Rapidly evolving protointrons in *Saccharomyces* genomes revealed by a hungry spliceosome

Jason Talkish, Haller Igel, Rhonda J. Perriman, Lily Shiue, Sol Katzman[1], Elizabeth M. Munding, Robert Shelansky[1], John Paul Donohue, and Manuel Ares, Jr.*

Center for Molecular Biology of RNA

Department of Molecular, Cell & Developmental Biology

[1]Department of Biomolecular Engineering

University of California, Santa Cruz

Santa Cruz, CA 95064

*Corresponding author

ares@ucsc.edu

(831) 459-4628

Abstract

Introns are a prevalent feature of eukaryotic genomes, yet their origins and contributions to genome function and evolution remain mysterious. In budding yeast, repression of the highly transcribed intron-containing ribosomal protein genes (RPGs) globally increases splicing of non-RPG transcripts through reduced competition for the spliceosome. We show that under these "hungry spliceosome" conditions, splicing occurs at more than 150 previously unannotated locations that do not overlap known introns, called protointrons. Protointrons use a less constrained set of splice sites and branchpoints than standard introns, including in one case AT-AC in place of GT-AG. Protointrons are not conserved in all closely related species, suggesting that they are not maintained by purifying selection. Some protointrons are found in CUTs and SUTs, where they may contribute to the creation of new genes. Others are found across boundaries between noncoding and coding sequences, or lie entirely within coding sequences, where they may contribute to the creation of new protein variants, or new regulatory controls for existing genes. We define protointrons as (1) present in one or at most a few very closely related species, (2) inefficiently spliced, and (3) not clearly understood to contribute to correct expression or regulation of a gene. A very few protointrons in *S. cerevisiae* challenge this classification, suggesting an evolutionary continuum that connects protointrons with standard introns. The snapshot of intron evolution we provide highlights the important role of the spliceosome in the expansion of transcribed genomic sequence space, leading to the birth and refinement of new eukaryotic genes and gene functions.

Introduction

Eukaryotic genes are often split by intervening sequences, or introns, that are removed during and after transcription by the spliceosome and associated splicing proteins. Although much is known about the biochemical mechanisms of intron recognition and splicing [1-3], a clear understanding of the events and processes that explain the appearance and persistence of introns during the evolution of eukaryotic genomes remains elusive [4, 5].

As a necessary step in the expression of most extant eukaryotic genes, splicing has been exploited by evolution in at least two main ways. One allows diversification of the structure and function of the RNA and protein products of a gene by producing multiple distinct mRNAs through alternative splicing [2]. A second allows changes in gene expression through nonsense-mediated decay (NMD), whereby alternative splicing can lead to either functional mRNA, or to transcripts with premature stop codons that are degraded, providing developmental on-off control, or stable homeostatic expression settings [2]. The complex gene architecture of multicellular organisms, as contrasted with the simpler gene architecture in single-celled eukaryotes, has prompted widespread speculation that splicing is responsible for emergent complexity in metazoans. Although it contributes in complex and critical ways to gene function and regulation in extant eukaryotes, how splicing came to reside so pervasively in the eukaryotic lineage remains to be explained [4-6].

Until recently, gain or loss of introns has been detected by comparing closely related genomes. Many such "presence/absence" variations are inferred to be intron loss, in which reverse transcription of a spliced RNA, followed by homologous recombination of the intronless cDNA back into the gene of origin, erases the intron [6, 7]. Several mechanisms for the gain of new introns have been proposed (for review see [8]). For example, single nucleotide changes that create new splice sites (and thus new introns) can lead to "intron sliding" or new alternative splicing events [9]. "Exonization" of an Alu sequence in a large intron can lead to inclusion of a new exon and splitting of an intron into two smaller

introns [10]. These intron gain mechanisms rely on pre-existing local splicing events, and represent intron diversification, rather than wholesale intron creation at sites where no intron previously existed. Several distinct mechanisms for *de novo* intron creation have been proposed. In a genetic drift or "intronization" mechanism [11], a genome sequence drifts, accumulating mutations that by chance allow its transcripts to be recognized by the splicing machinery and spliced. A recently recognized mechanism is "intron transposition" whereby an intron at one location is copied and inserted at a new location. First described in the marine alga *Micromonas* [12], unusual classes of apparently mobile introns called "Introner Elements" appear to have expanded in some lineages [13-17]. These elements preferentially reside in transcribed regions in the splicing-sense orientation, are removed from primary transcripts by the spliceosome, and do not appear to encode protein [12, 14, 17]. Models to explain this expansion invoke an "armed spliceosome" intermediate carrying an intron lariat RNA, which is then reverse spliced into an mRNA (which then must be converted to cDNA to return to the genome at a new location, [6, 8]) or into the single stranded DNA of an R-loop [15]. Other models suggest that new sequences that act as introns may be products of DNA damage repair [18], and still other repeated intron classes might be non-autonomous DNA transposons, based on putative target site duplications [19]. Recently, intron transposition through an RNA intermediate without target site duplication has been documented in *S. cerevisiae* [20], supporting the idea that some form of reverse splicing may operate to spread spliceosomal introns.

