Supporting Figures

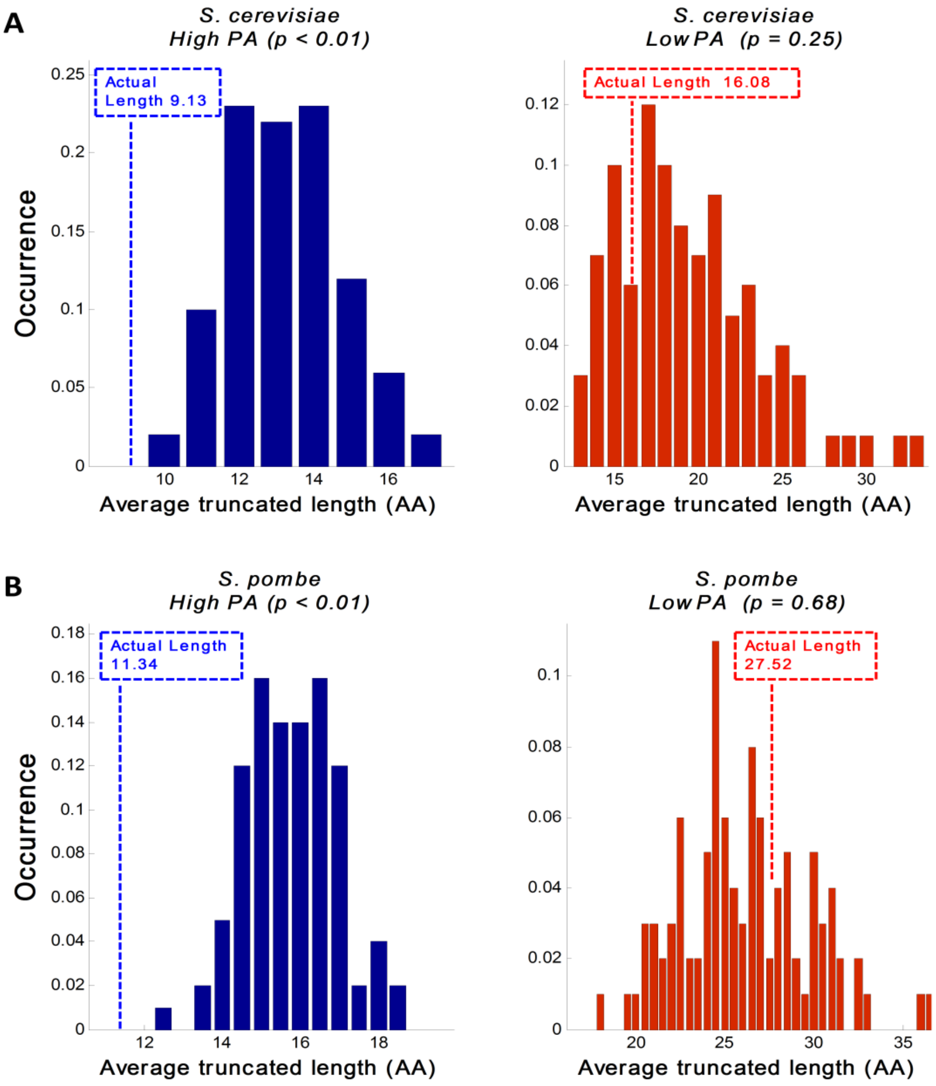


Figure S1: First intronic STOP codon analysis in *S. cerevisiae* and *S. pombe*, while separately permuting the first and second fragments of the introns. For highly and lowly expressed genes in *S. cerevisiae* we considered the first segment as 50nt after the donor consensus sequence; for *S. pombe* we considered the first segment to be 25nt, since in this organism introns are relatively very short (with median intron length of 56nt; see **Table S5**); results were similar to the original ones. A) Intronome analysis of highly *vs.* lowly expressed genes in *S. cerevisiae* demonstrates evidence of selection in highly expressed genes, but no significant selection in lowly expressed genes (empirical p<0.01 and empirical p=0.25; left and right, respectively; actual length displayed by broken line). B) Intronome analysis of highly *vs.* lowly expressed genes in *S. pombe* demonstrates evidence of selection in highly expressed genes but no significant selection in lowly expressed genes (empirical p<0.01 and empirical p=0.68, respectively; actual length displayed by broken line).

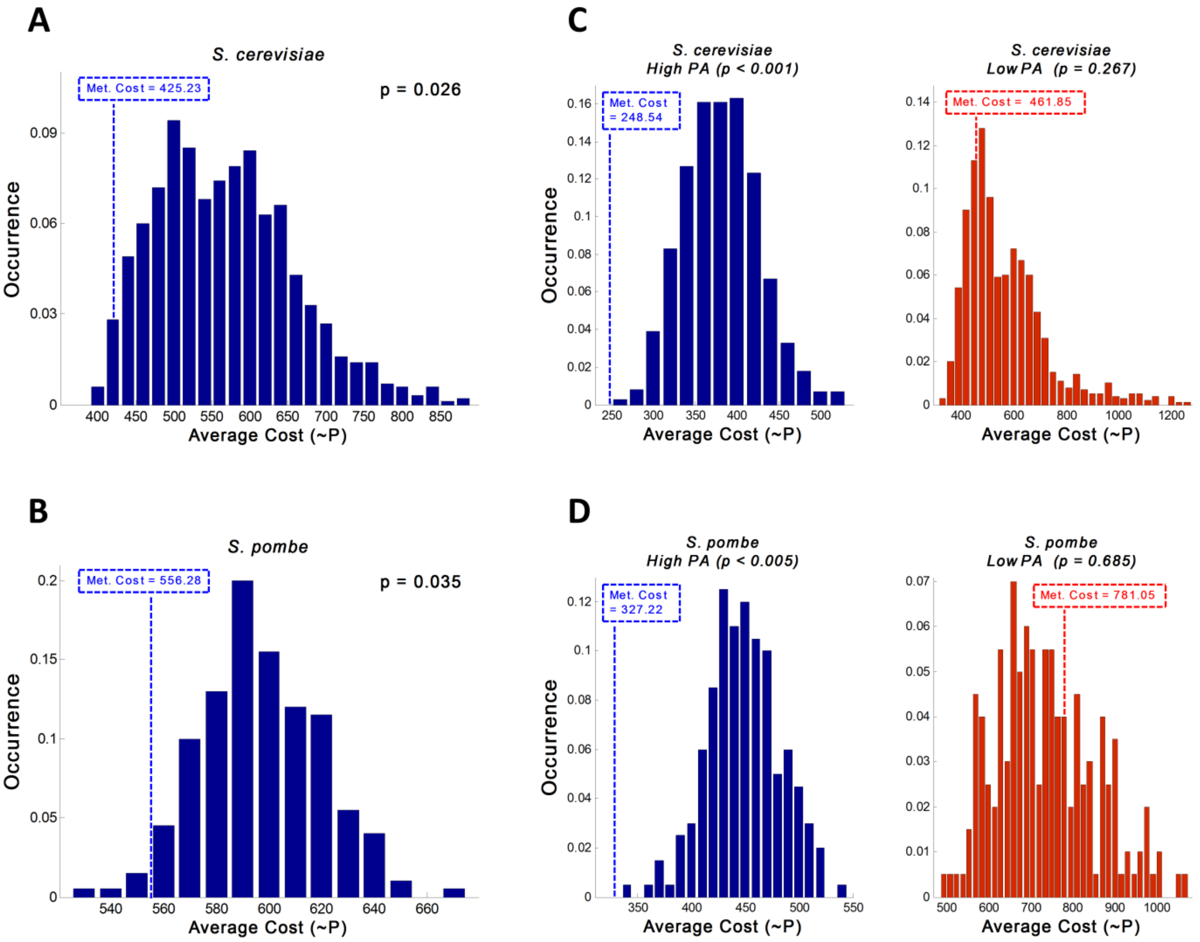


Figure S2: Metabolic costs analysis of the First intronic STOP codon in *S. cerevisiae* and *S. pombe*. Analysis of the sum of costs when translating the truncated peptides up to the STOP codon, over the entire transcriptome and in comparison to the randomized models: A-B) The average actual costs in *S. cerevisiae* (A) and in *S. pombe* (B) are 25% and 7% lower than is expected by the randomized models (425.2 *vs.* 567.5 and 556.3 *vs.* 596.8; *S. cerevisiae* and *S. pombe*, respectively). Likewise, distribution analysis suggests that there is a preference for lower metabolic costs in the actual intronome (empirical p=2.6·10^{-2} and empirical p=3.6·10^{-2} for *S. cerevisiae* and *S. pombe* respectively; actual length displayed in broken line). C-D) Analysis of highly *vs.* lowly expressed genes in *S. cerevisiae* (C) and in *S. pombe* (D) suggests that the signal is stronger and significant only for highly expressed genes (highly expressed: empirical p<1·10^{-3} and empirical p<5·10^{-3}; lowly expressed: empirical p=0.267 and empirical p=0.685, *S. cerevisiae* and *S. pombe*, respectively; actual length displayed in broken line). The cost units are in (~P) which is an estimation of the energy of the activated ATP phosphate.

