Figures and figure supplements

Quality control of transcription start site selection by nonsense-mediated-mRNA decay

Christophe Malabat, et al.
Figure 1. Transcription start site sequencing. (A) Schematic view of the methodology used to produce the transcription start site (TSS) sequence-tag libraries (RNA molecules are in blue, DNA molecules in red). (B) All protein-coding genes were aligned on the A of their annotated ATG start codon and the distribution of the TSSs read counts was computed for each position in a window from −200 to +200 nucleotides for samples treated (blue curve) or not (red curve) with tobacco acid pyrophosphatase (TAP). The insert within the figure shows a zoomed view of the −100 to +50 nucleotide region. (C) Correlation between two biologically independent replicates (replicate 1: library Lsp_03.WT, replicate 2: library Lsp_04.WT; see Table 1). The Pearson’s correlation value (ρ) between the read counts of the 174,151 TSSs identified in the two data sets is indicated. DOI: 10.7554/eLife.06722.003
Figure 1—figure supplement 1. Reproducibility of transcription start site (TSS) sequencing. Correlation between two data sets generated in our laboratory months apart (A) and between our data and those published by Pelechano et al. (2013) (B). The Pearson’s correlation value (ρ) between the two data sets is indicated. DOI: 10.7554/eLife.06722.004
Figure 1—figure supplement 2. Transcription start site (TSS) consensus sequences. (A) Consensus sequence around all the identified TSSs generated using the Web-LOGO algorithm (Crooks et al., 2004). The site of transcription initiation (the TSS) is labeled +1 while the preceding nucleotide position is labeled −1. (B) Percentage of mismatches between the actual genomic sequence and the first base (red squares) or all the other bases (blue dots) of the complementary DNAs (cDNAs) after alignment of the TSS reads on the genome. A, G, C, and T on the bottom line correspond to the genomic sequence. A strong percentage of mismatches (∼60%) was observed specifically for the first cDNA position when pyrimidines are encoded on the genome at the corresponding location. In 80% of the cases, these mismatches consisted of a pyrimidine to A mismatch. (C) As in (A) but taking into account only the TSSs for which a pyrimidine to A mismatch at the 5′-end of the cDNAs was observed. The vertical arrows pointing to an A indicate the fact that the +1 nucleotide position, a pyrimidine in the genome, corresponds to a mismatched A at the first position of the cDNA. Up to 30% of the cDNA 5′-ends mapping on a pyrimidine actually conform to this consensus. Note that the presence of a conserved A at position −7 while it is located at position −8 in the general consensus and the presence of the pyrimidine to A mismatch suggests that transcription may not start at the pyrimidine but rather at the following A. Indeed, if one would remove the base at the 5′-end of the cDNA reads and align these to the genome, they would then perfectly conform to the general consensus, including the A at position −8, suggesting that an extra A is added at the 5′-end of the transcript during its synthesis. (D) Schematic model depicting how a one nucleotide ‘backward-shift’ relative to the template occurring during the transcription of the first three consecutive As might lead to the incorporation of four As at the beginning of the transcript. This scenario is supported by the fact that for TSSs mapped at PyAAA sequences, the corresponding transcripts carried three or four As at about equal frequencies, which suggests that this ‘one nucleotide back-shifting’ phenomenon occurs about half of the time when transcription initiates on this particular motif.

DOI: 10.7554/eLife.06722.005
Figure 2. Differential effect of UPF1 and/or RRP6 deletion on mRNAs and pervasive transcripts. Frequency distribution of the ratios of transcription start site clusters (TSSCs) read counts in upf1Δ (red), rrp6Δ (blue), or upf1Δrrp6Δ (purple) compared with wild-type (WT) for mRNAs (A), cryptic unstable transcripts (CUTs) (B), Xrn1-sensitive transcripts (XUTs) (C) and stable unannotated transcripts (SUTs) (D). The dashed vertical lines mark a twofold increase in TSSC read counts in the various mutants relative to the wild-type. The number of identified TSSCs and of features to which they were assigned is indicated for each transcript class. Note that CUTs, XUTs and SUTs constituting overlapping transcript populations, when a TSSC was assigned to a pervasive transcript annotated in more than one of these classes, we arbitrarily associated the corresponding TSSC in priority to CUTs, then to XUTs and finally to SUTs.

