sequenced in anti-sense direction with a custom primer requiring the reverse complement of the reads to pinpoint the 3' end [14].

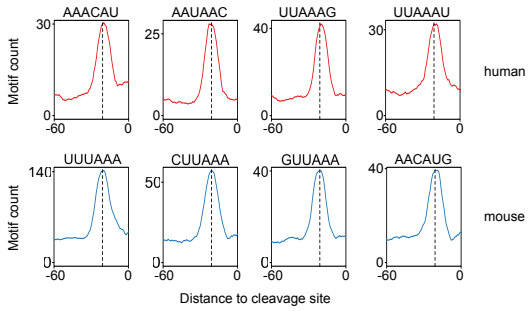
1.9 PolyA-seq

Library preparation for the PolyA-seq protocol includes the following steps: (1) Reverse transcription, primed with an oligo-dT sequence, (2) second strand synthesis with random hexamers linked to a second PCR primer, and (3) PCR amplification. Sequencing is accomplished in anti-sense orientation with a primer ending in 10 Ts and the resulting reads need to be reverse complemented to pinpoint the pre-mRNA cleavage site [15, 16].

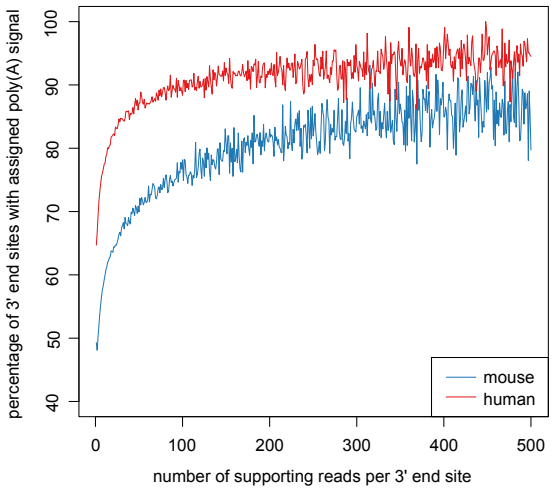
1.10 SAPAS

In the SAPAS protocol, reverse transcription is accomplished by an anchored oligo-dT primer. The products of reverse transcription and PCR amplification are expected to have the sequence AAAAAA-GAAAAAAGAAAAAA preceding the 3' adapter. Libraries are sequenced in anti-sense direction with a regular primer requiring to trim 20 nucleotides from the 5' end of reads and to reverse complement reads to pinpoint the 3' end [17, 18].

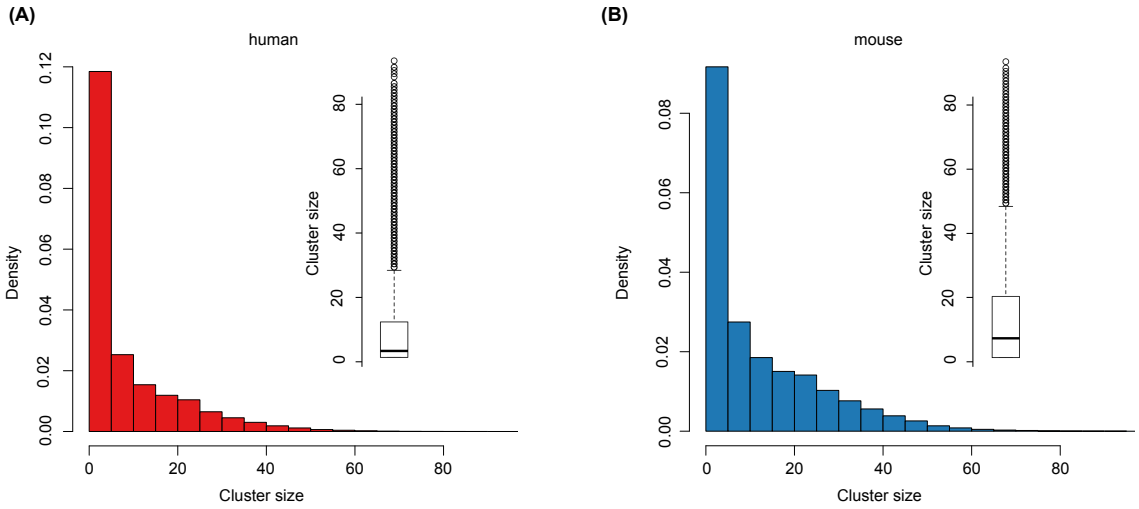
2 Supplementary Figures



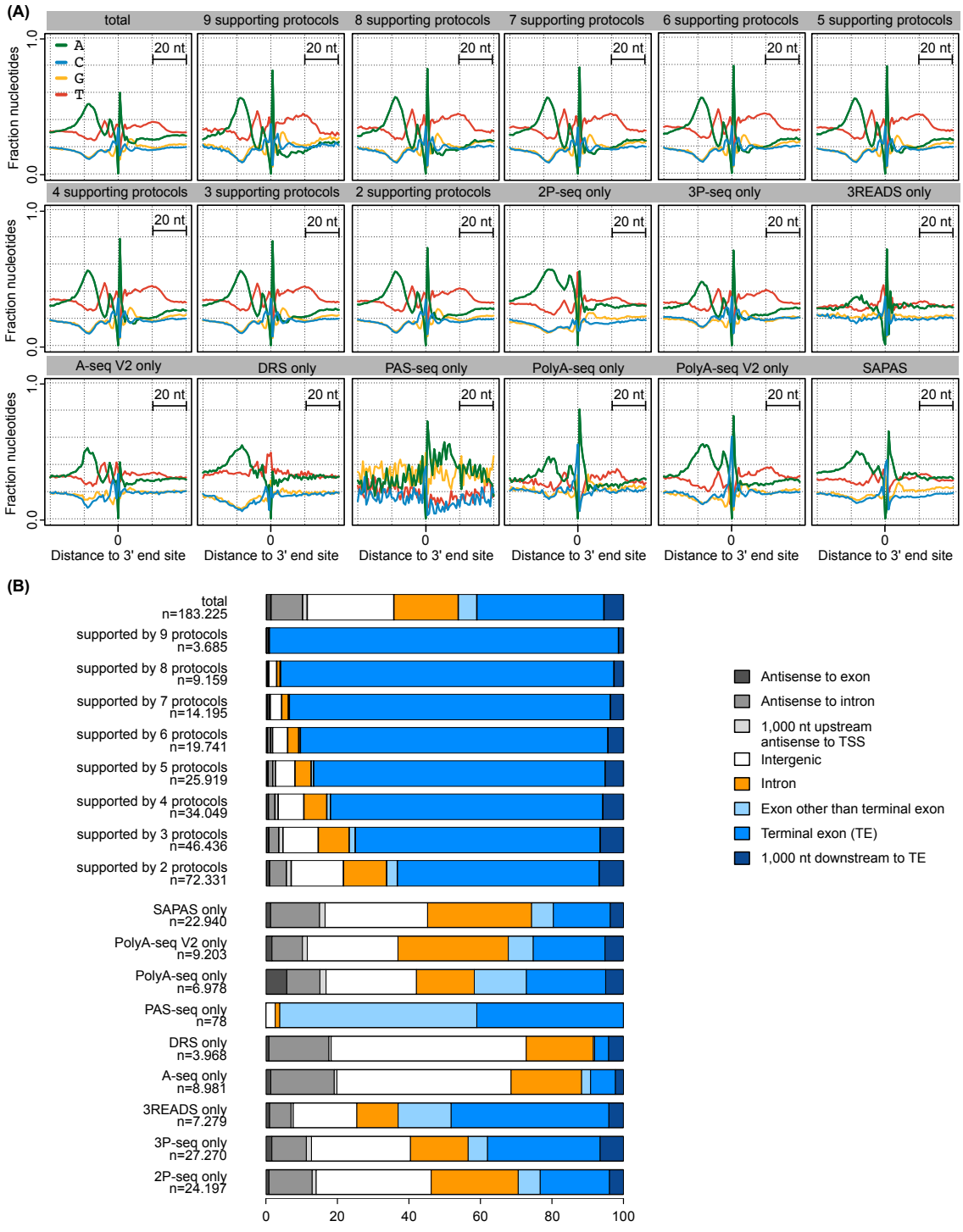
Supplementary Figure 1. Frequency profiles of the poly(A) signals that have been identified only in human (red) or mouse (blue).



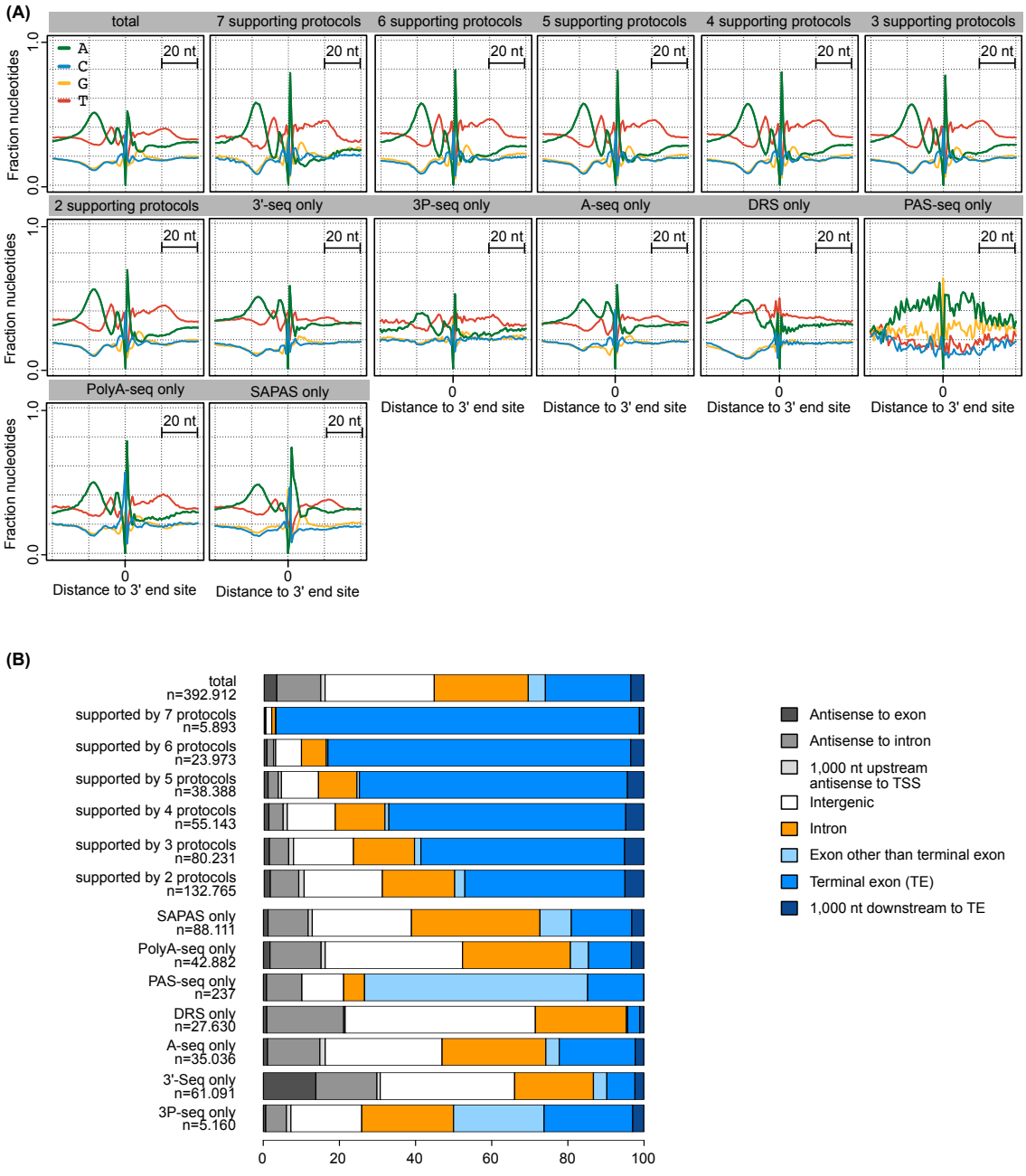
Supplementary Figure 2. Fraction of the putative 3' end sites with an assigned poly(A) signal in their upstream region (60 to 10 nucleotides upstream) as a function of the number of supporting reads per site (summed reads over all considered samples).