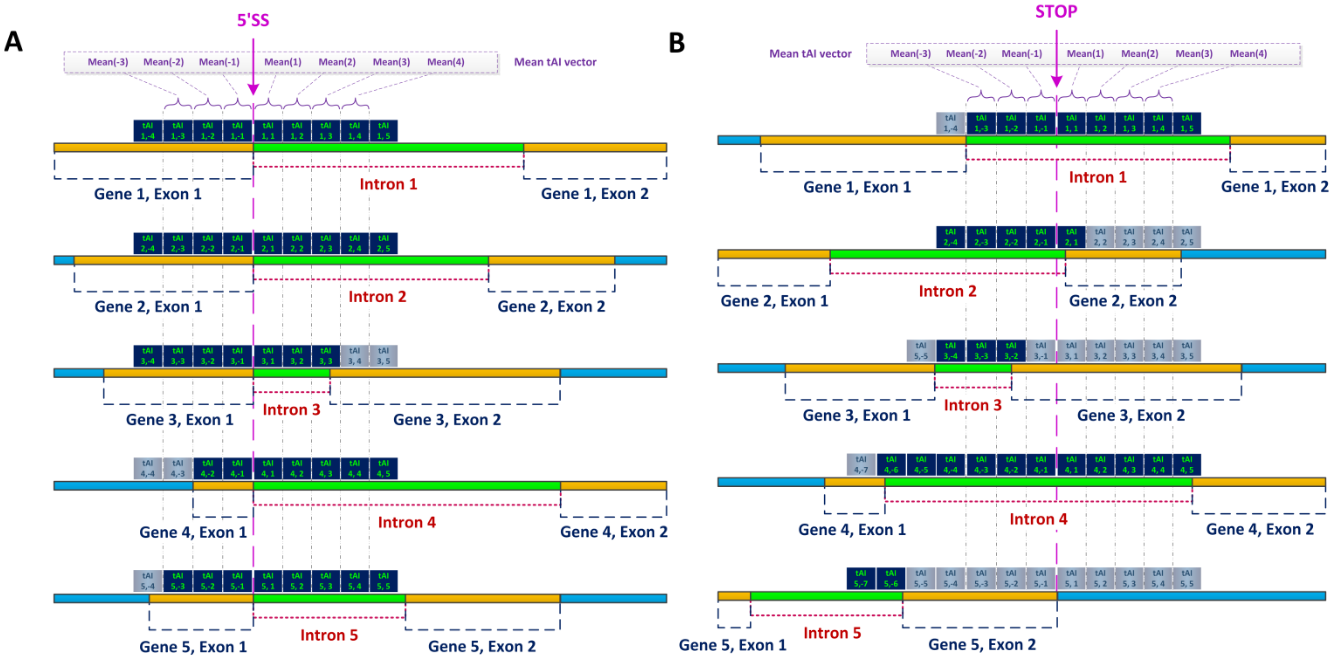


Figure S3: Codon-usage bias scheme illustrating the process of mean TDR/tAI profile generation. All intron-containing genes were aligned according to their donor site (5’SS, A), and according to their first intronic STOP codon (B). The TDR/tAI index was calculated per codon and then averaged over all genes in that position to form Mean vectors. Untranslated regions such as 5’UTRs and 3’UTRs were excluded, as well as downstream exons in 5’SS aligned profiles, and all exons in STOP aligned profiles.

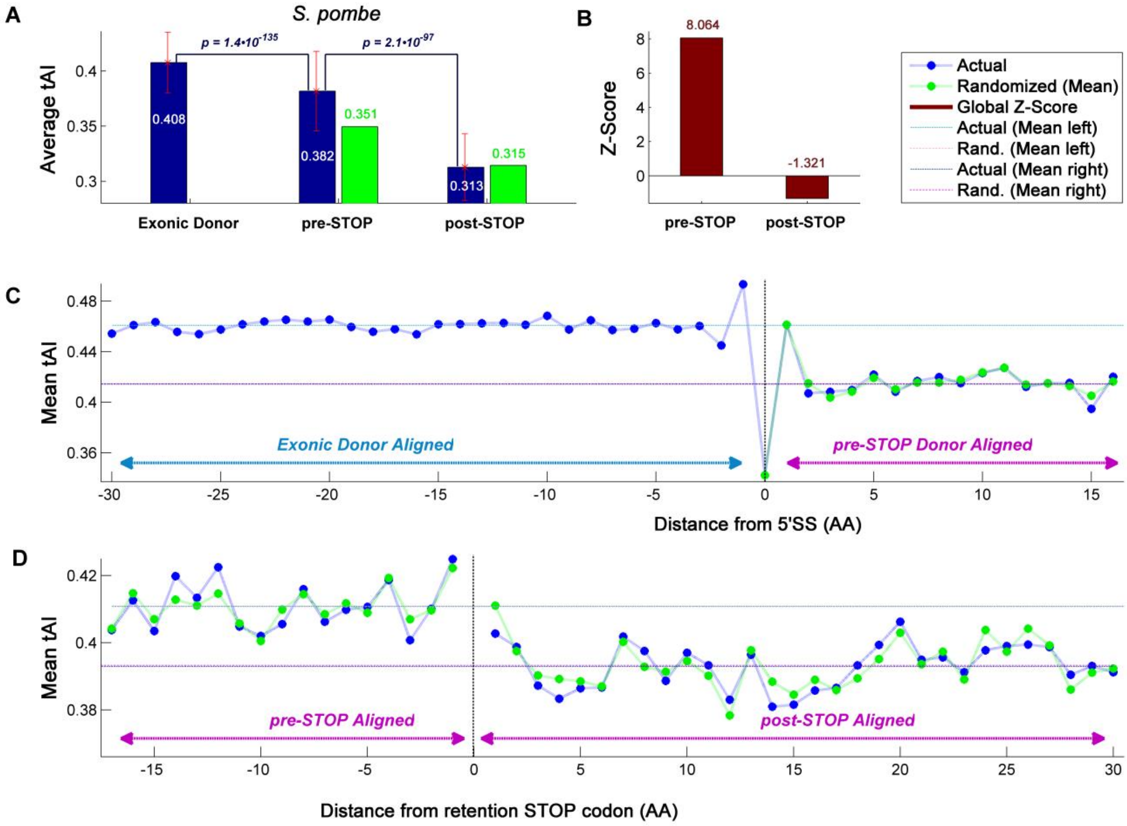


Figure S4: Codon-usage bias adaptation to the tRNA pool profiles for the *S. pombe* intronome. The profiles show that the average tRNA adaptation index (tAI) values upstream from the first intronic STOP codon is higher than the average of the tAI values downstream from it, supporting the conjecture that the beginning of introns undergo evolutionary selection for higher TE. A) Average actual and random tAI (blue and green, respectively) aligned to the beginning of the 5'SS (Exonic donor) and the first intronic STOP (pre-STOP, post-STOP) locations; exonic domain randomization is not shown. B) Standard normalization Z-score values (red) indicate global selection level. C) Mean tAI profile aligned to the beginning of the 5'SS; as expected, TDR values downstream from the 5' end of the exon/intron boundary (right side of the 5'SS) are lower than upstream from it (left side of the 5'SS; $p=1.43\cdot10^{-135}$, Wilcoxon signed-rank test that is a paired test); In addition, mean tAI downstream from the 5' end of the exon/intron boundary (blue) is significantly higher than in the case of the randomized models (green; empirical $p<5\cdot10^{-3}$). D) Mean tAI profile aligned to the beginning of the first Intronic STOP codon; tAI before the first intronic STOP codon (blue) is higher than in the post-STOP domain ($p=2.07\cdot10^{-97}$, Wilcoxon signed-rank test); genes with STOP codon positioned in the downstream exon and locations with less than 15% of the intronome were ignored.

