Landes Bioscience www.landesbioscience.com

Supplemental Material to:

Juan Mata

Genome-wide mapping of polyadenylation sites in fission yeast reveals widespread alternative polyadenylation

2013; 10(8)
http://dx.doi.org/10.4161/rna.25758

www.landesbioscience.com/journals/rnabiology/article/25758/

Addendum:

1. 2013RNABIOL0135R-Supdata.xlsx
2. 2013RNABIOL0135R-Suptables.xlsx

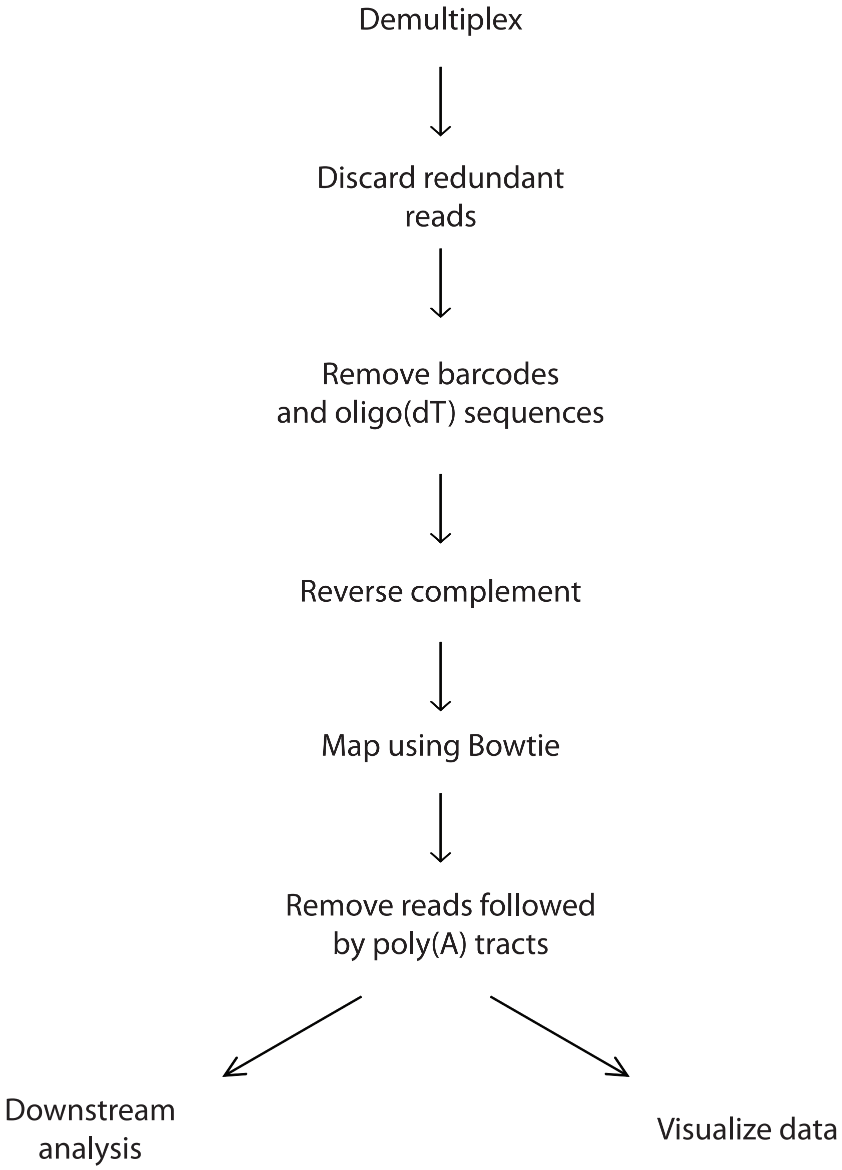
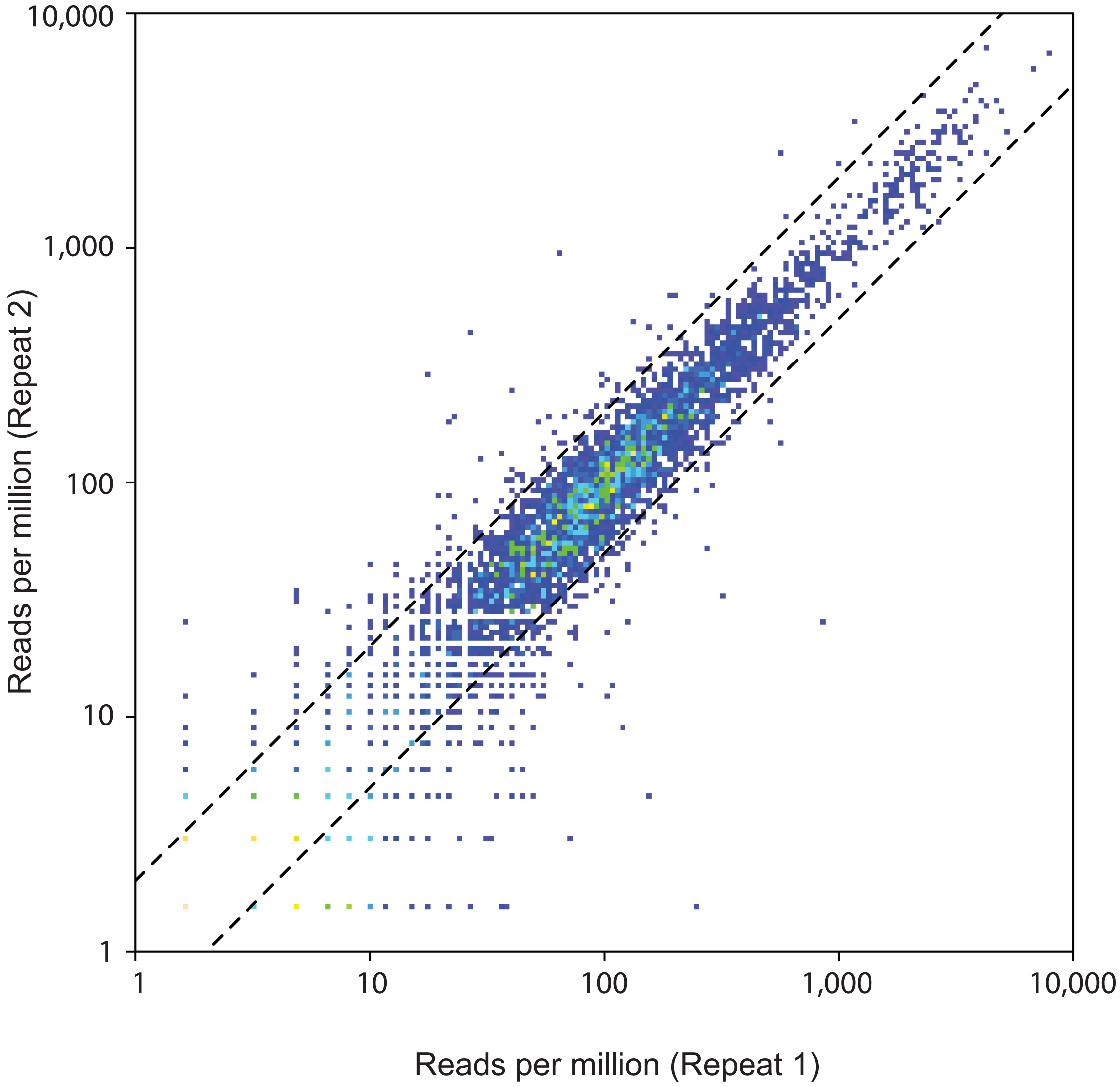


