Translational control of intron splicing in eukaryotes

Olivier Jaillon, Khaled Bouhouche, Jean-François Gout, Jean-Marc Aury, Benjamin Noel, Baptiste Saudemont, Mariusz Nowacki, Vincent Serrano, Betina M. Porcel, Béatrice Ségurens, et al.

LETTERS

Translational control of intron splicing in eukaryotes

Olivier Jaillon[1,2,3]*, Khaled Bouhouche[4,5,6,7,8]*, Jean-François Gout[9], Jean-Marc Aury[1,2,3], Benjamin Noel[1,2,3], Baptiste Saudemont[4,5], Mariusz Nowacki[4,5], Vincent Serrano[4,5], Betina M. Porcel[1,2,3], Béatrice Ségurens[1], Anne Le Mouël[4,5], Gersende Lepère[4,5], Vincent Schächter[1,2,3], Mireille Bétermier[6,7,8], Jean Cohen[6,7,8], Patrick Wincker[1,2,3], Linda Sperling[6,7,8], Laurent Duret[9] & Eric Meyer[4,5]

Most eukaryotic genes are interrupted by non-coding introns that must be accurately removed from pre-messenger RNAs to produce translatable mRNAs[1]. Splicing is guided locally by short conserved sequences, but genes typically contain many potential splice sites, and the mechanisms specifying the correct sites remain poorly understood. In most organisms, short introns recognized by the intron definition mechanism[2] cannot be efficiently predicted solely on the basis of sequence motifs[3]. In multicellular eukaryotes, long introns are recognized through exon definition[2] and most genes produce multiple mRNA variants through alternative splicing[4]. The nonsense-mediated mRNA decay[5,6] (NMD) pathway may further shape the observed sets of variants by selectively degrading those containing premature termination codons, which are frequently produced in mammals[7,8]. Here we show that the tiny introns of the ciliate *Paramecium tetraurelia* are under strong selective pressure to cause premature termination of mRNA translation in the event of intron retention, and that the same bias is observed among the short introns of plants, fungi and animals. By knocking down the two *P. tetraurelia* genes encoding UPF1, a protein that is crucial in NMD, we show that the intrinsic efficiency of splicing varies widely among introns and that NMD activity can significantly reduce the fraction of unspliced mRNAs. The results suggest that, independently of alternative splicing, species with large intron numbers universally rely on NMD to compensate for suboptimal splicing efficiency and accuracy.

With an average length of 25 nucleotides (nt), the spliceosomal introns of *P. tetraurelia* are among the shortest reported in any eukaryote[9]. Annotation of the somatic genome[10], which was based in part on the alignment of 78,110 expressed sequence tags (ESTs), predicted a total of 39,642 protein-coding genes containing 90,282 introns (2.3 introns per gene on average), 96.8% of which are between 20 and 34 nt in length. That such small introns are recognized through intron definition, as in other unicellular eukaryotes[11], is supported by our observation that introns inserted in the coding sequence of a green fluorescent protein reporter are efficiently spliced out (not shown). Alternative splicing is very limited: not a single case of exon skipping was observed, and fewer than 0.9% of the 13,498 introns covered by at least two ESTs were found to use alternative splice sites, usually closely spaced 3′ sites (results not shown). The compositional profiles of 5′ and 3′ splice sites revealed that only the first and last three bases of introns are highly constrained (Fig. 1); by comparison with short introns of other eukaryotes[3], these profiles seem to have a very low information content.

The size distribution of predicted introns shows a conspicuous deficit in introns whose length is a multiple of 3 (hereafter called $3n$ introns): these represent only 18.7% of the total, in contrast with 42.3% and 39.0% for $3n+1$ and $3n+2$ introns, respectively (Fig. 1c). Because intron prediction relies heavily on the reconstruction of open reading frames and is therefore more likely to overlook short $3n$ introns that do not contain in-frame stop codons, we extracted a high-confidence data set by selecting 6,137 gene models for which each of the predicted introns was confirmed by the alignment of at least one EST. Among the 15,286 confirmed introns, $3n$ introns are still strongly under-represented (Fig. 1d): 21.6% of the total, in contrast with 40.2% and 38.2% for $3n+1$ and $3n+2$ introns, respectively (significantly different from a random distribution; $\chi^2 = 956$, $P < 10^{-16}$). Thus, the under-representation of $3n$ introns is not attributable to annotation artefacts.