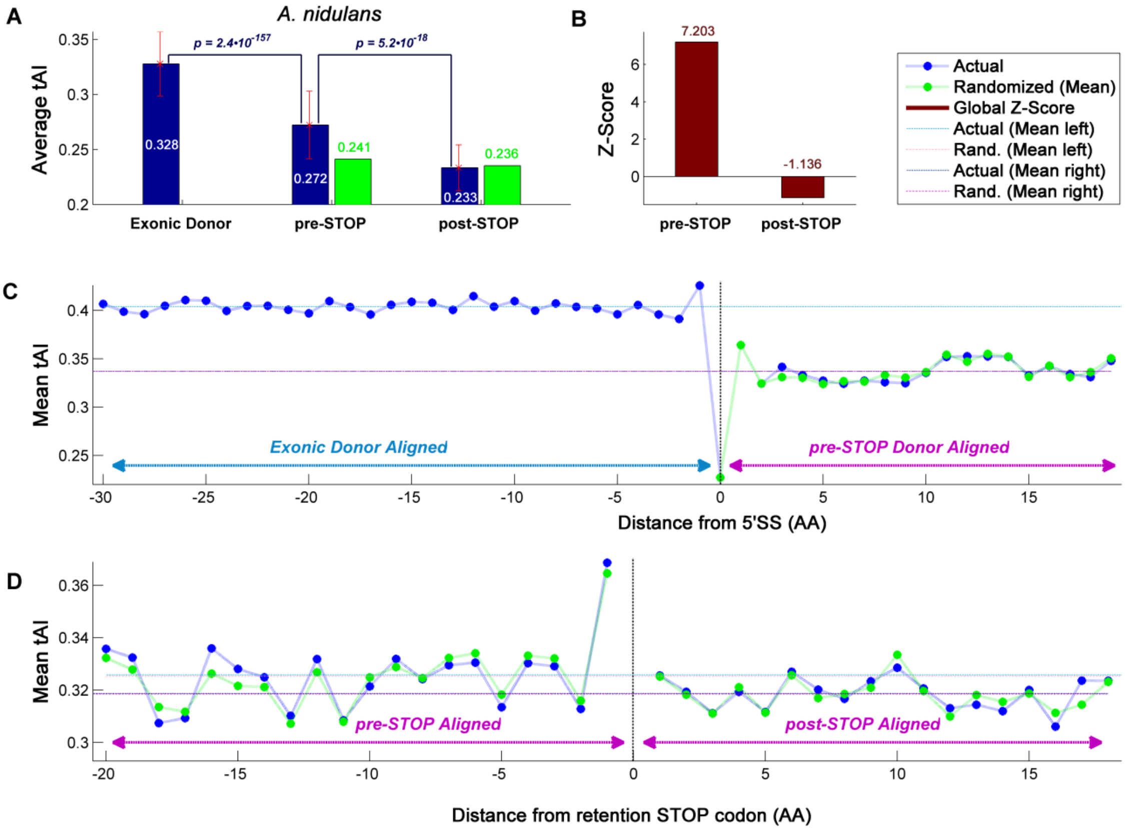


Figure S5: Codon-usage bias adaptation to the tRNA pool profiles for the *A. nidulans* intronome. The profiles show that the average tAI values upstream from the first intronic STOP codon are higher than the average of the tAI values downstream from it, supporting the conjecture that the beginnings of introns undergo evolutionary selection for higher TE. A) Average actual and random tAI (blue and green, respectively) aligned to the beginning of the 5’SS (Exonic donor) and the first intronic STOP (pre-STOP, post-STOP) locations; exonic domain randomization is not shown. B) Standard normalization Z-score values (red) indicate global selection level. C) Mean tAI profile aligned to the beginning of the 5’SS; as expected, TDR values downstream from the 5’ end of the exon/intron boundary (right side of the 5'SS) are lower than upstream from it (left side of the 5'SS; $p=2.41\cdot10^{-157}$, Wilcoxon signed-rank test that is a paired test); In addition, tAI downstream from the 5’ end of the exon/intron boundary (blue) is significantly higher than in the case of the randomized models (green; empirical $p<5\cdot10^{-3}$). D) Mean tAI profile aligned to the beginning of the first Intronic STOP codon; tAI before the first intronic STOP codon (blue) is higher than in the post-STOP domain ($p=5.24\cdot10^{-18}$, Wilcoxon signed-rank test); genes with STOP codon positioned in the downstream exon and locations with less than 15% of the intronome were ignored.

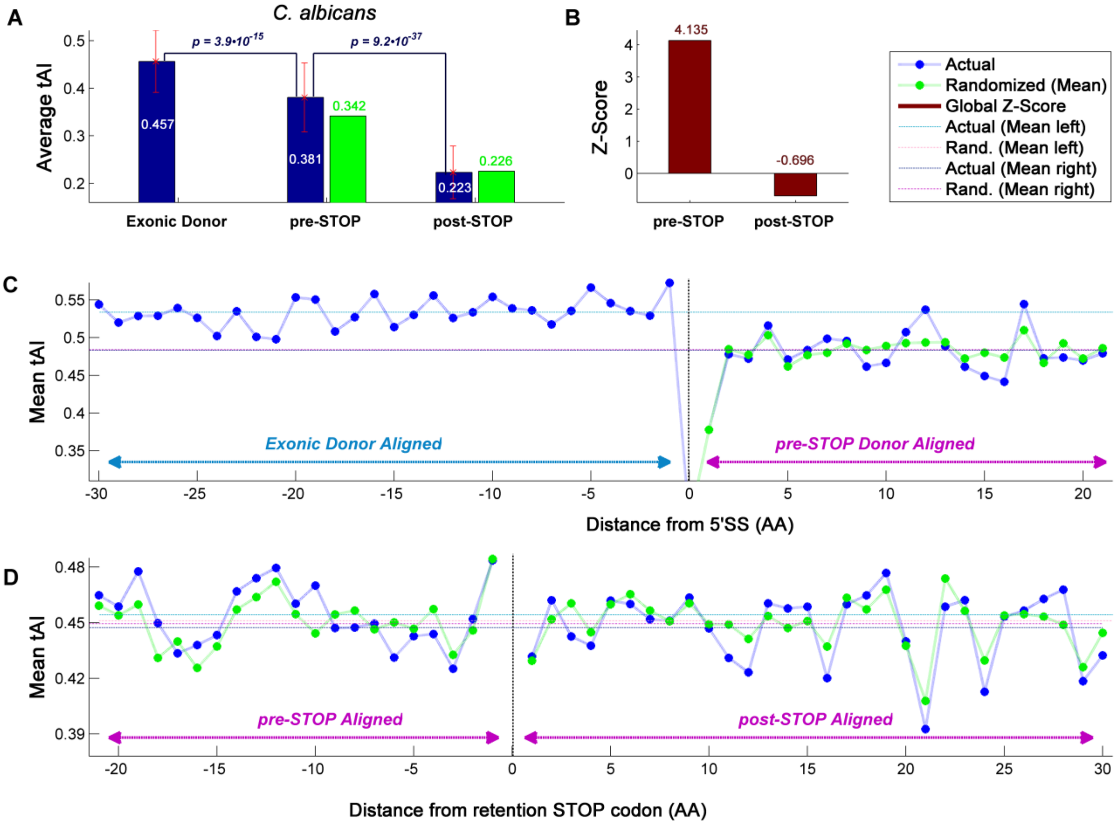


Figure S6: Codon-usage bias adaptation to the tRNA pool analysis for the *C. albicans* intronome. The profiles show that the average tAI values upstream from the first intronic STOP codon are higher than the average of the tAI values downstream from it, supporting the conjecture that the beginning of introns undergo evolutionary selection for higher TE. A) Average actual and random tAI (blue and green, respectively) aligned to the beginning of the 5’SS (Exonic donor) and the first intronic STOP (pre-STOP, post-STOP) locations; exonic domain randomization is not shown. B) Standard normalization Z-score values (red) indicate global selection level. C) Mean tAI profile aligned to the beginning of the 5’SS; as expected, TDR values downstream from the 5’ end of the exon/intron boundary (right side of the 5'SS) are lower than upstream from it (left side of the 5'SS; $p=1.07\cdot10^{-9}$, Wilcoxon signed-rank test that is a paired test); In addition, tAI downstream from the 5’ end of the exon/intron boundary (blue) is significantly higher than in the case of the randomized models (green; empirical $p<1\cdot10^{-3}$). D) Mean tAI profile aligned to the beginning of the first Intronic STOP codon; tAI before the first intronic STOP codon (blue) is higher than in the post-STOP domain ($p=9.21\cdot10^{-37}$, Wilcoxon signed-rank test); genes with STOP codon positioned in the downstream exon and locations with less than 15% of the intronome were ignored.

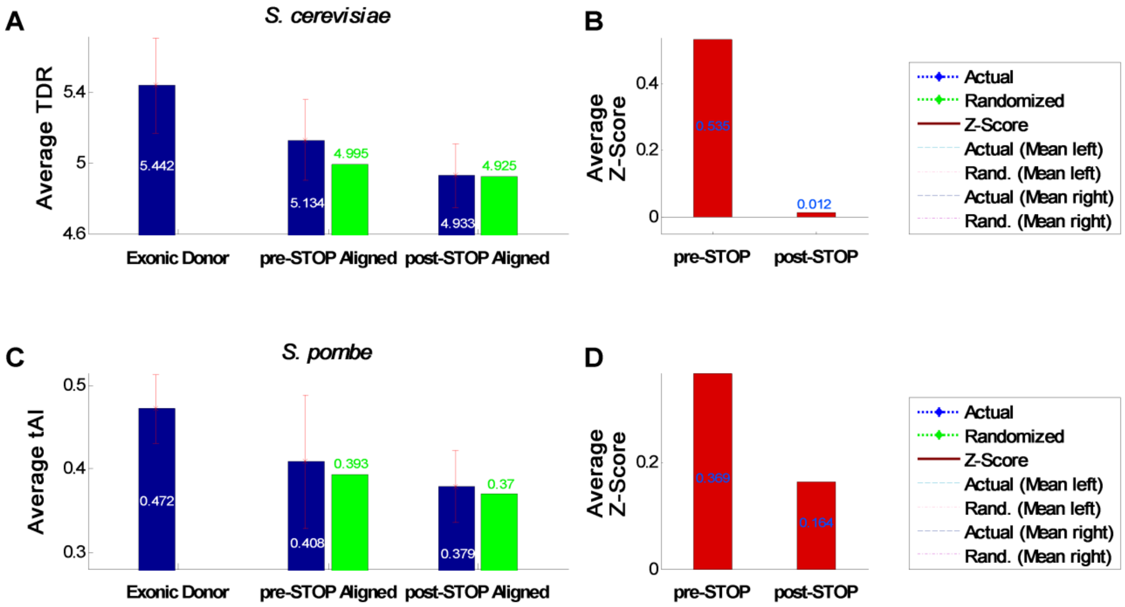


Figure S7: Codon-usage bias profiles for the intronome of *S. cerevisiae* and *S. pombe*, while permuting the first and second fragments of the introns separately. For *S. cerevisiae* we considered the first segment as 50nt after the donor consensus sequence; for *S. pombe* we considered the first segment to be 25nt, since in this organism introns are relatively very short (median intron length of 56nt); results were similar to the original ones. A) Average actual and random TDR (blue and green, respectively) in *S. cerevisiae* aligned to the beginning of the 5'SS (Exonic donor, Intronic donor) and the first intronic STOP (pre-STOP, post-STOP) locations; exonic domain randomization is not shown. B) Standard normalization Z-score values (red) indicate the selection level. C) Average actual and random tAI (blue and green, respectively) in *S. pombe* aligned to the beginning of the 5'SS (Exonic donor, Intronic donor) and the first intronic STOP (pre-STOP, post-STOP) locations; exonic domain randomization is not shown. D) Standard normalization Z-score values (red) indicate the selection level.