Figure S1. Analysis pipeline.

A



B

	Repeat 1	Repeat 2	Repeat 3
Repeat 1	1.0		
Repeat 2	0.94	1.0	
Repeat 3	0.92	0.96	1.0

Figure S2. Reproducibility of 3PC protocol.

Three independent biological replicates were performed. The number of reads that map to all 3’ UTRs was quantified. (A) Scatter plot comparing read numbers for replicates 1 and 2. Dashed lines correspond to 2-fold differences. (B) Pair-wise Pearson correlations among the three experiments.

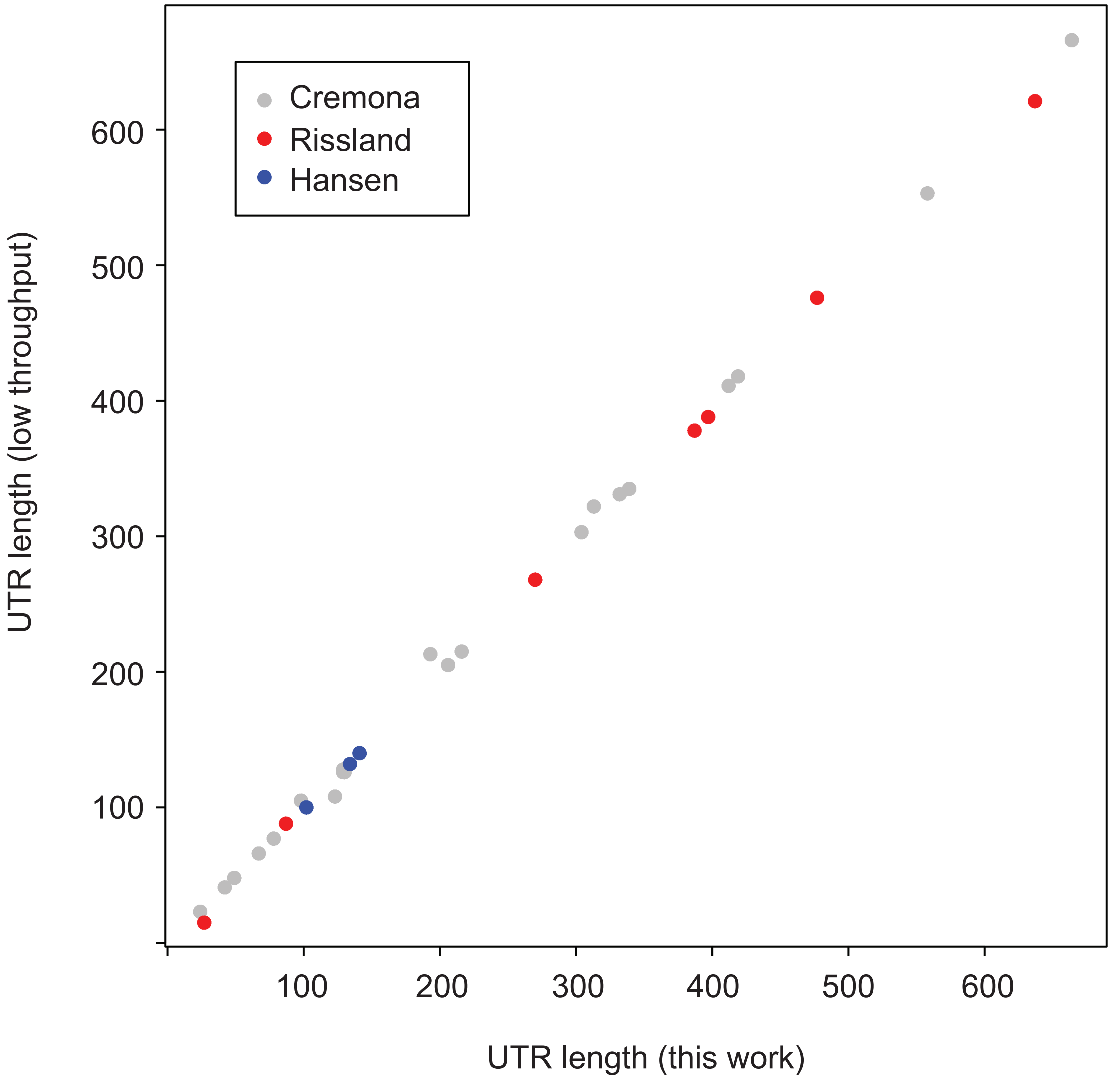


Figure S3. Comparison of 3PC results with published CS mappings.

The position of the CSs determined by low throughput methods was compared with the peak of the closest cluster. Cremona *et al*. and Hansen *et al*. amplified mRNA ends by RT-PCR and sequenced the resulting cDNAs. Rissland *et al.* used circularized rapid amplification of cDNA ends (cRACE) to isolate mRNA ends and analysed the results by sequencing.

