DNA Research, 2016, 23(4), 377–394
doi: 10.1093/dnares/dsw019
Advance Access Publication Date: 2 June 2016
Full Paper

OXFORD

Full Paper

Selection for reduced translation costs at the intronic 5′ end in fungi

Zohar Zafrir[1], Hadas Zur[1], and Tamir Tuller[1,2,*]

[1]Department of Biomedical Engineering, Tel Aviv University, Tel Aviv, Israel, and [2]The Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 69978, Israel

*To whom correspondence should be addressed. Tel. +972 3-640-5836. Fax. +972 3-640-8123.
E-mail: tamirtul@post.tau.ac.il

Edited by Dr Mikita Suyama

Received 15 December 2015; Accepted 26 April 2016

Abstract

It is generally believed that introns are not translated; therefore, the potential intronic features that may be related to the translation step (occurring after splicing) have yet to be thoroughly studied. Here, focusing on four fungi, we performed for the first time a comprehensive study aimed at characterizing how translation efficiency is encoded in introns and affects their evolution. By analysing their intronome we provide evidence of selection for STOP codons close to the intronic 5′ end, and show that the beginning of introns are selected for significantly high translation, presumably to reduce translation and metabolic costs in cases of non-spliced introns. Ribosomal profiling data analysis in *Saccharomyces cerevisiae* supports the conjecture that in this organism intron retention frequently occurs, introns are partially translated, and their translation efficiency affects organismal fitness. We show that the reported results are more significant in highly translated and highly spliced genes, but are not associated only with genes with a specific function. We also discuss the potential relation of the reported signals to efficient nonsense-mediated decay due to splicing errors. These new discoveries are supported by population-genetics considerations. In addition, they are contributory steps towards a broader understanding of intron evolution and the effect of silent mutations on gene expression and organismal fitness.

Key words: mRNA translation, intron evolution, transcript evolution, splicing, silent mutations

Introduction

RNA splicing is the process in which pre-mRNA transcripts mature and where introns, intervening fragments within transcripts, are recognized and removed, leaving retained regions termed exons that compose the mature mRNA. In eukaryotes spliceosomal introns are confined to the nucleus, and the splicing process is executed by the spliceosome, one of the largest molecular complexes in the cell.[1,2] Accurate processing of introns is a crucial regulatory step in determining the cell expression profile, and is required before protein translation can be initiated. As such, splicing efficiency (SE; efficient recognition and proper splicing by the spliceosome) and specificity are used to regulate growth, development, and overall response to external signals.[3,4] Extensive studies over the past two decades have revealed the core chemical reactions, several sequence determinants, and the major protein component interactions during intron splicing.[5–8] Nevertheless, the debate on intron origin and evolution has been ongoing intensively for decades.[9,10]

The dynamic nature of spliceosome assembly and the complex interactions of both its proteins and RNA components with the pre-mRNA, give rise to a range of intronic SEs among different genes and within the same gene that may lead to altered SE, and facilitates the creation of different mature mRNA products from identical pre-mRNA transcripts;[11] in this process particular exons of a gene may be included within or excluded from (i.e. 'skipped') the final

processed mRNA transcript. This phenomenon is termed alternative splicing (AS) and is the chief constituent of the increased number of proteins encoded by genomes of multicellular organisms.[12] Intron retention, in which an intron remains in the mature mRNA transcript, is widespread among plants and protozoa, but is infrequent in vertebrates and seldom found in fungi.[13–17] As such, it is often considered the earliest version of AS to have evolved and might reflect mis-splicing, as the splicing machinery may fail to recognize weak splice sites flanking short introns. Additional AS events include alternative splice sites selection, whilst the combination of two or more events can generate more complex mature mRNA products.[18–23]

It was suggested that small introns in diverse organisms (such as *Paramecium tetraurelia*, *Homo sapiens*, *Caenorhabditis elegans*, and *Schizosaccharomyces pombe*) are under selective pressure to cause premature translation termination in case an intron has not been spliced (i.e. 'retained'). In particular, it has been proposed that the nonsense-mediated mRNA decay (NMD) pathway may have a proofreading role in gene expression, eliminating transcripts that have not been properly spliced;[24] moreover, the core components of the NMD machinery are evolutionarily conserved in all eukaryotic organisms tested so far, and their deletion or silencing prevents NMD in eukaryotic cells.[25–29]

The amount of intron-containing genes, intron density per gene or per kilobase pair, and intron length, varies from one eukaryote to another.[30–33] In *Saccharomyces cerevisiae* there are <300 intron-containing genes (5% of ~6,000), of which few are mediated by two introns or more.[34,35] In fact, until now only two AS or intron retention events have been identified in this organism.[15,36] Nevertheless, in *S. cerevisiae* intron-containing genes are particularly highly expressed and account for more than 70% of its proteome. Additionally, several intron sequences are found duplicated within ribosomal protein paralogs.[37,38] Conversely, in *S. pombe* (an organisms with ~5,000 genes) there are ~5,000 introns spreading over a third of its genome, with up to 20 introns mediating a single gene[39] and with few known genes displaying AS.[17,40,41] Other fungi such as *Aspergillus nidulans* and *Candida albicans* show variations in intronic characteristics as well.[42,43] Nonetheless, even though eukaryotic evolution has been generally characterized by widespread intron gain and loss events,[10,44,45] the ubiquity of introns and the core of the spliceosome is conserved in all well-characterized eukaryotes.[1,46] Moreover, *Hemiselmis andersenii* is currently the only known eukaryotic organism without any introns or spliceosome subunit genes.[47]

The intron boundary signals at the donor splice site (or 5'SS) and acceptor splice site (or 3'SS), the branch site (BS), and the polypyrimidine tract, are canonical sequence elements which are essential for intron recognition and for splicing to occur. The factors that bind to these sequence motifs and the biochemical reactions they perform are relatively well known due to the extensive research in this field. Systematic investigations show that this process is highly regulated: from spliceosome assembly, through pre-mRNA recognition and binding, to the splicing reaction and complex disassembly.[23,48] Additionally, it has been suggested that in yeast, introns regulate ribosome biogenesis and function, and affect cell fitness under stress.[49]

Despite such comprehensive studies, very little is known regarding the intronic features that are related to translation. Specifically, introns are believed to be the non-translated parts of the transcript; however, since splicing is not a perfect mechanism, it is possible that high enough splicing errors, such as intronic retention, will trigger selective pressure for translation efficiency (TE) adaptation at the 5′ end of introns. Here, we conduct a novel large scale study of four fungal intronomes, to characterize novel genomic level intronic features in the vicinity of the intronic splice-sites, related to post-splicing/translation regulation, and understand their evolution.

Materials and Methods

The analysed organisms

The four fungi analysed here (*S. cerevisiae, S. pombe, A. nidulans*, and *C. albicans*) were chosen based on the following considerations: *S. cerevisiae* and *S. pombe* are well studied organisms with well-established databases that are known to have diverged 350–900 million years ago.[50] *A. nidulans* and *C. albicans* are two additional fungi known to have diverged from *S. pombe* about 650 million years ago[51] and from *S. cerevisiae* about 235 million years ago.[52] The genomes of these organisms are also fully sequenced and their introns are well annotated. In addition, *C. albicans* is a dimorphic fungus which can be a significant pathogen in humans.

Intronic sequence information

S. cerevisiae open reading frames (ORFs) and intron-containing gene sequences (strain 288c) were taken from the *Saccharomyces* Genome Database (SGD); [53] BS location information was obtained from the Ares lab database.[35] *S. pombe* genome information (Assembly 16) was taken from the PomBase database; [54] BS locations were calculated based on the position-specific scoring matrix information extracted from the Sanger Institute, and is based on the original full genome sequencing.[39] *A. nidulans* (FGSC A4) and *C. albicans* (SC5314 Assembly 21) genome information was taken from the *Aspergillus* Genome Database (AspGD)[55] and *Candida* Genome Database (CGD),[56] respectively; BS locations were calculated based on the fungal BS consensus sequence (***CURAY***). We used only introns taken from coding sequences (CDSs) and excluded 5′ untranslated region (UTR) and 3′ UTR introns. Introns associated with putative or alternatively spliced genes were also excluded; the full intron exclusion list can be found in Supplementary **Table S4**. Additional GC content and general information can be found in Supplementary **Table S5**; exon–intron GC content was calculated using intron sequences and their flanking exon sequences (100 bp upstream and downstream).

Protein abundance, mRNA levels, and metabolic cost

Protein abundance (PA) information for *S. cerevisiae* and *S. pombe* was taken from the PaxDb dataset,[57] which integrates information from various resources. The mRNA levels for *S. cerevisiae* are based on RNA-seq and DNA chip and were obtained by integrating three datasets[58–60] in the following manner: first, we normalized each dataset by its average mRNA levels; next, for each gene we averaged all its normalized measurements (dataset with no value was ignored). Protein per mRNA ratio (PPR), i.e. the number of proteins produced on average from an mRNA molecule, was calculated by dividing PA by mRNA levels. PA levels are measured in parts per millions. The mRNA is based on RNA-seq and DNA chip and thus is proportional to the mRNA levels (number of molecules in the cell). Since our analysis is based on spearman correlation, ranking of genes according to their expression levels is enough to provide the required results (i.e. the actual levels will not change the results). Data for metabolic cost of amino acids (AAs) for *S. cerevisiae* were taken from Ref. 61 and

were also used as an approximation for *S. pombe*, as no *S. pombe* data were available.

Randomization models

The randomized models were designed to conserve encoded protein information, in addition to the intronic properties. Specifically, we maintained codon frequencies [i.e. codon-usage bias (CUB)], canonical splicing signals, and GC content. To this end, the distribution of codons in the CDS was calculated for each gene, and the sequence was then randomized while maintaining these codon frequencies (i.e. for each AA in the gene, the probability of a certain codon was determined by its individual distribution in the CDS) and the amino acid content and order (i.e. the encoded protein); this was separately done for upstream and downstream exons. For each gene, intronic nucleotides were uniformly permutated, maintaining consensus sequences (5′SS, BS, and 3′SS) and the intronic GC content. We used the following randomization schemes to generate random sets: (i) Codon only and (ii) Intron only. Next, a combination of the basic schemes was applied, i.e. (i) + (ii), and is used throughout the study; the randomization models are also illustrated in Fig. 1C and D. In the models used in Figs 3 and 5, we randomized the intronic sequences while assuming introns as being translated. Specifically, we 'continued' the introns in the same frame of their upstream exons and considered the triplets of nucleotides in the introns as 'artificial' codons; next, we scrambled them while controlling for their CUB and GC content.

Evaluation of significance levels

For each genome, up to 1,000 random intronomes were generated (1,000 for *S. cerevisiae* and *C. albicans*, 200 for *S. pombe* and *A. nidulans*). The level of significance (i.e. empirical *P*-value) was determined based on a comparison between the random intronomes and the actual one as follows, when we expect the real to be significantly lower than the randomized version (e.g. in the case of the distance to STOP codon): Let S_0 be the value of the actual intronome and $S = S_1,\ S_2, \cdots, S_N$ a vector containing N random values sampled from the randomized models, where S_i is the value of the i_{th} sample; then $p = \frac{1}{N}\sum_{i=1}^{N} |\{S_i > S_0\}|$, $N = 1{,}000$. When we expect the real to be significantly higher than the randomized version (e.g. the typical decoding rate (TDR)/tAI) the *P*-valued was computed as follows: $p = \frac{1}{N}\sum_{i=1}^{N} |\{S_i < S_0\}|$, $N = 1{,}000$.

The mean distance to the first STOP codon (shown e.g. in Fig. 2) was calculated over all introns and then compared to the distribution of means in the randomized sequences, which maintain the length and nucleotide distribution per intron. Note that performing the *P*-value computation separately for each intron cannot 'mathematically' work in this case: since three of the codons are STOP codons, the probability to get an in-frame intronic STOP codon in a certain position in the case of uniform nucleotide distribution is ∼0.05 (1–61/64). Similarly, the probability to get a STOP codon in the first or second positions is ∼0.09 (1–(61/64)*(61/64)). Thus, a significant intron (with $P \sim 0.05$ or very close to this value) can only be observed if the STOP codon is exactly in the first 'codon' of the intron. Nevertheless, for a large set of introns (e.g. an entire intronome) we have enough 'degrees of freedom' to detect such signals.

CUB calculation and randomization

The mean of TDR scores were calculated using the modified Sharp and Li formulation: [62,63] $\text{TDR} = \exp\left(1/L\sum_{l=1}^{L} ln(1/\mu_i(l))\right)$, where L is the AA sequence length, and μ_i is the typical codon decoding time of the i_{th} codon; μ values for *S. cerevisiae* were taken from Ref. 64 and are based on experimental measurements.[60] The tRNA adaptation index (tAI) scores were calculated using the scheme described in Ref. 65; genomic tRNA copy numbers for *S. cerevisiae* and *S. pombe* were obtained from the genomic tRNA Database;[66] for *A. nidulans*, and *C. albicans*, we used the tRNA copy numbers reported in Ref. 67 while organism-specific S_{ij} values were taken from Ref. 68.