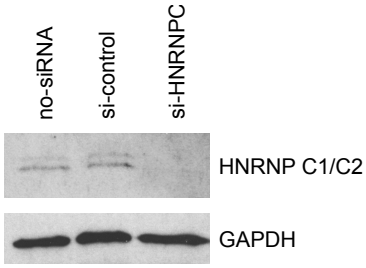
Supplementary Figure 3. Distribution of cluster sizes **(A)** human catalog **(B)** mouse catalog. The large majority of clusters has a short span (less than 20 nt) in both human and mouse.



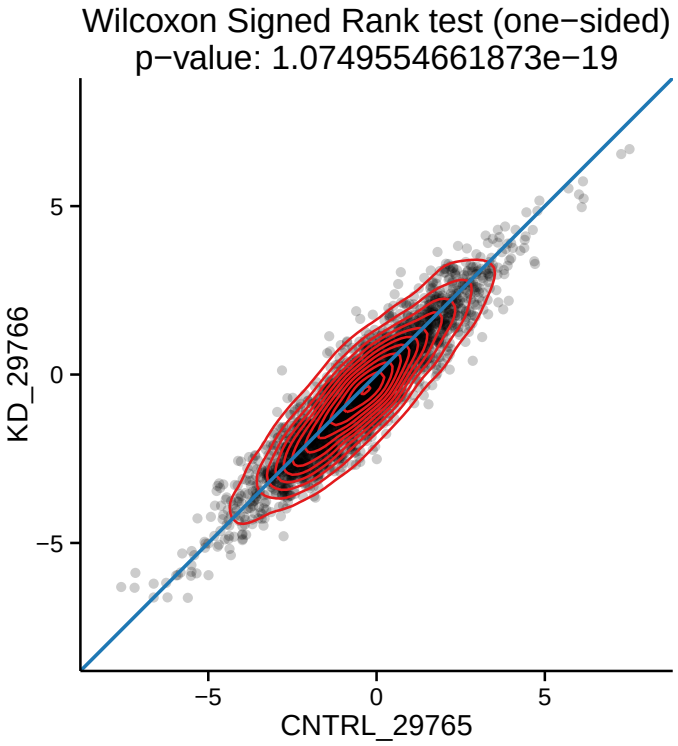
Supplementary Figure 4. Characteristics of mouse poly(A) clusters. **(A)** Nucleotide composition around cleavage sites supported by the indicated number of protocols or the name of the protocol for clusters that had a single protocol support. **(B)** Annotation of clusters supported by various types of protocols (n - number of poly(A) clusters in the indicated category).



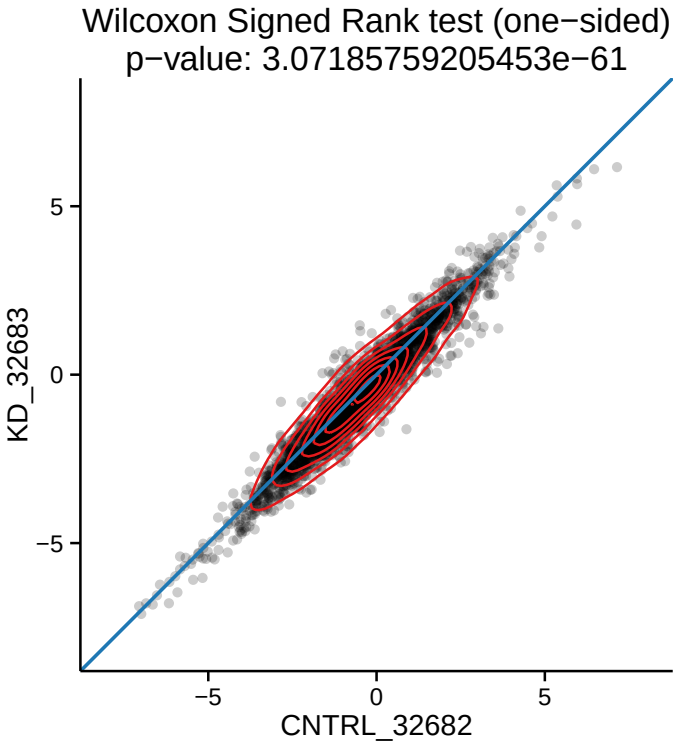
Supplementary Figure 5. Characteristics of human poly(A) clusters. **(A)** Nucleotide composition around cleavage sites supported by the indicated number of protocols or the name of the protocol for clusters that had a single protocol support. **(B)** Annotation of clusters supported by various types of protocols (n - number of poly(A) clusters in the indicated category).



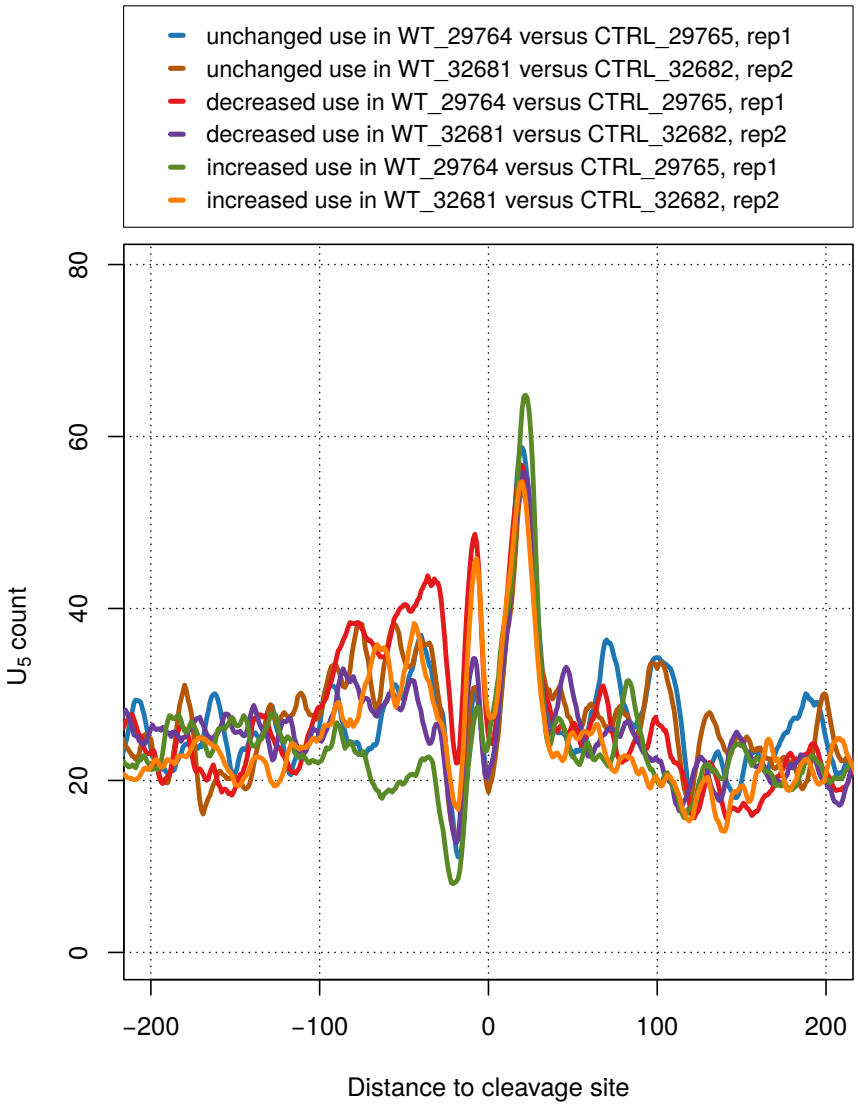
Supplementary Figure 6. Western blot showing the expression levels of HNRNP C1/C2 and GAPDH in cells that were either untreated, or treated with either a control siRNA or with si-HNRNPC (50 picomoles siRNA per well of a 6-well plate).



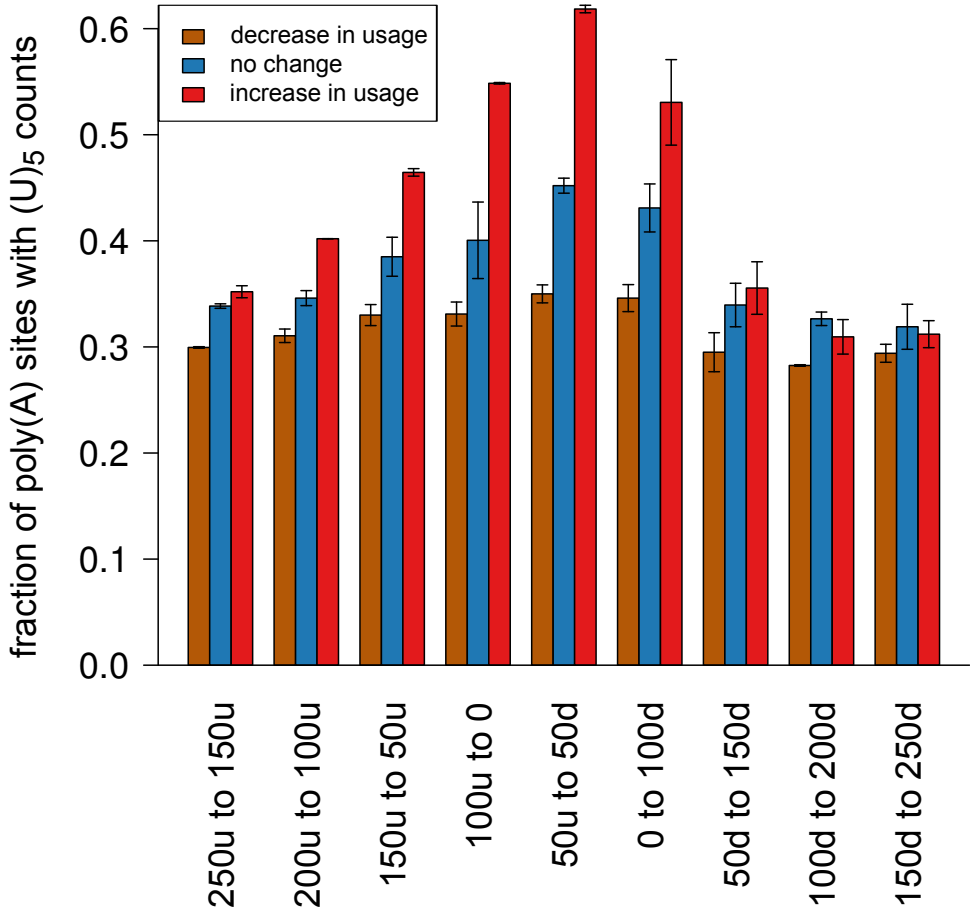
Supplementary Figure 7. Contour plot of the proximal-to-distal poly(A) site usage ratios in si-HNRNPC transfected versus si-Control transfected HEK 293 cells in replicate 1. For each plot only exons having exactly two expressed poly(A) sites were considered (2607 exons in total). The proximal-to-distal ratio is significantly higher in cells treated with the control siRNA indicating that on average 3'UTRs tend to be elongated, rather than shortened, upon knockdown of hnRNP C.



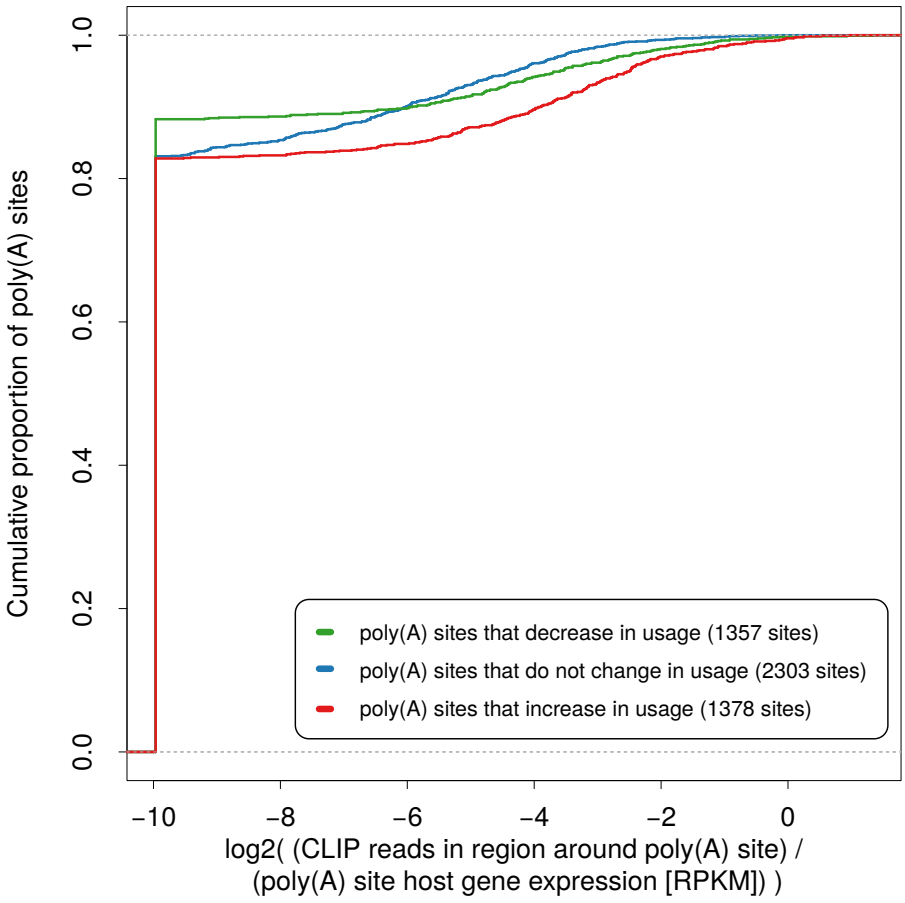
Supplementary Figure 8. Contour plot of the proximal-to-distal poly(A) site usage ratios in si-HNRNPC transfected versus si-Control transfected HEK 293 cells in replicate 2. For each plot only exons having exactly two expressed poly(A) sites were considered (2607 exons in total). The proximal-to-distal ratio is significantly higher in cells treated with the control siRNA indicating that on average 3'UTRs tend to be elongated, rather than shortened, upon knockdown of hnRNP C.