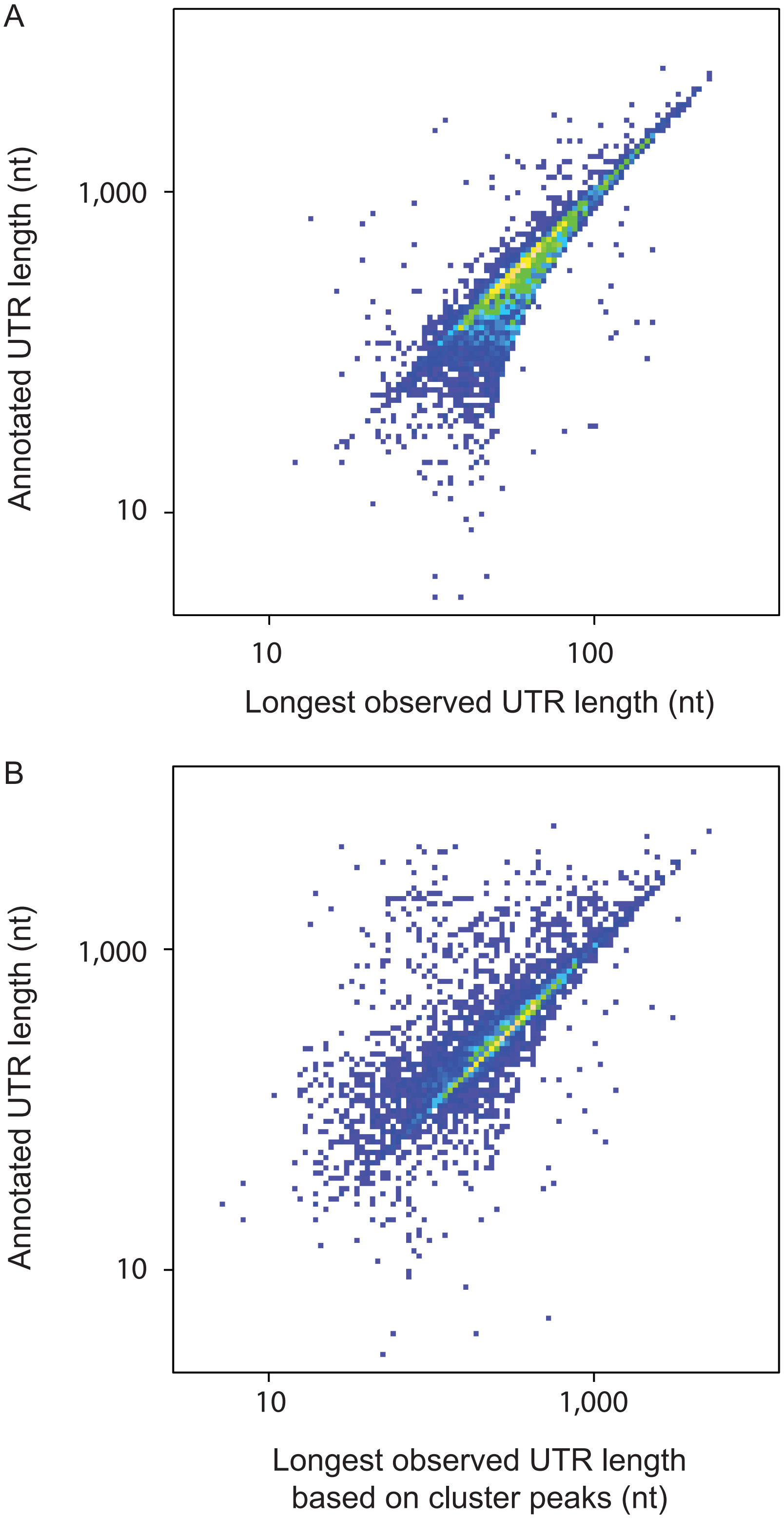


Figure S4. Comparison of annotated 3' UTR lengths to those mapped by 3PC.

Annotated 3' UTRs were obtained from Pombase (see Methods) and refer to the longest observed CS. For the purpose of comparison to annotated 3' UTR, the most distal CS mapped by 3PC was used. This was done in two different ways: (A) The most distal observed single CS within the region defined by the 3' UTR plus 200 nucleotides downstream was compared to the annotated 3' UTR. (B) The peak of the most distal cluster identified by 3PC in a 3' UTR plus 200 nucleotides downstream was used for the comparison. Note that the 3' UTRs defined in (B) represent a 'typical' 3' UTR rather than the longest, explaining why in many cases they are shorter than the annotated ones.