One particular feature of $3n$ introns is that they would not cause a frame shift during the translation of intron-retaining mRNAs, whereas the retention of most $3n+1$ or $3n+2$ introns (93.8% and 84.0% of those in the confirmed set, respectively) would introduce a premature termination codon (PTC) in the downstream exons. To

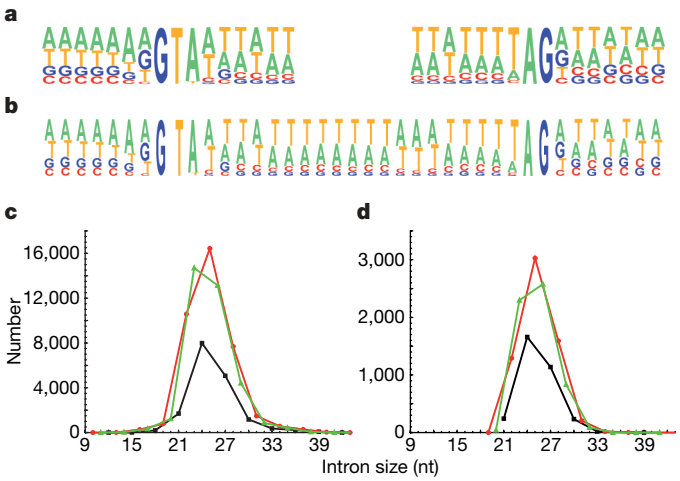

Figure 1 | Characteristics of *P. tetraurelia* introns. **a**, Compositional profiles of the 5′ (left) and 3′ (right) splice sites, including seven nucleotides outside and nine nucleotides inside the intron ($n = 15,286$ EST-confirmed introns). **b**, Compositional profile of the entire length of 25-nt introns (the most abundant size class), with seven nucleotides of the flanking exons on both sides ($n = 3,028$ EST-confirmed introns). **c**, Size distribution of the 90,282 annotated introns. $3n$, $3n+1$ and $3n+2$ introns are shown in black, red and green, respectively. **d**, Size distribution of the 15,286 EST-confirmed introns.

[1]Genoscope (CEA), 2 rue Gaston Crémieux CP5706, 91057 Evry, France. [2]CNRS, UMR 8030, 2 rue Gaston Crémieux CP5706, 91057 Evry, France. [3]Université d'Evry, 91057 Evry, France. [4]École Normale Supérieure, Laboratoire de Génétique Moléculaire, 46 rue d'Ulm, 75005 Paris, France. [5]CNRS, UMR 8541, 46 rue d'Ulm, 75005 Paris, France. [6]CNRS, Centre de Génétique Moléculaire, UPR 2167, 91198 Gif-sur-Yvette, France. [7]Université Paris-Sud, 91405 Orsay, France. [8]Université Pierre et Marie Curie – Paris 6, 75005 Paris, France. [9]CNRS, Laboratoire de Biométrie et Biologie Évolutive, UMR 5558, Université de Lyon, Université Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France.
*These authors contributed equally to this work.

confirm a possible link with translation, size distributions were plotted separately for introns that do or do not contain an in-frame UGA, the only stop codon used in *Paramecium* (Fig. 2). Strikingly, the fraction of 3*n* introns is only 19.1% in the stopless subset, but close to the expected one-third in the stop-containing subset (35.7%). As a consequence of the larger size of the stopless subset, in-frame UGAs are about twice as frequent in the whole set of 3*n* introns as in other size classes (Supplementary Table 1 and Supplementary Figs 1 and 2).