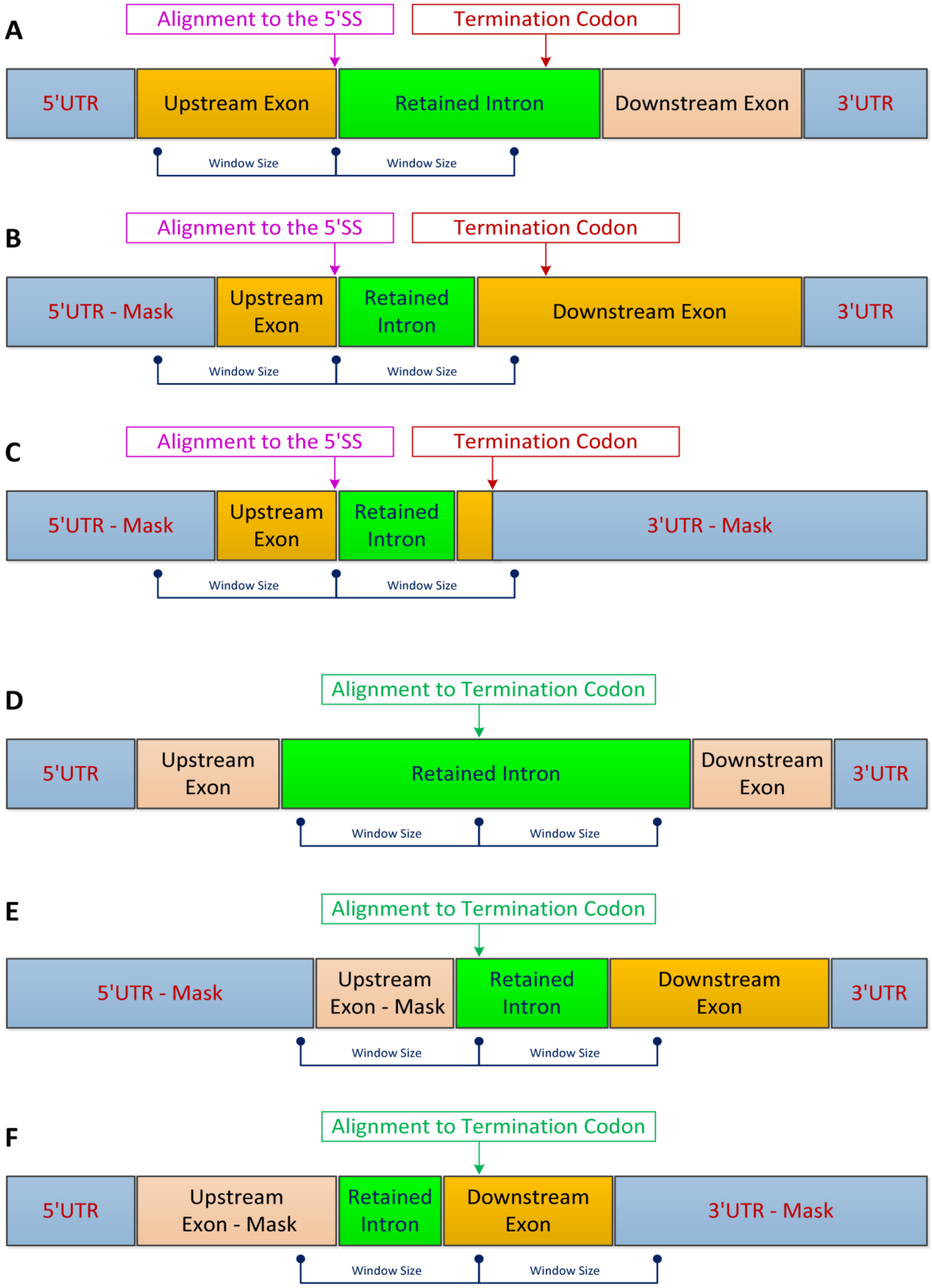


Figure S8: TDR/tAI profiles generation scheme showing some possible exon-intron boundary and termination point cases. TDR/tAI index was calculated per codon and all genes were aligned around the 5’SS (A-C) and premature translation-termination codon (PTC, D-F). As illustrated, relevant exonic information is added in 5'SS alignment, while UTRs are ignored. In the PTC alignment case, upstream exons and UTRs are ignored.

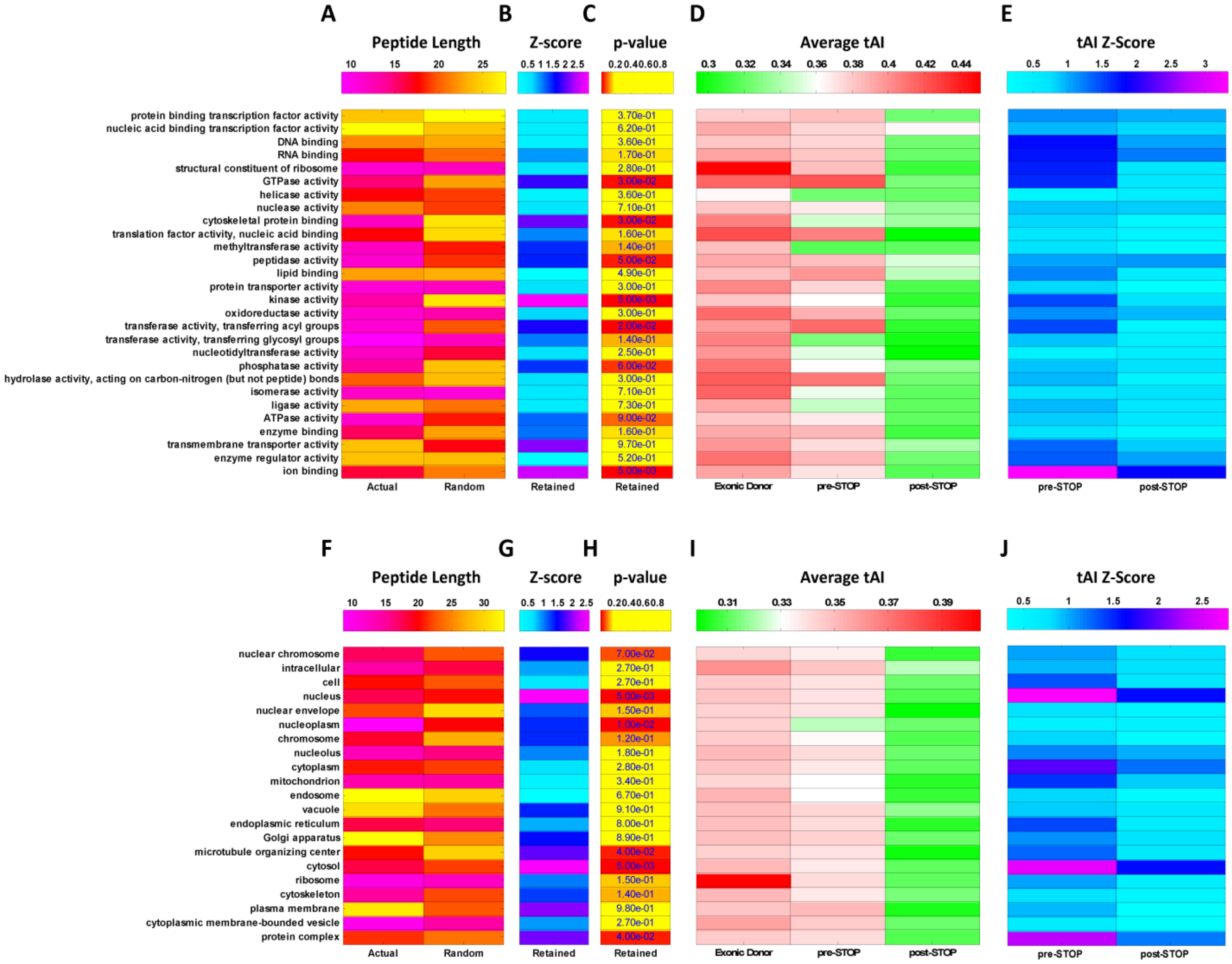


Figure S9: GO terms analysis in *S. pombe* reveals that retained introns translation preference is not function-specific. A-C) Summary of retained peptide length (A) and Z-score corresponding to peptide length relatively to randomized models (B) for molecular function demonstrates that in 79% of the gene groups there is preference for STOP codons closer to the 5'SS. C) Statistical analysis indicates that the observed signal is noteworthy in 22% of the examined gene groups (empirical p<0.1 for all cases; see Methods). D-E) Summary of average tAI (D) and Z-score corresponding to tAI levels relatively to randomized models (E; in absolute values) profiles in the pre-mRNA domains for molecular function demonstrates that in various gene groups there is preference for higher tAI values at the beginning of introns in comparison to the region downstream from the first intronic STOP codon; statistical analysis indicates that the observed signal is significant in 61% of the examined gene groups (p<0.05 for all cases, Wilcoxon rank-sum test; see Methods). F-H) Summary of retained peptide length (F) and Z-score corresponding to peptide length relatively to randomized models (G) for cellular component demonstrates that in 76% of the gene groups there is preference for STOP codons closer to the 5'SS. H) Statistical analysis indicates that the observed signal is noteworthy in 22% of the examined gene groups