In the preparation of the mean profiles we used a single codon resolution (i.e. $L = 1$; see an illustrative scheme in Supplementary Fig. S3). The profiles of the TDR (Fig. 3B) and tAI (Supplementary Figs S4–S6; C and D) were computed in an identical manner; hence, here we use the tAI notations to describe the procedure. Let $\text{tAI}(i)^j$ denote the tAI value of a single codon, located in the i_{th} AA of a gene j with length n. Gene tAI profile is defined as the vector of all the gene tAI values, i.e. $\text{tAI}_{\text{Gene}_j} = \left(\text{tAI}^j(1), \text{tAI}^j(2), \ldots, \text{tAI}^j(n)\right)$. For each organism, all intron-containing genes were aligned once according to their 5′SS location and once according to their first intronic STOP codon location. Let $i_{5'\text{ss}}$ and i_{stop} denote the positions of the 5′SS and first intronic STOP codon (for each of the analysed introns these indexes correspond to its 5′SS and STOP codon, respectively). The profiles of mean tAI were calculated based on window WS of 30 codons (i.e. 90 nt). Thus:

$$\overline{\text{tAI}_{5'\text{ss}}} = \left(\overline{\text{tAI}(i_{5'\text{ss}} - \text{WS} + 1)}, \ldots, \overline{\text{tAI}(i_{5'\text{ss}})}, \ldots, \overline{\text{tAI}(i_{5'\text{ss}} + \text{WS})}\right)$$

$$\overline{\text{tAI}_{\text{STOP}}} = \left(\overline{\text{tAI}(i_{\text{stop}} - \text{WS} + 1)}, \ldots, \overline{\text{tAI}(i_{\text{stop}})}, \ldots, \overline{\text{tAI}(i_{\text{stop}} + \text{WS})}\right),$$

where $\bar{\text{tAI}}(i) = \sum_j \frac{\text{tAI}^j(i)}{|\text{Gene}_j,}$ and Gene_j is the number of genes with translatable codons at their i_{th} location. Additionally, downstream exons in 5′SS profiles and all exons in STOP profiles were excluded, as well as UTRs such as 5′UTR and 3′UTR; analysis of STOP profiles when downstream exons are included (as illustrated in Supplementary **Fig. S8**) showed similar results. Values for the entire domains (shown e.g. in Fig. 3A) were calculated as the geometric mean per intron and then averaged over all introns. Wilcoxon signed-rank (paired) tests and paired *t*-tests were performed between the upstream/downstream average values per gene on the TDR/tAI profiles; additional statistics, including Spearman correlations of TDR/tAI with PA levels, can be found in Supplementary **Table S1**.

The randomized models used maintained the CDS codon frequencies (i.e. CUB) and uniformly permutated the intronic nucleotides, maintaining consensus sequences (as previously described; see also **Fig. 1C and D**). During randomization, the 5′SS alignment termination position (Supplementary **Fig. S8A–C**) was set according to the randomization results, i.e. randomized termination point location. In the case of termination point alignment (Supplementary **Fig. S8F**), the termination point alignment position was set to the original place of the retained intron in order to avoid location bias. The intronome fraction level which determines the edges of analysis was set to 15% of the overall intronome (the results are robust to reasonable changes in this cut-off).

Computing Z scores based on the randomized models

Standard normal distribution scoring (i.e. Z-score) is a statistical measure that can be used for quantitative selection level evaluation

via the comparison of the real signal to a randomized one; higher Z-score means higher chance that the selective pressure hypothesis is true and vice versa. Absolute values higher than 1.96 are typically considered to be significant, based on 95% confidence level (i.e. $P < 0.05$). Z-score values were calculated according to the following equation: $Z_{\text{score}} = \frac{\mu_{\text{real}} - \mu_{\text{rand}}}{\sigma_{\text{rand}}}$ where μ_{real} is the mean related to the actual intronome, μ_{rand} is the random model mean, and σ_{rand} is the random model standard deviation (STD). We used the following scheme related to the TDR/tAI profiles values: for each randomized model, a vector of average values was generated (i.e. an average for each random intronome) and the global Z-score was calculated.

Ribosomal profiling analysis

The ribosomal profiling (RP) method gives quantitative information of ribosome footprints and ribosomal density (RD) in a single nucleotide resolution.[60] The analysis was performed as follows: RP raw data were obtained from NCBI GEO database[69] (accession GSE13750) for two footprinting experiment replicates in rich media and mapped in a similar manner to the one described in Refs 60 and 70. It includes the following major steps: (i) Degenerate reads, i.e. reads in which 18 or more of the first 22 nt are A bases (meaning the entire read is composed of the homopolymer A tail), were eliminated. Reads mapped to ribosomal RNA genes and other non-coding genes, such as tRNA, snRNA, snoRNA, taken from Biomart (version R64-1-1),[71] were excluded. (ii) The entire pre-mRNA set of the *S. cerevisiae* genome was reconstructed utilizing BioMart (version R64-1-1); [71, 59] and each pre-mRNA included its 5′UTR, exons, introns, and 3′UTR, genes with no 5′UTR and 3′UTR annotations were supplemented with flanking genomic segments with the maximum length ensuring there is no overlap with existing coding regions. (iii) Reads of 36 nt were mapped using bowtie[72] to the pre-mRNA sequences allowing up to two mismatches on a seed length of 21 nt, to account for polyadenylation of footprint fragments. (iv) Non-uniquely mapped reads were extended from the mapped seed length of 21–28 nt to determine uniqueness, by first removing the poly-A tail of each 36 nt long read according to the following heuristic: we determine the poly-A tail from the read's end, by identifying the longest stretch of As allowing two mismatches and removing it, reads longer than 28 nt were cropped, and reads shorter than the 21 nt seed length after poly-A tail removal were discarded, now the read with the lowest mismatch score was selected if it uniquely exists; otherwise, the remaining contender multi-reads were distributed proportionally to the density of the uniquely mapped reads in their vicinity (30 nt upstream and downstream).

Based on the RP footprint read counts (RCs) of the two aforementioned experiment replicates, and the 5′UTR, exon, intron, 3′UTR annotations of 282 *S. cerevisiae* intron-containing genes, we performed the following analysis: (i) the position of the ribosome A-site was related to 15 nt from the beginning of the read; (ii) we averaged the profiles obtained for the two replicates; (iii) averaged profiles were generated in a similar way to the tAI profiles previously described, however in a nucleotide-based resolution with the following adjustments: downstream exons from the 5′SS alignment were excluded, as well as upstream exons from the STOP alignment; in addition, genes with a STOP codon positioned in the downstream exon were excluded; in the second intronic STOP analysis, RCs upstream from the first intronic STOP codon were excluded as well. Wilcoxon signed-rank test and paired *t*-test were performed between the upstream/downstream average values per gene on the RP (Supplementary **Table S1**). Analysis scheme with a possible gene example is presented in Supplementary **Fig. S10**. The estimation of the difference between the pre-STOP and post-STOP domain (shown in **Fig. 7G**) was done using 3,804 genes with observed protein levels taken from Ref. 73, while excluding putative genes, dubious ORFs, and genes with No RC values at all. List of RP genes with nonzero RC in their pre-STOP domain can be found in Supplementary **Table S6**. Additional A-site related offset alignments of 14 and 16 nt show similar results (Supplementary **Fig. S12**).

Subgroups analysis and bias control

Introns were divided into subgroup sets according to various criteria: ribosomal *vs*. non-ribosomal, and highly expressed *vs*. lowly expressed (based on PA, RD, mRNA, or PPR measurements). Subgroups of different size may cause bias in the strength/significance of the detected statistical signals; therefore (and in order to control for the effect of different group sizes on the Z-score), we ensured that the size of both subgroup pairs is identical (i.e. equal number of introns in each subgroup); in case of a difference, introns were chosen randomly form the larger group to match subgroup size (e.g. in *S. cerevisiae* 93 introns were randomly chosen from a set of 187 non-ribosomal introns in total to match the ribosomal ones). Two sided Wilcoxon rank-sum test and two sample *t*-test were performed on the TDR/tAI profiles.

Synthetic YiFP reporter library building and analysis

The synthetic YiFP reporter library was downloaded from Yofe *et al.*[74] Briefly, to create the synthetic intron reporter library, *S. cerevisiae* was transformed with a library of DNA transformation cassettes; each containing a different native yeast intron. The cassettes were assembled using the Y-operation[75,76] by which introns were embedded in a Yellow Fluorescent Protein (YFP) fragment and concatenated to a common selection marker in high throughput. In this manner 240 strains were created, termed YiFP strains, where the sole difference between all strains is the native *S. cerevisiae* intron intervening the YFP gene. The contribution of introns to the regulation of gene expression was assessed by dynamic measurements of YFP expression. Following normalization, the expression level of each intron strain was compared to that of the intron-less YFP strain to give a measure of its relative expression level, which is related to intronic SE. YiFP strains that had a signal to noise ratio (SNR) of 5 and above were classified as spliced (178 introns), while the others ($\text{SNR} < 5$; 62 introns) were classified as un-spliced.

Gene ontology function specific group analysis

We used the gene ontology (GO) annotation database (http://www.geneontology.org/, 13 May 2016, date last accessed) to derive all the functional gene groups related to the three ontology domains (Cellular component, Molecular function, and Biological process). We analysed a subset of the terms, i.e. GO Slim, evaluating 90 terms in the database with at least 50 introns (out of a total of 145 terms; each term contains at least 1% of the *S. pombe* intronome), and built corresponding Z-score profiles for each subgroup. Profiles were then vertically aligned in three ontology clusters, to present exonic donor, pre-STOP and post STOP intervals. Significance level for each term was calculated using a two sided Wilcoxon rank-sum test on the tAI upstream and downstream intervals, aligned to the STOP codon.

Considering cases when the first in-frame STOP codon is downstream of the introns

Here, we provide some additional specific details regarding the manner in which we consider the cases when the first in-frame STOP is downstream of the introns: the cases where the first in-frame STOP codon is in the 3′ UTR are very rare (<0.5%) and are ignored or do not exist; these cases are very probably related to sequencing/annotation errors (i.e. the main ORF does not include a STOP codon), and do not exist at all in well annotated organisms (e.g. *S. cerevisiae*). We ignore the very few cases (around ∼2%) where the first in-frame STOP is a main ORF STOP codon since, by the definition of our random model, in these cases the random and real sequences are identical. It is easy to see that any reasonable consideration of these cases does not affect any conclusion of the study.

Also note that when we plot TE (e.g. adaptation to the tRNA pool/TDR) and Ribo-seq profiles we ignore non-intronic parts since we do not want to bias the results (the TDR/tAI/Ribo-seq in exons is expected to be very high in comparison to introns), and to perform the most conservative and relevant test possible.

Results

In this study, we analyse the intronome of four fungi: *S. cerevisiae*, *S. pombe*, *A. nidulans*, and *C. albicans*; *S. cerevisiae*; and *S. pombe* are both well studied and diverged ∼350–900 million years ago; [50] *A. nidulans* and *C. albicans* are two additional fungi with fully sequenced genomes and well annotated introns (see further details regarding these organisms in Materials and methods). Specifically, the analysed data include 277 introns from *S. cerevisiae*, 4,747 introns from *S. pombe*, 2,427 introns from *A. nidulans*, and 391 introns from *C. albicans*.

Our objective was to evaluate systematically novel sequence features/patterns near the splice sites that promote splicing regulation and affect TE at the genomic level, towards a better understanding of intronic evolution. To this end, we defined three pre-mRNA exonic and intronic regions that will be used hereafter and are illustrated in **Fig. 1A and B**: Exonic Donor—exonic region 90 nt upstream from the 5′SS (excluding upstream 5′ UTR); pre-STOP—intronic region 90 nt upstream from the first STOP codon (excluding the STOP itself and upstream exons), aligned to the 5′SS or to the first STOP codon; post-STOP—intronic region 90 nt downstream from the first STOP codon (excluding the STOP itself, downstream exons, and 3′ UTR), aligned to the first STOP codon; the pre-mRNA exon–intron boundaries, consensus sequences, and STOP codon are illustrated as well. We focused on three related aspects: (i) The position of STOP codons in possibly retained parts of the introns. (ii) The possible adaptation of the intronic 'codons' to the translation-elongation step in potentially retained parts of the introns (i.e. the intronic beginning; in-frame of the preceding exon and before the first STOP codon). (iii) The density of ribosomes on possibly retained parts of the introns. In all cases, we performed a comparative adaption analysis of various intronic subsets and functions.

Evidence of selective pressure for intronic STOP codons near the 5′SS presumably to decrease translation and metabolic costs of retained introns

Splicing, like other biological processes, does not exhibit perfect fidelity. Thus, it is clear that there are cases where introns are not spliced, and translation-elongation continues through the retained beginning of the intronic sequence.[14] If such events are frequent enough, we presume that evolution will shape the beginning of retained introns for reduced translation and metabolic costs, and possibly for increased TE (see illustration of truncated peptide formation in **Fig. 2A**). In this subsection, we study one aspect related to this argument.

Assuming that retained introns' translation is usually deleterious to the cell, we expect to see preference for close STOP codons or premature termination codons (PTCs) at the beginning of introns in-frame of the exons preceding them. To test this hypothesis, we compared the mean distance to the first STOP codon in possibly retained introns in actual genomes, to the ones obtained in randomized genomes that controlled for their exonic CUB and GC content, maintained intronic consensus sequences, and controlled for intronic GC content (per gene; see details in Materials and methods and in **Fig. 1C and D**). Specifically, we calculated the average length of all truncated peptides originating from non-spliced transcripts (relative to the end of each preceding exon) and compared it with the average lengths obtained using the randomized models. As can be seen in **Fig. 2B and C**, the average truncated peptide length in the randomized models was 37% and 8% longer than the actual one for *S. cerevisiae* and *S. pombe*, respectively; evaluation of the average length distribution in the randomized models shows that it is indeed significantly longer than observed in the actual intronome (empirical *P*-value: $P=2.3{\cdot}10^{-2}$, $P=2.4{\cdot}10^{-2}$; $P=8.55{\cdot}10^{-8}$ and $P=1.25{\cdot}10^{-15}$, Wilcoxon signed-rank test between the actual and average randomized models values per gene, which is a paired test; *S. cerevisiae* and *S. pombe*, respectively; see details in Materials and methods and in Supplementary **Fig. S14A and B**).