The specific counter-selection of stopless 3*n* introns suggests that *Paramecium* introns are under strong selective pressure to cause premature translation termination in the event of intron retention. A similar bias would easily have been overlooked in other eukaryotes that have longer introns and use three stop codons, because most introns are expected to contain in-frame stops. We therefore examined separately the stopless and stop-containing subsets of complementary-DNA-confirmed introns from *Arabidopsis thaliana*, *Homo sapiens*, *Caenorhabditis elegans* and *Drosophila melanogaster* (Fig. 3 and Supplementary Fig. 3). In all species a highly statistically significant deficit in 3*n* introns is observed among stopless introns but not among stop-containing introns ($P < 10^{-12}$; Supplementary Table 2). The bias is observed only for short introns, suggesting that it may apply to those recognized by intron definition (Supplementary Table 2). In *Schizosaccharomyces pombe*, whose introns are all recognized by intron definition[11], the bias is obvious among annotated introns (Supplementary Fig. 3), and the same trend is observed in a small cDNA-confirmed subset (Supplementary Table 2). Thus, stopless 3*n* introns recognized through intron definition seem to be counter-selected in all intron-rich eukaryotic genomes.

The *P. tetraurelia* genome offers insight into the evolution of intron sequences, as the result of a well-preserved whole-genome duplication that has allowed the identification of 12,026 pairs of duplicated genes[10]. Alignment of the 1,112 pairs belonging to the EST-confirmed set revealed only a handful of cases of intron gains or losses and showed that in at least 37% of 2,774 intron pairs, at least one intron has changed size class since the duplication. The selective pressure that maintains 3*n* depletion in the face of such length variation must therefore be quite strong. In addition, 6,443 pairs of introns of identical sizes provide evidence for evolutionary conservation of stop codons in 3*n* introns. Indeed, 59% of in-frame UGAs in 3*n* introns are conserved in the duplicate, in contrast with 38% for out-of-frame UGAs in 3*n* introns and 37% for in-frame UGAs in non-3*n* introns ($P < 0.001$; see Supplementary Fig. 4).

Because no mechanism other than translation itself is currently known to recognize in-frame stop codons, the finding that eukaryotic short introns are under strong selective pressure to introduce PTCs implies that these introns are translated at a substantial frequency. If translation occurs only in the cytoplasm, this further implies that introns are frequently retained in exported mRNAs, which could be linked to the weakness of splicing signals. During the pioneer round of translation[12], the PTCs resulting from intron retention will trigger mRNA degradation by NMD, thereby protecting cells from possible dominant-negative effects of truncated proteins. Relying on NMD to compensate for inefficient splicing would make stopless 3*n* introns dangerous because their retention, which does not introduce any PTC, can still affect protein function.

As a first test of these hypotheses, we used the double-stranded RNA feeding technique[13] to knock down NMD activity in *P. tetraurelia*. Targeting either or both of the two *UPF1* paralogues consistently resulted in a modest but significant decrease in *UPF1* mRNA levels (more than twofold; Supplementary Fig. 5). This treatment reduced vegetative growth rate by about 30% and completely blocked meiosis (not shown). We then used an oligo(dT)-primed RT–PCR assay to monitor the fraction of unspliced mRNAs for different types of introns, focusing on introns that were found to be maintained in some ESTs or that had non-consensus bases at the third or third-before-last positions (Supplementary Table 3). Spliced and unspliced versions were amplified together in the same PCR reaction with primers flanking the introns, resolved by electrophoresis and quantified (Fig. 4). Even in normal NMD conditions, a variable fraction of unspliced mRNAs was detected for most of the 3*n* + 1, 3*n* + 2 or stop-containing 3*n* introns tested. Knocking down *UPF1* genes increased this fraction by 10–588% (Fig. 4 and Supplementary Fig. 6). Thus, splicing efficiency varies widely among these introns, and NMD can efficiently reduce the unspliced fraction, at least for some of them.

In contrast, all three stopless 3*n* introns tested seem to be very efficiently spliced: only intron 7 showed a small but detectable fraction of unspliced mRNAs, and as expected this was not altered by *UPF1* knockdown. This suggests that many of the stopless 3*n* introns present in the genome are tolerated because they happen to be so efficiently spliced that translational control of splicing is not required. In support of this idea, the analysis of introns occasionally retained in ESTs from wild-type cells shows that the retention rate of stopless introns is significantly lower for 3*n* introns than for 3*n* + 1 or 3*n* + 2 introns (0.55%, in contrast with 0.86% or 0.79%; see Supplementary Table 4). On average, stopless 3*n* introns also have stronger splicing signals than other types of introns (Supplementary Table 5).