(empirical $p<0.1$ for all cases). I-J) Summary of average tAI (I) and Z-score corresponding to tAI levels relatively to randomized models (J; in absolute values) profiles in the pre-mRNA domains for cellular component demonstrates that in various gene groups there is preference for higher tAI values at the beginning of introns in comparison to the region downstream from the first intronic STOP codon; statistical analysis indicates that the observed signal is significant in 85% of the examined gene groups ($p<0.05$ for all cases, Wilcoxon rank-sum test). Thus, the reported signal is universal and not related to any one functional gene group.

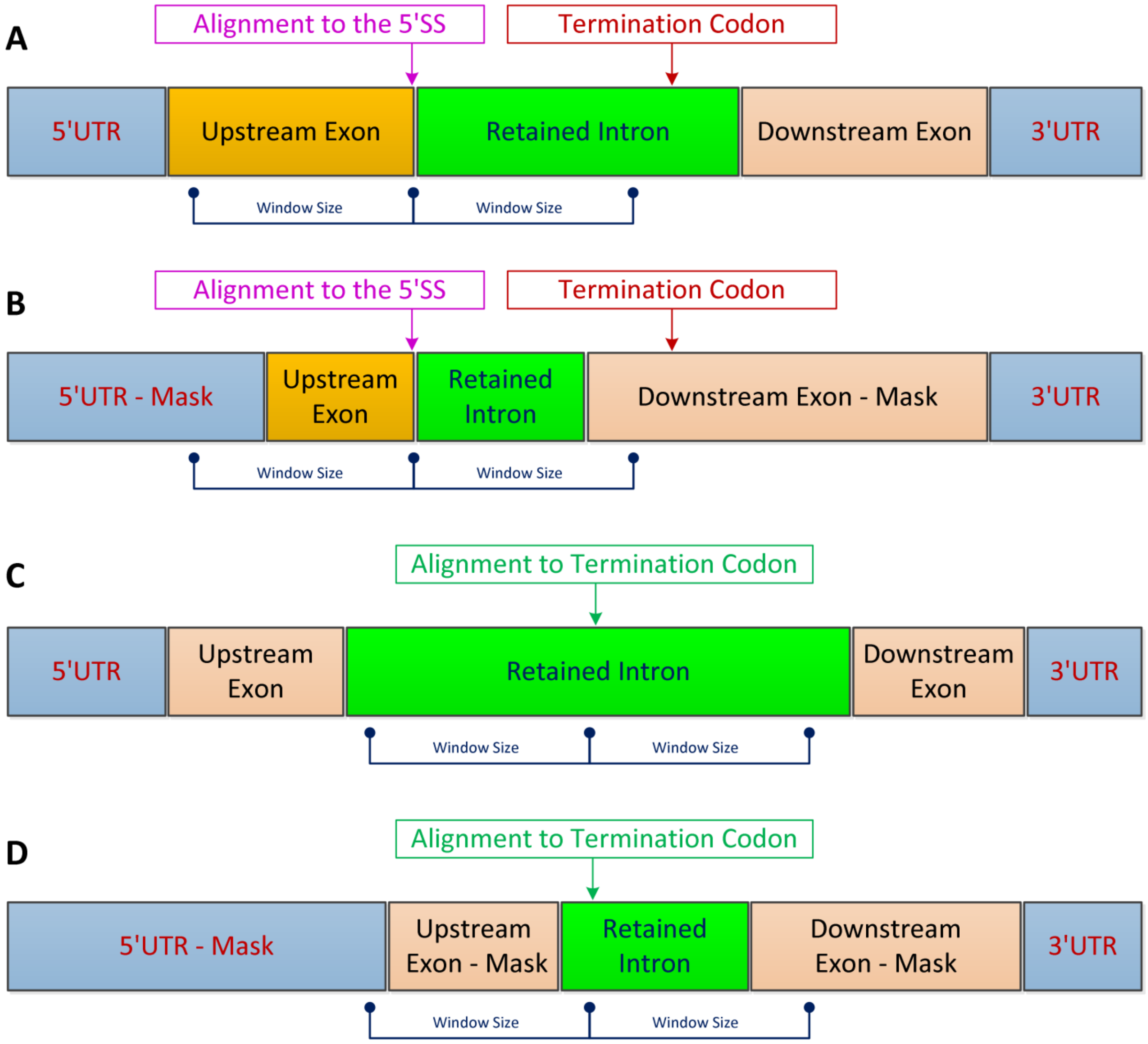


Figure S10: Ribosomal profiles (RP) generation scheme showing some possible exon-intron boundary and termination point cases. All intron-containing genes were aligned in nucleotide resolution around the 5’SS (A-B) and premature translation-termination codon location (PTC, C-D). As illustrated, relevant exonic information is added in the 5'SS alignment, while UTRs are ignored. In the PTC alignment case, both exons and UTRs are ignored. Window size is 50nt.

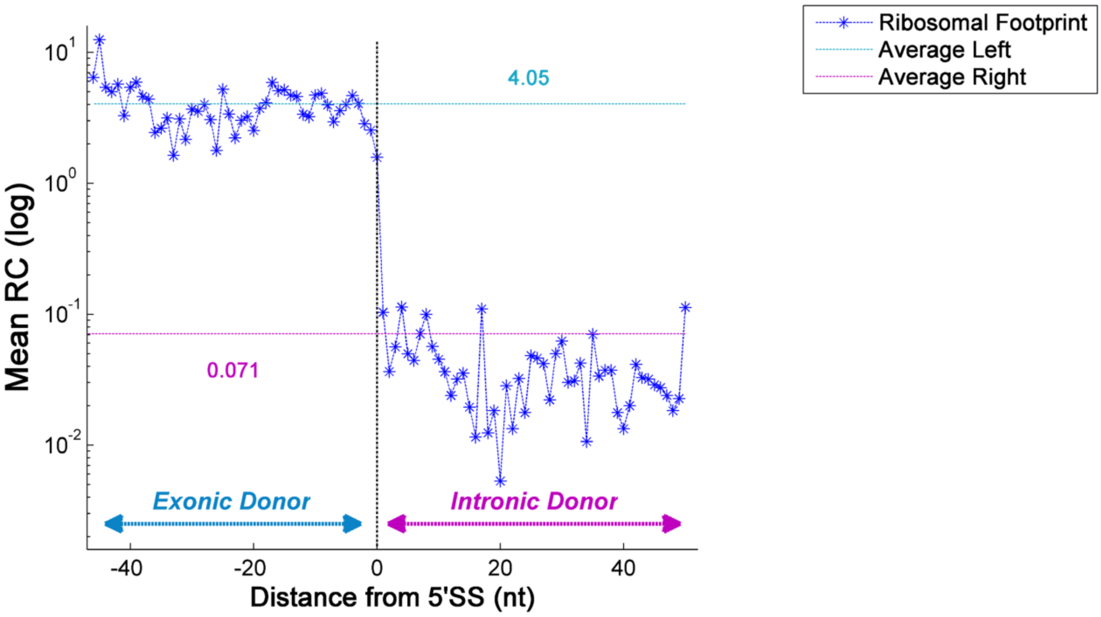


Figure S11: RP analysis of the beginning of the introns in *S. cerevisiae*. As expected, the mean RC values upstream from the Exon/Intron boundary (4.05; Exonic side, left of the 5'SS) are significantly higher than downstream mean RC values ($7.08 \cdot 10^{-2}$; Intronic side, right of the 5'SS; $p=2.93 \cdot 10^{-39}$, Wilcoxon signed-rank test that is a paired test); RC A-site offset is 15nt.