We expect selection for intronic STOP codons near the 5′ end of the intron to be higher in highly expressed genes, which potentially consume more intercellular translation resources. Thus, we further examined two intronic subgroups of *S. cerevisiae* and *S. pombe*, originating from highly and lowly expressed genes; 60 and 500 introns were included in each subgroup for *S. cerevisiae* and *S. pombe*, respectively, constituting 21% and 11% of their respective intronome, based on their PA (see Materials and methods). Analysis confirmed that the average retained peptide length in lowly expressed genes was 76% and 143% longer than in highly expressed genes for *S. cerevisiae* and *S. pombe*, respectively ($P=3.61{\cdot}10^{-4}$ and $P=1.54{\cdot}10^{-2}$, Wilcoxon rank-sum test between the highly and lowly expressed genes; **Fig. 2D and E** and Supplementary **Fig. S14C and D**). Evaluation of the average length distribution obtained using the randomized models showed that randomized highly expressed genes have significantly longer peptide length than observed in the actual intronome (empirical $P<1{\cdot}10^{-3}$ and empirical $P<5{\cdot}10^{-3}$ for *S. cerevisiae* and *S. pombe*, respectively). Contrary, lowly expressed genes did not exhibit evidence of significant selection (empirical $P>0.24$ and empirical $P>0.66$, for *S. cerevisiae* and *S. pombe*, respectively; Fig. 2D and E), as expected. Similar results and conclusions were obtained using randomized models that separately maintained GC content in the first and second intronic fragments (Supplementary Fig. S1).

Furthermore, we considered the metabolic biosynthesis costs of the truncated peptides originating from the non-spliced transcripts (relative to the end of each preceding exon and up to the STOP codon; see Materials and methods). The results shown in Supplementary Fig. S2A and B demonstrate that the metabolic costs

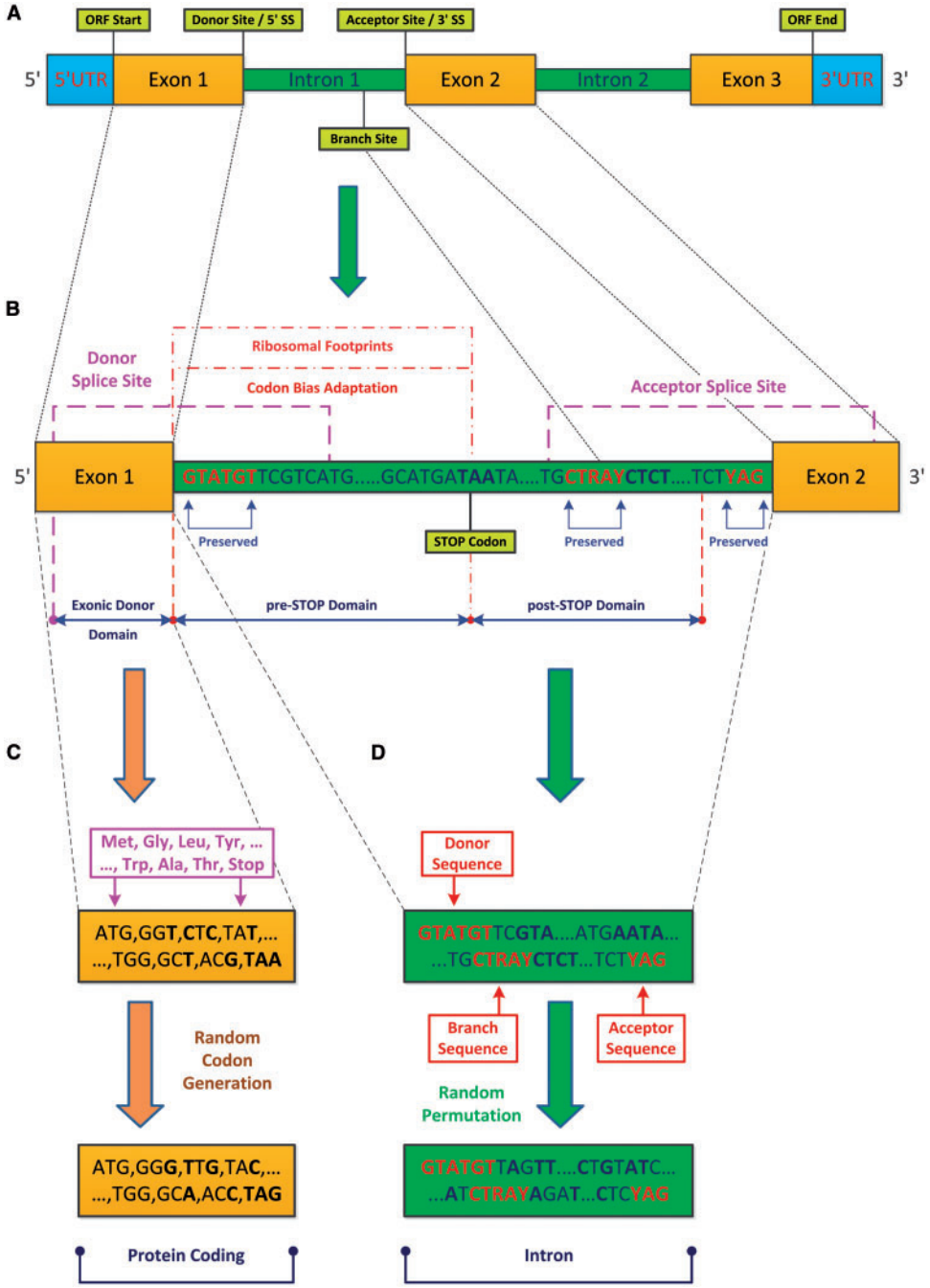

Figure 1. Pre-mRNA exonic and intronic regions, basic definitions, and randomization models. (A) The analysed fungal genes can be divided into three major regions: UTRs, exons, and introns (in the analysed organisms we did not consider UTR introns since they are not flanked by translated regions). The introns include three canonical consensus sequences: the donor (or 5′SS; subsequence *GTATGT*) and acceptor (or 3′SS; subsequence *YAG*) that define intronic boundaries, and the branch site (BS; subsequence *CTRAY*) that is required for the lariat formation. (B) In our analyses, exons and introns were divided into three domains: Exonic Donor (up to 90 nt upstream from the 5′SS), pre-STOP (up to 90 nt upstream from the first intronic STOP codon), and post-STOP (up to 90 nt downstream from the first intronic STOP codon); the features surrounding the boundaries of these regions are studied here. (C and D) In order to demonstrate that the reported features are under selective pressure, we compared the potentially translated sequences to the ones obtained by the following randomized models: (C) encoded protein information is maintained; synonymous codons frequencies of each gene (separately) are maintained; (D) uniform permutation of intronic nucleotides (per intron); the randomization models preserve these consensus sequences (5′SS/BS/3′SS), as well as additional exonic and intronic characteristics (see Materials and methods and Supplementary information). The results suggest adaptation to the translation process at the beginning of introns via preference for STOP codons close to the intronic 5′ end, and for CUB resembling the one appearing in annotated ORFs. This discovery is also supported by ribosomal profiling (RP) footprint measurements which are significantly higher in the 5′ end intronic region than in the intronic region downstream of the STOP codon.

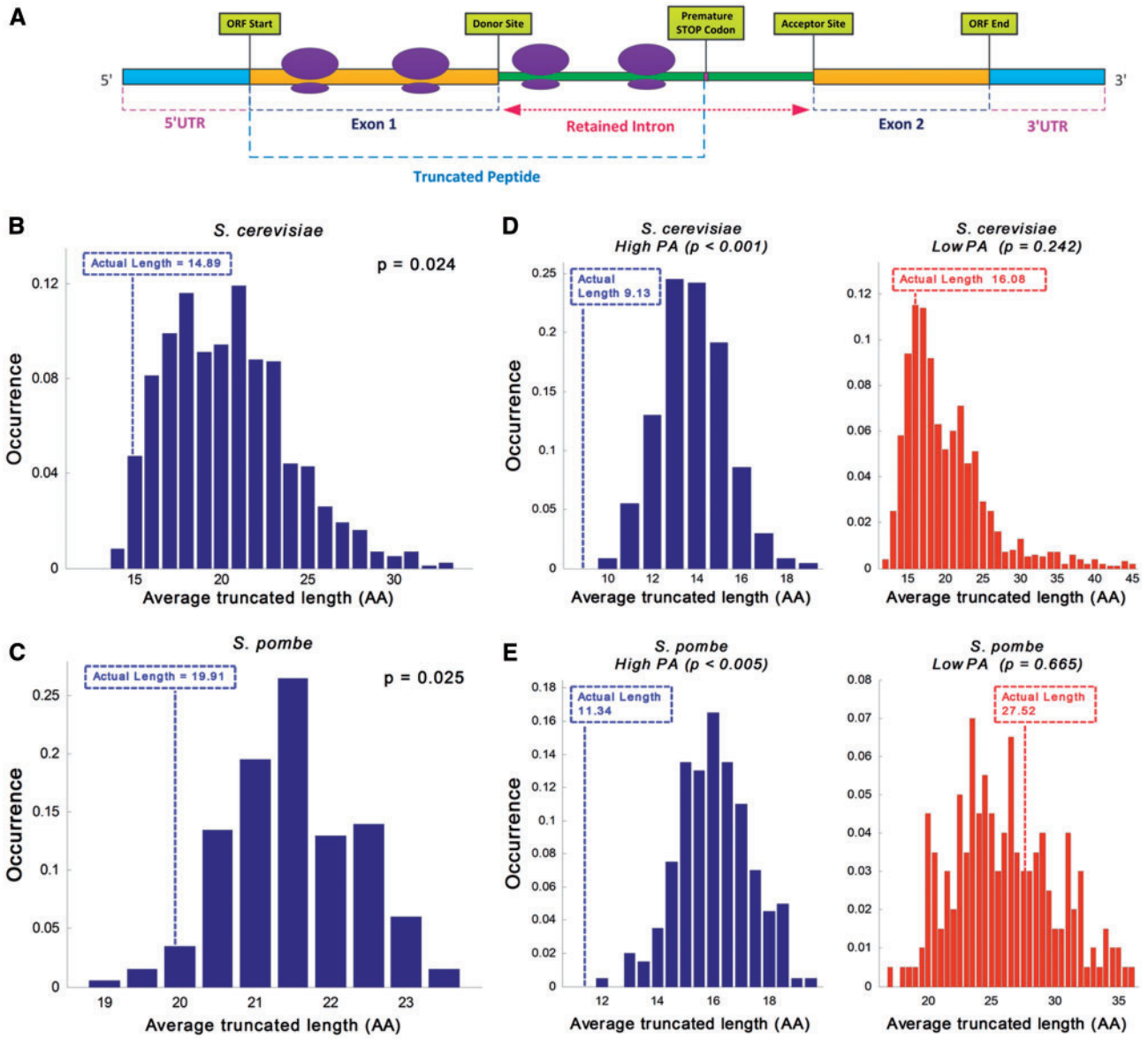

Figure 2. First intronic STOP codon analysis. The position of the first STOP codon relatively to the beginning of the intron in the actual retained intronome tends to be closer to the 5′SS than in randomized intronome models. (A) Illustration of the translation process showing ribosomes on a transcript with a retained intron; the generated protein contains amino acids (AAs) that are encoded in the intronic nucleotide composition; as demonstrated in this case translation is usually terminated by a PTC, resulting in a truncated peptide. (B and C) Analysis of the average potential truncated peptide length distribution over the entire transcriptome (transcripts with introns) in the randomized models compared to the actual ones: (B) Average truncated peptide length distribution in *S. cerevisiae* suggests that there is preference for shorter retained protein length; the average length of the randomized model is 37% longer than the actual intronome (20.37AA *vs.* 14.89AA, respectively; empirical $P=0.024$; see Materials and methods; actual length displayed in broken line). (C) Average truncated peptide length distribution in *S. pombe* suggests that there is preference for shorter retained protein length; the average length of the randomized model is 8% longer than the actual intronome (21.49AA *vs.* 19.91AA, respectively; empirical $P=0.025$; actual length displayed by broken line); for comparison, the average truncated peptide length in the case of uniform nucleotide distribution is 21.33AA. (D and E) Average potential truncated peptide length over subsets of the actual transcriptome in comparison to the randomized models: (D) Intronome analysis of highly *vs.* lowly expressed genes in *S. cerevisiae* exhibits 76% longer truncated protein length in lowly expressed genes (9.13AA *vs.* 16.08AA; left and right, respectively); distribution analysis demonstrates evidence of selection in highly expressed genes but no significant selection in lowly expressed genes (empirical $P< 1{\cdot}10^{-3}$ and empirical $P=0.242$; left and right, respectively; actual length displayed by broken line). (E) Intronome analysis of highly *vs.* lowly expressed genes in *S. pombe* exhibits 143% longer truncated protein length in lowly expressed genes (11.34AA *vs.* 27.52AA; left and right, respectively); distribution analysis demonstrates evidence of selection in highly expressed genes but no significant selection in lowly expressed genes (empirical $P< 5{\cdot}10^{-3}$ and empirical $P=0.665$, respectively; actual length displayed by broken line).

of the truncated peptides are 25% and 7% lower than in the randomized models for *S. cerevisiae* and *S. pombe*, respectively; analysis of the randomized models average cost shows that it is indeed significantly higher than observed in the actual intronome (empirical *P*-value: $P=2.6{\cdot}10^{-2}$, $P=3.5{\cdot}10^{-2}$, $P=9.36{\cdot}10^{-7}$, and $P=8.5{\cdot}10^{-15}$, Wilcoxon signed-rank test between the actual and average randomized models values per gene; *S. cerevisiae* and *S. pombe*, respectively; Supplementary **Fig. S14E and F**). Further analysis confirmed that the