Splicing in the *S. cerevisiae* genome appears to have been streamlined by evolution, such that about 5% of genes have introns, and most that do have only one. Despite their scarcity in genes, introns appear in about 25% of transcripts when cells are growing in rich medium [21, 22]. More than a third of annotated introns are found in genes for ribosome biogenesis (ribosomal protein genes, RPGs), and their mRNAs account for 90% of the splicing performed in rapidly growing cells [22, 23]. This unusual distribution of introns in a highly expressed class of genes with shared function presents both challenges

and opportunities for studying integration of splicing into core cellular regulation. For example, repressing transcription of the RPGs increases the efficiency of splicing for the majority of non-RPG introns and can suppress temperature-sensitive spliceosomal protein mutations [24, 25]. Based on this we proposed that relieving pre-mRNA competition by reducing RPG expression frees the spliceosome to process less competitive, splicing substrates it normally ignores. This phenomenon has been shown to contribute to the efficiency of meiotic splicing [25, 26], as well as regulation of Coenzyme Q_6 synthesis [27], whereby regulated repression of RPGs potentiates splicing control mechanisms at other genes.

In this study, we use rapamycin to repress RPGs and create hungry spliceosome conditions in three related yeast species, revealing locations in the genomes where protointrons reside. These protointrons are found in mRNAs and noncoding RNAs such as promoter-associated transcripts, CUTs, SUTs, XUTs and other noncoding RNAs [28-36]. Other studies have found undocumented splicing events [37-44], including alternative splicing of known introns. Here we discover, collect, and reclassify splicing events into two main categories, standard introns and protointrons, based on efficiency, conservation, location, and relevance to extant gene expression, in order to understand intron creation. Related yeasts *S. bayanus* and *S. mikatae* also have protointrons, but most are species-specific, indicating that protointrons appear and disappear during evolution. We propose that protointrons are created through genetic drift and enable genomic fluidity through a mechanism that is distinct from gene duplication or intron transposition. We suggest that the small subset of more efficiently spliced protointrons may be advanced in the process of becoming clade- or species-specific standard introns. This intermediate class of introns tend to occur in 5'UTRs where they might buffer the negative effects of RNA secondary structure or micro-ORFs (uORFs) on translation, however the adaptive value of any of these emerging introns remains to be determined. This and the accompanying study [43] reveal the extent to which the spliceosome recognizes and splices intron-like sequences, thus expanding the information contained in the genome. This contribution of the spliceosome to the information content of

the transcriptome may enhance the rate at which new genes and regulatory mechanisms appear in eukaryotes.

Results

Deeper RNA sequencing of a nonsense-mediated decay deficient strain confirms a class of rare splicing events in *Saccharomyces cerevisiae*

In experiments where we observed increases in splicing efficiency of standard introns after repression of RPG expression [25], we also observed unannotated splicing events, distinct from known introns, whose splicing efficiency improved. In addition, inventive new RNAseq methods designed to capture branchpoints [37, 39] have provided evidence for splicing events elsewhere in the transcriptome. Some of these events appear to be activated in response to stress [37], which down-regulates RPG expression and creates hungry spliceosome conditions. Still others are more readily detected during meiosis [37, 44], or when cells are deleted for RNA decay pathway components that degrade unstable transcripts [37, 38, 43]. Our interest in understanding the evolution of splicing prompted us to focus on these distinct new introns. To capture more of them, we obtained additional RNAseq data that included non-polyadenylated RNAs and RNAs sensitive to nonsense-mediated decay. We made four libraries, one each from rRNA-depleted RNA from untreated (0 min) and rapamycin treated (blocks nutrient signaling and represses RPGs, 60 min) replicate cultures of a yeast strain deficient in NMD (*upf1Δ*, [45]). We obtained more than 300 million reads that show excellent between-replicate coherence in gene expression changes (Fig. S1A). These data confirm our previous observation [25] that splicing of a majority of standard introns in non-RPGs increases after rapamycin treatment. A splicing index relating change in the ratio of junction/intron reads over time (SJ index = log2[junctions-t60/intron-t60] – log2[junctions-t0/intron-t0]) increases for most introns in transcripts whose total transcript levels change less than two-fold during the experiment (Fig 1A, blue circles, Table S1, NB: RPGs are repressed >2

fold and are excluded). In addition to the standard introns, we observe more than 600 splicing events that result from use of alternative 3' or 5' splice sites overlapping the standard introns (see also the accompanying study by Douglass et al. [43]). Some of these splicing events have been previously characterized [38, 41, 42], and produce out of frame mRNAs that are more easily detected in the *upf1Δ* strain due to the loss of NMD (Table S2). Because our interest here is in introns appearing at novel locations, we have not studied the splicing events that overlap the standard introns any further. At this sequencing depth, a set of splicing events that do not overlap any standard intron is evident (Table 1 and Table S3). These are supported by reads that span sequences with known characteristics of introns but occur at much lower frequencies than those for standard introns. As is the case for standard introns, splicing of the majority of these nonstandard introns increases upon rapamycin treatment (Fig 1A, orange circles, positive values indicate increased splicing), suggesting they are also in competition with RPG pre-mRNAs for the spliceosome.