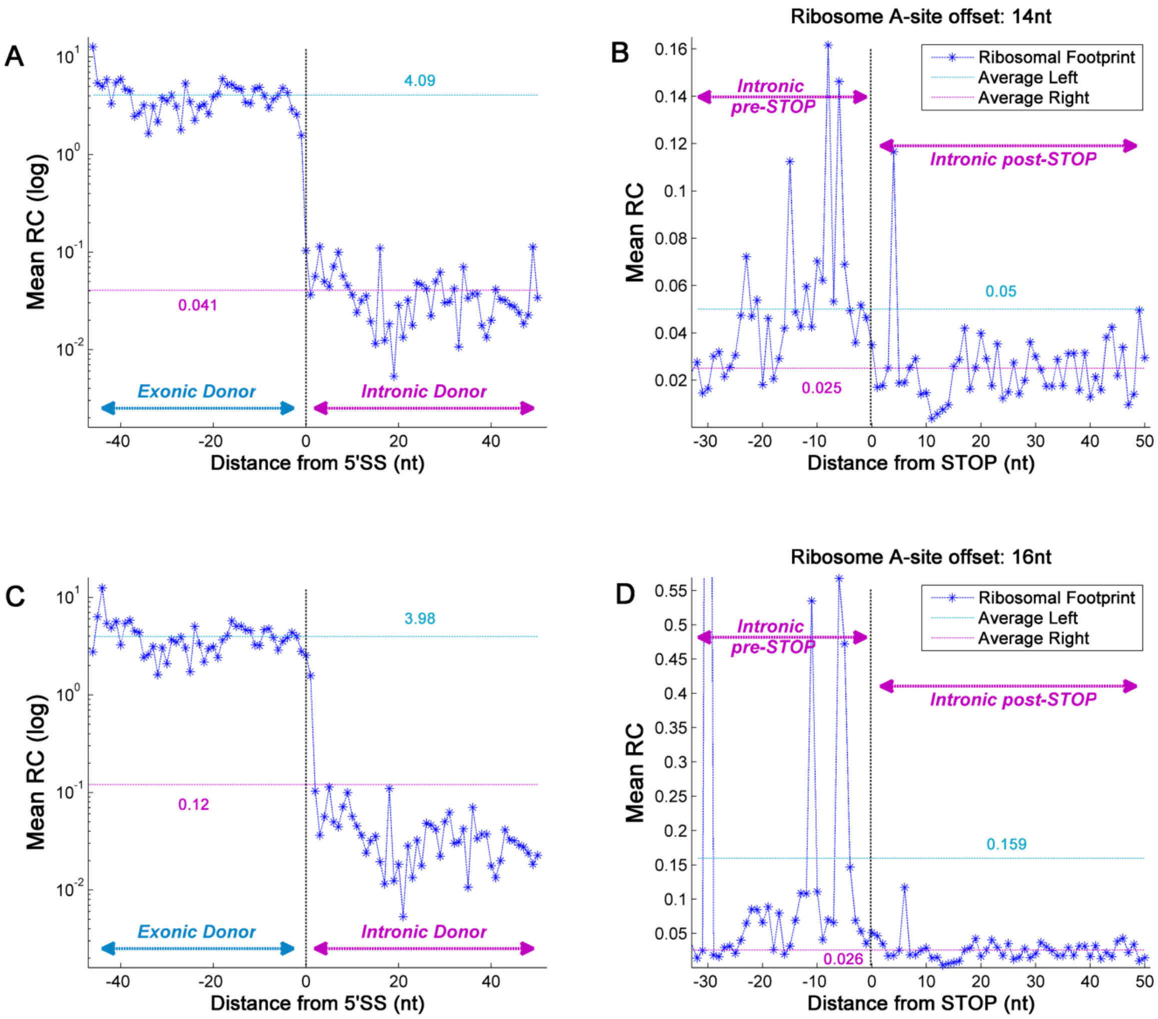


Figure S12: Ribosomal profiling (RP) analysis of *S. cerevisiae*. A) The mean read count (RC) values upstream from the Exon/Intron boundary (Exonic side, left of the 5'SS) are two orders of magnitude higher than downstream mean RC values (Intronic side, right of the 5’SS; p=1.55·10^{-40}, Wilcoxon signed-rank test that is a paired test). B) Mean RC values upstream from the first intronic STOP codon (pre-STOP side, left of the Termination Point) show higher values compared with downstream mean RC values (post-STOP side, right of the Termination Point; 5·10^{-2} before *vs.* 2.51·10^{-2} after the first intronic STOP codon; p=8.17·10^{-7}, Wilcoxon signed-rank test); RC A-site offset is 14nt. C) The mean RC values upstream from the Exon/Intron boundary (Exonic side, left of the 5'SS) are one and a half orders of magnitude higher than downstream mean RC values (Intronic side, right of the 5’SS; p=1.33·10^{-37}, Wilcoxon signed-rank test). D) Mean RC values upstream from the first intronic STOP codon (pre-STOP side, left of the Termination Point) show higher values compared with downstream mean RC values (post-STOP side, right of the Termination Point; 1.59·10^{-1} before *vs.* 2.56·10^{-2} after the first intronic STOP codon; p=1.16·10^{-10}, Wilcoxon signed-rank test); RC A-site offset is 16nt.

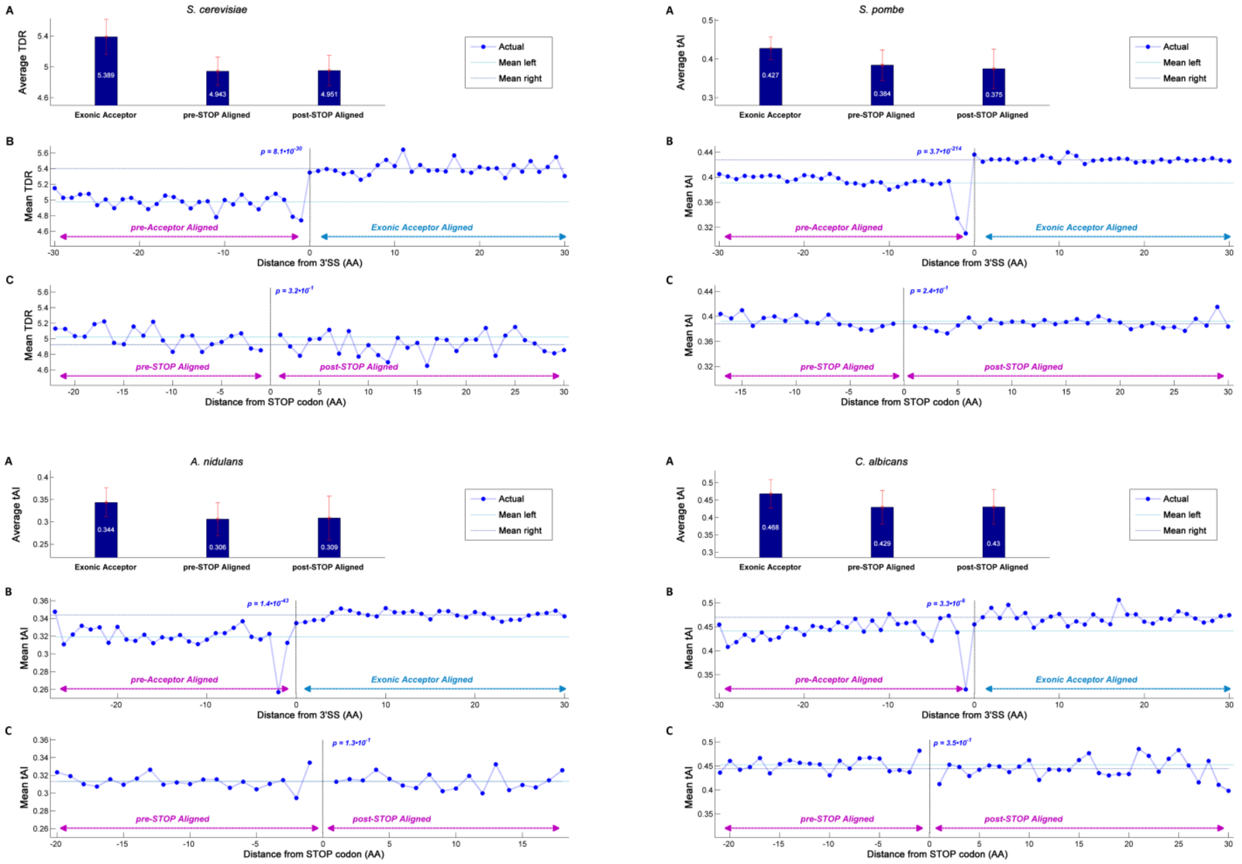


Figure S13: Codon-usage bias analysis for all studied organisms, alignment to the acceptor site. The profiles show that the average TDR/tAI values upstream and downstream from the last intronic STOP codon (*i.e.*, closer to the acceptor site) are similar to the average of the TDR/tAI values downstream from it. A) Average actual and random TDR (S*. cerevisiae*) and tAI (*S. pombe*, *A. nidulans*, and *C. albicans*) aligned to the beginning of the 3'SS (Intronic acceptor, Exonic acceptor) and the last intronic STOP (pre-STOP, post-STOP) locations. B) Mean TDR/tAI profiles aligned to the beginning of the 3'SS; as expected, the values downstream from the 3' end of the exon/intron boundary (right side of the 3'SS) are higher than upstream from it (left side of the 3'SS; $p<3.3\cdot10^{-6}$, Wilcoxon signed-rank test that is a paired test). C) Mean TDR/tAI profiles aligned to the beginning of the last Intronic STOP codon, closer to the acceptor site; tAI before the last intronic STOP codon is similar to the post-STOP domain ($p>0.13$, Wilcoxon signed-rank test).