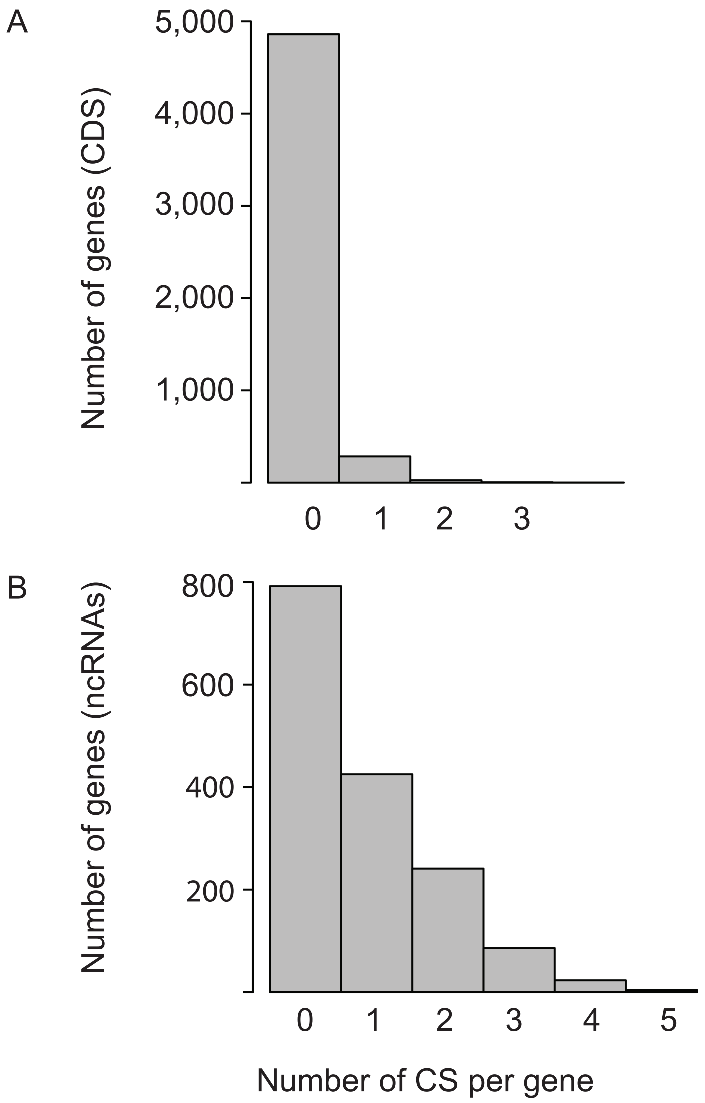


Figure S5. Alternative polyadenylation in coding sequences and ncRNAs.

Histograms displaying the distribution of the number of CSs within coding sequences (A) and in ncRNAs (B).