Identification and validation of protointrons

To extract and validate nonstandard splicing locations from the RNAseq data, we inspected reads that span and are missing genomic sequences bounded by GT or GC on the 5' side and AG on the 3' side, aggressively filtering out those that were unlikely to have been generated by the spliceosome (Table S4, see also Methods). For example, we ignored reads that are abundant in only one library due to spurious PCR-derived "jackpot" amplification, or that are incorrectly mapped as spliced over naturally repeated sequences. We also filtered out reads that appeared fewer than three times, spanned sequences with no discernable match to a relaxed and appropriately positioned branchpoint consensus sequence (RYURAY, >45 from a 5'ss, >7 from a 3'ss), or that use a GAG 3' ss (although some of these may be true). Finally, in order to focus on new intron locations, we separated out reads that overlap known standard introns (these are found in Table S2). Finally, we merged overlapping alternative splice

sites observed at each new intron location into one. From this sequencing experiment we identify 226 splicing events at 162 intron locations in the *Saccharomyces cerevisiae* genome that do not overlap with standard introns. We call these protointrons (Table 1 and Table S3, NB: we have reclassified some previously annotated introns as protointrons based on their less efficient splicing and lack of conservation in close relatives, see below).

To characterize protointrons as a distinct class of splicing events, we inspected the alignments of many individual protointrons (for example Fig 1B) and validated more than a dozen of them by RT-PCR, cloning, and Sanger sequencing (Fig 1B and C, Table S5). These events map to a diversity of transcribed locations, including mRNAs, a variety of non-coding RNAs (CUTs, SUTs, XUTs, etc.), and within telomeric Y' repeats. Examples can be found entirely within a coding region (e. g. *TAF13* Fig 1B, C, Fig S1D), or in a noncoding RNA antisense to a standard gene (e. g. *antiASH1*, Fig 1B, C, Fig S1B). Often alternative splice sites are observed, for example in the noncoding RNA *XUT12R-370* antisense to *TUS1* (Fig 1B, note only the product derived from use of the downstream 3' ss is visible on the gel in Fig 1C, Fig S1C). An internally initiating RNA from the meiotic *SPO1* gene expressed only during vegetative growth has a protointron (Fig 1C, Fig S1E). Protointrons can also be found crossing boundaries from the coding region to either UTR in mRNAs, and still others can be found completely within a UTR. In many cases, excision, cloning and sequencing the faint band near the size predicted by the RNAseq reads identifies additional alternatively spliced forms not observed by RNAseq. For the four events shown in Fig 1C, we successfully confirmed splice junctions indicated by RNAseq (Table S3).

Several of the protointrons predicted by the RNAseq data appear to use unusual 5' splice sites (5' ss) not anticipated by examination of standard introns. Standard intron 5' ss show strong conservation of the G residue at position 5 (G5, underlined here: GUAY<u>G</u>U), which contributes to intron recognition through interactions first with U1 snRNA and later with U6 snRNA [46, 47]. The 5' ss consensus in

mammals has a less strongly conserved G5, and a subset of mammalian introns use G at position 6 instead [48]. Validation tests of several protointrons whose 5' ss lack G5 show that they are authentic products of splicing (Fig 1D, Fig S1F). During this effort we also detected splicing at an AT-AC junction in the noncoding RNA *SUT635* (Fig 1E). Although yeast does not have a minor (U12) spliceosome, some major (U2) spliceosomal introns use AT-AC junctions [49], and mutation of a standard yeast intron shows that AT-AC is the most efficient splice junction dinucleotide combination after GT-AG [50]. Alternative AT-AC junction use has been reported for the standard intron (a normal GT-AG intron) in *RPL30* [37], however no standard yeast intron uses AT-AC junctions normally. The appearance of an AT-AC protointron in *SUT635* suggests that AT-AC introns may represent an alternative path to evolution of standard introns.

Distinct features of protointrons: sequence, conservation, size, and splicing efficiency

To begin contrasting the features of protointrons and standard introns, we evaluated evolutionary conservation, the most obvious difference. To analyze conservation around splicing signals of both intron classes, we extracted a sequence window surrounding the 5'ss, predicted branchpoint, and 3'ss from each standard intron and protointron and compared them. Position-specific weight matrix-based logos of the splicing signals (Fig 2A) reveal that protointrons use a more divergent collection of splicing signals than do standard introns. As noted above, G5 of the 5'ss is a prominent feature of the standard yeast intron 5'ss but is less well represented among protointron 5'ss. Similarly, the predicted branchpoint sequences of protointrons lack strong representation of bases to either side of the core CUAA of the UACUAAC consensus for standard introns. Branchpoints of standard introns are enriched for Us upstream of the consensus and for A just downstream, and neither of these context elements are prominent in the predicted branchpoints of protointrons. The 3'ss are more similar, but as with the branchpoint, the U-rich context detectable upstream of the standard intron 3'ss is not observed in

protointrons. The more diverse collection of splicing signals used by protointrons is similar to that of the overlapping alternative splice sites of standard introns [38], and a mixed set of overlapping and distinct potential introns detected using a branchpoint sequencing approach [37]. We conclude that protointrons use a wider variety of branchpoints and splice sites than do standard introns in *S. cerevisiae*. This suggests that as protointrons evolve toward standard intron status they may acquire mutations that enhance the context and match to the consensus of the core splicing signals, in part to increase their ability to compete with RPG pre-mRNAs [25].