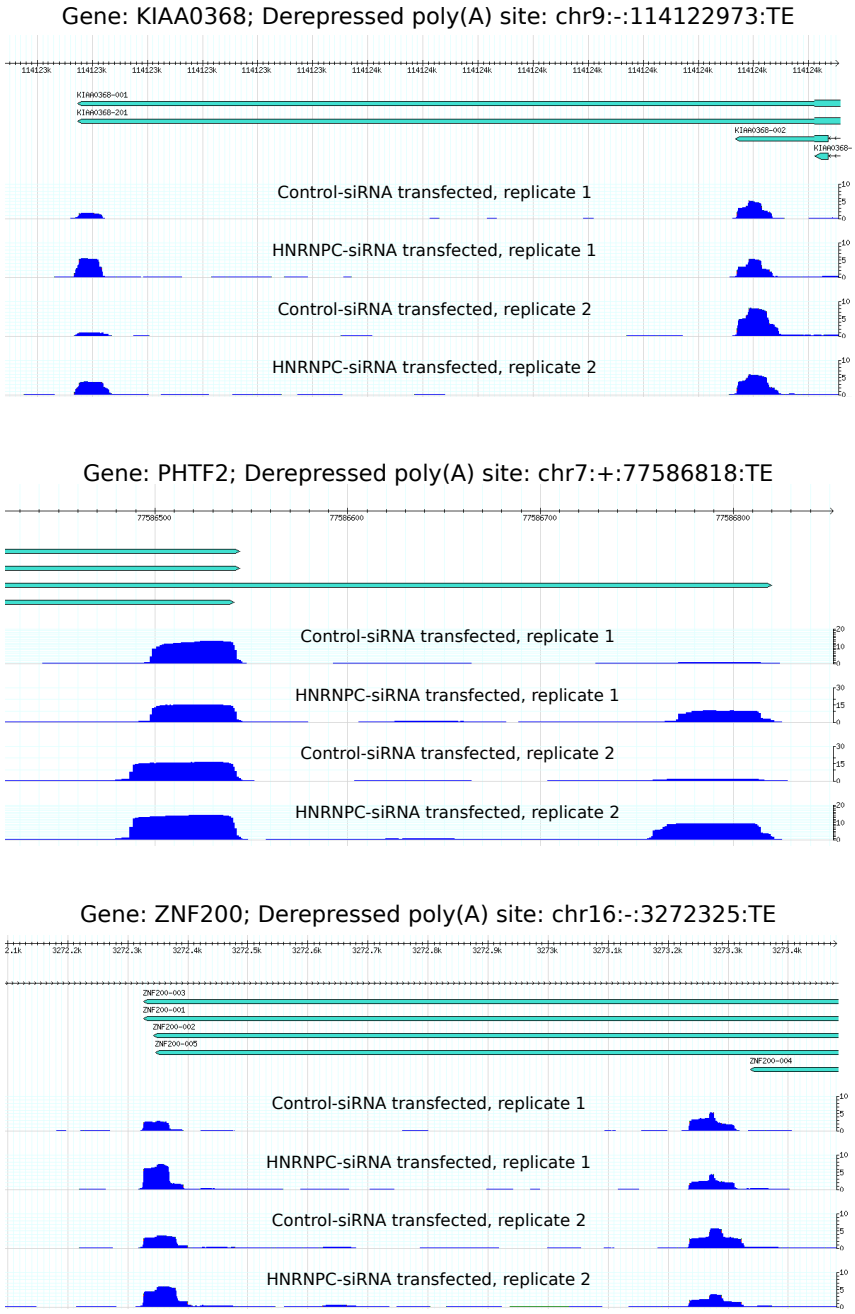
Supplementary Figure 9. Smoothened (+/-5 nt) density of non-overlapping $(U)_5$ tracts in the vicinity of sites with a consistent behavior (increased, unchanged, decreased use) in untransfected (wild type, WT) compared to the si-Control transfected (CTRL) HEK 293 cells.



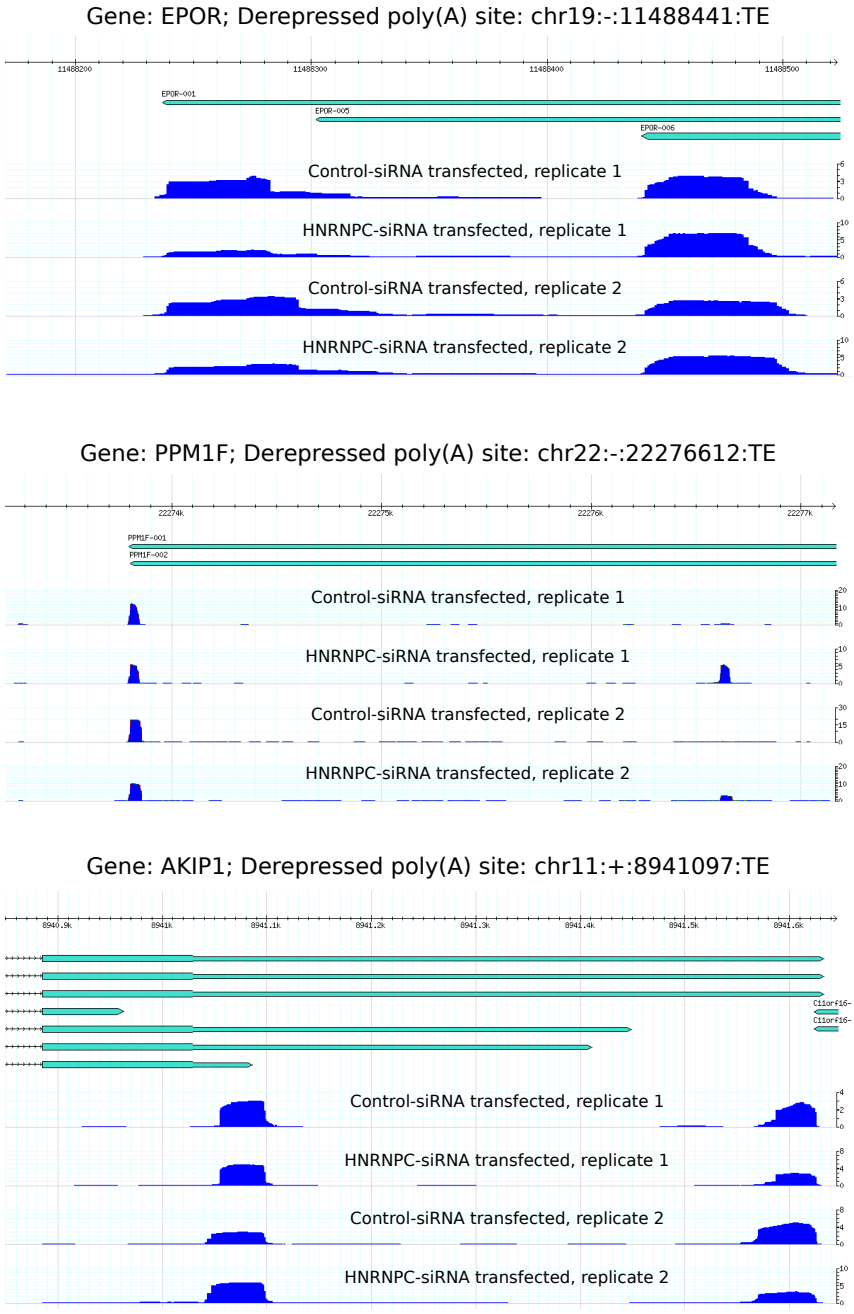
Supplementary Figure 10. Relationship between the $(U)_5$ content around poly(A) sites and their behavior upon HNRNPC knock-down. 1000 poly(A) sites that increased most, decreased most or changed least (and reproducibly, between the two replicate experiments) in usage upon HNRNPC knock-down were extracted, and the fractions of each of these types of sites that had at least one occurrence of the $(U)_5$ motif at the indicated distance from the poly(A) site were calculated. 'u' and 'd' indicate upstream and downstream of poly(A) sites and the numbers indicate the boundaries (in nt) of the windows relative to poly(A) sites.



Supplementary Figure 11. Number of HNRNPC CLIP reads that intersect with a region of +/-50 nucleotides around poly(A) sites belonging to different categories (consistently decreased/unchanged/increased poly(A) site usage upon HNRNPC knock-down). The number of HNRNPC CLIP reads was normalized by the expression ([RPKM]) of each poly(A) site's host gene. Poly(A) sites that increase in usage have a significantly higher CLIP read support compared to poly(A) sites that do not change in usage upon HNRNPC knock-down (p-value < 0.0007, two-sided Kolmogorov-Smirnov test).

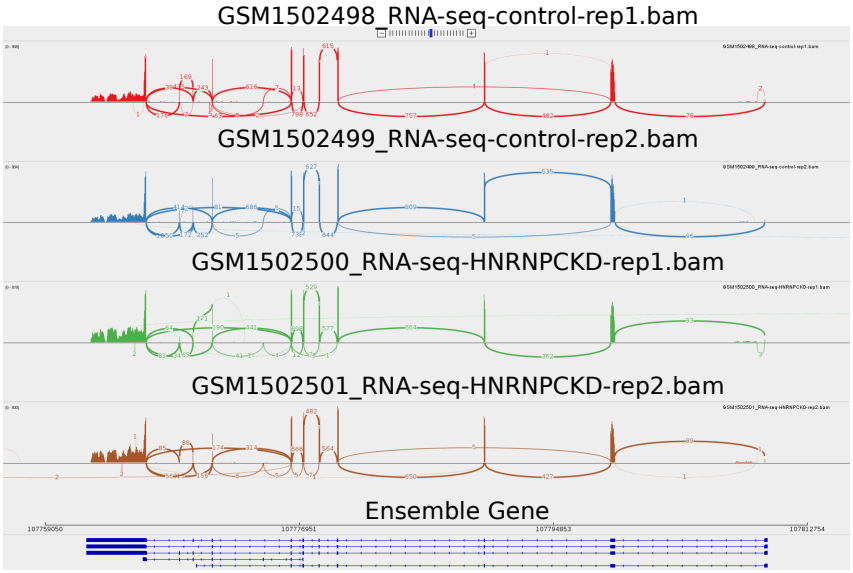


Supplementary Figure 12. Browser shots of A-Seq read densities within 3' UTRs with **distal poly(A) sites that are derepressed upon knock-down of HNRNPC**. The y-axis shows library size normalized read counts per nucleotide.