DOI: 10.7554/eLife.06722.006
Figure 3. XUTs are primarily targeted for degradation by the nonsense-mediated mRNA decay (NMD). (A) Northern blot analysis of three different Xrn1-sensitive transcripts (XUTs) in wild-type (WT), upf1Δ, xrn1Δ, and upf1Δxrn1Δ cells. ACT1 was used as a loading control. (B) Frequency distribution of the ratios of transcription start site clusters (TSSCs) read counts between upf1Δ and wild-type cells obtained with two different methods for library preparation and normalization procedures. ‘TSSCs’ refers to data obtained using the protocol developed to identify TSSs and ‘RNAseq’ to the protocol used to obtain the whole transcriptome (see ‘Materials and methods’). Only reads corresponding to annotated mRNAs, XUTs and SUTs were included in the analysis (see Supplementary file 2). (C) and (D) Frequency distribution of the ratios of read counts for upf1Δ (red), xrn1Δ (blue) or upf1Δxrn1Δ (purple) compared with wild-type for mRNAs and XUTs and SUTs respectively. The vertical dashed lines mark a twofold increase in read counts.

DOI: 10.7554/eLife.06722.007
Figure 3—figure supplement 1. The presence of short open reading frames (ORFs) and long 3’-UTRs is a hallmark of natural nonsense-mediated mRNA decay (NMD) substrates. (A) Comparison of ORF and 3’-UTR sizes for mRNAs (blue), stable unannotated transcripts (SUTs) (red), and Xrn1-sensitive transcripts (XUTs) (green). The 5’- and 3’-ends were determined for all classes of transcripts using previously published data (Xu et al., 2009; van Dijk et al., 2011; Pelechano et al., 2013). For SUTs and XUTs, we took into account for the analysis the size of the first potential ORF found after the 5’-end and the distance between the end of the transcript and the stop codon of this potential ORF. The median size of these ORFs was 14 amino acids for XUTs and SUTs, compared with 401 for annotated ORFs. The rightmost and uppermost panel displays the frequency distribution of the different features according to the size of their 3’-UTRs and to the size of their ORF respectively. (B) Box plots illustrating the distribution of the size of 3’-UTRs for XUTs and SUTs sensitive (red) or not (blue) to NMD.
Figure 4. Novel transcripts revealed upon deletion of UPF1 and/or RRP6. (A–C) Frequency distribution of the ratios of transcription start site clusters (TSSCs) read counts in upf1Δ (red), rrp6Δ (blue) or upf1Δrrp6Δ (purple) compared with wild-type for transcripts initiating within intergenic regions—‘intergenic TSSCs’ (A), from within an mRNA transcribed region but antisense to the mRNA—A-TSSCs (B) or within an mRNA transcribed region, but in the sense orientation with respect to the mRNA—B-TSSCs (C). (D) Schematic representation of the various classes of TSSCs described above. The small blue and orange vertical bars represent individual TSSs within TSSCs (dashed lines). (E) Northern blot analysis of poly(A)+ RNA from wild-type, upf1Δ, rrp6Δ and upf1Δrrp6Δ. The category to which the transcripts belong is indicated on the right. ACT1 was used as a loading control.