With the exception of the protointrons found entirely within protein coding sequences, typically at least one (and often all three) of the splicing signals for a given protointron in *S. cerevisiae* is imbedded in sequence that is not conserved in closely related species (Fig 1B, E, Fig S1C, F). To analyze whether this is a distinguishing characteristic of protointrons, we recorded the average phastCons score (range between 0, evolving as not conserved, and 1, evolving as highly conserved, [51]) of the nucleotides within sequence windows containing the 5'ss, the branchpoint, and the 3'ss of the standard introns and protointrons and plotted them (blue bars, standard introns; orange bars, protointrons, Fig 2B). Many standard introns are embedded in conserved protein coding sequences, and thus the splice sites and their immediate exon context are also conserved, such that the average phastCons scores for both the 5' and 3'ss windows rise above 0.7 (Fig 2B). The branchpoints of standard introns are also conserved but have a broader and lower score distribution, because constraining protein coding exon sequences are not usually found near the branchpoints of standard introns. In contrast, the distributions of phastCons scores for each of the protointron splicing signals is clearly bimodal, meaning that protointron splicing signals are either highly conserved, falling within protein coding sequence, or are poorly conserved, falling in UTRs or intergenic regions (Fig 2B). This distribution illuminates the sequence landscape within which most protointrons arise. Since the strongly transcribed regions of the genome code mostly for protein, and transcription is a prerequisite for

splicing, protointrons tend to appear in or span less well conserved noncoding RNA sequences such as UTRs and ncRNAs. This distinguishes protointrons from standard introns and suggests that most of the protointrons we detect have appeared only recently in the *S. cerevisiae* genome, when transcribed non-protein coding regions acquire intron-like features by mutation.

Standard introns show a bimodal length distribution, with peaks at about 100 nt and 400 nt ([52], Fig 2C). In contrast, the distribution of protointrons is on average shorter, with a single main peak at around 100 nt in length, and few larger than 300 nt. These distributions are significantly different (Kolmogorov-Smirnoff test, $p \leq 10^{-4}$), suggesting that if protointrons evolve into standard introns, they acquire additional sequence features. Many of the larger standard introns are found in RPGs [52], where secondary structures and other long distance RNA-RNA interactions promote efficient and accurate splicing [53, 54]. The elaboration of such structures during evolution of increased splicing efficiency may explain the increased intron length that characterizes the large intron class in yeast.

Protointrons are less efficiently spliced than standard introns (Fig 2D, Tables 1, S1 and S3). The vast majority of standard introns are spliced at greater than 80% efficiency by comparison of splice junction reads to intron base coverage. Exceptions include meiotic introns whose efficient splicing may require repression of RPGs or the expression of a meiosis-specific splicing factor like Mer1 [24-26]. In contrast most protointrons have splicing efficiencies below 20% at best. A few protointrons, such as the introns in the *S. cerevisiae MTR2*, *USV1*, and *MCR1* genes or the *S. bayanus YTA12* gene, are uncharacteristically well spliced, suggesting that they may be transitional intron forms or species-specific standard introns (see below). Splicing improves for most protointrons and standard introns after rapamycin treatment, however some standard introns appear to show reduced splicing, suggesting splicing repression in response to rapamycin at those introns.

Coding regions are depleted of sequences required for splicing

Protointrons that emerge within coding regions might disrupt gene expression by reducing mRNA levels or creating toxic proteins. Since protointrons emerge readily from nonconserved sequence (Fig 2B), we wondered whether the appearance of protointrons within ORFs most often reduces fitness, and thus whether the frequency of splice site and branchpoint sequences might be lower than would be expected by chance in protein coding regions. In rare cases such introns might allow advantageous mRNA regulation through NMD to emerge (as appears to have been the case with a recently evolved standard intron in *PRP5*, [55]) or make mRNA for beneficial alternative proteins (as may be the case for *PTC7*, [27, 56, 57]; and *MRM2* [42], see below), which have in-frame conserved coding sequences through their introns. To test whether *S. cerevisiae* coding sequences might be depleted of splicing signals, we compared their frequency in the ORF set of the extant *S. cerevisiae* genome with that in 10,000 synthetic ORF sets derived by randomizing the order of codons for each ORF. This approach maintains ORF length, integrity, GC content, and codon usage in the permuted ORF sets while generating partially randomized nucleotide sequences that can be used as a background sequence set for comparison, and has been used to evaluate co-evolution of RNA processing signals within coding sequence [58].