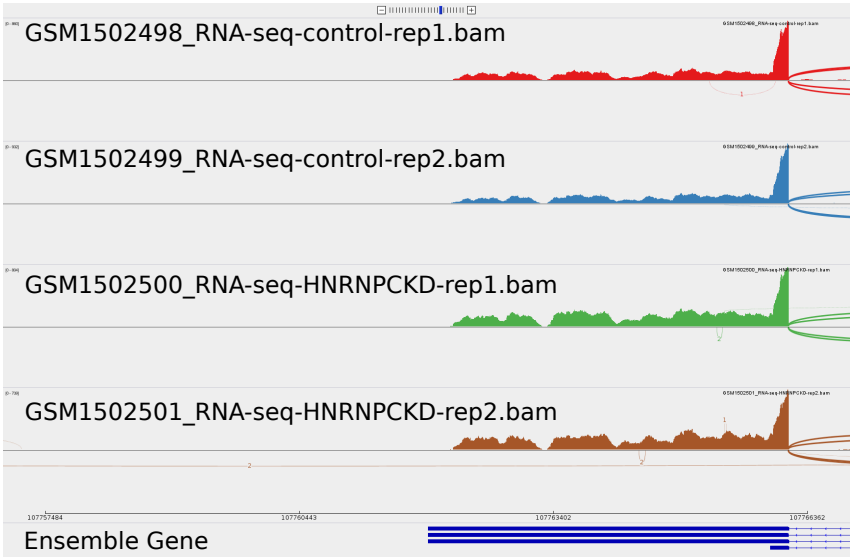


Supplementary Figure 13. Browser shots of A-Seq read densities within 3' UTRs with **proximal poly(A) sites that are derepressed upon knock-down of HNRNPC**. The y-axis shows library size normalized read counts per nucleotide.

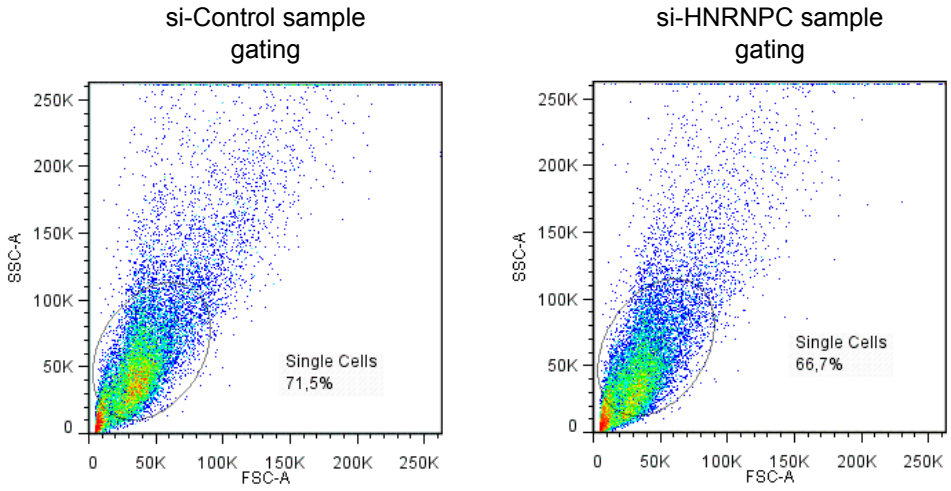
(A) Sashimi plots of the CD47 locus as derived from mRNA-Seq data
region: chr3:107756068-107815808 (human genome version hg19)



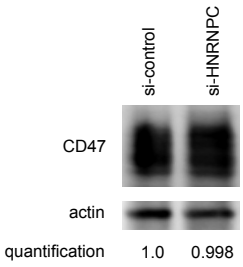
(B)
Sashimi plots of the CD47 3'UTR locus as derived from mRNA-Seq data
region: chr3:107756992-107766867 (human genome version hg19)



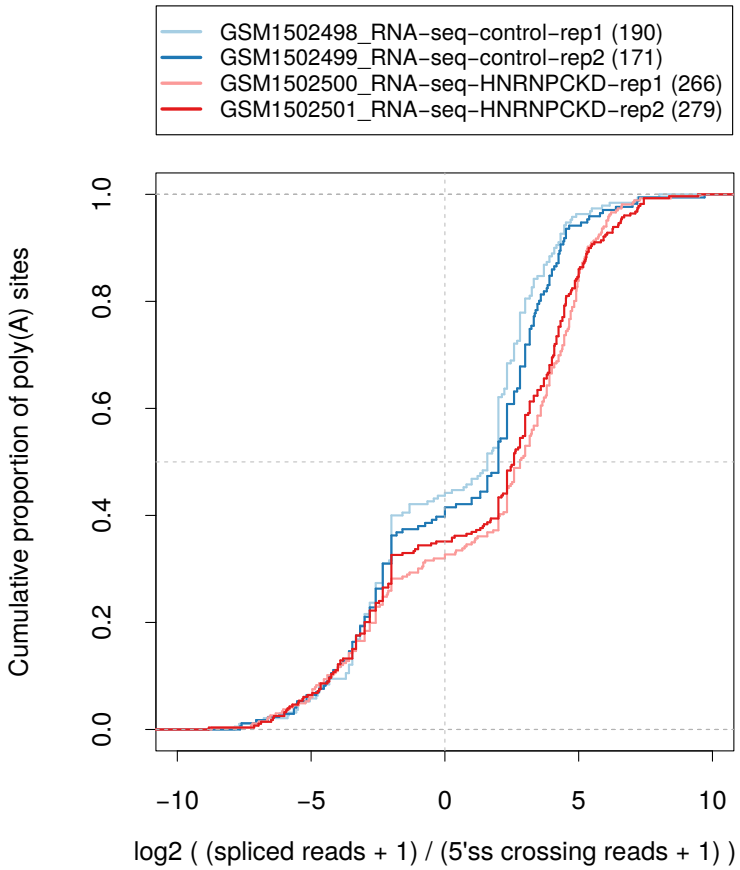
Supplementary Figure 14. "Sashimi" plots constructed from previously published (see [19]) mRNA-Seq data (2 replicates of 2 experiments) obtained from HEK 293 cells that have been transfected with si-Control or si-HNRNPC, respectively. After adaptor removal, paired-end reads were mapped applying the STAR aligner with default settings [20]. The mappings were visualized (Sashimi plots) using the Integrative Genomics Viewer (IGV) software [21].



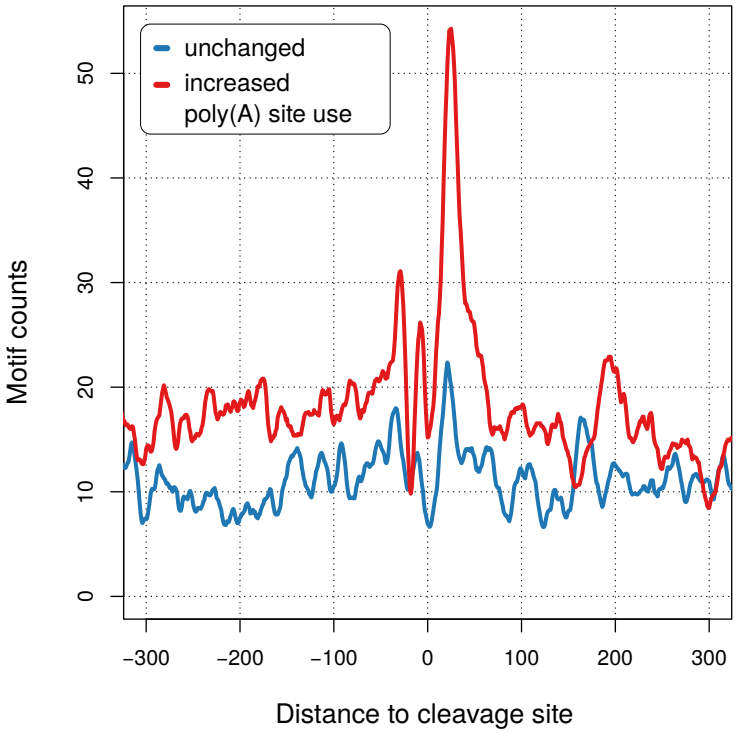
Supplementary Figure 15. For indirect immunophenotyping of membrane CD47 levels in HEK 293 cells that were either treated with a control siRNA (left panel) or with si-HNRNPC (right panel) a minimum of 10000 gated events was considered. The gate is indicated.



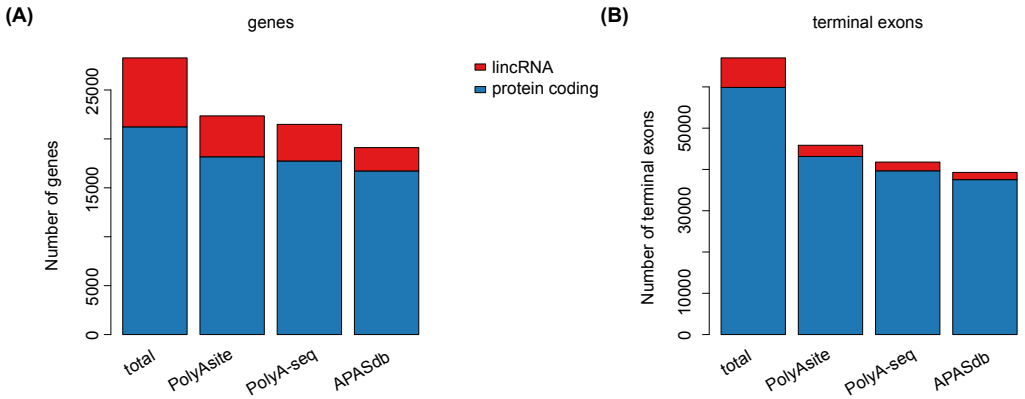
Supplementary Figure 16. Western blots of CD47 and actin proteins in cells treated with either a control siRNA or with si-HNRNPC for 72 hrs. Signals were quantified with the ImageJ software and relative CD47 levels are reported with respect to actin and control siRNA = 1.



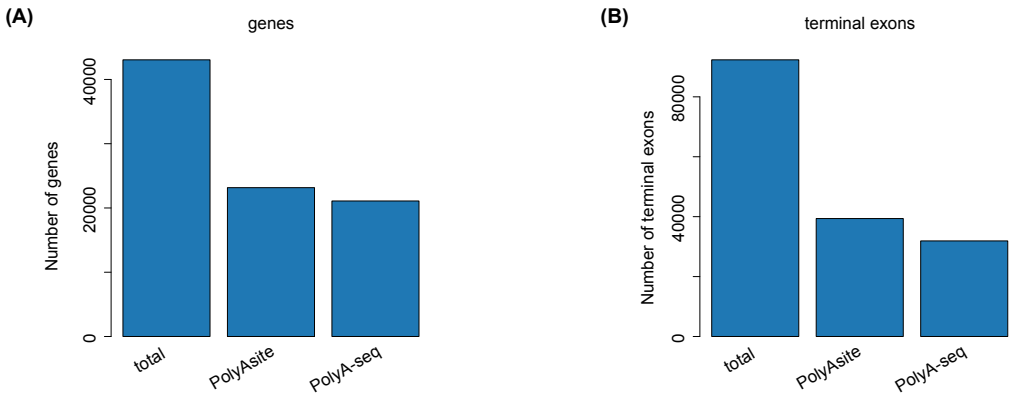
Supplementary Figure 17. Cumulative distribution functions of the log2 ratios of spliced reads to reads that map beyond the 5' splice site (5'ss) of the closest, upstream located exon of each consistenly derepressed, intronic poly(A) site. Intronic poly(A) sites are associated predominantly with the emergence of new exons relative to the extension of internal exons, in both si-Control and si-HNRNPC transfected cells. The HNRNPC knock-down causes a further significant shift towards novel terminal exons created by splicing rather than by internal exon extension (replicate 1 p-value: 4.0e-06, replicate 2 p-value: 8.6e-03, two-sided Mann-Whitney U test). The numbers shown in the legend (written in brackets) indicate the number of intronic poly(A) sites that were used to construct this plot (for more details, see the Methods section).