The prominent role of NMD in shaping the observed bias is further supported by knockdown of the *Paramecium UPF2* gene (Supplementary Fig. 6) by RNA-mediated interference (RNAi), and by an analysis of the last introns of genes across species. Mammals are

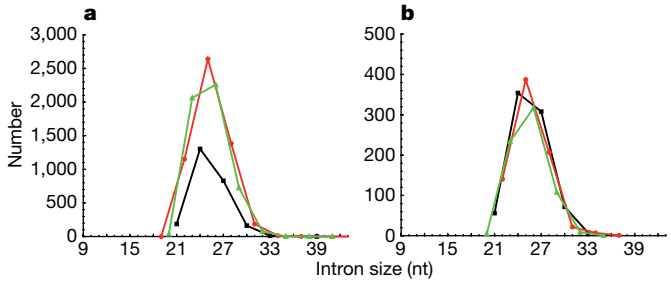

Figure 2 | Size distributions of the 13,050 stopless and 2,236 stop-containing introns from the EST-confirmed set. **a**, Stopless introns; **b**, stop-containing introns. 3*n*, 3*n* + 1 and 3*n* + 2 introns are shown in black, red and green, respectively.

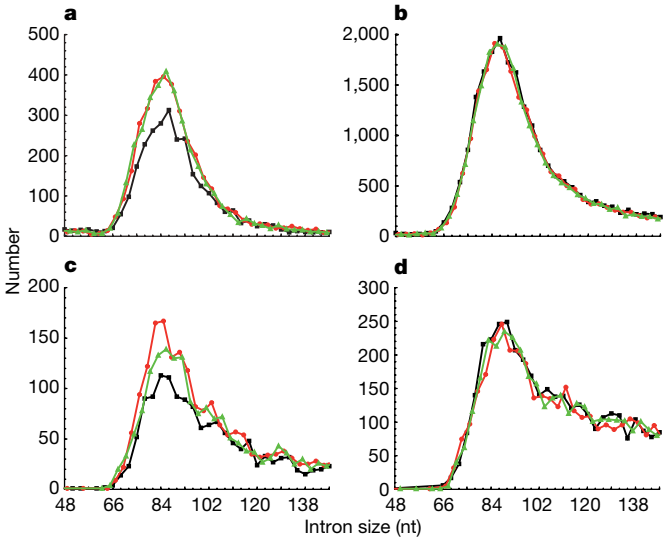

Figure 3 | Size distributions of introns in other eukaryotes. The graphs show the lower modes of the distributions of stopless (**a**, **c**) and stop-containing (**b**, **d**) confirmed introns from *A. thaliana* (**a**, **b**; $n = 10{,}482$ and 87,440, respectively) and *H. sapiens* (**c**, **d**; $n = 6{,}835$ and 123,915, respectively). 3*n*, 3*n* + 1 and 3*n* + 2 introns are shown in black, red and green, respectively.