To assess the representation of 5' ss and branch point sequences, we chose two 6-mers as proxies – one for the 5' ss (GTATGT), and one for the branch point (ACTAAC). We counted the number of times each appeared in the extant S. cerevisiae ORF set, and in each of the 10,000 permuted ORF sets, and plotted them. For both 6-mers, their counts in the extant ORF set are less than 3 standard deviations below the mean of their respective counts from the 10,000 permuted ORF sets (vertical lines), indicating that these 6-mers are significantly underrepresented in the natural *S. cerevisiae* coding sequences as compared to the randomized coding sequences (Fig 3A).

To determine whether these proxy 6-mers were unusually depleted as compared to other 6-mers, we calculated a Z-score for each of the 4096 6-mer sequences in turn, comparing the counts of each in

the extant *S. cerevisiae* ORF set with its mean counts in the 10,000 permuted ORF sets, and plotted the distribution (Fig 3B, grey bars). The intron branchpoint 6-mer "ACTAAC" is found comparatively much less frequently in the extant genome than are other 6-mers (Fig 3B), and much less frequently than the average stop codon-containing 6-mer (blue bars). The 5' ss 6-mer "GTATGT" is also more depleted than the average 6-mer, especially when compared to the subset of ATG containing 6-mers (maroon bars, Fig 3B). Because there is a large amount of information in coding sequences, we cannot be certain the depletion of the splicing signal 6-mers is due to splicing. Numerous other features are being randomized by the process of codon permution, for example 6-mer representation may be influenced by di-codon frequencies that affect translation [59]. Even so, these observations are consistent with the idea that splicing signals within ORFs carry a risk of reduced fitness. This suggests that a robust level of spliceosome activity may be sufficient to lead to loss of correct mRNA should genetic drift create splice sites within ORFs, providing a rationale for tight regulation of a splicing activity limited to pre-mRNAs that can compete [25].

Protointrons are idiosyncratic to closely related species

S. cerevisiae protointrons are not conserved in closely related yeasts, suggesting that they have appeared in *S. cerevisiae* in the time since these lineages diverged. To explore the hypothesis that protointrons arise and disappear differently in the other *Saccharomyces* lineages, we treated cultures of *S. mikatae* and *S. bayanus* with rapamycin for 0 or 60 minutes, isolated RNA, depleted rRNA, and made cDNA libraries for sequencing. These strains are NMD competent (*UPF1*), so our ability to detect transcripts subject to NMD is limited. Regardless, nearly all annotated introns in *S. cerevisiae* are present in the *S. mikatae* and *S. bayanus* genomes, and *S. bayanus* has an additional standard intron in a *CHA4*-like gene that has no ortholog in *S. cerevisiae* (Tables S6 and S7). Some annotated *S. cerevisiae* introns are missing in these close relatives and based in part on their locations and splicing efficiencies,

we propose reclassifying them as protointrons (see below). In contrast to the high conservation of standard intron locations, we detect distinct sets of protointrons in each species (Tables S6 and S7). We validated a subset of these (Fig 4) and find that most all of the *S. mikatae* protointrons are not present in either *S. cerevisiae* or *S. bayanus*, and that the *S. bayanus* protointrons are not found in *S. cerevisiae* or *S. mikatae* (Tables S3, S6, and S7). An exception to this is *YIL048W/NEO1*, in which the same protointron is observed in both *S. cerevisiae* (Table S3) and *S. mikatae* (Fig 4, Table S6). High sequence conservation in the coding region of *YIL048W/NEO1* has preserved the splicing signals in all of the *Saccharomyces* yeasts. The intron is in frame with the coding sequence (Table S8), thus although the splicing of this protointron is not efficient, it remains possible this intron could generate a functional alternative protein that is under selection. We conclude that protointrons are idiosyncratic in closely related yeast species. This is evidence for rapid evolutionary appearance and disappearance of sequences that can be functionally recognized by the spliceosome. The dynamics of creation of protointrons thus appears consistent with genetic drift, primarily in the rapidly evolving nonconserved sequences of recently diverged genomes.

Introns with features of both protointrons and standard introns may be intermediates in de novo intron creation

Based on the above analysis and those described elsewhere that note "novel" splicing [37-39, 42-44, 57], we propose defining standard introns as (1) conserved in related organisms or clades, (2) efficiently spliced under appropriate physiological conditions, and (3) established in the pathway for production or regulation of a functional gene product. Furthermore, we propose defining protointrons as (1) present in one or at most a few very closely related species, (2) inefficiently spliced, and (3) not clearly understood to contribute to correct expression or regulation of a gene. This simple definition allows classification of most all of the observed splicing events in the yeast genome as either protointron

or standard intron, with only a few exceptions. The vast majority of protointrons arise by neutral drift and likely provide no fitness advantage. A few introns do not neatly fall into one or the other category and may be transitioning from protointron to standard intron status. Like protointrons these intermediate class introns are clade- or species-specific, but unlike typical protointrons they have increased splicing efficiency, and may appear positioned to influence expression of the gene that carries them.