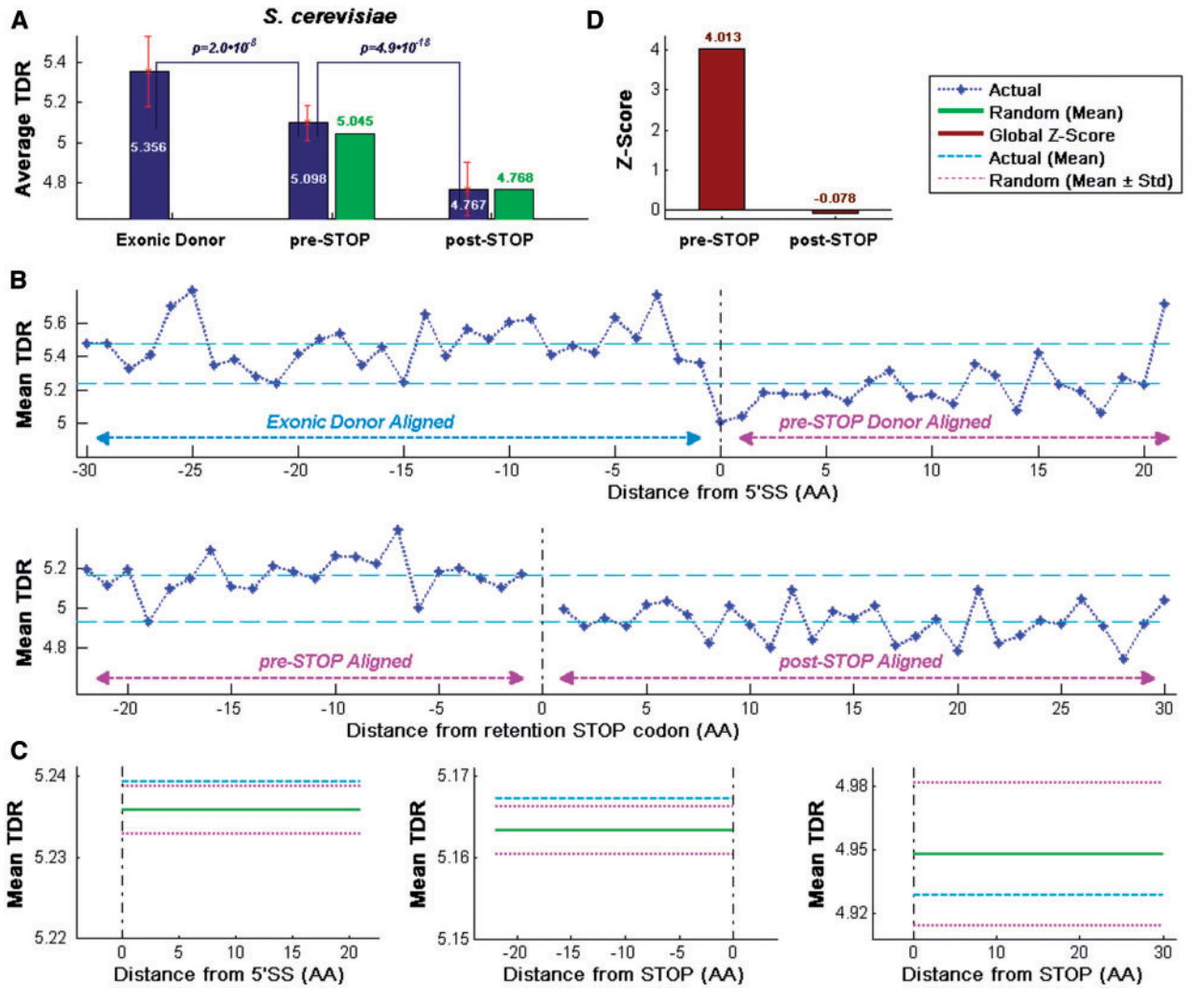

Figure 3. CUB profiles for the *S. cerevisiae* intronome. The profiles show that the average of TDRs index values upstream from the first intronic STOP codon is higher than the average of the TDR values downstream of it, supporting the conjecture that the beginnings of introns undergo evolutionary selection for higher TE. (A) Actual and random average TDR values (blue and green, respectively; see Materials and methods) aligned to the beginning of the 5′SS (Exonic Donor) and the first intronic STOP codon (pre-STOP, post-STOP) locations. (B) Mean TDR profile aligned to the beginning of the 5′SS (top) and the first Intronic STOP codon (bottom). As expected, TDR values downstream from the 5′ end of the exon/intron boundary (right side of the 5′SS) are lower than upstream of it (top; $P = 2.04{\cdot}10^{-8}$, Wilcoxon signed-rank test between the exonic and pre-STOP values, per gene). Mean TDR profile of the first Intronic STOP codon demonstrates that the TDR upstream from the first intronic STOP location is higher than the TDR downstream of it (bottom; $P = 4.91{\cdot}10^{-18}$, Wilcoxon signed-rank test between the pre-STOP and post-STOP values, per gene). (C) Mean values over all regions of the actual intronome and randomized version (blue and green, respectively). The STD of the randomized genomes for these regions (light magenta) shows a significant signal in the pre-STOP domain; mean values are 5.239/5.167/4.929 for the actual intronome and 5.236/5.163/4.948 for the randomized intronome (left, middle, and right; respectively). (D) High standard normalization global Z-scores (red, over all intronic 'codons'; see Materials and methods) in the pre-STOP domain and very low global Z-scores in the post-STOP domain support the conjecture that the intronic pre-STOP codon selective pressure hypothesis is true; genes with STOP codons positioned in the downstream exon and locations with <15% of the intronome were ignored.

signal is stronger for highly expressed genes (highly: empirical $P < 1{\cdot}10^{-3}$ and empirical $P < 5{\cdot}10^{-3}$; lowly: empirical $P > 0.26$ and empirical $P > 0.68$, *S. cerevisiae* and *S. pombe*, respectively; Supplementary **Fig. S2C and D**).

Together, these results support the conjecture that introns are selected for a STOP codon close to the 5′SS in order to decrease the length and corresponding metabolic costs of undesired translated peptides in the case of intron retention, presumably to reduce the global cost of translating such peptides; not surprisingly, highly expressed genes show evidence of significantly higher levels of selection for these intronic features.

Evidence of selective pressure for elevated intronic TE *near the 5′SS and upstream from the first STOP codon presumably to decrease translation costs and improve* TE *of retained introns*

It was suggested that selection for certain synonymous codons improves TE (in terms of rate and fidelity), reduces the cost of the translation process (e.g. via improving ribosomal allocation), and affects organismal fitness.[77–81] Thus, a preference for codons that are translated in a more efficient manner may possibly reduce translational costs of retained introns, even if their product is not functional or entirely functional. Therefore, if a relatively large fraction of the

intronome is occasionally retained, we can expect to see preference for codons with translation rates or TE relatively similar to typical ORFs in the intronic region upstream from the first intronic STOP codon.[82,83] Hence, and in order to provide evidence supporting this hypothesis, we analysed two measures that estimate the TE or speed of codons: the first is the tAI of a gene (or other genomic sequence), which estimates its adaptation to the tRNA pool.[84] In the case of *S. cerevisiae*, we used an additional measure: the TDR (see Materials and methods), which is expected to be more accurate since it incorporates direct experimental RP data; [63,64] TDR is currently not available for the other studied organisms. To this end, the tAI/TDR profiles of all intron-containing genes were computed both in the exonic and the intronic regions (see details in Materials and methods and Supplementary **Fig. S3**). The profiles were aligned around: (i) the 5′SS of the introns and (ii) the first intronic STOP codon, and a mean profile was generated. A summary of the analysis for *S. cerevisiae* is presented in **Fig. 3**. As anticipated, the average TDR values in the exonic regions upstream of the 5′SS were higher than in the intronic regions downstream from it (**Fig. 3A and B**; $P = 2.04 \cdot 10^{-8}$; Wilcoxon signed-rank test between the upstream and downstream values per gene). Furthermore, the average TDR values of the intronic sequences tend to be higher before the first intronic STOP codon relatively to the intronic sequence downstream from it ($P = 4.91 \cdot 10^{-18}$, Wilcoxon signed-rank test between the upstream and downstream values per gene). Similarly, when the mean intronic TDR before the first intronic STOP codon was compared with the randomized models it was found to be significantly higher (**Fig. 3C**). Finally, comparison to randomized models shows that the average Z-score value in the ***pre-STOP*** domain is significant, while not significant in the ***post-STOP*** domain (**Fig. 3D**; 4.013 *vs.* −0.078, respectively), supporting the suggested hypothesis that the region before the first intronic STOP codon undergoes evolutionary selection for higher TE (see details on Z-score calculation in Materials and methods). Similar results were obtained in *S. pombe*, *A. nidulans*, and *C. albicans* (using tAI; see Supplementary **Figs S4–S6**, respectively). A summary of the average TDR/tAI and corresponding Z-scores shown in **Fig. 4A and B** demonstrates that in all studied organisms the ***pre-STOP*** domain exhibits higher translation adaptation levels in comparison to the ***post-STOP*** domain; full analysis and statistical information, including average random values and empirical *P*-value information, can be found in Supplementary **Table S1**. Similar results were obtained when using randomized models that separately maintained GC content in the first and second intronic fragments (Supplementary **Fig. S7**); profiles aligned around the 3′SS can be seen in Supplementary **Fig. S13**.

We suggest that the beginning of introns tend to be translated frequently enough to trigger TDR/tAI preference resembling typical fungal ORFs. Specifically, when we compared the average TDR/tAI of the intronic upstream region to all of *S. cerevisiae* and *S. pombe* genes, we found that in these organisms 23.87% and 31.02% of the genes, respectively, have lower values, as can be seen in **Fig. 4C and D**. This result demonstrates that the TDR/tAI at the beginning of introns is similar to the TDR/tAI for typically expressed genes.

Evidence of higher selective pressure on CUB *to improve* TE *at the beginning of introns for introns located in highly translated genes in S. cerevisiae*

In budding yeast, almost all introns were lost during evolution.[85,86] However, the ones that were preserved are known to be generally found in highly expressed genes; moreover, about a third of its intronome originates from genes encoding ribosomal proteins.[37] It is also known that highly expressed genes tend to undergo selection to include codons that are translated more efficiently;[63,64,84,87,88] thus, we also expect higher selection pressure for 'codons' that are translated more efficiently at the 5′ end of introns in highly translated genes. Therefore, in order to determine whether there is a significant difference in the selection for TE between highly and lowly expressed genes at the beginning of introns, intron-containing genes in *S. cerevisiae* were divided into two subgroups: 93 introns originating from the very highly expressed ribosomal genes and 93 introns originating from non-ribosomal genes (arbitrarily selected out of 187 in total to control for the effect of different group sizes; see Materials and methods), and TDR profiles were generated for each group. We further analysed the randomized model that combines randomizations of codons and introns for the aforementioned subgroups to identify evidence of selection based on Z-score profiles calculations. Moreover, we examined the corresponding TDR Z-score profiles of four additional subgroups of the *S. cerevisiae* intronome based on their PA, mRNA levels, RD measurements,[60] and the PPR, which is an estimation of the translation rate (271, 273, 279, and 266 introns, respectively; see Materials and methods). This was accomplished via sorting the intronome according to each measure, and analysing its upper tertiary and lower tertiary, i.e. 85 introns in each subgroup. A summary of the TDR Z-score profiles shown in **Fig. 5A** demonstrates that both ribosomal and non-ribosomal introns exhibit significant selective pressure in the pre-STOP domain ($Z > 3.5$; see Materials and methods), but no substantial selection in the post-STOP domain. Likewise, introns with high PPR/PA/mRNA/RD exhibit higher and significant selective pressure in comparison to introns with low PPR/PA/mRNA/RD, in the pre-STOP domain, and no substantial selection in the post-STOP domain ($|Z| < 1.34$). This suggests a higher level of retained intron adaptation to the translation process in those subgroups (e.g. in the pre-STOP domain $Z = 5.3$ for high PPR *vs.* $Z = -0.99$ for low PPR; see Materials and methods; additional information can be found in Supplementary **Table S2**). Furthermore, the selection signal is better related to PA rather than to mRNA levels ($Z = 4.74$ for high PA *vs.* $Z = 3.79$ for high mRNA): the expected intronic TE selection pressure is related to [*mRNA concentration* • *translational initiation rate per mRNA* • *splicing error rate*], whereas the protein level (PA) is related to [*mRNA concentration* • *translational initiation rate per mRNA*] and is expected to have higher correlation with selection levels.[60] Thus, the fact that the TE signal better correlate with PA rather than with mRNA supports the conjecture that indeed the selection for codons with high TDR at the 5′ end of introns is related to translation and not to pre-translation gene expression stages.