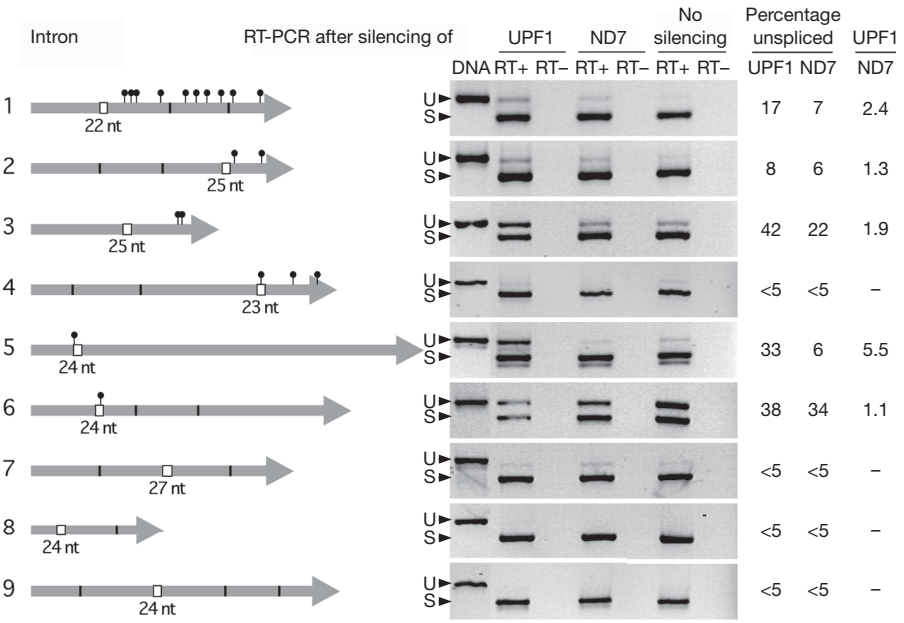

Figure 4 | Accumulation of unspliced mRNAs after *UPF1* knockdown. Tested introns (boxes), their positions within coding sequences (thick grey arrows; see Supplementary Table 3), PTCs introduced by their retention (black pinheads) and exon–exon junctions resulting from the splicing of other introns (vertical black lines) are shown schematically. RT–PCR products from spliced (S) and unspliced (U) mRNAs from wild-type cells (no silencing), or after silencing of *UPF1A* and *UPF1B* genes or of the unrelated *ND7* gene, were resolved on agarose gels (negative of ethidium bromide stain). RT−, control reactions without reverse transcriptase. The fraction of unspliced mRNAs was quantified with ethidium bromide signals.

peculiar in that a PTC will trigger NMD only if it is located at least 50 nt upstream of the last exon–exon junction[5,6]. The retention of the last intron of a gene therefore cannot be detected by NMD in mammals, even if it introduces a PTC. Accordingly, stopless 3*n* introns are not under-represented among the last introns of genes in *H. sapiens*, whereas they are in non-mammalian species (Supplementary Table 2).

It has been proposed that the appearance of genome-wide alternative splicing during the evolution of multicellular organisms was linked to the weakening of strong splice sites of ancestral introns recognized through intron definition[4]. We found that alternative splicing does not occur to any significant degree in *Paramecium*, an organism in which a relatively large intron number is associated with very weak splice sites and with the strong counter-selection of those introns that cannot be detected by NMD. The finding that the latter features are common to various intron-rich eukaryotes suggests that, independently of alternative splicing, it may be more advantageous to rely on NMD surveillance than to evolve a more efficient splicing system. Supporting this view, the rare species that seem to have lost NMD are almost entirely devoid of introns[14].

Finally, we note that the observed bias is also compatible with the controversial proposal, based on studies of the nonsense-associated alternative splicing[15–17] and suppression of splicing[18,19] effects, that the translatability of pre-mRNA sequences can influence splice site choice[20,21]. Although this idea was revived by the finding that a substantial fraction of mammalian NMD events occurs in the nucleus[22,23] and by the controversial possibility of nuclear translation[24–27], it should be emphasized that it does not necessarily imply nuclear translation before splicing. RNA interference can regulate many different steps of gene expression, and introducing a frameshift in a *Paramecium* coding sequence can trigger RNAi[28]. Together with the genetic link that has been uncovered between the RNAi and NMD pathways in *C. elegans*[29] and in *A. thaliana*[30], this raises the theoretical possibility that a translation test in the cytoplasm, which will trigger NMD in many cases of intron retention, couples mRNA degradation with the formation of RNA signal molecules that can feed back to the nucleus to modulate the splicing of homologous pre-mRNAs. Whether NMD simply allows the selection of correctly spliced transcripts or whether it has some more active function in the choice of splice sites, our results suggest that this ancient mechanism may have evolved together with spliceosomal introns and the need to control splicing patterns.

METHODS SUMMARY

Bioinformatic analyses. Intron sets from all species were confirmed by the alignment of cDNAs or ESTs, except for *S. pombe*. Only GT/AG or GC/AG introns shorter than 5,000 nt were considered. A minor fraction of gene models containing in-frame stop codons in the coding sequences were excluded from all data sets. When cDNA sequences revealed alternative splicing, each intron form was counted only once.