DOI: 10.7554/eLife.06722.009
Figure 4—figure supplement 1. Distribution of read counts for different classes of transcripts. Box plots illustrating the distribution of read counts for the different classes of transcription start site clusters (TSSCs) identified in this study for wild-type (WT) (grey), upf1Δ (red), rrp6Δ (blue), and upf1Δrrp6Δ (purple) cells.
DOI: 10.7554/eLife.06722.010
Figure 4—figure supplement 2. Transcription start site (TSS) consensus sequences for the different classes of transcripts. Consensus sequences around the TSSs generated using the Web-LOGO algorithm (Crooks et al., 2004) for the different categories of TSS clusters (TSSCs) identified in this study: 

Figure 4—figure supplement 2. continued on next page
mRNAs (A), cryptic unstable transcripts (CUTs) (B), Xrn1-sensitive transcripts (XUTs) (C), stable unannotated transcripts (SUTs) (D), ‘intergenic’ transcripts (E), A intragenic (F) and B intragenic (G) transcripts. The numbers of reads and TSSs are indicated.

DOI: 10.7554/eLife.06722.011

Figure 5—effect of the absence of SET2 on transcription start sites (TSSs) identified inside open reading frames (ORFs). (A) Frequency distribution of the ratios of TSS clusters (TSSCs) read counts in upf1Δset2Δ compared with upf1Δ cells for the intragenic A-TSSCs. The comparison was performed for all the TSSCs (blue line) and for the ones identified in a single upf1Δ mutant (red line). (B) As in (A) but for B-TSSCs.

DOI: 10.7554/eLife.06722.012
Deletion of SET2 specifically increases the expression level of intragenic transcription start site clusters (TSSCs). (A-G)

Figure 5—figure supplement 1. Deletion of SET2 specifically increases the expression level of intragenic transcription start site clusters (TSSCs). (A-G)

Frequency distribution of the ratios of TSSC read counts in upf1Δ (red), set2Δ (green) or upf1Δ set2Δ (orange) compared with wild-type for mRNAs (A), cryptic unstable transcripts (CUTs) (B), Xrn1-sensitive transcripts (XUTs) (C), stable unannotated transcripts (SUTs) (D), ‘intergenic’ (E), ‘A-’ (F) and B-TSSCs (G). The dashed vertical lines mark a twofold (log2 = 1) increase in TSS counts in the mutants relative to the wild-type. The numbers of identified TSSCs and of features to which they were assigned is indicated for each class.

DOI: 10.7554/eLife.06722.013
Figure 6. Transcription start site (TSSs) identified inside open reading frames (ORFs) show a differential sensitivity to the absence of SET2 according to their position along the mRNA. (A) Visualization of the TSS reads at the YIL136w and YMR114c loci in the wild-type (WT; black), upf1Δ (red), set2Δ (green), and upf1Δset2Δ (orange) cells. The blue arrows represent ORFs and the horizontal red bar the position of the probes used for the Northern blots displayed on the right. Arrowheads in the right panel indicate the position of the full-length (F.L.) and internally initiated (B1, B2, and B) transcripts. (B) and (C) Frequency distribution of A-TSSCs and B-TSSCs read counts respectively in upf1Δset2Δ vs upf1Δ according to the distance from their associated mRNA TSS. Blue and red lines are for TSSCs sensitive and insensitive to the deletion of SET2 respectively. (D) Frequency distribution of read counts for the A- (blue) and B- (red) TSSCs in upf1Δset2Δ compared to set2Δ cells. DOI: 10.7554/eLife.06722.014
Figure 6—figure supplement 1. Consensus sequences around the transcription start sites (TSSs) for (A) A-TSSCs and (B) B-TSSCs identified in the upf1Δset2Δ mutant generated using the Web-LOGO algorithm (Crooks et al., 2004). The numbers of reads and TSSs are indicated.