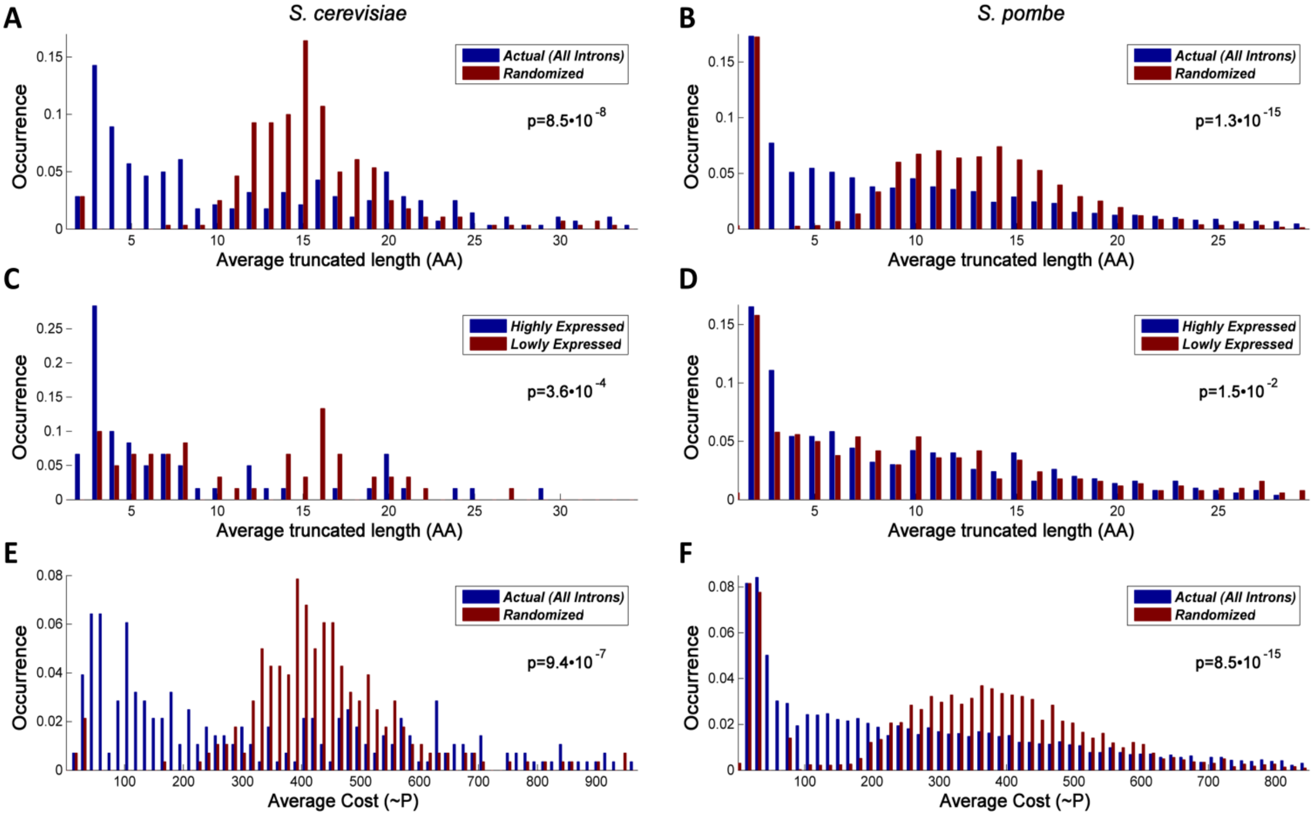


Figure S14: Detailed statistical analysis of the truncated peptide length and metabolic cost distributions in *S. cerevisiae* and *S. pombe*. A) Analysis of the Intronome of *S. cerevisiae* demonstrates that the average truncated length distribution in the randomized models is significantly longer than observed in the actual intronome (20.37AA *vs.* 14.89AA, respectively; p=8.55·10^{-8}, Wilcoxon signed-rank test that is a paired test). B) Analysis of the Intronome of *S. pombe* demonstrates that the average truncated length distribution in the randomized models is significantly longer than observed in the actual intronome (21.49AA *vs.* 19.91AA, respectively; p=1.25·10^{-15}, Wilcoxon signed-rank test). C) Intronome analysis of highly *vs.* lowly expressed genes in *S. cerevisiae* demonstrates that there is a preference for shorter retained protein length in highly expressed genes (9.13AA *vs.* 16.08AA, highly and lowly, respectively; p=3.61·10^{-4}, Wilcoxon rank-sum test). D) Intronome analysis of highly *vs.* lowly expressed genes in *S. pombe* demonstrates that there is a preference for shorter retained protein length in highly expressed genes (11.34AA *vs.* 27.52AA, highly and lowly, respectively; p=1.54·10^{-2}, Wilcoxon rank-sum test). E) Analysis of the Intronome of *S. cerevisiae* demonstrates that the average metabolic cost distribution in the randomized models is significantly higher than observed in the actual intronome (567.5 *vs.* 425.2, respectively; p=9.36·10^{-7}, Wilcoxon signed-rank test). F) Analysis of the Intronome of *S. pombe* demonstrates that the average metabolic cost distribution in the randomized models is significantly higher than observed in the actual intronome (596.8 *vs.* 556.3, respectively; p=8.5·10^{-15}, Wilcoxon signed-rank test).

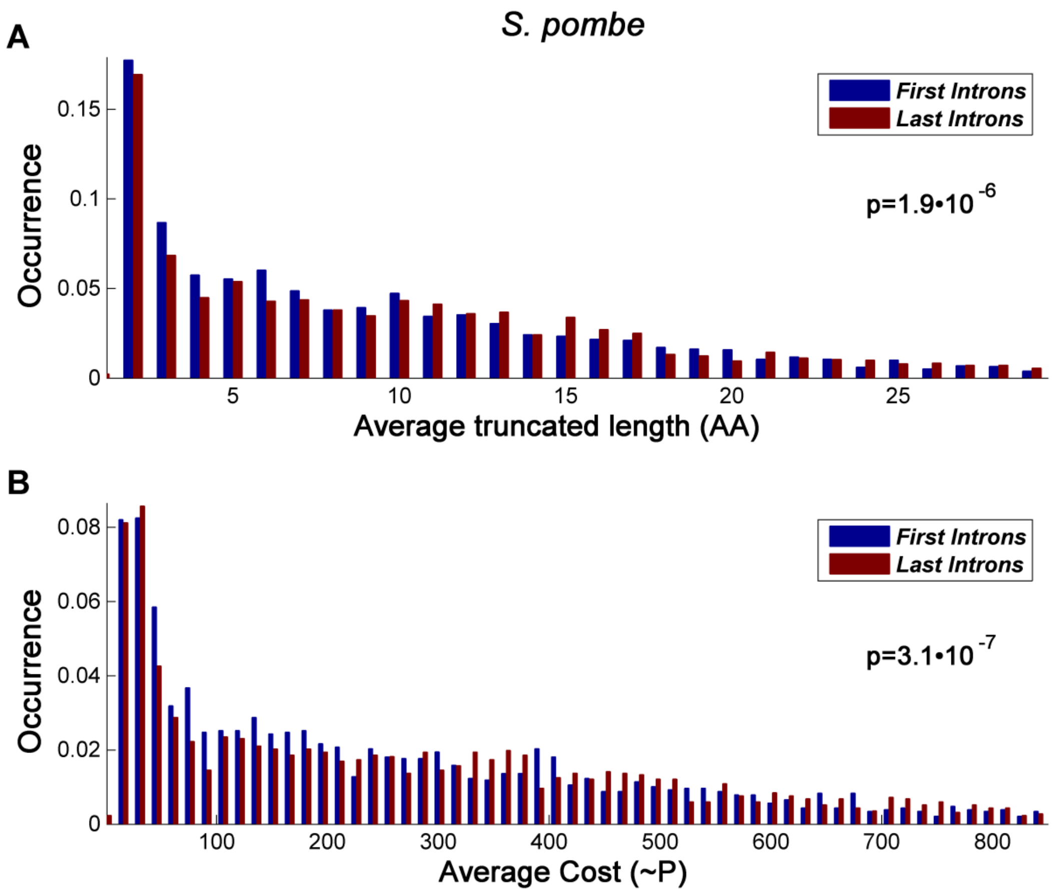


Figure S15: Detailed statistical analysis of the truncated peptide length and metabolic cost distributions of the first introns in *S. pombe*. A) Analysis of the first introns, located close to the beginning if the ORF, and the rest of the introns demonstrates that there is a preference for shorter retained protein length in the first introns (17.65AA *vs.* 21.98AA, first and last introns, respectively; $p=1.94\cdot10^{-6}$, Wilcoxon rank-sum test). B) Analysis of the first introns, located close to the beginning if the ORF, and the rest of the introns demonstrates that there is a preference for lower average metabolic cost in the first introns (486.6 *vs.* 620.1, first and last introns, respectively; $p=3.11\cdot10^{-7}$, Wilcoxon rank-sum test). Similar results were obtained for the first introns *vs.* only the second introns (17.65AA vs. 20.58AA for the average truncated length and 486.6 *vs.* 581.2 for the average metabolic cost; $p=6.99\cdot10^{-4}$ and $p=4.64\cdot10^{-4}$, respectively; histograms not shown).