Examples of protointrons that may be on an evolutionary path toward standard intron status include introns in the 5' UTRs of *S. cerevisiae MTR2, USV1, YEL023C*, and *MCR1*, and in the 5' UTR of *S. bayanus YTA12* (Fig 5). Using the unrooted tree describing relationships between the genomes of the sensu stricto yeasts [60], we map the appearance of these high efficiency protointrons as predicted by their presence in extant genomes. The *MTR2* intron contains essential sequences [61] and can diversify the N-terminal sequence of the mRNA export protein Mtr2 [62]. When sequences of related yeasts became available [63], it became clear that the *MTR2* intron is unique to *S. cerevisiae* (Fig 5). High efficiency protointrons are found in the 5'UTRs of *USV1* and *MCR1*. The *USV1* intron is efficiently spliced after rapamycin treatment in *S. cerevisiae* and is also functional in *S. mikatae* (Fig 5, Table S6). However, *S. kudiravzevii* and *S. bayanus* have different sets of nucleotide changes that eliminate splice sites and branchpoints required for this intron. The *MCR1* intron is spliced at about 30%, and appears to be shared by *S. paradoxus*, but is absent in *S. mikatae*, *S. bayanus*, and *S. kudriavzevii*. None of these splicing events alter the N-terminus of Usv1 (a stress induced transcription factor) or Mcr1 (a mitochondrial NADH-cytochrome b5 reductase), however both introns remove uORFs from the 5' UTRs of these genes, suggesting that splicing could affect 5'UTR function in mRNA translation or stability for both genes. Finally, a very efficiently spliced (>95%) intron is found in the 5'UTR of the *S. bayanus YTA12* gene, as well as in *S. kudriavzevii* (Fig 5). Removal of this intron does not alter the N-terminus of Yta12, a mitochondrial protein complex assembly factor [64], but does lead

to removal of a uORF. Interestingly the *S. cerevisiae YTA12* gene matches at 88 of 117 (75%) positions in the *S. bayanus* intron and has neither an intron nor any uORF (see below).

The apparent relationship between introns and uORFs leads to the idea that 5'UTR introns may be adaptive by protecting mRNAs with long 5'UTRs from the general negative effect of uORFs [36, 65]. To test this idea, we asked whether uORFs are present more frequently in 5'UTRs that contain introns, as compared to similarly sized 5'UTRs that do not. There are 22 yeast genes with standard introns in their 5'UTRs (this number does not count the annotated introns in *MCR1*, *MTR2*, or *USV1*, which are not conserved aross the *sensu stricto* group). The size range of the (unspliced) 5'UTRs for these 22 is from ~240 to 950 nucleotides, and there are 91 intronless genes with 5'UTRs in this size range. We counted uORFs longer than 4 codons (including the AUG, but not the stop codon) within 5'UTRs in this size range. Among the 91 genes without 5' UTR introns, 53 lack any uORFs, whereas 38 have at least one uORF. All 22 genes with 5'UTR introns have at least one uORF, and for 20 of these all the uORFs in the 5' UTR are removed by splicing. This distribution of uORFs in 5'UTRs with introns is unlikely to have been generated by chance (Fisher's exact test, $p < 10^{-5}$). One hypothesis to explain this is that by removing much of the 5' UTR RNA, an intron may protect a gene that has a long 5'UTR from genetic drift that creates uORFs, or other translational inhibitory features like RNA secondary structure [36]. Additional experiments will be needed to determine which if any of these splicing events promotes gene expression and whether or not the effect contributes to fitness.

A second evolutionary scenario whereby protointrons may be adaptive concerns in frame splicing, which would produce an alternative polypeptide. This appears to be the case for the standard intron in *PTC7* [27, 56]. A similar intron is found in *MRM2*, where as many as three different proteins may be produced (Tables S1 and S2, see also [42], NB: not annotated in SGD, but this fits the standard intron definition). Despite forces that seem to deplete splicing signals within ORFs (Fig 3), we find 20 protointrons within ORFs that do not interrupt the reading frame and thus could produce functional

proteins, particularly under stress or other conditions where RPG transcription is reduced (Table S8). We suggest that such protointrons provide evolutionary opportunities to create new protein isoforms from existing genes.

Telomeric Y' repeats

Intron-like sequences have been noted in the telomeric Y' family of repeat sequences for more than 25 years and continue to be annotated in the Saccharomyces Genome Database (SGD). So far, molecular tests for splicing of these annotated introns have been negative [52, 66]. Intron predictions allow some Y' repeat element copies to encode a large (1838 amino acid) protein (e. g. YNL339C, Fig 6A), that carries an N-terminal Sir1 domain and a central DExD helicase domain. Other Y' elements differ in sequence and can only encode fragments of the open reading frame that may nonetheless produce smaller functional proteins, for example the helicase overexpressed in telomerase-deficient "escaper" colonies [66, 67]. The function(s) of any of the Y' element predicted proteins in normal cells are not known.