DOI: 10.7554/eLife.06722.015
Figure 6—figure supplement 2. Transcription start sites (TSSs) are associated within nucleosome-free regions. Average nucleosome density retrieved from Kaplan et al. (2009) plotted as a function of the distance of individual TSSs for TSSs belonging to the different categories of TSS clusters (TSSCs) identified in this study: mRNAs (A), cryptic unstable transcripts (CUTs) (B), Xrn1-sensitive transcripts (XUTs) (C), stable unannotated transcripts (SUTs) (D), ‘intergenic’ transcripts (E), A intragenic (F) and B intragenic (G) transcripts.
DOI: 10.7554/eLife.06722.016
Figure 7. Deletion of UPF1 reveals numerous minor mRNA-associated transcription start sites (TSSs). (A) Genes were aligned by their start codon and the log2 of ratios of TSS reads for upf1Δ vs wild-type was plotted for TSSs upstream or downstream (iTSSs) the annotated ATG start codons, as depicted in the right panel. The main open reading frames (ORFs) are represented by large blue arrows and upstream ORFs (uORFs) or small internal out-of-frame ORFs by small orange arrows. The thin blue arrows indicate the TSSs. (B) The curve represents the probability at each nucleotide position that the distributions of reads corresponding to the red and blue curves shown in A are the same. The dashed red line marks the 0.05 p-value. (C) Genes were aligned by their annotated start codons (A of the ATG at position +1) and the cumulative TSS read counts per nucleotide (smoothed over 11 nucleotides) was plotted for the wild-type (black) and upf1Δ (red) cells. Inset: Magnification of the +1 to +50 region. The p-value < 2.2 x 10^{-16} is the probability (ANOVA test) that the distributions of the values, per nucleotide, for the red and black curves are the same within the +1 to +50 nucleotides interval.

DOI: 10.7554/eLife.06722.017
Figure 7—figure supplement 1. Initiation of iTSS at purines reflects a bias in codon composition. (A) As for Figure 7C but without smoothing. (B) Genes were aligned by their annotated start codons and the total number of transcription start sites (TSSs) per nucleotide was plotted for the wild-type (black) and upf1Δ (red) cells. (C) Zoom of panel (B) for the region −20 to +50 nucleotides around the annotated start codons (upper panel). Note that A and G of the start codon are frequently used for transcription initiation and iTSSs display periodicity of three nucleotides. This periodicity can be explained by the fact that RNA polymerase II transcription initiates preferentially at purines (R) (Zhang and Dietrich, 2005), which are enriched at the first position of codons as previously reported (Mackiewicz et al., 1999) and depicted in the middle and bottom panels.

DOI: 10.7554/eLife.06722.018
Figure 7—figure supplement 2. TSSs) consensus sequences. Transcription start sites downstream of annotated ATGs (Consensus sequences around the iTSSs generated using the Web-LOGO algorithm (Crooks et al., 2004). To avoid a potential bias due to the presence of the ATG start codon in all genes, only iTSSs mapping at least four nucleotides from the start codon were used for this analysis. The numbers of reads and TSSs are indicated.

DOI: 10.7554/eLife.06722.019
Figure 8. Distribution of in-frame vs out-of-frame ATG codons and methionines for genes with or without transcription start sites downstream of annotated ATGs (iTSSs). (A) Cumulative total number of transcription start sites (TSSs) per nucleotide in upf1Δ cells for TSSs directly followed by the annotated ATGs (blue line) or an upstream open reading frame (uORF) (red line) and for TSSs downstream the annotated ATG (iTSSs) and followed by an out-of-frame ATG (red dashed line) or by an in-frame ATG (blue dashed line). (B) Proportion of in-phase ATGs (+1 frame) following an annotated start codon, binned over nine nucleotides, for protein-coding genes with (red; 3327 genes) or without (blue; 1904 genes) iTSS reads in upf1Δ, as defined in the text. The dashed line indicates the expected value for a random distribution. Genes with iTSS reads are significantly depleted of in-frame ATGs relative to all ATGs in the first 100 nucleotides when compared with genes without iTSS reads (p-value = 3.13 × 10^{-5}; ANOVA). (C) Proportion (log2) along the ORFs of in-frame codons, normalized over the regions downstream the first 300 nucleotides of genes, for all codons (binned over nine nucleotides) for the two sets of genes defined in B. The ATG codon is in red. (D) Frequency of methionines per amino acid position along the ORFs for the two sets of genes defined in B. The p-values (ANOVA method) for the difference in methionine composition over the regions between 0–100 and 100–1000 nucleotides are indicated on the figure.

DOI: 10.7554/eLife.06722.020