***Paramecium* strain, cultivation, and RNAi treatment.** The entirely homozygous strain 51 was grown in a wheatgrass-powder infusion medium bacterized with *Klebsiella pneumoniae* the day before use, and supplemented with 0.8 mg l^{-1} β-sitosterol. RNAi treatment was conducted by the feeding technique: cells were cultured for seven days on the same medium containing ampicillin at 0.1 mg ml^{-1} and bacterized with the HT115 *E. coli* strain, which produces double-stranded RNA from any sequence cloned into plasmid L4440 after induction with isopropyl β-D-thiogalactoside (IPTG). Sequences used for silencing of the *UPF1A*, *UPF1B*, *UPF2*, *ND7* and *ICL7a* genes were segments 1,885–2,289, 1,887–2,285, 1,143–1,546, 870–1266 and 1–580 of the genes (from the ATG), respectively. These genes can be accessed with ParameciumDB (http://paramecium.cgm.cnrs-gif.fr/) under accession numbers GSPATG00034062001, GSPATG00037251001, GSPATG00017015001, GSPATG00002403001 and GSPATG00021610001, respectively.

Northern blot analyses and RT–PCR quantification of unspliced mRNAs. Total RNA was extracted from cells grown on *K. pneumoniae* or the relevant feeding *E. coli* strains with the use of the TRIzol (Invitrogen) procedure, modified by the addition of glass beads. Northern blots, reverse transcription and PCR were performed with standard procedures. In the RT–PCR assay, the small length difference between the spliced and unspliced versions is unlikely to have biased the PCR reaction. Any possible bias would be the same in all samples, so that it would not affect the ratio of unspliced fractions between the *UPF1A*/*UPF1B* and *ND7* silencing conditions.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 September; 21 November 2007.

1. Roy, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Rev. Genet.* **7**, 211–221 (2006).

2. Berget, S. M. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270,** 2411–2414 (1995).
3. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA* **98,** 11193–11198 (2001).
4. Ast, G. How did alternative splicing evolve? *Nature Rev. Genet.* **5,** 773–782 (2004).
5. Conti, E. & Izaurralde, E. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol.* **17,** 316–325 (2005).
6. Maquat, L. E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature Rev. Mol. Cell Biol.* **5,** 89–99 (2004).
7. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA* **100,** 189–192 (2003).
8. Pan, Q. *et al.* Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.* **20,** 153–158 (2006).
9. Russell, C. B., Fraga, D. & Hinrichsen, R. D. Extremely short 20–33 nucleotide introns are the standard length in *Paramecium tetraurelia*. *Nucleic Acids Res.* **22,** 1221–1225 (1994).
10. Aury, J. M. *et al.* Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444,** 171–178 (2006).
11. Romfo, C. M., Alvarez, C. J., van Heeckeren, W. J., Webb, C. J. & Wise, J. A. Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. *Mol. Cell. Biol.* **20,** 7955–7970 (2000).
12. Ishigaki, Y., Li, X., Serin, G. & Maquat, L. E. Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. *Cell* **106,** 607–617 (2001).
13. Galvani, A. & Sperling, L. RNA interference by feeding in *Paramecium*. *Trends Genet.* **18,** 11–12 (2002).
14. Lynch, M. The origins of eukaryotic gene structure. *Mol. Biol. Evol.* **23,** 450–468 (2006).
15. Mohn, F., Buhler, M. & Muhlemann, O. Nonsense-associated alternative splicing of T-cell receptor beta genes: no evidence for frame dependence. *RNA* **11,** 147–156 (2005).
16. Wang, J., Chang, Y. F., Hamilton, J. I. & Wilkinson, M. F. Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay. *Mol. Cell* **10,** 951–957 (2002).
17. Wang, J., Hamilton, J. I., Carter, M. S., Li, S. & Wilkinson, M. F. Alternatively spliced TCR mRNA induced by disruption of reading frame. *Science* **297,** 108–110 (2002).
18. Miriami, E., Sperling, R., Sperling, J. & Motro, U. Regulation of splicing: the importance of being translatable. *RNA* **10,** 1–4 (2004).
19. Wachtel, C., Li, B., Sperling, J. & Sperling, R. Stop codon-mediated suppression of splicing is a novel nuclear scanning mechanism not affected by elements of protein synthesis and NMD. *RNA* **10,** 1740–1750 (2004).
20. Maquat, L. E. NASty effects on fibrillin pre-mRNA splicing: another case of ESE does it, but proposals for translation-dependent splice site choice live on. *Genes Dev.* **16,** 1743–1753 (2002).
21. Wilkinson, M. F. & Shyu, A. B. RNA surveillance by nuclear scanning? *Nature Cell Biol.* **4,** E144–E147 (2002).
22. Buhler, M., Wilkinson, M. F. & Muhlemann, O. Intranuclear degradation of nonsense codon-containing mRNA. *EMBO Rep.* **3,** 646–651 (2002).
23. Iborra, F. J., Escargueil, A. E., Kwek, K. Y., Akoulitchev, A. & Cook, P. R. Molecular cross-talk between the transcription, translation, and nonsense-mediated decay machineries. *J. Cell Sci.* **117,** 899–906 (2004).
24. Brogna, S., Sato, T. A. & Rosbash, M. Ribosome components are associated with sites of transcription. *Mol. Cell* **10,** 93–104 (2002).
25. Dahlberg, J. E. & Lund, E. Does protein synthesis occur in the nucleus? *Curr. Opin. Cell Biol.* **16,** 335–338 (2004).
26. Iborra, F. J., Jackson, D. A. & Cook, P. R. The case for nuclear translation. *J. Cell Sci.* **117,** 5713–5720 (2004).
27. Nathanson, L., Xia, T. & Deutscher, M. P. Nuclear protein synthesis: a re-evaluation. *RNA* **9,** 9–13 (2003).
28. Garnier, O., Serrano, V., Duharcourt, S. & Meyer, E. RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*. *Mol. Cell. Biol.* **24,** 7370–7379 (2004).
29. Domeier, M. E. *et al.* A link between RNA interference and nonsense-mediated decay in *Caenorhabditis elegans*. *Science* **289,** 1928–1931 (2000).
30. Arciga-Reyes, L., Wootton, L., Kieffer, M. & Davies, B. UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in *Arabidopsis*. *Plant J.* **47,** 480–489 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank V. Wood, P. Mooney and A. Tivey for providing gff files for *S. pombe* data, and D. Gogendeau and J. Beisson for the gift of the *ICL7a* feeding plasmid. This work was funded by the CNRS and by the Agence Nationale de la Recherche. K.B. was supported by a postdoctoral contract from the CNRS. Experimental work was supported by grants from the Ministère de la Recherche and the Association pour la Recherche sur le Cancer.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.M. (emeyer@biologie.ens.fr).