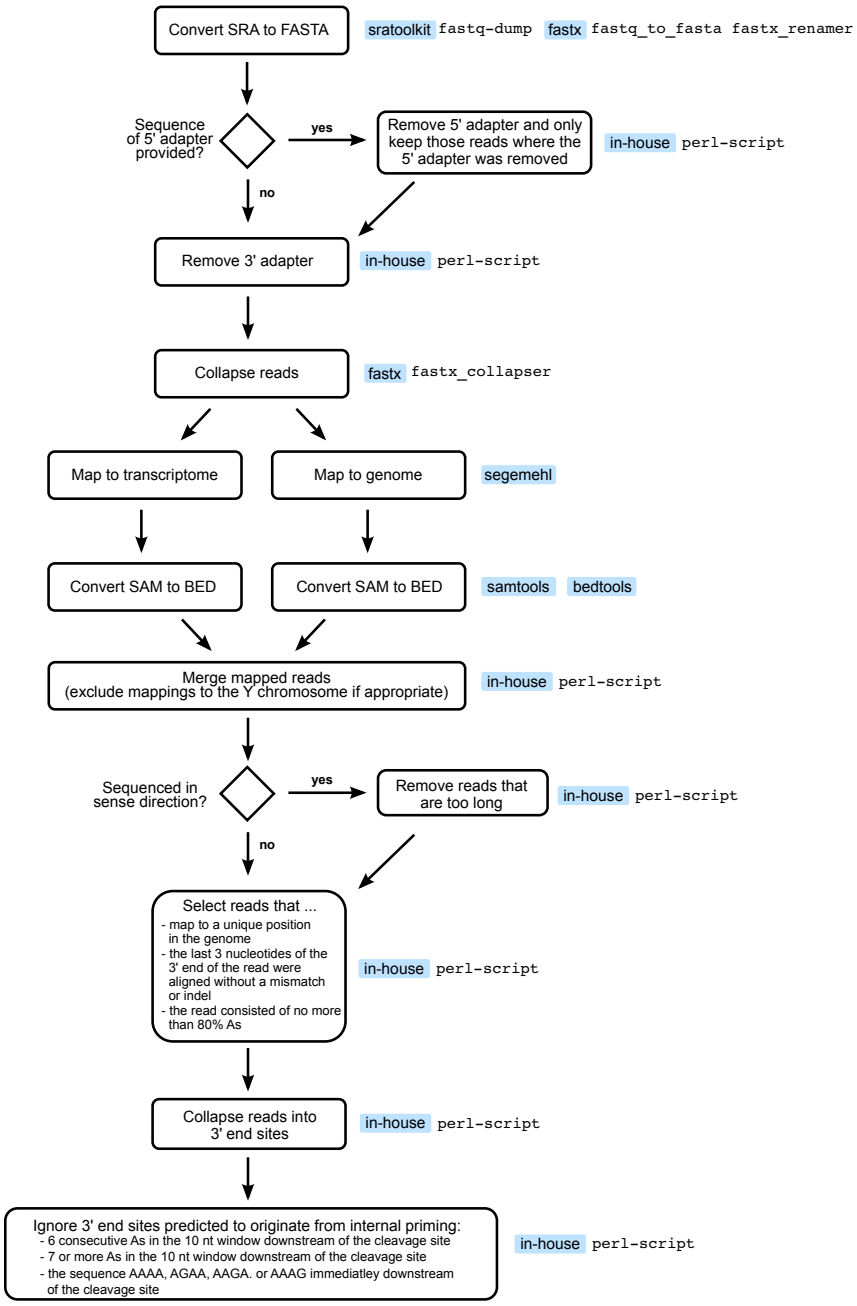
Supplementary Figure 18. Smoothened (+/-5 nt) density of non-overlapping $(U)_5$ tracts in the vicinity of intronic poly(A) sites with a consistent behavior (increased or unchanged use) in the two HNRNPC knock-down A-seq2 experiments.



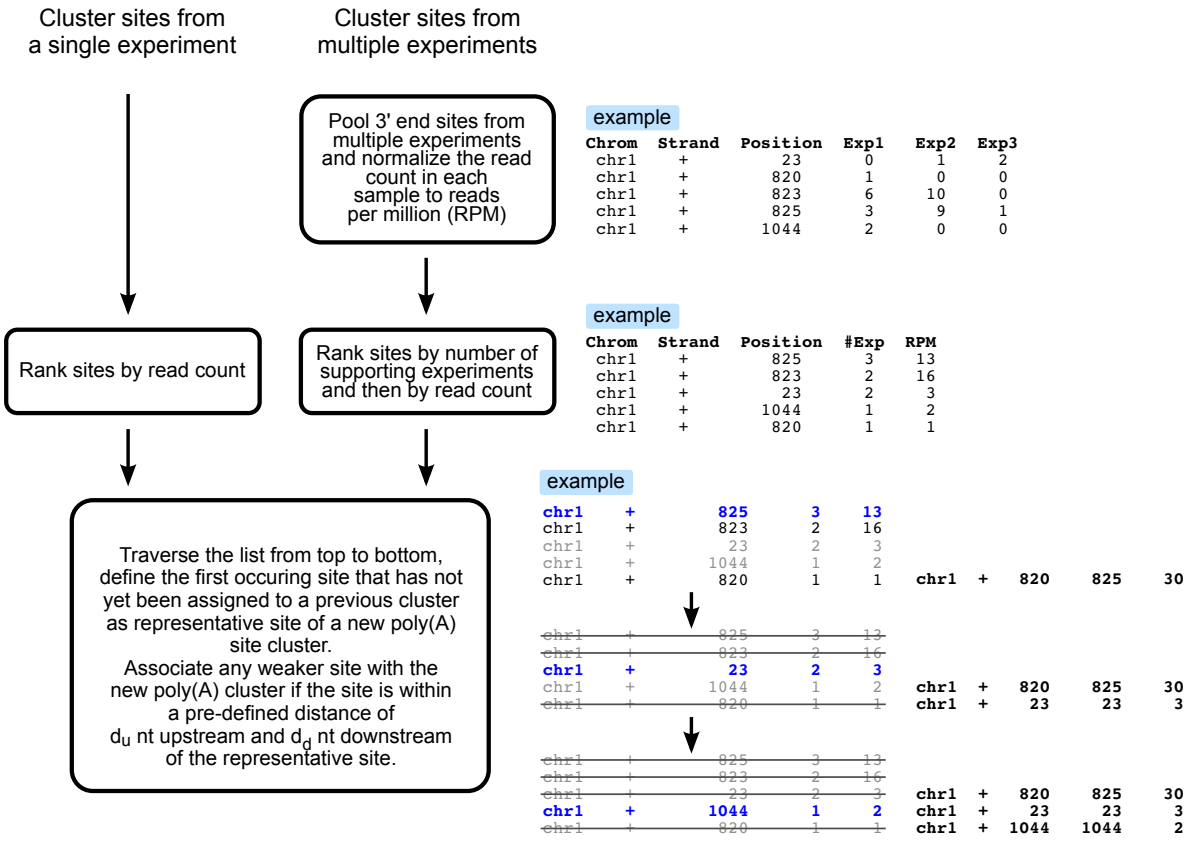
Supplementary Figure 19. Number of annotated features (based on the UCSC Basic Table of the GENCODE v19 human (hg19) annotation) that are covered by sites from different atlases. **(A)** Coverage of genes by sites from PolyAsite (present manuscript), PolyA-seq [15] and APASdb [18]. A gene was considered covered if the genomic position of at least one poly(A) site was within the genomic range of the gene. **(B)** Same as **(A)** but for the terminal exons from the annotation.



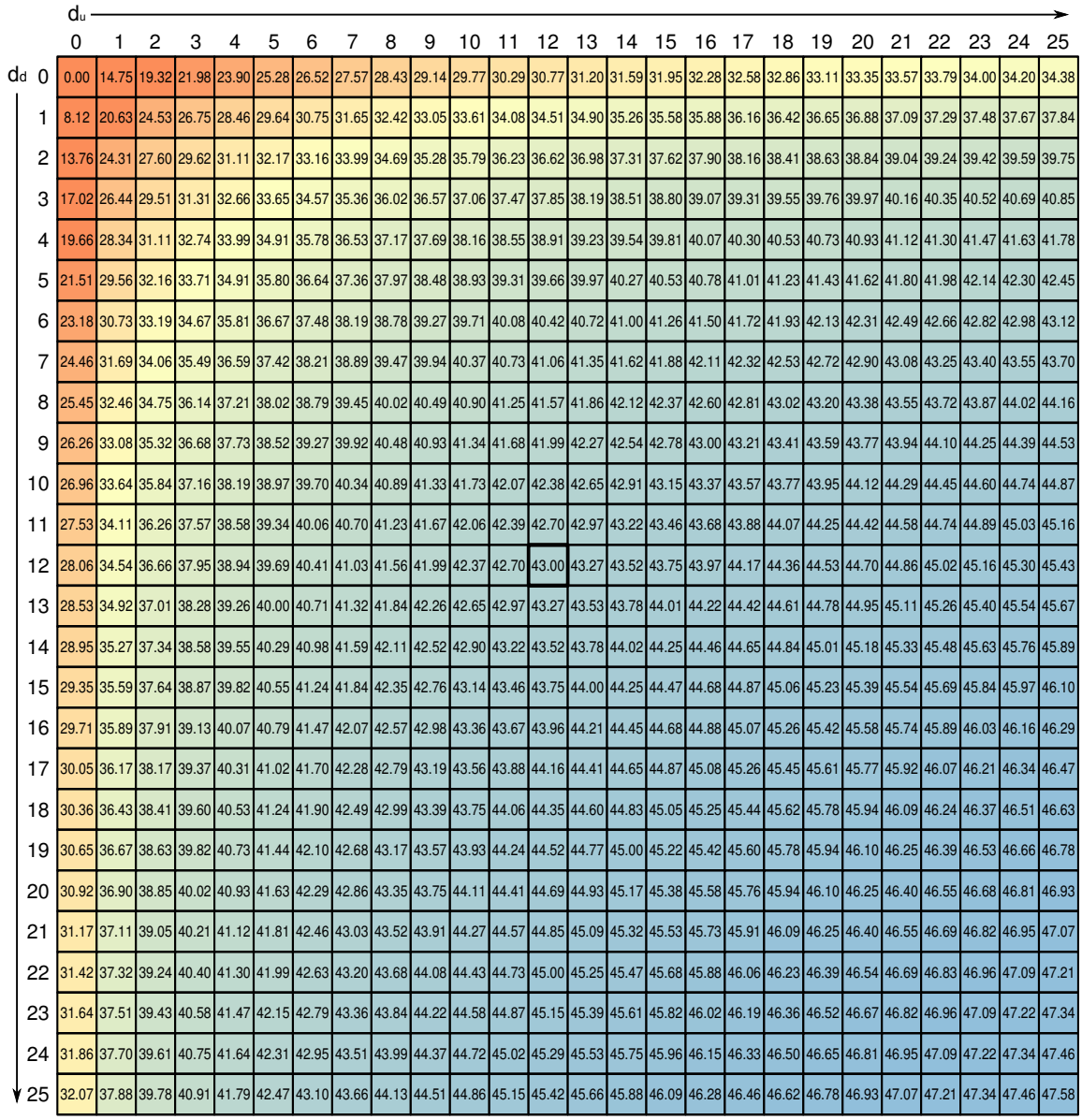
Supplementary Figure 20. Number of annotated features (based on the ENSEMBL mouse (mm10) annotation from UCSC) that are covered by sites from different atlases. **(A)** Coverage of genes by sites from PolyAsite and PolyA-seq [15]. A gene was considered to be covered if the genomic position of at least one poly(A) site was within the genomic range of the gene. **(B)** Same as **(A)** but for the terminal exons from the annotation.



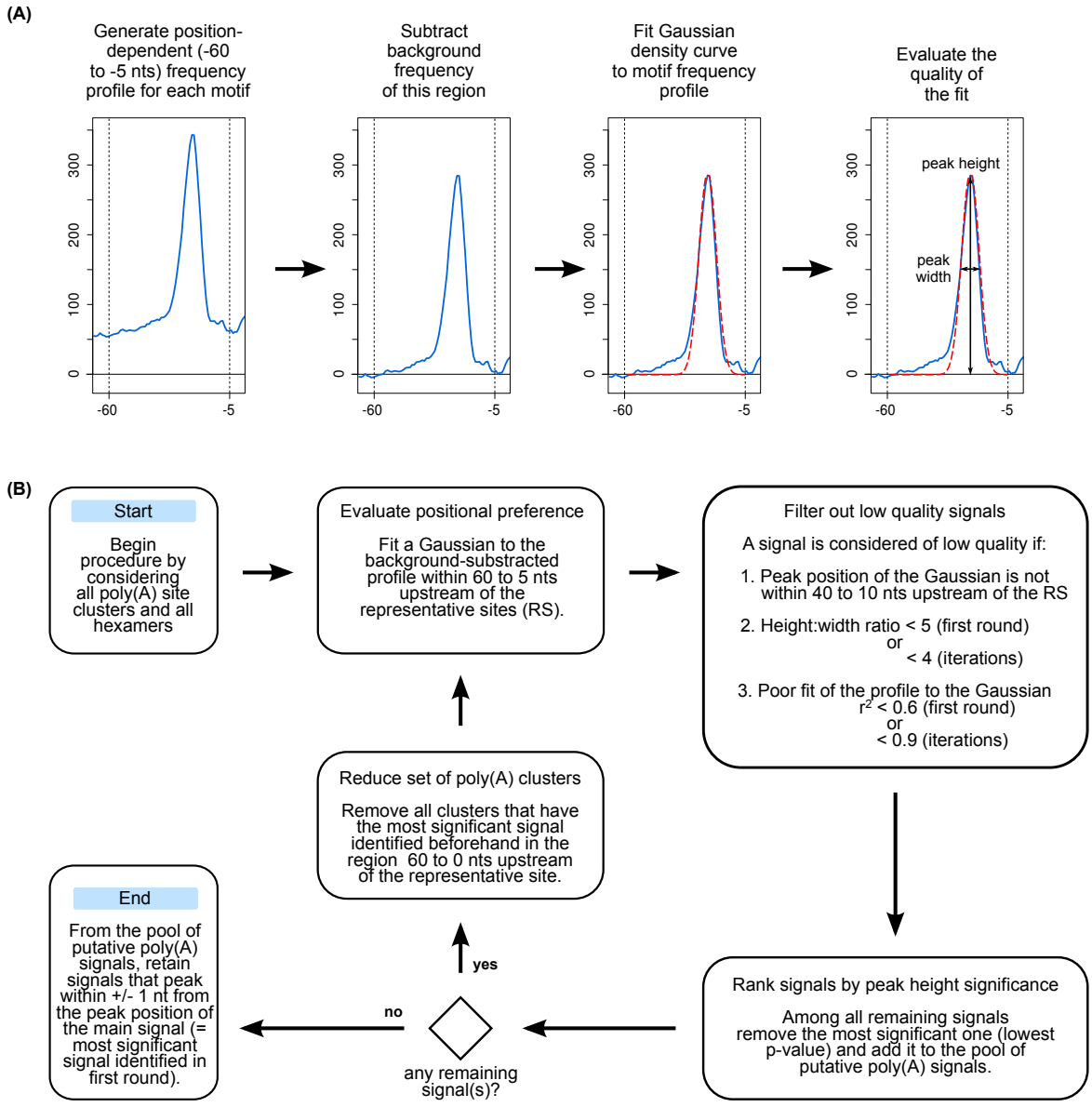
Supplementary Figure 21. Outline of the computational pipeline for processing 3' end sequencing data.