To determine if Y' element transcripts are spliced, we allowed RNAseq reads to map to the repetitive Y' elements (i. e. without masking). Although the mapped locations may not be the precise origin of the RNA that created the read, this allows us to identify spliced reads and assign them to possible members of the Y' repeat family. We find two introns within the Y' repeat family (Fig 6), one of which lies on the far left of the repeat and is required to create the open reading frame for the longest predicted protein (exemplified by YNL339C near TEL14L, Fig 6A). The other is in the center of the Y' repeat (exemplified by YLR464W near TEL12R, Fig 6B). Neither of these introns matches the annotations at SGD, and instead in both cases, downstream 3' ss are used (see also [42]). It is not uncommon for the yeast spliceosome to skip proximal 3' ss in favor of a distal 3' ss, in some cases due to secondary structure of the pre-mRNA [68]. The Y' elements in *S. cerevisiae* differ from each other;

not all can express a protein as large as YNL399C after removal of intron 1 using the distal 3' ss. Intron 2 splicing does not greatly extend the open reading frame of YLR464W. To confirm that the reads arise from splicing rather than from a deleted copy of the Y' element precisely lacking the intron, we searched the genome using the "spliced" sequence produced for intron 1 or intron 2 and found that there is no such contiguous genomic sequence. We conclude that Y' element transcripts can carry at least two introns that are distinct from current annotations in SGD (Fig 6).

To evaluate the sequence relationships of the Y' repeat element introns we aligned them with each other, after merging identical copies into one. All the predicted intron 1 sequences have the second most common 5' ss in the yeast genome (GUACGU, followed by the preferred A at position 7, [52]. The most distal 3' ss of several possible creates the large ORF, and is UAG for all except YPR202W, which has a CAG. Several other potential 3' ss (including the one annotated in SGD) are skipped or used alternatively. Interestingly only YPR202W, YRF1-3, YRF1-6, and YRF1-7 have the canonical UACUAAC branch point sequence, whereas most of the others have UAUUAAC, a variant found in some standard introns (Fig 6A). The remaining group (YEL075C, YRF1-2, and YRF1-4) are deleted for the region containing the branch point, suggesting that they are unable to be spliced. Intron 2 has the most common 5'ss GUAUGU, and the most common branch point UACUAAC, and uses the first AAG (a less commonly used but standard 3' ss) downstream from the branch site (Fig 6B). Most copies also contain an alternative 5' ss which is used less frequently. We have not estimated the efficiency of splicing of these introns because we cannot reliably assign reads to specific repeat elements with confidence. The current genome assemblies of *S. mikatae*, *S. paradoxus*, and *S. kudriavzevii*, but not the *S. bayanus* assembly include at least one Y' element related to the *S. cerevisiae* elements [60], but the precise numbers and arrangements of the Y' elements in those genomes await refinement of the genome assemblies for those organisms.

The *S. bayanus YTA12* protointron functions in *S. cerevisiae*

The finding of a highly efficient protointron in the *YTA12* 5' UTR of *S. bayanus* (and putatively in *S. kudriavzevii,* Fig 5) that is not observed in the alignable syntenic sequence of *S. cerevisiae* prompted us to test (1) whether this *S. bayanus*-specific intron can be spliced in *S. cerevisiae*, and (2) whether the intron might confer some advantage for growth on glycerol, given the function of Yta12 in assembly of mitochondrial protein complexes [64]. Fig 7A shows an alignment of the region including and upstream of the Yta12 start codon from *S. bayanus* (sacBay), *S. cerevisiae* (sacCer) and *S. cerevisiae* in which the 117 bp of the *S. cerevisiae* genome corresponding to the *S. bayanus* intron have been replaced in *S. cerevisiae* with the *S. bayanus* intron (Sc-SbI). This replacement was made using CRISPR/Cas9 guided cleavage of an *S. cerevisiae*-specific target sequence within the syntenic region and a repair fragment containing the *S. bayanus* intron (Fig 7B). As controls, we created *S. cerevisiae* strains precisely deleted for the syntenic region aligning with the *S. bayanus* intron, as well as versions of the *S. bayanus* intron with 5' ss mutations (Fig 7C). We isolated RNA and evaluated the expression of these modified *YTA12* genes by extension of a labeled primer complementary to *YTA12* mRNA with reverse transcriptase (Fig 7C). The major transcription start sites for *YTA12* in *S. cerevisiae* map about 300 nt from the 5' end of the primer (Fig 7A and C, lane 1). These start sites are unaffected by the 117 bp deletion (the same collection of cDNAs are shorter by 117 residues, lane 2). The migration of cDNAs from the deletion strain are useful to mark the expected position of spliced RNAs, and indeed replacement of the 117 bp with the *S. bayanus* intron sequence (lane 3) results in the appearance of the same collection of cDNAs with the disappearance of the signal from pre-mRNA, indicating efficient splicing (lane 3, compare with lane 1).