The connection between elevated intronic TE *near the 5′SS and upstream from the first STOP codon and SE measured via a synthetic intron library system*

The aim of the current section is to provide evidence supporting the conjecture that the reported signals at the intronic 5′ are associated with SE. To this end, we analysed previously reported measurements of different introns transformed into a Yellow Fluorescence intron Library (YiFP) as a system for estimating SE in *S. cerevisiae* (Ref. 74; higher YiFP is related to higher SE; see **Fig. 5B**, Materials and methods, and Supplementary **Note S2** for additional details). First, we extracted the intronic SE library related measurements of 215 introns. Next, we looked at the TDR and TDR Z-score profiles of two subgroups of endogenous introns in correspondence to their YiFP expression levels: 80 genes with the highest levels and 80

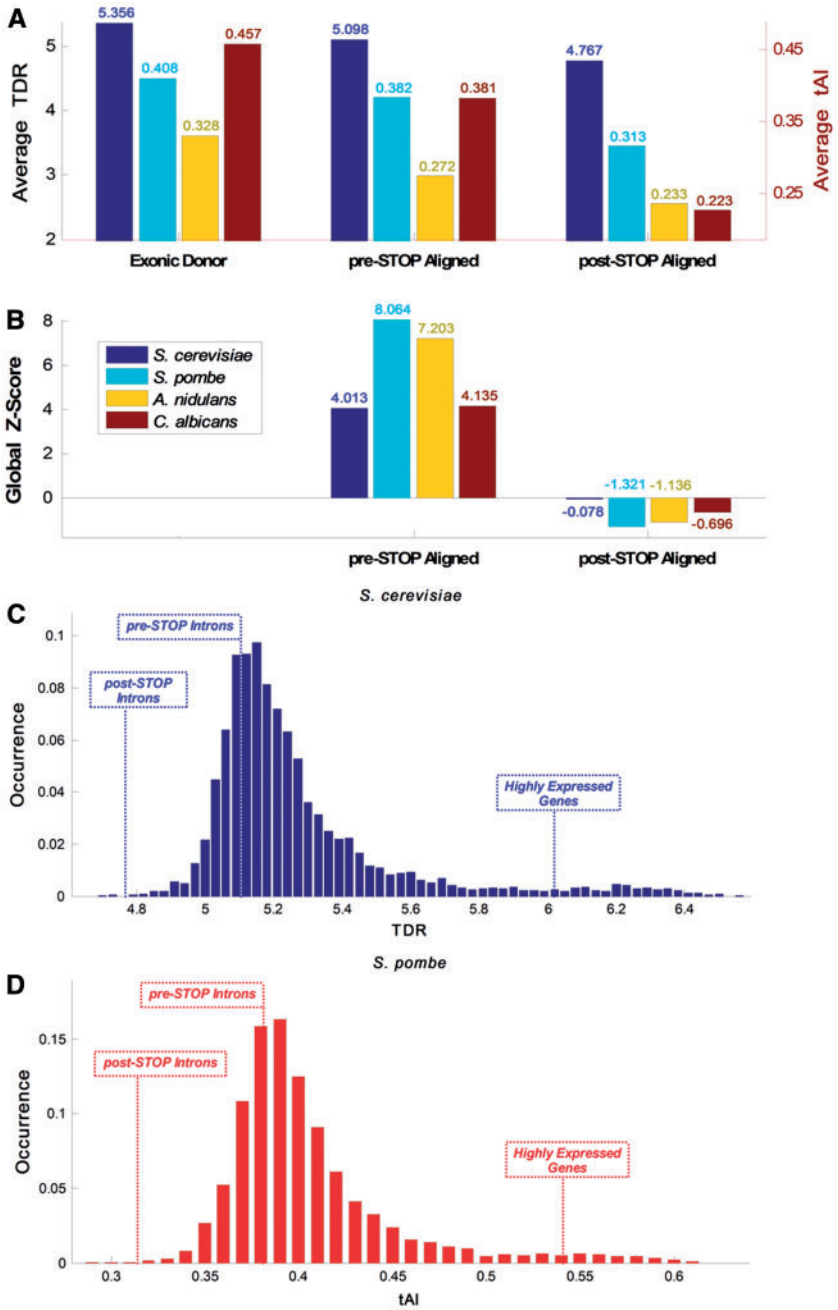

Figure 4. CUB and translation efficiency summary. (A) Average TDR (*S. cerevisiae*) and tAI (*S. pombe*, *A. nidulans*, and C. *albicans*) in the Exonic Donor, pre-STOP and post-STOP domains of the examined organisms exhibit monotonic decrease in CUB over these three domains. (B) In all of the examined organisms the global Z-score at the pre-STOP domain is >2 whereas the Z-score post-STOP domain is negative, supporting the hypothesis that there is selection for codons similar to some typical ORFs in the pre-STOP but not in the post-STOP domain, probably to optimize translation. (C) Distribution analysis of the *S. cerevisiae* whole genome demonstrate that the average TDR at the beginning of its introns is higher than in 23.87% of the genes (average value of 5.1); downstream from the first intronic STOP codon the values drop to be higher than only 0.07% of the genes (average value of 4.77); in comparison, average values of the top 300 highly expressed genes is 6.17. (D) Distribution analysis of the *S. pombe* whole genome demonstrate that the average tAI at the beginning of its introns is higher than in 31.02% of the genes (average value of 0.382); downstream from the first intronic STOP codon the values drop to be higher than only 0.1% of the genes (average value of 0.313); in comparison, the average value of the top 300 highly expressed genes is 0.542.

genes with the lowest levels; each group accounts for ∼33% of the synthetic library intronome. As can be seen in Fig. 5C, in the pre-STOP domain there is evidence of significant relation between selection for codons with high TE (i.e. TDR) and high SE ($Z=3.49$ for high YiFP *vs*. $Z=-0.29$ for low YiFP; additional information can be found in Supplementary Table S2). This and the previous results demonstrate the coupling between the different gene expression steps and aspects: highly expressed genes are expected to undergo stronger selection for all gene expression aspects, including translation cost/efficiency and splicing cost/efficiency. Specifically, introns of highly expressed genes are expected to be more optimal in terms of their SE and their TE.

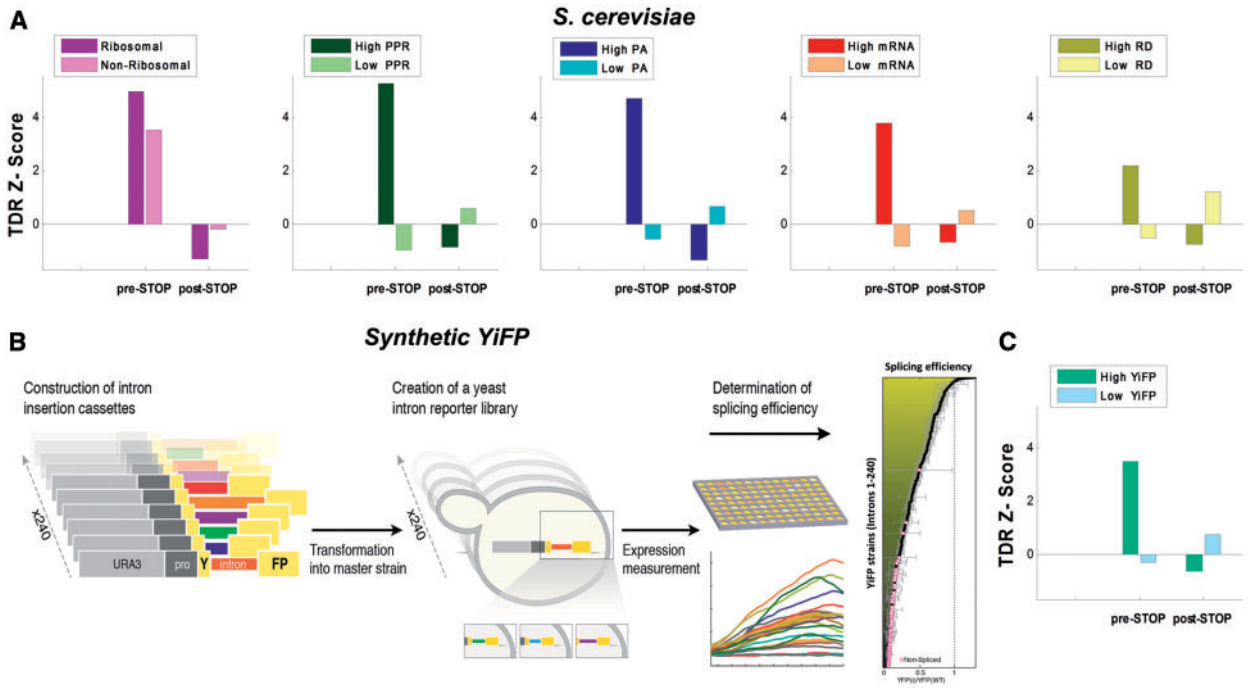

Figure 5. CUB analysis in various subgroups of *S. cerevisiae*. (A) Average TDR Z-scores for various intronome subgroups examined in *S. cerevisiae*: ribosomal *vs.* non-ribosomal, high levels of PPR *vs.* low levels of PPR, high PA *vs.* low PA, high mRNA levels *vs.* low mRNA levels, and high RD *vs.* low RD.[60] Introns originating from highly translated genes (e.g. high PPR/PA; $Z > 4.74$) exhibit higher TDR Z-score values than introns originating from lowly translated genes (e.g. low PPR/PA; $|Z|<0.99$) in the pre-STOP domain (see Materials and methods); results in the post-STOP domain were not significant ($|Z|<1.34$); see full information in Supplementary **Table S2**. (B and C) Standardized reporter approach for studying SE in *S. cerevisiae*: (B) Overview of a previously reported approach for studying splicing mediated gene expression of a reported gene as an assessment of SE (see details in Materials and methods and Ref. 74). (C) Average TDR Z-scores for high/low levels of SE in the synthetic library: results were significant in the pre-STOP domain for highly spliced genes ($Z = 3.49$), but not significant in the post-STOP domain ($|Z|<0.74$; see full information in Supplementary **Table S2**). These results support the hypothesis that there is higher selective pressure for TE in introns originating from highly expressed/translated and highly spliced genes (*vs.* introns originating from lowly expressed/translated and lowly spliced genes); thus, they are consistent with the accepted hypothesis that highly expressed genes are generally more adapted to the tRNA pool, the translation process, and to various gene expression steps in general.[62,81,87,88]

Evidence that selective pressure on TE *near the 5′SS and upstream from the first STOP codon in S. pombe is not function specific*

An intriguing question is whether the previously shown intronic translation features are not related to the regulation of specific functional gene groups, and can be detected in various gene groups. To this end, we used GO and analysed the retained peptide length, tAI profiles, and tAI Z-scores mentioned in the previous sections in *S. pombe* for 90 different GO terms separately (see Materials and methods). As can be seen in **Fig. 6** for genes of various biological processes, analysis of the STOP codon distance from the 5′SS demonstrates that the average retained peptide length is lower in 61% of the gene functions for the actual intronome in comparison to the randomized ones (**Fig. 6A;** 79%/76% for molecular function and cellular component, respectively; Supplementary **Fig. S9**; 22% of the terms exhibit selection levels with empirical *P*-value $P < 0.1$; **Fig. 6B and C;** 29% for both molecular function and cellular component, Supplementary **Fig. S9**). In addition, and consistent with previous results, the highest tAI values are found upstream from the 5′SS in most cases; naturally and as expected, the tAI values drop in the pre-STOP domain, however, these values are significant higher than the values in the post-STOP domain (which have the smallest tAI values). This result suggests a partial adaptation to the translation process at the pre-STOP domain and minimum or no adaptation to the translation process at the post-STOP domain. In addition, the Z-score profiles in the pre-STOP domain exhibit higher selection levels for most cases (**Fig. 6D and E**), supporting the conjecture above. Finally, 71%/61%/85% of the gene functions in biological process, molecular function, and cellular component, respectively, show significantly higher values in comparison of the pre-STOP and post-STOP domains ($P < 0.05$ for all cases, Wilcoxon rank-sum test; results for molecular function and cellular component can be seen in Supplementary **Fig. S9**; see details in Materials and methods, and also a list of the included and excluded terms in Supplementary **Table S3**). Altogether, these results suggest that the signal reported in the previous sections is not function specific or limited to a small group of genes with specific function, but rather tends to appear in genes with various functions.

RP *analysis in budding yeast demonstrates higher ribosome density at the beginning of introns compared to downstream from the first intronic STOP codon*

To further demonstrate that the beginning of introns tends to be partially but considerably translated, we utilized *S. cerevisiae* RP experimental information. The RP method is currently one of the most promising approaches for studying gene translation; [60] it provides a quantitative measure of ribosome translation status for every nucleotide in the genome at any given moment (see details in **Fig. 7A–D**

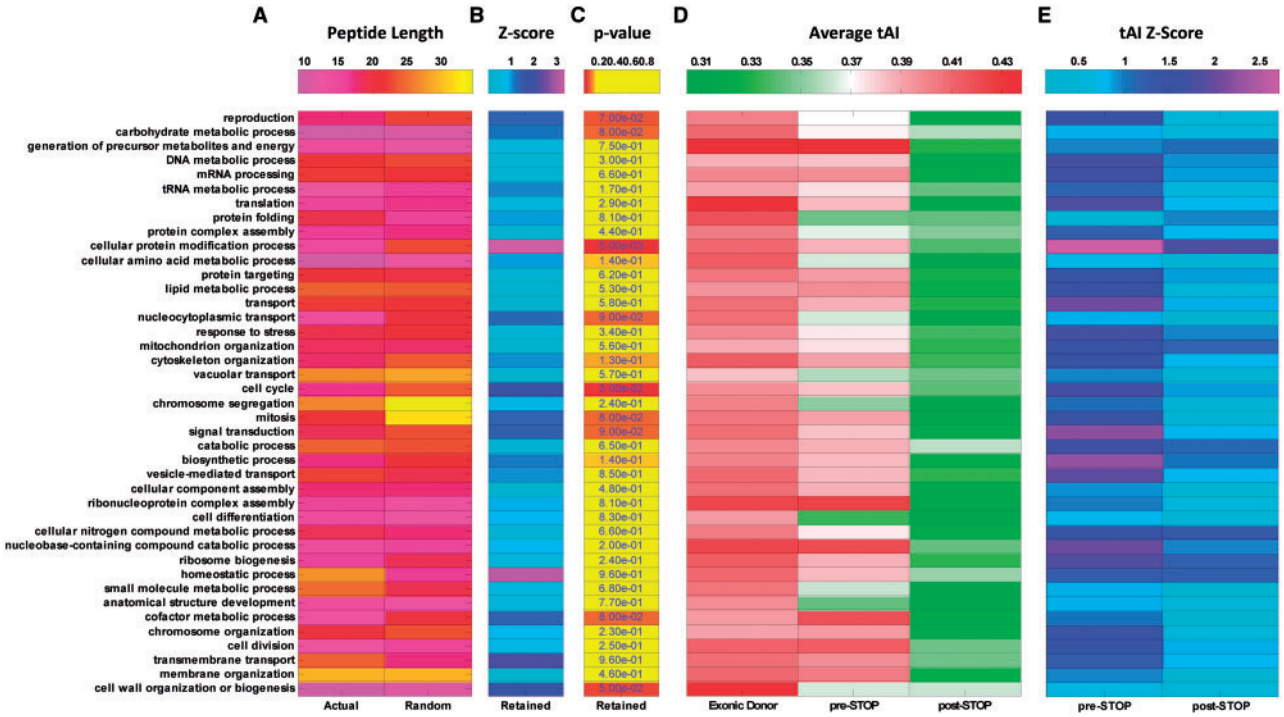

Figure 6. GO terms analysis in *S. pombe* reveals that retained introns translation preference is not unique to genes with a very specific function. (A–□C) Summary of retained peptide length (A) and Z-score corresponding to peptide length relatively to randomized models (B) for biological process demonstrates that in 61% of the gene groups there is preference for a STOP codon closer to the 5′SS. (C) Statistical analysis indicates that the empirical *P*-value related to the observed signal is <0.1 in 22% of the examined gene groups (see Materials and methods). (D and 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 biological process 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 71% of the examined gene groups ($P < 0.05$ for all cases, Wilcoxon rank-sum test; see Materials and methods). Thus, the reported signal is universal and not related to any one functional gene group.