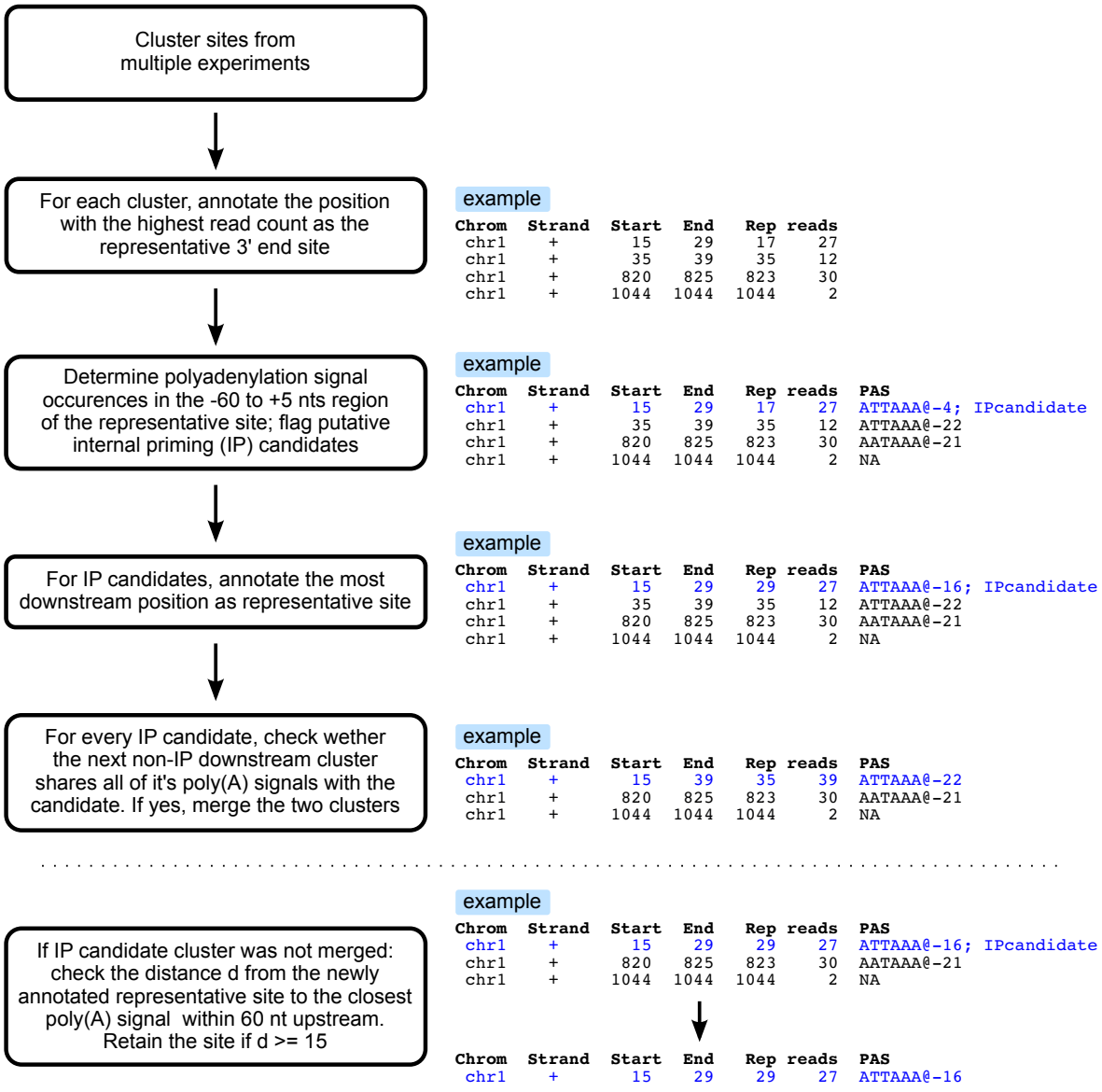
Supplementary Figure 22. Outline of the computational pipeline for clustering closely spaced 3' end sites into 3' end processing regions. A toy example data set is used to illustrate the procedure.



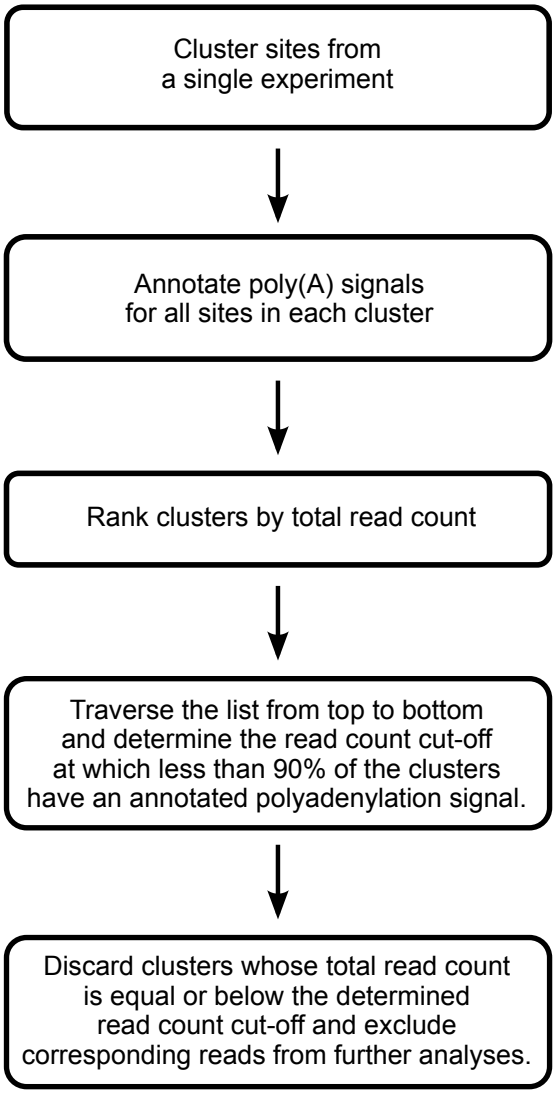
Supplementary Figure 23. Evaluation of the distance parameters for clustering closely spaced, putative 3' end processing sites. d_u and d_d refer to the distance upstream and downstream of the representative site, respectively. Values in the plot denote the percentage of 3' end processing sites that were part of a multi-site cluster when a particular set of distance parameters was applied to cluster individual sites. While initially there is a steep increase in the proportion of reads in clusters, a plateau is soon reached. Distances $d_u = 12$ and $d_d = 12$ were chosen in this study.



Supplementary Figure 24. Outline of the computational procedure that we used to identify poly(A) signals from poly(A) site clusters obtained from high-throughput sequencing of pre-mRNA 3' ends.



Supplementary Figure 25. Outline of the strategy to evaluate poly(A) clusters potentially originating from internal priming.



example

Chrom	Strand	Start	End	Rep reads	PAS
chr1	+	23	23	23	3	ATTAAA
chr1	+	820	825	823	30	AATAAA
chr1	+	1044	1044	1044	2	NA

example

Chrom	Strand	Start	End	Rep reads	PAS
chr1	+	820	825	823	30	AATAAA
chr1	+	23	23	23	3	ATTAAA
chr1	+	1044	1044	1044	2	NA

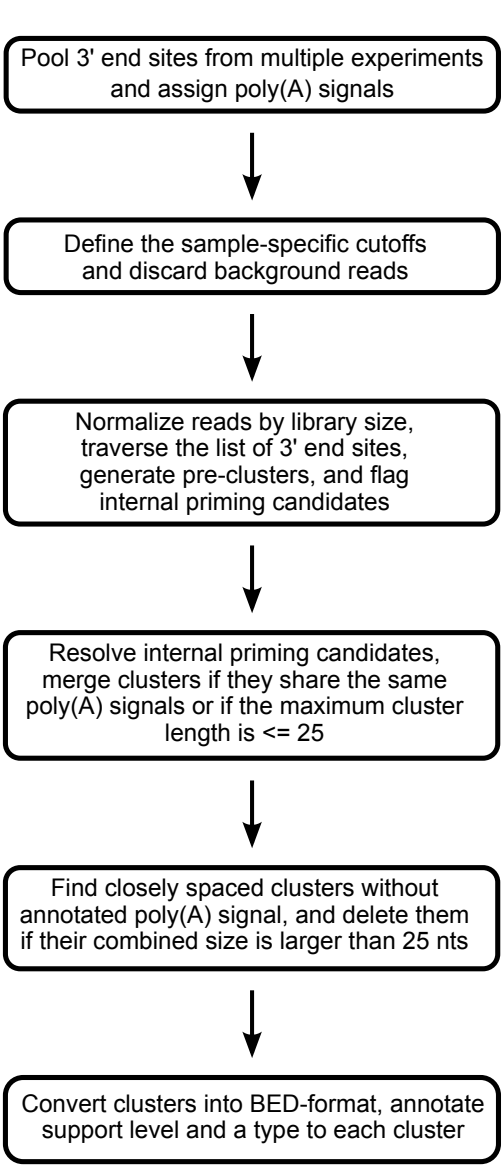
example

Chrom	Strand	Start	End	Rep reads	PAS
chr1	+	820	825	823	30	AATAAA
chr1	+	23	23	23	3	ATTAAA
chr1	**+**	**1044**	**1044**	**1044**	**2**	**NA**

example

Chrom	Strand	Start	End	Rep reads	PAS
chr1	+	820	825	823	30	AATAAA
chr1	+	23	23	23	3	ATTAAA

Supplementary Figure 26. Outline of the procedure that we used to filter out clusters that do not have sufficient experimental support (sample-specific cut-off of read counts).



example

Chrom	Strand	Position	Exp1	Exp2	Exp3	PAS
chr1	+	23	0	1	2	ATTAAA@-8
chr1	+	820	1	0	0	AATAAA@-18
chr1	+	823	6	10	0	AATAAA@-21
chr1	+	825	3	9	1	AATAAA@-23
chr1	+	840	2	0	0	NA

example

Chrom	Strand	Position	Exp1	Exp2	Exp3	PAS
chr1	+	23	0	0	2	ATTAAA@-8
chr1	+	820	0	0	0	AATAAA@-18
chr1	+	823	6	10	0	AATAAA@-21
chr1	+	825	3	9	0	AATAAA@-23
chr1	+	843	0	0	0	AATAAA@-41

example

Chrom	Strand	Start	End	Rep	reads	PAS
chr1	+	23	23	23	0.56	ATTAAA@-8
chr1	+	820	825	823	19.23	AATAAA@-21
chr1	+	843	851	850	4.88	AATAAA@-48

example

Chrom	Strand	Start	End	Rep	reads	PAS
chr1	+	23	23	23	0.56	ATTAAA@-8
chr1	+	820	851	823	24.11	AATAAA@-21
chr1	+	934	937	937	1.32	NA
chr1	+	951	964	959	2.43	NA

example

Chrom	Strand	Start	End	Rep	reads	PAS
chr1	+	23	23	23	0.56	ATTAAA@-8
chr1	+	820	851	823	24.11	AATAAA@-21

example

Chrom	Start	End	ID	Sup	Strand	PAS
chr1	23	23	chr1:+:23:TE	3	+	ATTAAA@-8
chr1	820	851	chr1:+:823:DS	6	+	AATAAA@-21

Supplementary Figure 27. Outline of the computational procedure that we used to combine 3' end processing sites from multiple experiments into a comprehensive catalog of 3' end processing clusters.