Supporting Tables' Description

Supporting Table S1: TDR and tAI Statistical Information

This table summarizes the statistical analyses of the examined organisms: for *S. cerevisiae* we used the mean of typical decoding rate (TDR) measure; for *S. pombe*, *A. nidulans*, and *C. albicans* we used the average tRNA adaptation index (tAI) measure. The scores were aligned to the 5'SS and the first intronic STOP codon; the randomized model average TDR/tAI and corresponding Z-score values are also displayed. Wilcoxon signed-rank test and paired t-test were performed between the upstream/downstream average values per gene on the TDR/tAI profiles and ribosomal profiling (RP). Empirical p-values were calculated based on up to 1000 randomized models (1000 for *S. cerevisiae* and *C. albicans*, 200 for *S. pombe* and *A. nidulans*). Spearman correlation was performed on the TDR/tAI and protein abundance (PA) levels.

Supporting Table S2: Subgroups TDR and Selection Level Analysis

This table summarizes the average TDR and Z-scores values for various gene subgroups in *S. cerevisiae*: ribosomal introns *vs.* non-ribosomal introns, PA/PPR/RD highly expressed *vs.* PA/PPR/RD lowly expressed, mRNA highly expressed *vs.* mRNA lowly expressed, and YiFP highly spliced *vs.* YiFP lowly spliced. The data is aligned to the 5'SS and the first intronic STOP codon.

Supporting Table S3: Go Terms Inclusion/Exclusion List

This table includes a list of GO functional groups in *S. pombe* that were include in or excluded from the analysis, based on their introns number (threshold was 50 introns). Two sided Wilcoxon rank-sum was performed on the average TDR values per gene for matching subgroups between the pre-STOP and post-STOP domains.

Supporting Table S4: Exclusion List

This table includes lists of genes that were excluded from the analysis, the analyzed organism, and the exclusion reason.

Supporting Table S5: General Information

This table includes general information regarding the four analyzed organisms, including number of introns, GC content, and STOP codon information.

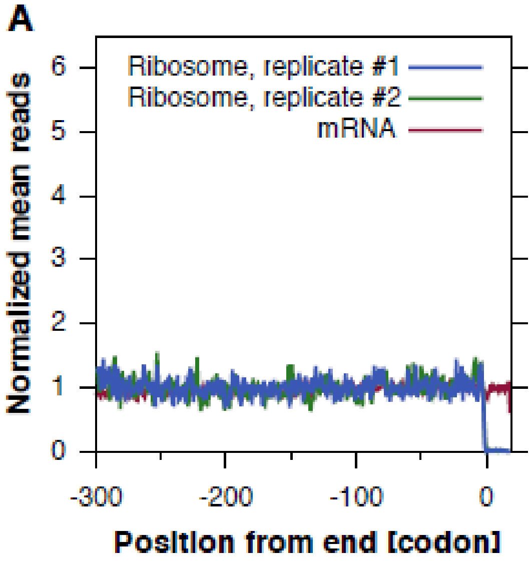
Supporting Table S6: Ribosomal Profiling gene List

This table includes a list of *S. cerevisiae* genes and their corresponding mean RC values for the pre-STOP domain; only genes with nonzero RC are shown.

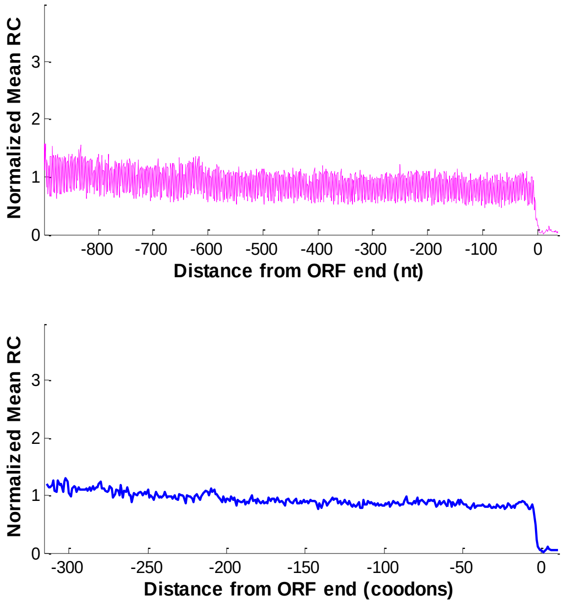
Supporting Notes

Supporting Note S1: About the shape of the intronic STOP Ribo-seq profile

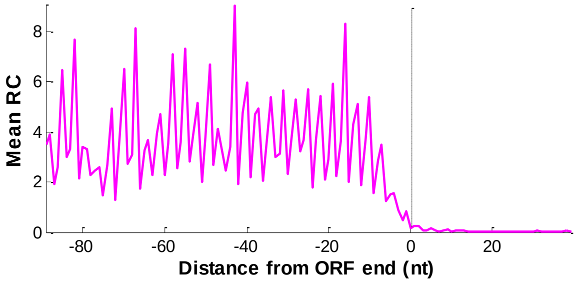
The purpose of this text is to demonstrate that our analysis is in agreement with previously published pre-STOP Ribo-seq read count profiles in yeast (55). We compared our results with the reported profile near the STOP codon that appears to be uniform at the ORF's end, with drop in reads downstream of it (supplemental figure S11A).

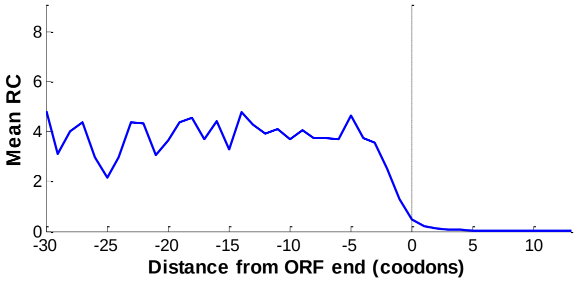


First, we plotted the profile of the same genes used in (55) under the same normalizations that were performed in the study; we obtained the following figures that are very similar and demonstrated that our analysis is sound.



Next, we plotted the Ribo-seq RC profile near the ORF annotated STOP codon corresponding only to the 275 genes with introns (the ones used in the Ribo-seq figure that appear in the paper).





As can be seen, the resultant graph is similar to the reported plot of Ribo-seq read counts surrounding the first intronic STOP codon (the peaks before the STOP codon can be observed). The differences between this graph and the previous one are due to the different number of genes (less genes look “noisier”), the smaller area under observation, and the normalization/filtering performed (in (55) the read count of each gene was normalized by the RC at a region at the beginning of the transcript; in addition, a set of 1525 genes with highest RC and other properties was chosen for the graph). Thus, in our opinion the more plausible explanation for the RC pattern at the intron 5’ end (*e.g.* the peak at -10) is the “noise” and “missing values” of the Ribo-seq data.

Supporting Note S2: About splicing error rate estimation and intron retention

In the paper we mention various measures related to, or expected to be correlated with, splicing efficiency (SE), intron retention, and intron translation. In this note we briefly discuss these measures and the relations between them.

The natural definition of SE is the percentage of transcripts that are spliced correctly, while splicing error is the percentage of incomplete splicing; it is related but not identical to intron retention. Therefore, we would like to mention a few points that make this definition even more complicated (or less trivial) in general and specifically in our case: the SE values measured via the synthetic system shown in Figure 5B and also the Ribo-Seq ratio are *not* exactly the SE values related to endogenous genes (but it is expected to be highly/significantly correlated with it).

The 'intuitive' gene's SE for gene *i* is defined as:

$$(1)\quad \frac{mRNA\ molecules\ of\ gene\ i}{pre{-}mRNA\ molecules\ of\ gene\ i}$$

The SE which is related to the Ribo-Seq data is defined as:

$$(2)\quad 1-\frac{non{-}spliced\ mRNA\ molecules\ of\ gene\ i\ transported\ out\ of\ the\ nucleuses}{mRNA\ molecules\ of\ gene\ i\ transported\ out\ of\ the\ nucleuses}$$

see also (117) regarding nuclear pre-mRNA degradation.

The synthetic SE system (shown in Fig 5B) for strain *i* (with the i_{th} intron mediating its YFP; see (69)) is defined as:

$$(3A)\quad \frac{protein\ molecules\ related\ to\ strain\ i}{protein\ molecules\ related\ to\ a\ strain\ without\ an\ intron}$$

or we can similarly look at:

$$(3B)\quad \frac{protein\ molecules\ related\ to\ strain\ i}{protein\ molecules\ related\ to\ strain\ j}$$

It is easy to see that these measures are not identical (but all are relevant and expected to be correlated).

It is important to mention that in the YFP experiment (described in (69)) we introduced most of the yeast introns into identical YFP genes and measured the corresponding YiFP levels. Thus, the differences in YiFP levels among the strains are expected to be related to different (*a*) transcription rates, (*b*) splicing efficiencies, (*c*) pre-mRNA transport / degradation efficiencies.

Furthermore, it should be remembered that the intron has two 'ends'. A problem in splicing of one end is defined as splicing error; however, our study is related and relevant only in the case of 5'SS error/retention.

Finally, SE in the synthetic system is not expected to be identical to the one obtained in the corresponding endogenous genes. For example, we and others have shown that some aspects of the splicing signals are partially encoded in the exon-intron boundaries, and particularly in the interaction between the exon and the intron. Thus, we expect to see a significantly higher splicing error rate in the case of the synthetic library Nevertheless, we would like to emphasize here that the points above should not contradict the fact that the YiFP library SE estimations are expected to correlate (not perfectly) with the actual SEs (The correlation between the two measures of SE, synthetic and endogenous, is $r = 0.233$ $p = 5.86 \cdot 10^{-4}$); thus, it makes sense to include it in the study.