and in the Materials and methods). In order to estimate the translation levels of retained introns we compared the RD before and after the first intronic STOP codon. Specifically, the analysis was performed based on the RP footprint RC information in these regions in a per nucleotide resolution, through alignments around the 5′SS and first intronic STOP codon, averaging over the entire intronome (see Materials and methods and Supplementary **Fig. S10**). As expected, the mean RC values in the exons upstream of the 5′SS are around two orders of magnitude higher than the values at the intronic downstream side (Supplementary **Fig. S11**). Analysis of the region surrounding the first intronic STOP codon shows that mean RC values upstream of the termination point were 2.56 times significantly higher than those downstream from it (**Fig. 7E**; mean RC of $6.39 \cdot 10^{-2}$ before *vs.* $2.54 \cdot 10^{-2}$ after the first intronic STOP codon; $P = 1.42 \cdot 10^{-9}$, Wilcoxon signed-rank test between the upstream and downstream values per gene; see also Supplementary **Table S1**). Interestingly, analysis of the mean RC in the 3′ UTR ($3.47 \cdot 10^{-2}$; after the annotated STOP and the end of the coding region), found similar levels to the ones detected in the intronic post-STOP domain ($2.54 \cdot 10^{-2}$). This result supports the conjecture that the intronic STOP codons function in a similar manner as the annotated STOP, in the case where introns are properly spliced; thus, the level of RC detected at the intronic post-STOP domain is similar to the 'background noise' of the RP method. Furthermore, examination of RD in alignment to the second intronic STOP codon does not show any significant difference (**Fig. 7F**; mean RC of $2.47 \cdot 10^{-2}$ before *vs.* $2.35 \cdot 10^{-2}$ after the second intronic STOP codon; $P = 0.21$, Wilcoxon signed-rank test between the upstream and downstream values per gene; see Materials and methods), further supporting the conjecture that indeed the observed signal near the first intronic STOP codon is related to genuine intronic translation.

Further, we evaluated the mean TDR at the pre-STOP domain assuming that due to splicing error rate consequently 1% of the introns are translated (e.g. see Ref. 89) To this end, we compared the mean TDR at the pre-STOP domain of the introns to the mean TDR of ORFs with RC that are $[0.01 \pm 0.005] \cdot$[mean RC of ORFs with introns] and found them to be very similar (5.165 *vs.* 5.148; in comparison the average and STD of the mean TDR over all annotated ORFs are 5.27 and 0.274, respectively). We also compared the mean TDR at the pre-STOP of the introns to the mean TDR of ORFs with similar RC (±10%), to the RC that appears in the pre-STOP domain and found them to be very similar (5.165 *vs.* 5.231); see also Supplementary **Note S2**.

Finally, as can be seen in **Fig. 7G**, analysis of the whole *S. cerevisiae* genome demonstrates that 40.8% of the genes have lower mean RC values than introns in the pre-STOP domain (the relatively high number can be explained by the fact that genes with introns are extremely highly expressed in *S. cerevisiae*). These findings are in agreement with our previous results, as they support the conjecture that the intronic regions upstream of the first intronic STOP undergo significantly higher translation than the intronic regions downstream of it. In addition, the intronic translation rate is comparable to

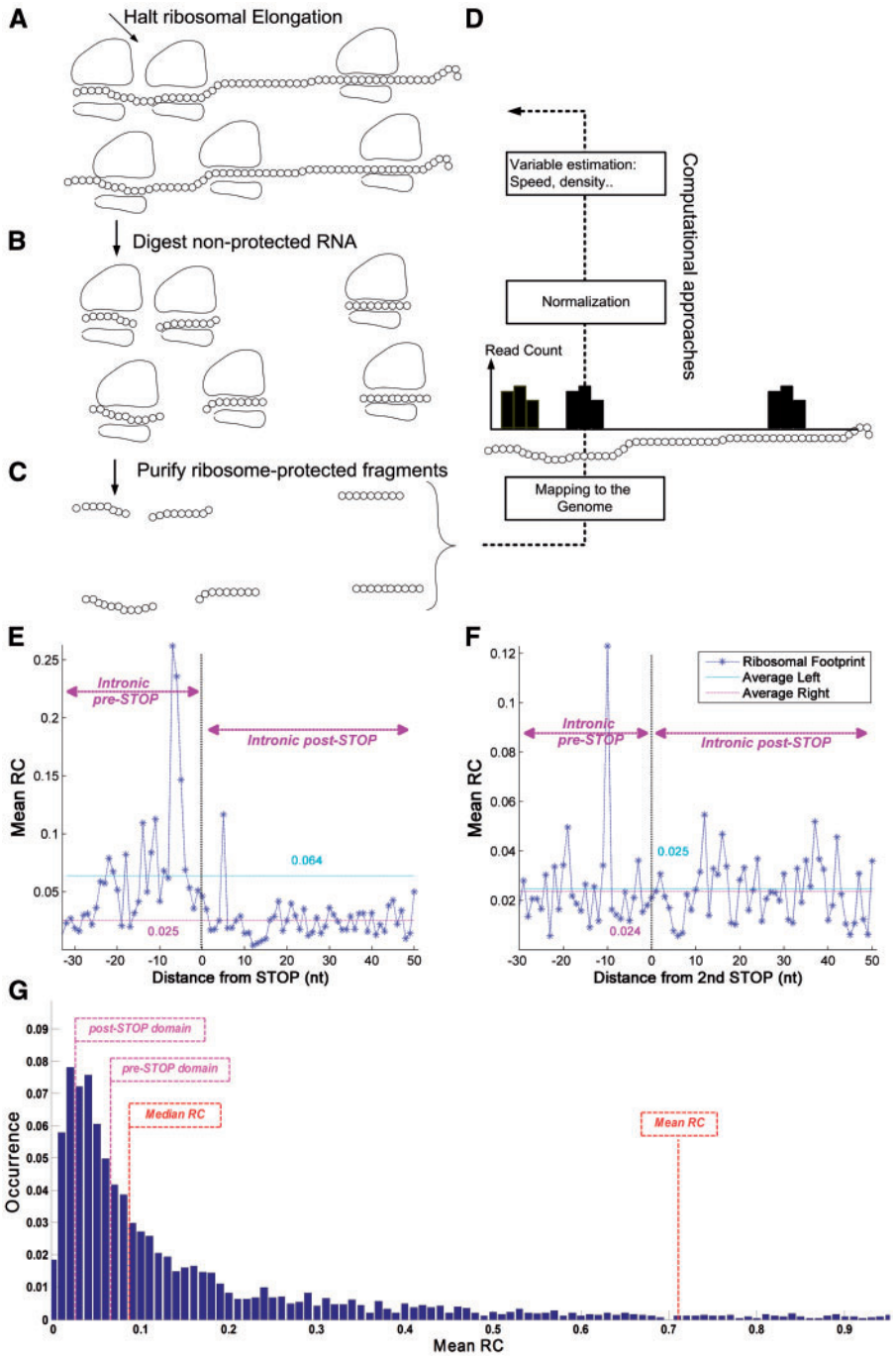

Figure 7. RP analysis of *S. cerevisiae*. The RP method gives quantitative information of RD in a single nucleotide resolution. (A) Cells are treated (for example) with cycloheximide to arrest translating ribosomes; (B and C) RNA fragments protected from RNases by the ribosome are isolated and processed for Illumina high-throughput sequencing; (D) Based on the sequence reads, it is possible to computationally infer various biophysical properties related to the translation-elongation process. Ribosomal footprint RCs of a certain codon are generated when the codon is covered by ribosomes. Thus, highly translated genes tend to create a higher number of reads. (E) The mean RC values upstream from the first intronic STOP codon (pre-STOP side, left of the Termination Point) shows higher values compared with downstream mean RC values (post-STOP side, right of the Termination Point; 6.39×10^{-2} before *vs.* 2.54×10^{-2} after the first intronic STOP codon; $P = 1.42 \times 10^{-9}$, Wilcoxon signed-rank test). (F) The mean RC profile shows no significant difference between the intronic pre-STOP and post-STOP domains (2.47×10^{-2} before *vs.* 2.35×10^{-2} after second intronic STOP codon; $P = 0.21$, Wilcoxon signed-rank test); RC A-site offset is 15 nt. (G) RC distribution analysis of the *S. cerevisiae* genome demonstrates that as much as 40.8% of the yeast genes have lower mean RC values than introns in the pre-STOP domain, while merely 15.77% of the genes have lower mean RC values than introns in the post-STOP domain (see details in Materials and methods).

translation rates of known ORFs. Thus, these results support the conjecture that the reported intronic sequence features are indeed related to translation.

The reported results from the population-genetics point of view

The aim of this section is to demonstrate that the reported results are supported by population-genetics considerations. We will start with the conjecture that the 5′ end of introns undergo selection for 'codons' with better TE, and continue with the conjecture that there is selection for an intronic STOP codon in-frame with the upstream exon and close to the intron 5′ end. It was estimated in yeast[78] that the CUB selective coefficient is $S=2.6$ for highly expressed genes (100 copies of the mRNA). This means that the expected frequency P of the optimal codon is: $P=\left(1+\frac{u}{v}e^{-S}\right)^{-1}$, where u and v are the mutation rates from c_1 to c_2 and from c_2 to c_1, respectively (c_1 and c_2 are two synonymous codons of the same AA). In addition, $S=4{\cdot}Ne{\cdot}s$, where Ne is the effective population size and s is the selective pressure. If we assume that $u=v$ we get $P=\left(1+e^{-S}\right)^{-1}=0.93$, and since in *S. cerevisiae* $Ne=10^7$ (see in Ref. 78) we get a selective pressure of $s=6.5{\cdot}10^{-8}$ for highly expressed genes.

In Ref. 78 they also provide a relation between expression levels and the selection coefficient (see Fig. 2 in this article), and suggest that 'genes expressed at over one transcript per cell show noticeable, albeit small, signs of selection on codon-usage'. Since ribosomal proteins have the highest expression levels in the cell (e.g. Y over 100 copies), we expect that an error rate of a few percentages x will be comparable to the selection coefficient for codons in genes with expression levels $x{\cdot}y$. As can be seen in Ref. 78, for $x\sim 5\%$ (similarly to the RC observed at the intronic 5′ end of *S. cerevisiae*) the estimated S is around 0.5, and thus the corresponding selective pressure is around $s=1.25\times10^{-8}$. According to them, this S value is close to the ones estimated for genes in *H. Sapiens*, *Encephalitozoon cuniculi*, and *Plasmodium falciparum*, and is higher than the S value of *Mus musculus*, where selection for CUB was observed (i.e. the S value is significantly higher than 0 according to Ref. 78).

Regarding the second signal, the conclusion is similar: The selection coefficient for a close STOP codon is expected to be higher than for CUB due to the following reasons: (i) Based on TDR/tAI/Metabolic-cost in *S. cerevisiae* the mean TE or metabolic cost difference between all pairs of (different) codons is 1.4/0.299/19.78, respectively, while the mean improvement in terms of translation rate/metabolic cost when removing a codon is expected to be equal to the mean translation rate/metabolic cost of one codon, which is 5.22/0.387/29.08, respectively. (ii) Mutations that decrease the distance to a STOP codon can decrease it by more than one codon.

The relation of the reported results to efficient NMD regulation

The fact that the reported signals correspond to a STOP codon close to the intronic 5′SS may also be (partially) related to selection for improved NMD. Specifically, recent studies suggested that longer 3′ UTRs following PTC is one factor related to the improvement of NMD efficiency.[90–93] Nevertheless, various lines of evidence support the conjecture that the reported result is at least partially related to translation rather than NMD:

(i) According to our analyses of reduced metabolic costs and TE adaption in the beginning of introns, the signal of more 'efficient' amino acids and codons before the first intronic STOP is expected to be related to translation and is not expected to be related to NMD.

(ii) The average distance between the PTC and the annotated STOP codons is 934.41 nt for *S. cerevisiae* and 932.36 for *Sch. pombe*. These results indeed detect a significant signal related to longer 3′ UTRs that may be related to improved NMD ($P=8.55\times10^{-8}$ and $P=1.25\times10^{-15}$; Wilcoxon signed-rank test between the actual and average randomized models values per gene). However, analyses and comparisons to the randomized models gave an average value of 918.38 nt for *S. cerevisiae* and 926.55 nt for *S. pombe*, i.e. 1.75% and 0.63% higher, respectively. We believe that the relative small change in the 3′ UTR length is not expected to affect the NMD efficiency (in comparison, evaluation of the ratios in the case of average truncated translated intronic length gave 36.78% and 7.95%, respectively; see **Fig. 2B and C**).