3 Supplementary Tables

Supplementary Table 1. Comparison of poly(A) sites that were reported by Derti et al. [15] and You et al. [18] for different human tissues. Both of these studies reported only one genomic position per poly(A) site cluster. To be more permissive in evaluating the overlap of these data sets, we first extended the poly(A) sites from these data sets by 25 nt up- and downstream. A poly(A) site from one study was considered to overlap if there was at least one cluster in the other data set such that both clusters overlapped each other by at least one nucleotide. For each tissue we report both the number of poly(A) site clusters that overlapped as well as those that were unique to a specific data set. In parentheses, the average number of reported reads for the underlying poly(A) sites of the corresponding set of clusters is indicated.

	PolyA-seq clusters overlapping with APASdb clusters	**APASdb clusters overlapping with PolyA-seq clusters**	**PolyA-seq unique clusters**	**APASdb unique clusters**
brain	31,356 (58.47)	30,856 (90.04)	57,754 (19.25)	23,827 (10.83)
kidney	23,793 (104.27)	23,090 (121.53)	71,152 (29.39)	12,006 (19.78)
liver	25,923 (175.45)	25,152 (116.98)	62,317 (16.23)	10,741 (7.26)
muscle	21,910 (151.16)	21,227 (123.36)	90,888 (17.03)	10,743 (37.56)
testes	34,810 (117.72)	34,057 (66.84)	80,258 (11.61)	34,860 (18.47)

Supplementary Table 2. Overview of the samples used to build the genome-wide catalog of 3' end processing site in human

series ID	**sample ID**	**protocol**	**tissue/cell type**	**gender**	**publication**
GSE40859	GSM1003590	"DRS"	"HeLa"	F	[10]
GSE40859	GSM1003591	"DRS"	"HeLa"	F	[10]
GSE40859	GSM1003592	"DRS"	"HeLa"	F	[10]
SRP025988	SRX388391	"DRS"	"HeLa"	F	[12]
SRP022151	SRX275752	"DRS"	"K562"	F	[11]
SRP022151	SRX275753	"DRS"	"K562"	F	[11]
SRP022151	SRX275806	"DRS"	"K562"	F	[11]
SRP022151	SRX275827	"DRS"	"K562"	F	[11]
SRP003483	SRX026582	"SAPAS"	"MDA-MB-231"	F	[17]
SRP003483	SRX026583	"SAPAS"	"MCF-10A"	F	[17]
SRP003483	SRX026584	"SAPAS"	"MCF-7"	F	[17]
GSE25450	GSM624686	"PAS-Seq"	"HeLa"	F	[14]
GSE30198	GSM747470	"PolyA-seq"	"Brain"	NA	[15]

Continued on next page

Supplementary Table 2 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE30198	GSM747471	"PolyA-seq"	"Kidney"	NA	[15]
GSE30198	GSM747472	"PolyA-seq"	"Liver"	NA	[15]
GSE30198	GSM747473	"PolyA-seq"	"MAQC Brain"	NA	[15]
GSE30198	GSM747474	"PolyA-seq"	"MAQC Brain"	NA	[15]
GSE30198	GSM747475	"PolyA-seq"	"MAQC UHR"	NA	[15]
GSE30198	GSM747476	"PolyA-seq"	"MAQC UHR"	NA	[15]
GSE30198	GSM747477	"PolyA-seq"	"Muscle"	NA	[15]
GSE30198	GSM747479	"PolyA-seq"	"Testis"	NA	[15]
GSE30198	GSM747480	"PolyA-seq"	"UHR"	NA	[15]
GSE37037	GSM909242	"A-seq"	"HEK293"	F	[22]
GSE37037	GSM909243	"A-seq"	"HEK293"	F	[22]
GSE37037	GSM909244	"A-seq"	"HEK293"	F	[22]
GSE37037	GSM909245	"A-seq"	"HEK293"	F	[22]
GSE40137	GSM986133	"A-seq"	"HEK293"	F	[8]
GSE40137	GSM986134	"A-seq"	"HEK293"	F	[8]
GSE40137	GSM986135	"A-seq"	"HEK293"	F	[8]
GSE40137	GSM986136	"A-seq"	"HEK293"	F	[8]
GSE40137	GSM986137	"A-seq"	"HEK293"	F	[8]
GSE40137	GSM986138	"A-seq"	"HEK293"	F	[8]
SRP029953	SRX351949	"3'-Seq"	"native B cells"	NA	[3]
SRP029953	SRX351950	"3'-Seq"	"native B cells"	NA	[3]
SRP029953	SRX351952	"3'-Seq"	"brain"	NA	[3]
SRP029953	SRX351953	"3'-Seq"	"breast"	F	[3]
SRP029953	SRX359328	"3'-Seq"	"embryonic stem cells (H9)"	F	[3]
SRP029953	SRX359329	"3'-Seq"	"ovary"	F	[3]
SRP029953	SRX359330	"3'-Seq"	"skeletal muscle"	NA	[3]
SRP029953	SRX359331	"3'-Seq"	"testis"	NA	[3]
SRP029953	SRX359332	"3'-Seq"	"MCF10A"	F	[3]
SRP029953	SRX359333	"3'-Seq"	"MCF10A"	F	[3]
SRP029953	SRX359334	"3'-Seq"	"MCF7"	F	[3]
SRP029953	SRX359335	"3'-Seq"	"HeLa"	F	[3]
SRP029953	SRX359336	"3'-Seq"	"HEK293"	F	[3]
SRP029953	SRX359337	"3'-Seq"	"NTERA2"	M	[3]
SRP029953	SRX359339	"3'-Seq"	"B-LCL cells"	NA	[3]
SRP029953	SRX359340	"3'-Seq"	"MCF10A"	F	[3]
SRP029953	SRX359341	"3'-Seq"	"MCF10A"	F	[3]
GSE52527	GSM1268942	"3P-Seq"	"HeLa"	F	[5]
GSE52527	GSM1268943	"3P-Seq"	"HEK293"	F	[5]
GSE52527	GSM1268944	"3P-Seq"	"Huh7"	NA	[5]
GSE52527	GSM1268945	"3P-Seq"	"IMR90"	F	[5]
GSE56657	GSM1366428	"DRS"	"neuroendocrine tumor"	F	[13]
GSE56657	GSM1366429	"DRS"	"neuroendocrine tumor"	M	[13]
GSE56657	GSM1366430	"DRS"	"Pituitary"	M	[13]

Continued on next page

Supplementary Table 2 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
SRP041182	SRX517334	"SAPAS"	"testis"	M	[18]
SRP041182	SRX517333	"SAPAS"	"ovary"	F	[18]
SRP041182	SRX517332	"SAPAS"	"skeletal muscle"	M	[18]
SRP041182	SRX517331	"SAPAS"	"adipose"	M	[18]
SRP041182	SRX517330	"SAPAS"	"thymus"	M	[18]
SRP041182	SRX517329	"SAPAS"	"small intestine"	M	[18]
SRP041182	SRX517328	"SAPAS"	"pancreas"	F	[18]
SRP041182	SRX517327	"SAPAS"	"liver"	M	[18]
SRP041182	SRX517326	"SAPAS"	"prostate"	M	[18]
SRP041182	SRX517325	"SAPAS"	"breast"	F	[18]
SRP041182	SRX517324	"SAPAS"	"bladder"	F	[18]
SRP041182	SRX517323	"SAPAS"	"uterus"	F	[18]
SRP041182	SRX517322	"SAPAS"	"lung"	M	[18]
SRP041182	SRX517321	"SAPAS"	"placenta"	F	[18]
SRP041182	SRX517320	"SAPAS"	"lymph node"	M	[18]
SRP041182	SRX517319	"SAPAS"	"heart"	M	[18]
SRP041182	SRX517318	"SAPAS"	"cervix"	F	[18]
SRP041182	SRX517317	"SAPAS"	"kidney"	M	[18]
SRP041182	SRX517316	"SAPAS"	"stomach"	M	[18]
SRP041182	SRX517315	"SAPAS"	"spleen"	M	[18]
SRP041182	SRX517314	"SAPAS"	"thyroid"	F	[18]
SRP041182	SRX517313	"SAPAS"	"brain"	F	[18]

Supplementary Table 3. Overview of the samples used to build the genome-wide catalog of 3' end processing site in mouse

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE30198	GSM747481	"PolyA-seq"	"Brain"	NA	[15]
GSE30198	GSM747482	"PolyA-seq"	"Kidney"	NA	[15]
GSE30198	GSM747483	"PolyA-seq"	"Liver"	NA	[15]
GSE30198	GSM747484	"PolyA-seq"	"Muscle"	NA	[15]
GSE30198	GSM747485	"PolyA-seq"	"Testis"	NA	[15]
GSE54950	GSM1327166	"A-seq V2"	"T cells"	NA	[9]
GSE54950	GSM1327167	"A-seq V2"	"T cells"	NA	[9]
GSE54950	GSM1327168	"A-seq V2"	"T cells"	NA	[9]
GSE54950	GSM1327169	"A-seq V2"	"T cells"	NA	[9]
GSE46433	GSM1130096	"2P-Seq"	"embryonic stem cells"	NA	[1]
GSE46433	GSM1130097	"2P-Seq"	"embryonic stem cells"	NA	[1]
GSE46433	GSM1130098	"2P-Seq"	"embryonic stem cells"	NA	[1]
GSE46433	GSM1130099	"2P-Seq"	"embryonic stem cells"	NA	[1]

Continued on next page

Supplementary Table 3 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE46433	GSM1130100	"2P-Seq"	"embryonic stem cells"	NA	[1]
GSE46433	GSM1130101	"2P-Seq"	"embryonic stem cells"	NA	[1]
SRP025988	SRX304982	"DRS"	"embryonic stem cell line E14Tg2a"	M	[12]
SRP025988	SRX304983	"DRS"	"embryonic stem cell line E14Tg2a"	M	[12]
GSE44698	GSM1089085	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089086	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089087	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089088	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089089	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089090	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089091	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089092	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089093	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089094	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089095	"2P-Seq"	"3T3"	NA	[2]
GSE44698	GSM1089096	"2P-Seq"	"3T3"	NA	[2]
GSE52528	GSM1268946	"3P-seq"	"heart"	NA	[5]
GSE52528	GSM1268947	"3P-seq"	"muscle"	NA	[5]
GSE52528	GSM1268948	"3P-seq"	"liver"	NA	[5]
GSE52528	GSM1268949	"3P-seq"	"lung"	NA	[5]
GSE52528	GSM1268950	"3P-seq"	"wat"	NA	[5]
GSE52528	GSM1268951	"3P-seq"	"kidney"	NA	[5]
GSE52528	GSM1268952	"3P-seq"	"heart"	NA	[5]
GSE52528	GSM1268953	"3P-seq"	"muscle"	NA	[5]
GSE52528	GSM1268954	"3P-seq"	"liver"	NA	[5]
GSE52528	GSM1268955	"3P-seq"	"lung"	NA	[5]
GSE52528	GSM1268956	"3P-seq"	"wat"	NA	[5]
GSE52528	GSM1268957	"3P-seq"	"kidney"	NA	[5]
GSE52528	GSM1268958	"3P-seq"	"embryonic stem cells"	NA	[5]
GSE25450	GSM624687	"PAS-Seq"	"ES"	NA	[14]
GSE60487	GSM1480973	"PolyA-seq V2"	"MEF"	NA	[16]
GSE60487	GSM1480974	"PolyA-seq V2"	"MEF"	NA	[16]
GSE60487	GSM1480975	"PolyA-seq V2"	"MEF"	NA	[16]
GSE60487	GSM1480976	"PolyA-seq V2"	"MEF"	NA	[16]
GSE60487	GSM1480977	"PolyA-seq V2"	"MEF"	NA	[16]
GSE60487	GSM1480978	"PolyA-seq V2"	"MEF"	NA	[16]
GSE60487	GSM1480979	"PolyA-seq V2"	"MEF"	NA	[16]
GSE60487	GSM1480980	"PolyA-seq V2"	"MEF"	NA	[16]
GSE62001	GSM1518105	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518106	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518107	"3READS"	"NA"	NA	[7]