METHODS

***P. tetraurelia* data set.** The EST-confirmed set was constituted by selecting from the published annotation[10] the 6,513 gene models for which each of the predicted introns was confirmed by the alignment of at least one EST. EST alignment was as described in ref. 10. Then, 376 gene models were excluded because ESTs revealed introns that were not predicted by the annotation, a possible source of error in the identification of the reading frame. Exclusion of the 376 problematic gene models did not significantly alter the size distribution, because $3n$ introns are also under-represented in these genes (21.6% of the total, including introns that had been overlooked in the annotation, in contrast with 40.3% and 38.0% for $3n + 1$ and $3n + 2$ introns, respectively).

***H. sapiens* data set.** The genome sequence and the Known Genes set[31] were downloaded from the UCSC (University of California Santa Cruz) genome browser (http://genome.ucsc.edu)[32,33]. The version of the genome sequence is NCBI hg17 (May 2004). The Known Genes set is based on data from UniProt (SWISS-PROT and TrEMBL) and mRNA data from the NCBI (National Center for Biotechnology Information) reference sequences collection (RefSeq)[34] and GenBank. Observations were confirmed with genes manually annotated from the Vega consortium on chromosomes 6, 7, 9, 10, 13, 14, 20, 22 and X (http://vega.sanger.ac.uk)[35] (data not shown).