(iii) The definition of what is considered a 'long' 3′ UTR is somewhat ambiguous and there is no agreement on the relation between the 3′ UTR length and the NMD pathway efficiency. Thus, the mechanism by which PTCs are detected and promote assembly of the UPF-SMG complex is still poorly understood, though seems to be related to the 3′ UTR composition, and the relation between the PTC recognition and the actual mRNA decay is not fully understood.[28,29,94,95]

(iv) If NMD is the dominant mechanism we expect to see a stronger signal of an intronic STOP near intronic 5′SS in introns further away from the beginning of the coding region; in these introns the increased ratio of the 3′ UTR due to a closer intron is expected to be stronger, promoting stronger selection. However, comparison of the average retained peptide length and metabolic costs in *S. pombe* between the first intron in each gene to the rest of the introns (2,269 *vs.* 2,478 introns, respectively) showed that the last ones are 25% longer and that the average metabolic cost in the first ones is 22% lower ($P=1.94\times10^{-6}$ and $P=3.11\times10^{-7}$, respectively; Wilcoxon rank-sum test; Supplementary **Fig. S15**). This result demonstrates that the signal is significantly stronger for the first introns in comparison to the rest, supporting the conjecture that NMD is not the most dominant factor here.

(v) Finally, we would like to emphasize that as in many cases reported in the past, a signal or phenomena may be related to multiple rules or reasons.[95] Thus, it is clearly possible that both explanations (translation and NMD) are related to the suggested selection for intronic STOP close to the 5′SS.

Discussion

Intron retention is seldom found in fungi, and has not been reported in the organisms analysed in this study. However, here, we analyse four fungal genomes and report various novel evidence supporting the conjecture that novel sequence features found near the intronic splice sites are under selective pressure, presumably to reduce the translation costs of retained introns. Specifically, we report the following major results that are observed at the intronome level: (i) We demonstrate that the intronic sequences are selected for a close STOP codon near the 5′SS, presumably to reduce the cost of translating undesired peptides encoded in retained introns. (ii) In association with this outcome, the beginning of introns are selected for codons that improve TE, which resemble the ones that appear in annotated fungal ORFs, and thus probably undergo co-evolution with the tRNA pool. Again, this signal probably improves the fitness of the organism via reducing the translation and metabolic costs of peptides encoded in retained introns. In all of these cases the reported signals were compared to the ones obtained in randomized genomes and

were shown to be significantly stronger, supporting the conjecture that they are indeed under selective pressure. (iii) Our analyses demonstrate that in *S. cerevisiae* there is higher RD at the beginning of introns, before the first intronic STOP codon, rather than downstream of it, supporting the conjecture that this region tends to undergo partial translation. (iv) The reported patterns occur in genes with various cellular functions supporting the conjecture that they are not function specific. (v) We show that the reported patterns are stronger in highly translated/expressed genes and in introns located in the beginning of the ORF, and specifically related to measures directly associated with TE, supporting the conjecture that they are indeed related to TE selection, and not to alternative signals. (vi) We show that the reported signals do not appear at the intronic 3′ end or in the region surrounding the second intronic STOP codon. (G) Finally, the level of codons' TE at the 5′ pre-STOP intronic domain is very similar to the one appearing in annotated ORFs with similar ribosome densities/expression levels as in those intronic regions.

While there is an ongoing debate regarding the relation between codon-usage and TE based on the analyses of RP data,[63,96–105] it is important to emphasize that the most recent studies clearly demonstrate the relation between codon-usage and translation-elongation.[63,96,100,103–105] The current understanding is that previous studies did not find such a relation due to various problems and biases, including the nature of RP data, inaccurate analysis of the data, low coverage of the data, and more.[63,100,103–106]

The reported signals are observed at the genomic level, but are harder to detect for single introns due to the nature of the studied signals (see further details in the Materials and methods). Therefore, a comparison of the entire intronome to its randomized versions should enable evaluating such a selection. Likewise, other important recent studies in the field have performed analogous analyses.[65,107–111] We believe that the relatively rare translation events accumulate over all introns to yield a phenomena substantial enough to trigger a selective pressure.

It is important to mention that the reported patterns cannot be completely (or trivially) explained by alternative mechanisms related to splicing instead of translation. For example, it was suggested that in some organisms the intronic boundaries include alternative/tandem splice sites;[112] thus, it is possible that the increased TE and RD at the intronic 5′ end is due to this phenomenon. However, the study did not include Fungi, which are different in terms of their splicing regulation and effective population size than the analysed organisms;[14,113–116] thus, it is not known whether tandem splice sites tend to occur also in Fungi. In addition, the fact that we did not find similar patterns surrounding the second STOP codon or at the intronic 3′SS (Fig. 7F1 and Supplementary S13) supports the conjecture that the reported signals are at least partially related to translation. These results and the fact that we used CUB measures that are directly related to translation (TDR/tAI) support the conjecture that the reported CUB at the intronic 5′ end cannot be fully explained via non-adaptive mechanisms or aspects not related to translation (e.g. minimizing splicing errors or nucleotide synthesis costs). Furthermore, by any means the selection for STOP codons near intronic 5′ end cannot be related to tandem splice site signals.

Some studies have suggested that the process of intron splicing is quite efficient (the probability of splicing error is ∼1%), that introns in highly expressed genes are spliced more accurately, and that splicing errors trigger NMD which trigger degradation of the mRNA.[89,117] However, only few genes were examined and only in mammals, and it is not clear whether the results are consistent in different conditions or tissues. Since there are changes at the range and order of magnitude in various aspects of gene expression,[118–120] we expect that also in terms of SE the differences may be very high (up to several orders of magnitudes). Here, because many of the analysed genes are highly expressed it is possible that even a translation ratio of 1% (that usually results in NMD) may result in substantial selection pressure that will enable the TE adaptation. Additionally, since the NMD pathway is based on ribosome halting at the premature STOP codon,[121] when considering very large numbers of introns it will 'cost' less if the STOP is be closer to the 5′ end of the intron. Moreover, our analyses demonstrate that the RD at the *S. cerevisiae* intronic pre-STOP 5′ end (but not in other parts of the introns) is in agreement with the ribosomal densities in a large fraction of the annotated ORFs in this organism; we also show that the mean TDR at the intronic pre-STOP 5′ end is very similar to the one obtained for annotated ORFs with ribosomal densities and expression levels (when considering the 1% splicing error rate) at the intronic pre-STOP 5′ end. Thus, our results make sense and are in very good agreement also with the experimental evidence suggesting that the splicing error rate is 1% (see also Supplementary **Note S1**). An intriguing question for further research is whether there is a difference between wild type and NMD mutants with regard to ribosome footprinting patterns and translation of retained introns.

One may also wonder why evolution does not optimize splicing such that there will be no splicing errors at all, and therefore no need for translation optimization of introns; specifically, this solution may sound energetically cheaper and more 'elegant'. A possible explanation to this "quandary" includes the fact that 'evolutionary selection' works more like a 'tinkerer' (mutation → selection → mutation → selection, etc.) that converge to local solutions rather than like an 'engineer' (that can find the globally optimal solution), due to various constraints and multiple variables involved in splicing and translation. Therefore, in our opinion it should not be surprising that evolution will 'converge' to the suggested 'solution'.[106,122–124] It is easy to see that various aspects in biological systems are not optimal or can be further optimized. Furthermore, the higher splicing error rate may not 'be solved' (in a cheaper manner than the current 'solution') due to biophysical constraints and limitations. For example, it is possible that there is a trade-off(s) (as in the case of many other aspects of gene expression) between the 'speed' of splicing and its accuracy (a lower error rate may be related to lower splicing 'speed', which may eventually be more deleterious than the current state). This fact could be the 'selective force' facilitating the partial codon-usage/translation cost optimization.

Specifically, we suggest that the improved translation cost of introns contributes towards better organismal fitness. Indeed, translating a larger amount of useless proteins from the aberrantly spliced mRNAs is not expected to be beneficial; however, translating it in a more efficient manner is 'better'/'cheaper' than translating it using non-efficient codons; on the other hand, completely eliminating these cases is also impossible due to possible higher (in terms of organismal fitness) negative SE or other/additional costs (and the nature of the evolutionary process which tend to converges to local optima).

Conclusions

One central conclusion of this study is that the classification to exons and introns in fungi is continuous rather than discrete, i.e. introns may also be translated to a certain degree. Moreover, it could make sense to classify many of the fungal intron-containing transcripts as alternatively spliced; while in many cases introns are spliced as

defined, there are considerably abundant examples (enough to trigger translational preference on introns) where an intron is retained and subsequently a shorter peptide or protein is translated. Therefore, it will be interesting to study various properties of these putative shorter proteins.

Another principle conclusion is that functional silent mutations related to translation may appear in regions traditionally not classified as translated regions (e.g. introns). Our results support various recent studies in the field that have suggested that synonymous and silent mutations are functional.[83,125–127] However, we emphasize for the first time, the role of CUB also in non-coding parts of the transcript such as introns.

Insights from the results reported here can be used for developing future models of intronic evolution that will consider the effect of mutations both on SE and the possibility of intronic translation.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. Funding to pay the Open Access publication charges for this article was provided by Tel-Aviv university and Minerva ARCHES award.

References

1. Nilsen, T.W. 2003, The spliceosome: the most complex macromolecular machine in the cell? *BioEssays*, **25**, 1147–9.
2. Hoskins, A.A. and Moore, M.J. 2012, The spliceosome: a flexible, reversible macromolecular machine. *Trends Biochem. Sci*., **37**, 179–88.
3. Le Hir, H., Nott, A. and Moore, M.J. 2003, How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci*., **28**, 215–20.
4. Wang, G.S. and Cooper, T.A. 2007, Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet*., **8**, 749–61.
5. Kramer, A. 1996, The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu. Rev. Biochem*., **65**, 367–409.
6. McKee, A.E. and Silver, P.A. 2007, Systems perspectives on mRNA processing. *Cell Res*., **17**, 581–90.
7. Toor, N., Keating, K.S., Taylor, S.D. and Pyle, A.M. 2008, Crystal structure of a self-spliced group II intron. *Science*, **320**, 77–82.
8. Wahl, M.C., Will, C.L. and Lührmann, R. 2009, The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701–18.
9. Rodríguez-Trelles, F., Tarrío, R. and Ayala, F.J. 2006, Origins and evolution of spliceosomal introns. *Annu. Rev. Genet*., **40**, 47–76.
10. Rogozin, I., Carmel, L., Csuros, M. and Koonin, E. 2012, Origin and evolution of spliceosomal introns. *Biol. Direct*, **7**, 11.
11. Maniatis, T. and Tasic, B. 2002, Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–43.
12. Black, D.L. 2003, Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem*., **72**, 291–336.
13. Wang, B.B. and Brendel, V. 2006, Genomewide comparative analysis of alternative splicing in plants. *Proc. Natl. Acad. Sci. USA*., **103**, 7175–80.
14. Kim, E., Magen, A. and Ast, G. 2007, Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*., **35**, 125–31.
15. Juneau, K., Nislow, C. and Davis, R.W. 2009, Alternative splicing of PTC7 in *Saccharomyces cerevisiae* determines protein localization. *Genetics*, **183**, 185–94.
16. Rhind, N., Chen, Z., Yassour, M., et al. 2011, Comparative functional genomics of the fission yeasts. *Science*, **332**, 930–6.
17. Marshall, A.N., Montealegre, M.C., Jiménez-López, C., Lorenz, M.C. and van Hoof, A. 2013, Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. *PLoS Genet*., **9**, e1003376.
18. Ast, G. 2004, How did alternative splicing evolve? *Nat. Rev. Genet*., **5**, 773–82.
19. Ner-Gaon, H., Halachmi, R., Savaldi-Goldstein, S., Rubin, E., Ophir, R. and Fluhr, R. 2004, Intron retention is a major phenomenon in alternative splicing in *Arabidopsis*. *Plant J*., **39**, 877–85.
20. Kim, E., Goren, A. and Ast, G. 2008, Alternative splicing: current perspectives. *BioEssays*, **30**, 38–47.
21. Barbazuk, W.B., Fu, Y. and McGinnis, K.M. 2008, Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res*., **18**, 1381–92.
22. Keren, H., Lev-Maor, G. and Ast, G. 2010, Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet*., **11**, 345–55.
23. McManus, C.J. and Graveley, B.R. 2011, RNA structure and the mechanisms of alternative splicing. *Curr. Opin. Genet. Dev*., **21**, 373–9.
24. Jaillon, O., Bouhouche, K., Gout, J.F., et al. 2008, Translational control of intron splicing in eukaryotes. *Nature*, **451**, 359–62.
25. Conti, E. and Izaurralde, E. 2005, Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol*., **17**, 316–25.
26. Rodríguez-Gabriel, M.A., Watt, S., Bähler, J. and Russell, P. 2006, Upf1, an RNA helicase required for nonsense-mediated mRNA decay, modulates the transcriptional response to oxidative stress in fission yeast. *Mol. Cellular Biol*., **26**, 6347–56.
27. Morozov, I.Y., Negrete-Urtasun, S., Tilburn, J., Jansen, C.A., Caddick, M.X. and Arst, H.N. 2006, Nonsense-mediated mRNA decay mutation in *Aspergillus nidulans*. *Eukaryot. Cell*, **5**, 1838–46.
28. Brogna, S. and Wen, J. 2009, Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol*., **16**, 107–13.
29. Peccarelli, M. and Kebaara, B.W. 2014, Regulation of natural mRNAs by the nonsense-mediated mRNA decay pathway. *Eukaryot. Cell*, **13**, 1126–35.
30. Kriventseva, E.V. and Gelfand, M.S. 1999, Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. *J. Biomol. Struct. Dyn*., **17**, 281–8.
31. Lim, L.P. and Burge, C.B. 2001, A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci. USA*., **98**, 11193–8.
32. Yu, J., Yang, Z., Kibukawa, M., Paddock, M., Passey, D.A. and Wong, G.K.S. 2002, Minimal introns are not "junk". *Genome Res*., **12**, 1185–9.
33. Alexander, K., Anatoliy, I. and Alexander, B. 2011, Statistical analysis of exon lengths in various eukaryotes. *Open Access Bioinformatics*, **2011**, 1–15.
34. Spingola, M., Grate, L., Haussler, D. and Ares, M. 1999, Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA*, **5**, 221–34.
35. Grate, L. and Ares Jr, M. 2002, Searching yeast intron data at Ares lab web site. In: Christine, G. and Gerald, R.F. (eds), *Methods in Enzymology*, pp. 380–92. Academic Press, imprint of Elsevier Science: San Diego, CA.
36. Mishra, S.K., Ammon, T., Popowicz, G.M., et al. 2011, Role of the ubiquitin-like protein Hub1 in splice-site usage and alternative splicing. *Nature*, **474**, 173–8.
37. Ares, M., Grate, L. and Pauling, M.H. 1999, A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA*, **5**, 1138–9.
38. Juneau, K., Miranda, M., Hillenmeyer, M.E., Nislow, C. and Davis, R.W. 2006, Introns regulate RNA and protein abundance in yeast. *Genetics*, **174**, 511–8.
39. Wood, V., Gwilliam, R., Rajandream, M.A., et al. 2002, The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–80.