Continued on next page

Supplementary Table 3 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE62001	GSM1518108	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518109	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518110	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518111	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518112	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518113	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518082	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518089	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518090	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518102	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518103	"3READS"	"NA"	NA	[7]
GSE62001	GSM1586365	"3READS"	"NA"	NA	[7]
GSE62001	GSM1586366	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518096	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518097	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518098	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518072	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518073	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518074	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518075	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518076	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518077	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518078	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518079	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518080	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518081	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518083	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518084	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518085	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518086	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518087	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518088	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518091	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518092	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518093	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518094	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518095	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518099	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518101	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518104	"3READS"	"NA"	NA	[7]
GSE62001	GSM1586367	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518071	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518114	"3READS"	"NA"	NA	[7]
GSE62001	GSM1586368	"3READS"	"NA"	NA	[7]
GSE62001	GSM1518100	"3READS"	"NA"	NA	[7]
GSE62001	GSM1586363	"3READS"	"NA"	NA	[7]

Continued on next page

Supplementary Table 3 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE62001	GSM1586364	"3READS"	"NA"	NA	[7]
SRP039327	SRX480169	"SAPAS"	"thymus"	NA	[18]
SRP039327	SRX480179	"SAPAS"	"thymus"	NA	[18]
SRP039327	SRX480205	"SAPAS"	"thymus"	NA	[18]
SRP039327	SRX480212	"SAPAS"	"thymus"	NA	[18]
SRP039327	SRX480221	"SAPAS"	"thymus"	NA	[18]
SRP039327	SRX480227	"SAPAS"	"thymus"	NA	[18]
SRP039327	SRX480229	"SAPAS"	"thymus"	NA	[18]
SRP039327	SRX480250	"SAPAS"	"thymus"	NA	[18]
SRP039327	SRX480287	"SAPAS"	"thymus"	NA	[18]

Supplementary Table 4. The 100 most significantly enriched hexamers (binomial test relative to what is expected given the mononucleotide composition of the region from -60 to 0 nt relative to poly(A) site) in the human poly(A) site catalog

hexamer	-log p-value
AATAAA	122788.1
AAATAA	42670.49
AAAAAA	33960.3
ATAAAA	33379.19
TAAAAA	24249.76
AAAATA	21755.03
AAAAAT	19162.31
TTAAAA	16451.96
ATAAAT	14493.43
AAAAAG	14079.72
TTTTTT	13455.43
ATTAAA	12302.28
TAAAAT	11913.92
GCCTGG	11751.91
ATAAAG	11628.45
CCTGGG	11165.77
TTTTCT	10964.83
TGTTTT	10879.94
CCAGCC	10729.18
AAAATG	9002.596
CAGCCT	8279.236
CTTTTT	8043.175
AGAAAA	7959.7
TTTCTT	7707.476
CTGGGC	7594.283
AAAGAA	7535.008
AAGAAA	7519.484
AAATGT	7297.44

Continued on next page

Supplementary Table 4 – continued from previous page

hexamer	-log p-value
GAAAAA	7156.527
AGCCTG	7106.297
TTTAAA	7019.924
TTTTTC	6929.253
TTTTGT	6754.398
CCTCCC	6622.351
TTGTTT	6515.799
TTCTTT	6484.465
TTTTAA	6444.964
TTTCTG	6351.61
CAATAA	6137.289
TAAATG	5913.602
TTTTTG	5750.779
AAAAAC	5741.94
TAAATA	5719.061
TCTTTT	5691.07
ATTTTT	5690.314
CTCCAG	5609.213
CAAAAA	5564.294
TTTGTT	5252.513
TTTTTA	5163.368
CTGTCT	5128.945
TGTGTG	5124.415
AAAACA	5094.2
CCCAGC	5042.282
TTCTGT	5016.795
CTCTGT	4984.282
ATAAAC	4984.15
CTCCCC	4866.824
TATTTT	4738.292
AAAAGA	4679.872
TTTCCT	4662.104
CTGCTG	4550.984
TTTTCC	4286.656
CCTGGC	4259.37
CCTGCC	4236.644
CTGCCT	4207.258
CTGTTT	4086.569
CCCTCC	4082.152
GGAAAA	4078.892
ACAGAG	4074.031
CTGTGT	4001.796
TCTGTG	3969.594
GTTTTT	3911.444
CCCAGG	3869.135
TGTCTC	3865.269

Continued on next page

Supplementary Table 4 – continued from previous page

hexamer	-log p-value
GCCTCC	3851.923
TGCTTT	3843.789
TGCCTG	3713.514
CTTCCC	3708.302
CCCCAG	3686.223
TAATAA	3629.887
TTTCTC	3577.619
TGGAAA	3574.17
TAAAAG	3557.743
TGCTGT	3532.84
TTTATT	3526.132
CCCCCA	3524.531
TCCAGC	3520.258
GAATAA	3458.727
GCTGTG	3405.909
TCTCTG	3392.311
CCACTG	3378.823
CCTCTG	3304.089
TTTCCC	3297.584
GGGAGG	3271.045
CATTTT	3270.061
TTCCTG	3266.088
CTGCCC	3236.691
CTTTCT	3230.07
CAGAGC	3226.857
CTGTGG	3207.589

Supplementary Table 5. The 100 most significantly enriched hexamers (binomial test relative to what is expected given the mononucleotide composition of the region from -60 to 0 nt relative to poly(A) site) in the mouse poly(A) site catalog

hexamer	-log p-value
AATAAA	78344.66
AAAAAA	33032.07
AAATAA	28932.12
ATAAAT	17302.62
ATAAAA	14803.36
TAAAAA	12938.72
TTAAAA	10366.85
TAAATA	10122.15
AAAAAG	8097.119
ATTAAA	7668.254
CAGTGT	6974.536
ATAAAG	6855.813

Continued on next page

Supplementary Table 5 – continued from previous page

hexamer	-log p-value
AAAATA	6839.607
ACAGTG	6185.573
CTGCCT	5763.978
TGTTTT	5692.668
TGTCTG	5583.763
CCTCCC	5520.302
TTTAAA	5008.553
GTGTAC	4968.018
GTGTGT	4958.019
GACAGC	4933.256
TAAAAT	4914.887
AAAAAT	4852.199
CCTCTG	4693.22
TAATAA	4460.155
CTTCTG	4436.615
TGTGTG	4411.729
CTGAAG	4159.753
TGTACT	4135.415
TTGTTT	3858.373
TTTTGT	3721.03
ATAAAC	3683.916
CCTGCC	3667.125
GTGTCT	3663.924
TTTTCT	3652.31
TGCCTC	3617.359
CTACAG	3575.848
AAAGAA	3570.49
GCTACA	3527.289
TTCTGG	3512.262
CTGTCT	3499.525
TTTGTT	3488.113
CTCCCC	3386.621
AGACAG	3353.467
TCTGAA	3231.828
ACAGCT	3161.227
CTGGTG	3148.898
AAATCT	3076.442
TCTGCC	3032.614
AAATGT	3023.56
CTGTGT	2979.327
CTCTGC	2974.548
AGTGTA	2935.839
CAATAA	2867.629
TTTCCT	2843.454
GGTGTG	2836.151
TGTGTC	2810.496

Continued on next page

Supplementary Table 5 – continued from previous page

hexamer	-log p-value
CCTGTC	2803.988
TTTTTT	2748.095
CCCTGT	2719.253
TGAAGA	2718.407
CTTCCT	2690.973
AAGAAA	2651.799
AAAAGA	2636.556
CCCTCC	2573.799
CTGCTG	2560.113
TTTCTT	2559.386
GCTGGG	2522.802
AAAAAC	2519.491
TCTCTG	2486.791
TCTGTG	2482.156
TTTCTG	2480.577
AAACCC	2460.335
AGCTAC	2456.855
TTTTAA	2438.885
TGCTGG	2436.94
CCTGGG	2436.371
GTCTGA	2414.336
TGCTGT	2412.297
CTCTGT	2361.324
TTCTGT	2360.056
GTGCTG	2358.721
AAAATG	2341.729
CAGCTA	2295.836
CCCTCT	2275.77
TACAGT	2265.152
TGTCTC	2255.793
TAAATG	2252.428
CTCCTG	2230.726
TTCTTT	2206.821
AAAACA	2176.917
CTGGGA	2176.094
TGCCTG	2171.784
CTCTTC	2161.823
GCCTCC	2150.538
GCTGTG	2141.131
TAAATC	2138.624
ACCCTG	2131.258
CCTGTG	2111.563

Supplementary Table 6. Summary statistics of 3' end sequencing libraries (A-Seq2 protocol [9]) for control-siRNA and HNRNPC-siRNA transfected HEK 293 cells.

	control-siRNA replicate 1 (ID: 29765)	**HNRNPC-siRNA replicate 1 (ID: 29766)**	**control-siRNA replicate 2 (ID: 32682)**	**HNRNPC-siRNA replicate 2 (ID: 32683)**
Number of reads sequenced	55,274,416	47,917,208	68,650,218	78,065,144
considered high-confidence reads that mapped to a unique position in the genome	6,836,446	9,265,965	13,818,252	15,319,388
Number of reads assigned to tandem poly(A) site clusters having >1 protocol support	2,991,716	4,115,507	6,989,361	8,601,510
Number of reads assigned to sample-specific clusters	2,976,577	4,107,667	6,893,361	8,529,512

Supplementary Table 7. Overview of the number and the proportion of features annotated in the human genome that are covered by poly(A) sites from different atlases.

		total	**PolyAsite**		**PolyA-seq**		**APASdb**	
			covered sites	percentage	covered sites	percentage	covered sites	percentage
genes	protein coding	21,232	18,139	85.43 %	17,742	83.56 %	16,724	78.77 %
	lincRNA	7,048	4,160	59.02 %	3,745	53.14 %	2,387	33.87 %
terminal exons	protein coding	59,869	42,579	71.12 %	39,670	66.26 %	37,533	62.69 %
	lincRNA	7,153	2,689	37.59 %	2,115	29.57 %	1,753	24.51 %

Supplementary Table 8. Overview of the number and the proportion of features annotated in the mouse genome that are covered by poly(A) sites from different atlases.

	total	**PolyAsite**		**PolyA-seq**	
		covered sites	percentage	covered sites	percentage
genes	43,054	22,988	53.39 %	21,088	48.98 %
terminal exons	92,351	38,529	41.72 %	31,903	34.55 %

References

1. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).

2. Spies, N., Burge, C. B. & Bartel, D. P. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* **23**, 2078–2090 (2013).

3. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**, 2380–2396 (2013).

4. Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of caenorhabditis elegans 3'UTRs. *Nature* **469**, 97–101 (2011).

5. Nam, J.-W. *et al.* Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell* **53**, 1031–1043 (2014).

6. Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Methods* **10**, 133–139 (2013).

7. Li, W. *et al.* Systematic profiling of poly(a)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet.* **11**, e1005166 (2015).

8. Gruber, A. R., Martin, G., Keller, W. & Zavolan, M. Cleavage factor im is a key regulator of 3' UTR length. *RNA Biol.* **9**, 1405–1412 (2012).

9. Gruber, A. R. *et al.* Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat. Commun.* **5**, 5465 (2014).

10. Yao, C. *et al.* Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18773–18778 (2012).

11. Ji, X., Wan, J., Vishnu, M., Xing, Y. & Liebhaber, S. A. αcp Poly(C) binding proteins act as global regulators of alternative polyadenylation. *Mol. Cell. Biol.* **33**, 2560–2573 (2013).

12. Lackford, B. *et al.* Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J.* **33**, 878–889 (2014).

13. Rehfeld, A. *et al.* Alternative polyadenylation of tumor suppressor genes in small intestinal neuroendocrine tumors. *Front. Endocrinol.* **5**, 46 (2014).

14. Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761–772 (2011).

15. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**, 1173–1183 (2012).

16. Batra, R. *et al.* Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. *Mol. Cell* **56**, 311–322 (2014).

17. Fu, Y. *et al.* Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* **21**, 741–747 (2011).

18. You, L. *et al.* APASdb: a database describing alternative poly(a) sites and selection of heterogeneous cleavage sites downstream of poly(a) signals. *Nucleic Acids Res.* **43**, D59–67 (2015).

19. Liu, N. *et al.* N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* **518**, 560–564 (2015).

20. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

21. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

22. Martin, G., Gruber, A. R., Keller, W. & Zavolan, M. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep.* **1**, 753–763 (2012).