***A. thaliana* data set.** We used the genome sequence and annotation from the TIGR5 release. The sequence (filename ATH1_chr_all.5con.gz) was downloaded from ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES. The gene annotations were retrieved with the biomart web server (www.biomart.org) at www.gramene.org. Predicted introns were confirmed by alignment with cDNA sequences. In all, 98,490 mRNA sequences from NCBI (excluding ESTs) were aligned with the genome sequence with blat[36]. For each mRNA sequence we selected the best genomic locus on the basis of blat scores. Alignments were filtered by selecting those with a score greater than 80% of the highest and greater than 50. Each mRNA was then realigned with the corresponding genomic region with est2genome[37]. The cDNA-confirmed set contained 21,233 gene models from the TIGR5 annotation, containing 110,629 introns, all confirmed by the alignment of at least one mRNA (same splice sites). A total of 386 gene models were excluded because cDNAs revealed introns that were not annotated (395 introns). The final set contained 20,847 genes and 108,783 introns.

***D. melanogaster* data set.** We used release 4 (April 2004) of the genome assembly distributed by the UCSC genome browser, and the FlyBase annotation (release 4.2, September 2005). We established that 99.7% of intron annotations are validated by the alignment of at least one mRNA from Refseq[34].

***C. elegans* data set.** We used the March 2004 genome assembly distributed by the UCSC genome browser, which is based on sequence version WS120 deposited into WormBase (www.wormbase.org) as of 1 March 2004, and the WormBase gene annotation. The WormBase genes correspond to gene predictions from the WormBase WS120 files downloaded from the Sanger Institute FTP site (ftp://ftp.sanger.ac.uk/pub/wormbase/FROZEN_RELEASES/WS120/CHROMOSOMES/).

***S. pombe* data set.** Genome assembly and gene annotations were obtained from the Sanger Centre (http://www.sanger.ac.uk/Projects/S_pombe/). The set of EST- or cDNA-confirmed introns was built by extracting intron annotations having the tag 'confirmed'.

General treatment of data sets. Only GT/AG or GC/AG introns shorter than 5,000 nt were considered in all species. Gene models containing in-frame stop codons in the coding sequences were excluded from all data sets where they occurred: *H. sapiens*, 820 models; *D. melanogaster*, 42 models; *C. elegans*, 12 models; *S. pombe*, 34 models. When cDNA sequences from these species revealed alternative splicing, each intron form was counted only once.

Reference genes. As negative controls for the RNAi experiments, we used two genes that are not involved in NMD: ND7 and ICL7a. ND7 is involved in the control of exocytosis[38]. ICL7a encodes a cytoskeletal protein[39].

RT–PCR quantification of unspliced mRNAs. Total RNA was extracted from cells grown on *K. pneumoniae* or the relevant feeding *Escherichia coli* strains with the TRIzol (Invitrogen) procedure, modified by the addition of glass beads. mRNA reverse transcription was performed with the SuperScript II kit (Invitrogen) and the anchor-oligo(dT) primer 5′-GCTCGGACCGTGGCTA-GCATTAGTGAGTTTTTTTTTTTTTTTT-3′. After alkaline lysis of RNA and removal of the oligo(dT) primer with Microcon YM-100 centrifugal devices (Millipore), short segments containing the introns of interest were amplified by PCR with the primers listed in Supplementary Table 3, either directly from the reverse transcriptase products or, if necessary, after a first amplification with a primer corresponding to the anchor sequence and the upstream primer. For those samples in which both bands were clearly visible, the fraction of unspliced mRNAs was calculated by quantification of the ethidium bromide signal from each of the two bands, using unsaturated exposures of the agarose gels shown in Fig. 4 and the TINA software. Quantification of RT–PCR products by extension of ^{32}P-labelled primers (Supplementary Fig. 6) was performed with Sequencing Grade *Taq* DNA polymerase (Promega). Radioactive signals were quantified with the ImageGauge software.

31. Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22,** 1036–1046 (2006).
32. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31,** 51–54 (2003).
33. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
34. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33,** D501–D504 (2005).
35. Ashurst, J. L. *et al.* The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* **33,** D459–D465 (2005).
36. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
37. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13,** 477–478 (1997).
38. Skouri, F. & Cohen, J. Genetic approach to regulated exocytosis using functional complementation in *Paramecium*: identification of the ND7 gene required for membrane fusion. *Mol. Biol. Cell* **8,** 1063–1071 (1997).
39. Gogendeau, D. *et al.* Functional diversification of centrins and cell morphological complexity. *J. Cell Sci.* **121,** 65–74 (2007).