40. Habara, Y., Urushiyama, S., Tani, T. and Ohshima, Y. 1998, The fission yeast prp10+ gene involved in pre-mRNA splicing encodes a homologue of highly conserved splicing factor, SAP155. *Nucleic Acids Res*., **26**, 5662–9.
41. Okazaki, K. and Niwa, O. 2000, mRNAs encoding zinc finger protein isoforms are expressed by alternative splicing of an in-frame intron in fission yeast. *DNA Res*., **7**, 27–30.
42. Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., et al. 2004, Introns and splicing elements of five diverse fungi. *Eukaryot. Cell*, **3**, 1088–100.
43. Mitrovich, Q.M., Tuch, B.B., Guthrie, C. and Johnson, A.D. 2007, Computational and experimental approaches double the number of known introns in the pathogenic yeast *Candida albicans*. *Genome Res*., **17**, 492–502.
44. Stajich, J., Dietrich, F. and Roy, S. 2007, Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol*., **8**, R223.
45. Carmel, L., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. 2007, Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol. Biol*., **7**, 192.
46. Collins, L. and Penny, D. 2005, Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol*., **22**, 1053–66.
47. Lane, C.E., van den Heuvel, K., Kozera, C., et al. 2007, Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proc. Natl. Acad. Sci. USA*., **104**, 19908–13.
48. Warf, M.B. and Berglund, J.A. 2010, Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci*., **35**, 169–78.
49. Parenteau, J., Durand, M., Morin, G., et al. 2011, Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell*, **147**, 320–31.
50. Berbee, M.L. and Taylor, J.W. 2001, Fungal molecular evolution: gene trees and geologic time. *Systematics Evolution*, pp. 229–45. Springer-Verlag: Berlin Heidelberg.
51. Berbee, M.L. and Taylor, J.W. 2010, Dating the molecular clock in fungi—how close are we? *Fungal Biol. Rev*., **24**, 1–16.
52. Taylor, J.W. and Berbee, M.L. 2006, Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia*, **98**, 838–49.
53. Cherry, J.M., Adler, C., Ball, C., et al. 1998, SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res*., **26**, 73–9.
54. Wood, V., Harris, M.A., McDowall, M.D., et al. 2012, PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res*., **40**, D695–9.
55. Arnaud, M.B., Cerqueira, G.C., Inglis, D.O., et al. 2012, The *Aspergillus* Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources. *Nucleic Acids Res*., **40**, D653–9.
56. Inglis, D.O., Arnaud, M.B., Binkley, J., et al. 2012, The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. *Nucleic Acids Res*., **40**, D667–D674.
57. Wang, M., Weiss, M., Simonovic, M., et al. 2012, PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics*, **11**, 492–500.
58. Wang, Y., Liu, C.L., Storey, J.D., Tibshirani, R.J., Herschlag, D. and Brown, P.O. 2002, Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. USA*., **99**, 5860–5.
59. Nagalakshmi, U., Wang, Z., Waern, K., et al. 2008, The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–9.
60. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S. and Weissman, J.S. 2009, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–23.
61. Wagner, A. 2005, Energy constraints on the evolution of gene expression. *Mol. Biol. Evol*., **22**, 1365–74.
62. Sharp, P.M. and Li, W.H. 1987, The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*., **15**, 1281–95.
63. Dana, A. and Tuller, T. 2014, The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res*., **42**, 9171–81.
64. Dana, A. and Tuller, T. 2015, Mean of the typical decoding rates: a new translation efficiency index based on the analysis of ribosome profiling data. *G3*, **5**, 73–80.
65. Tuller, T., Waldman, Y.Y., Kupiec, M. and Ruppin, E. 2010, Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. USA*., **107**, 3645–50.
66. Chan, P.P. and Lowe, T.M. 2009, GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res*., **37**, D93–7.
67. Man, O. and Pilpel, Y. 2007, Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet*., **39**, 415–21.
68. Sabi, R. and Tuller, T. 2014, Modelling the efficiency of codon–tRNA interactions based on codon usage bias. *DNA Res*., **21**, 511–26.
69. Edgar, R., Domrachev, M. and Lash, A.E. 2002, Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30**, 207–10.
70. Ingolia, N.T. 2010, Chapter 6—genome-wide translational profiling by ribosome footprinting. In: Jonathan, W., Christine, G. and Gerald, R.F. (eds), *Methods in Enzymology*, Vol. 470, pp. 119–42. Elsevier inc.
71. Kasprzyk, A. 2011, BioMart: driving a paradigm change in biological data management. *Database*, **2011**, bar049.
72. Langmead, B. and Salzberg, S.L. 2012, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*., **9**, 357–9.
73. Ghaemmaghami, S., Huh, W.K., Bower, K., et al. 2003, Global analysis of protein expression in yeast. *Nature*, **425**, 737–41.
74. Yofe, I., Zafrir, Z., Blau, R., et al. 2014, Accurate, model-based tuning of synthetic gene expression using introns in *S. cerevisiae*. *PLoS Genet*., **10**, e1004407.
75. Linshiz, G., Yehezkel, T.B., Kaplan, S., et al. 2008, Recursive construction of perfect DNA molecules from imperfect oligonucleotides. *Mol. Syst. Biol*., **4**, 191–200.
76. Shabi, U., Kaplan, S., Linshiz, G., et al. 2010, Processing DNA molecules as text. *Syst. Synth. Biol*., **4**, 227–36.
77. Kudla, G., Murray, A.W., Tollervey, D. and Plotkin, J.B. 2009, Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, **324**, 255–8.
78. dos Reis, M. and Wernisch, L. 2009, Estimating *Translational Selection in Eukaryotic Genomes*. *Mol. Biol. Evol*., **26**, 451–61.
79. Tuller, T., Carmi, A., Vestsigian, K., et al. 2010, An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–54.
80. Cannarozzi, G., Schraudolph, N.N., Faty, M., et al. 2010, A role for codon order in translation dynamics. *Cell*, **141**, 355–67.
81. Novoa, E.M. and Ribas de Pouplana, L. 2012, Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet*., **28**, 574–81.
82. Hershberg, R. and Petrov, D.A. 2008, Selection on codon bias. *Annu. Rev. Genet*., **42**, 287–99.
83. Plotkin, J.B. and Kudla, G. 2011, Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet*., **12**, 32–42.
84. Reis, M.D., Savva, R. and Wernisch, L. 2004, Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res*., **32**, 5036–44.
85. Jeffares, D.C., Mourier, T. and Penny, D. 2006, The biology of intron gain and loss. *Trends Genet*., **22**, 16–22.
86. Cohen, N.E., Shen, R. and Carmel, L. 2012, The role of reverse transcriptase in intron gain and loss mechanisms. *Mol. Biol. Evol*., **29**, 179–86.
87. Ikemura, T. 1985, Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol*., **2**, 13–34.
88. Kurland, C.G. 1991, Codon bias and gene expression. *FEBS Lett*., **285**, 165–9.
89. Pickrell, J.K., Pai, A.A., Gilad, Y. and Pritchard, J.K. 2010, Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet*., **6**, e1001236.
90. Eberle, A.B., Stalder, L., Mathys, H., Orozco, R.Z. and Mühlemann, O. 2008, Posttranscriptional gene regulation by spatial rearrangement of the 3? Untranslated region. *PLoS Biol*., **6**, e92.

91. Rebbapragada, I. and Lykke-Andersen, J. 2009, Execution of nonsense-mediated mRNA decay: what defines a substrate? *Curr. Opin. Cell Biol.*, **21**, 394–402.
92. Kurosaki, T. and Maquat, L.E. 2013, Rules that govern UPF1 binding to mRNA 3′ UTRs. *Proc. Natl. Acad. Sci. USA.*, **110**, 3357–62.
93. Hurt, J.A., Robertson, A.D. and Burge, C.B. 2013, Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res.*, **23**, 1636–50.
94. Toma, K.G., Rebbapragada, I., Durand, S. and Lykke-Andersen, J. 2015, Identification of elements in human long 3′ UTRs that inhibit nonsense-mediated decay. *RNA*, **21**, 887–97.
95. Hug, N., Longman, D. and Cáceres, J.F. 2016, Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res.*, **44**, 1483–95.
96. Supek, F. and Šmuc, T. 2010, On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*. *Genetics*, **185**, 1129–34.
97. Ingolia, N.T., Lareau, L.F. and Weissman, J.S. 2011, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.
98. Li, G.W., Oh, E. and Weissman, J.S. 2012, The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, **484**, 538–41.
99. Qian, W., Yang, J.-R., Pearson, N.M., Maclean, C. and Zhang, J. 2012, Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.*, **8**, e1002603.
100. Dana, A. and Tuller, T. 2012, Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLoS Comput. Biol.*, **8**, e1002755.
101. Charneski, C.A. and Hurst, L.D. 2013, Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol.*, **11**, e1001508.
102. Artieri, C.G. and Fraser, H.B. 2014, Evolution at two levels of gene expression in yeast. *Genome Res.*, **24**, 411–21.
103. Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S. and Futcher, B. 2014, Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*, **3**.
104. Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B. and Bartel, D.P. 2015, Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep.*, **14**, 1787–99.
105. Ben-Yehezkel, T., Atar, S., Zur, H., et al. 2015, Rationally designed, heterologous *S. cerevisiae* transcripts expose novel expression determinants. *RNA Biol.*, **13**, 972–84.
106. Tuller, T. and Zur, H. 2015, Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Res.*, **43**, 13–28.
107. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. 2002, Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–7.
108. Hershberg, R., Yeger-Lotem, E. and Margalit, H. 2005, Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet.*, **21**, 138–42.
109. Gatbonton, T., Imbesi, M., Nelson, M., et al. 2006, Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet.*, **2**, e35.
110. Gu, W., Zhou, T. and Wilke, C.O. 2010, A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.*, **6**, e1000664.
111. Mi, G., Di, Y., Emerson, S., Cumbie, J.S. and Chang, J.H. 2012, Length bias correction in gene ontology enrichment analysis using logistic regression. *PLoS One*, **7**, e46128.
112. Hiller, M., Szafranski, K., Sinha, R., et al. 2008, Assessing the fraction of short-distance tandem splice sites under purifying selection. *RNA*, **14**, 616–29.
113. Fox, A., Tuch, B. and Chuang, J. 2008, Measuring the prevalence of regional mutation rates: an analysis of silent substitutions in mammals, fungi, and insects. *BMC Evol. Biol.*, **8**, 186.
114. Gossmann, T.I., Keightley, P.D. and Eyre-Walker, A. 2012, The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol. Evol.*, **4**, 658–67.
115. De Conti, L., Baralle, M. and Buratti, E. 2013, Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip. Rev.*, **4**, 49–60.
116. Lanfear, R., Kokko, H. and Eyre-Walker, A. 2014, Population size and the rate of evolution. *Trends Ecol. Evol.*, **29**, 33–41.
117. Fox-Walsh, K.L. and Hertel, K.J. 2009, Splice-site pairing is an intrinsically high fidelity process. *Proc. Natl. Acad. Sci. USA.*, **106**, 1766–71.
118. Ramsköld, D., Wang, E.T., Burge, C.B. and Sandberg, R. 2009, An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, **5**, e1000598.
119. Schwanhausser, B., Busse, D., Li, N., et al. 2011, Global quantification of mammalian gene expression control. *Nature*, **473**, 337–42.
120. Vogel, C. and Marcotte, E.M. 2012, Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.*, **13**, 227–32.
121. Baker, K.E. and Parker, R. 2004, Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Curr. Opin. Cell Biol.*, **16**, 293–9.
122. Mayr, E. 2001, *What Evolution Is*. Basic Books, Perseus Books Group: New York.
123. Jacob, F. 1977, Evolution and tinkering. *Science*, **196**, 1161.
124. Pigliucci, M. and Kaplan, J. 2006, *Making Sense of Evolution: The Conceptual Foundations of Evolutionary Theory*. University of Chicago Press: Chicago.
125. Chamary, J.V., Parmley, J.L. and Hurst, L.D. 2006, Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.*, **7**, 98–108.
126. Sauna, Z.E. and Kimchi-Sarfaty, C. 2011, Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.*, **12**, 683–91.
127. Zur, H. and Tuller, T. 2013, New universal rules of eukaryotic translation initiation fidelity. *PLoS Comput. Biol.*, **9**, e1003136.