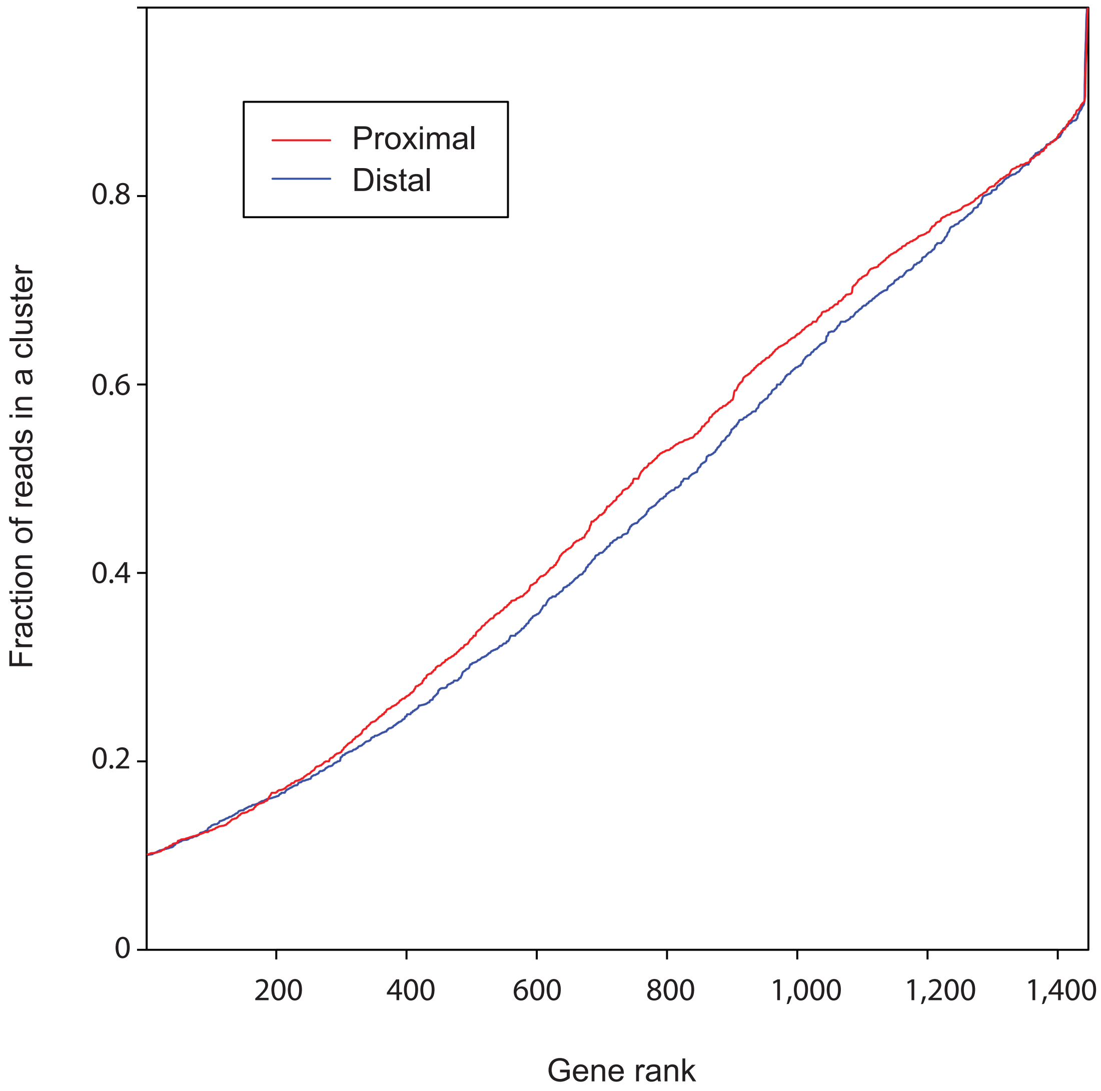


Figure S6. Comparison of the distribution of reads in proximal and distal CSs.

The analysis was performed with 3' UTRs containing exactly two clusters of CSs. For each 3' UTR, the fraction of reads mapping to the proximal (red line) or distal (blue line) cluster were sorted in increasing order. The fraction for each cluster is plotted against its rank. The distribution for proximal and distal clusters is very similar, indicating that there is no major preference in the use of CSs based on their relative position on the mRNA.

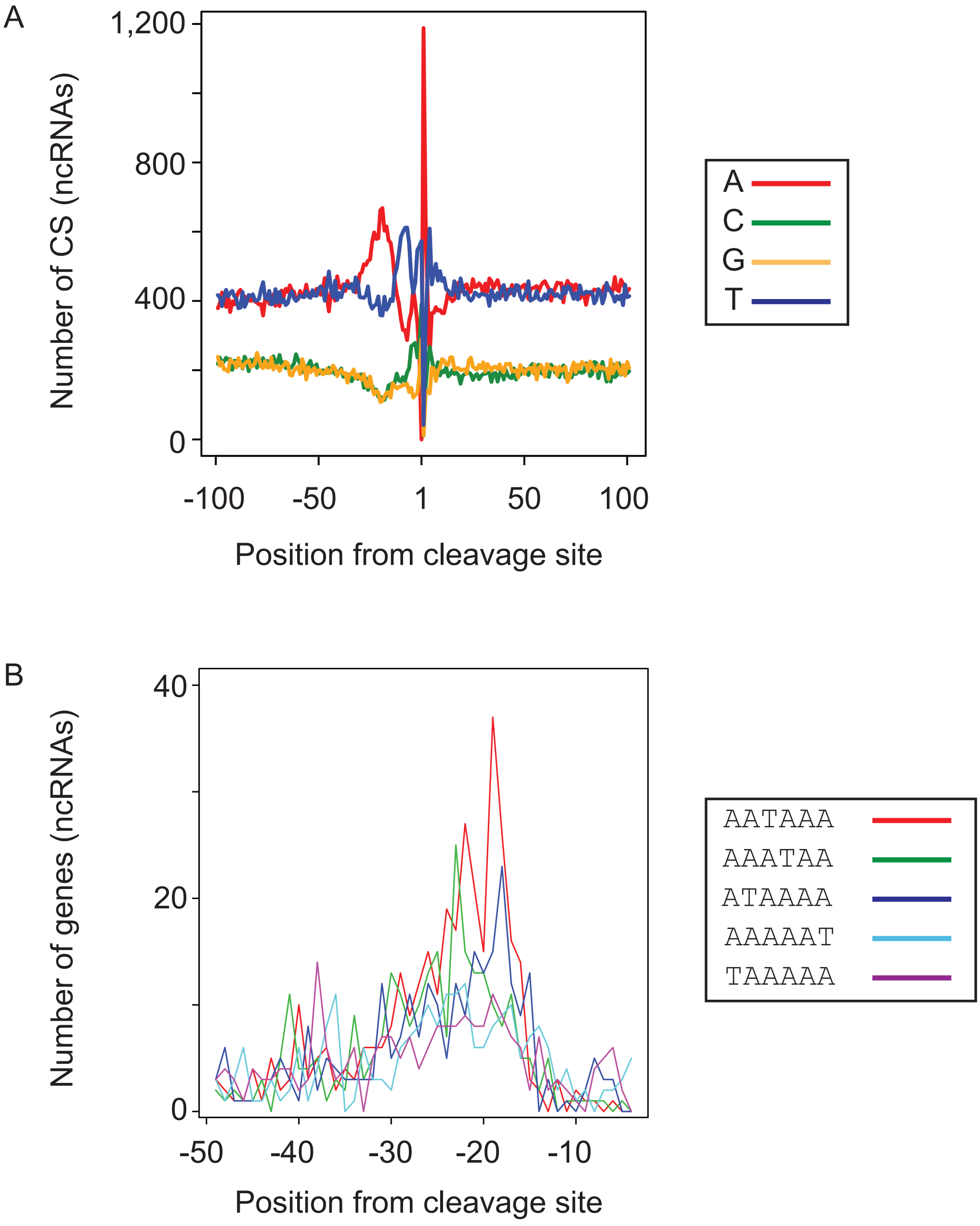


Figure S7. Cleavage and polyadenylation regulatory sequences in 3' UTRs.

The y axes show the number of CSs in ncRNAs. Only the peak of each cluster was used for the analyses. (A) Nucleotide composition of sequences around the CS. (B) Location of the most enriched motif relative to the position of the CS. The numbers refer to the location of the first nucleotide of the hexamer.

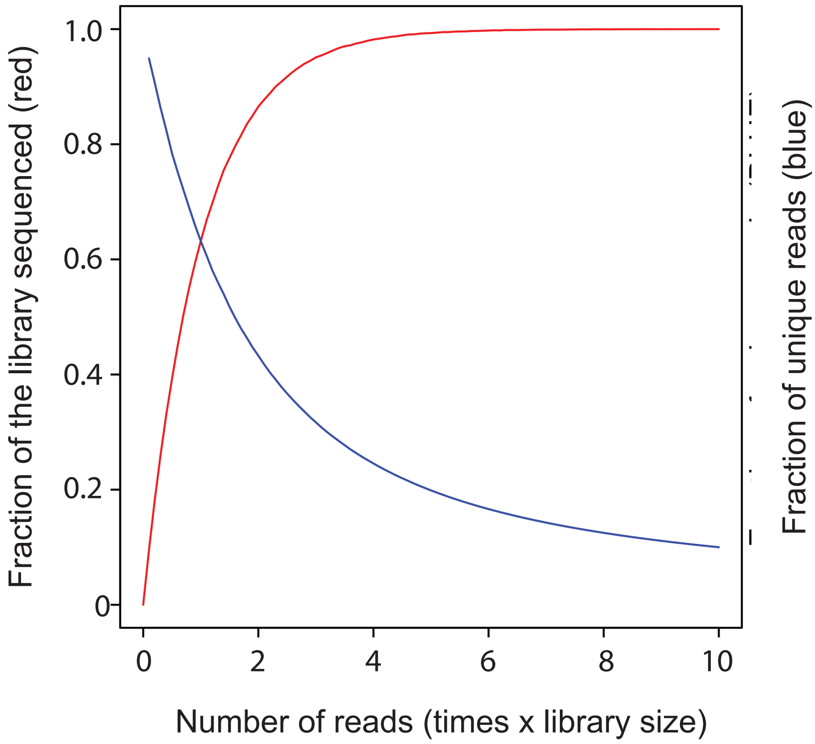


Figure S8. Relationships between sequence depth, library size and number of unique sequences.

The data mimic sequencing from a PCR-amplified library of arbitrary size. This was simulated by random sampling from a library of 100,000 unique elements with replacement. The probability of choosing any element (equivalent to the probability of selecting a given clone for sequencing) was set to be identical, which is equivalent to assuming that PCR amplification is similar for all clones.

The x axis shows the number of sequences obtained as a multiple of the number of unique clones in the library (in a real experiment, this would represent the number of unique cDNAs before PCR amplification). The blue line shows the fraction of unique reads as a function of the total number of reads. Assuming that there are no PCR amplification artifacts, the observed fraction of unique reads (obtained using UMIs) can be used to estimate the size of the original library. The red line shows the fraction of the library that has been sequenced as a fraction of the number of reads. Using the fraction of unique reads obtained as described above, it is possible to estimate the fraction of a library that has been sequenced.

```
                                                                  ------>
5Phos/AGATCGGAAGAGCGGTTCAG/iSp18/CACTCA/iSp18/ACACGACGCTCTTCCGATCTNNXXXNTTTTTTTTTTTTTTTVN
```

Figure S9. Structure of the RT primer.

The location of the anchored oligo(dT) sequence is underlined. This is preceded by a sequence consisting of three random nucleotides (N) that serve as unique molecular identifier, and three additional nucleotides (X) that are used as a barcode to allow the sequencing of multiple samples in a single Illumina lane. The spacers are indicated as /iSP18/. The sequences recognised by the PCR primers are shown in grey. Sequencing starts at the position marked by the arrow, and proceeds across the oligodT sequence into the cleavage site.

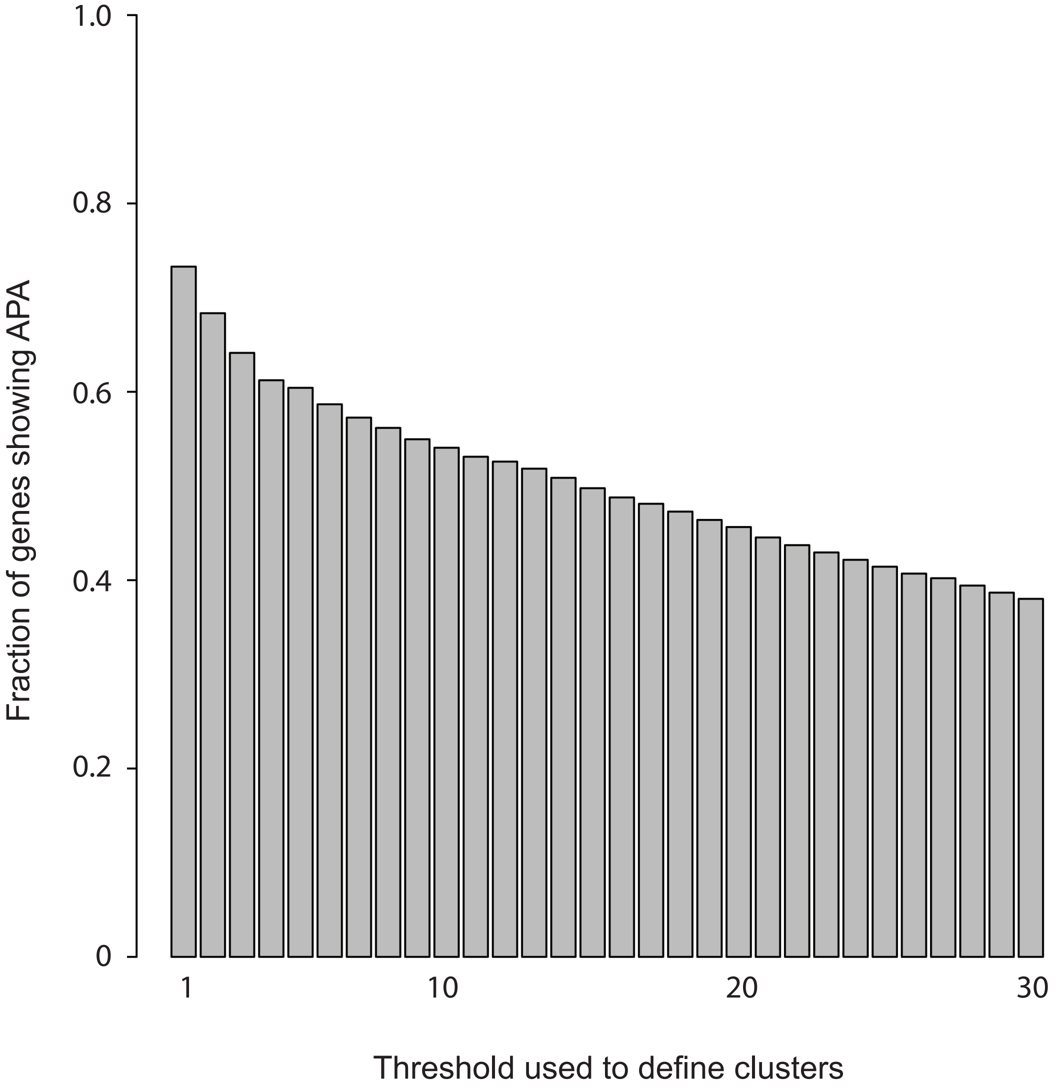


Figure S10. Effect of the threshold distance used to define clusters on the number of called APA.

Cluster generation was performed using increasing distances to merge reads into clusters (see Materials and Methods). The y axis shows the fraction of genes showing APA when clusters were generated using the distances indicated in the x axis. A large fraction of genes displays APA with all distances, indicating that the conclusion that APA is widespread is robust and independent of the way in which clusters are defined.