Mutation of the 5' ss from GUAUGU to GaAUGU or GaAcGU results in the reduction of the spliced mRNA cDNAs, and the appearance of cDNAs corresponding to pre-mRNA (lanes 4 and 5), indicating that splicing is inhibited by these mutations. Changing the 5' ss from GUAUGU to the less

commonly used GUAcGU reduces the efficiency of splicing but does not block it, as judged by the slight accumulation of unspliced RNA (lane 6). Unexpectedly, the *S. bayanus* intron sequence activates a set of cryptic start sites in the *S. cerevisiae* sequences downstream of the major start site and the *S. bayanus* intron (Fig 7C, lanes 2-6). These start sites are inefficiently used in the wild type *S. cerevisiae YTA12* promoter (lane 1). One consequence is that new mRNAs are made that initiate downstream of the intron and thus do not require splicing for expression of Yta12. This interpretation is supported by the observation that all the strains grow on YP glycerol plates as well as wild type BY4741 (not shown). This result highlights the challenge of anticipating the effect of mutations in 5' UTRs where transcription, splicing, and translation operate together on the same sequence. This experiment measures changes in splicing due only to differences in the intron, and not due to any differences in exonic sequences or trans-acting factors between *S. bayanus* and *S. cerevisiae*. We conclude that the efficiently spliced protointron from *S. bayanus* is equally at home in *S. cerevisiae*. This intron appears to have formed in *S. bayanus* after *S. bayanus* and *S. cerevisiae* last shared a common ancestor, but before the divergence of *S. bayanus* from *S. kudriavzevii*.

Discussion

A second class of splicing events exposes roles of the spliceosome in evolution

Here we characterize a class of splicing events in yeast we call protointrons. Many previous studies have noted "novel" introns in yeast under a variety of experimental conditions and genetic backgrounds [37-44, 57]. Here we distinguish protointrons by several criteria, most importantly that they reside at locations not overlapping known standard introns. We first recognized protointrons while studying how the abrupt disappearance of RPG pre-mRNA during early nutrient deprivation signaling frees the spliceosome to increase splicing of other pre-mRNAs [25, 26]. RNAseq analysis of NMD-deficient yeast cells treated with rapamycin revealed that protointrons are found throughout the

transcriptome in both coding and non-coding regions of pre-mRNAs, ncRNAs, and antisense transcripts, such as CUTs, SUTs, and XUTs (Fig 1, Fig S1). Protointrons contain all of the splicing signals necessary for recognition by the spliceosome (5'SS, BP, and 3'SS), however the sequences of these signals are more variable than those of standard introns (Fig 2A and B). Whereas standard introns are conserved in related organisms, efficiently spliced, and established for production or regulation of a functional gene product, protointrons are present in one or a few closely related species, not efficiently spliced, and do not clearly contribute to correct expression or regulation of a gene. Given this redefinition, we propose a revised intron annotation, including the addition of a standard intron in *MRM2*, and molecular evidence for the correct location of expected but not demonstrated splicing of Y' repeat element transcripts. We provide this and related data on a publicly accessible genome browser with several *Saccharomyces* species genomes at http://intron.ucsc.edu/.

Splicing events that occur outside our expectation of what is needed to make a protein or a structural RNA have attracted labels like "splicing noise" or "splicing error". But viewing the spliceosome as an enzyme able to catalyze a complex series of pre-mRNA binding, refolding, and release operations, including two cleavage-ligation reactions, or even just the first one [40, 69], on a very diverse set of substrates (for review see [2]) suggests that such terms should be more carefully defined. The protointrons described here, as well as for example similar newly evolved splicing events observed in mammalian lncRNAs [70, 71] reveal the outer edges of the substrate repertoire of this enzyme in sequence space, and do not represent either splicing noise or splicing errors. We suggest the term "splicing noise" should refer to fluctuations due to stochastic events affecting particular splicing events, just as the term "transcriptional noise" refers to the stochasticity of transcription events (see [72] and references therein). We also suggest the term "splicing error" should refer to events within the spliceosome that lead to spliceosome assembly or catalysis that is incompatible with successful completion of the two splicing reactions, spliced product release, and recycling. In order for splicing to

contribute to rapid evolution of multicellular organisms it seems likely that a variety of sequences besides highly evolved introns would need to be recognized and spliced, including those that appear in genomes by genetic drift. The extent to which these spliceosome-generated spliced RNA sequences contribute to fitness would eventually determine their evolutionary fate. The protointron class of splicing substrates represents opportunity to create new genes, create new proteins from existing genes, or impose new regulatory controls on existing genes.

Some protointrons show greater splicing efficiency and may be adaptive

The forces and mechanisms that drive intron evolution in eukaryotic genomes are still largely unknown. If protointrons represent raw material for intron creation by the process of intronization, then perhaps the most efficiently spliced protointrons represent intermediates in standard intron formation that are advancing by selection of improving mutations. Our data provide evidence for rapid and complete intronization in the *YTA12* 5' UTR between now and the time *S. bayanus* and *S. cerevisiae* last shared a common ancestor (~ 20 Mya, [60]). Over the 117 bp intron sequence, the *S. cerevisiae* 5' UTR differs at 29 positions (Fig 7A). Replacement of this region with the *S. bayanus* sequence produces an efficiently spliced intron in *S. cerevisiae* (Fig 7C). This intron transplantation experiment shows that no species-specific barrier prevents this sequence from serving as an efficient intron in *S. cerevisiae*. Although this intron appears fixed in the *S. bayanus* and *S. kudravzevii* branch of the *Saccharomyces* tree, there is currently no evidence for fitness effects, and thus this intron could be a product of neutral evolution.

In some cases, a protointron might provide increased fitness that would explain its evolutionary persistence. We suggest three specific ways that protointrons may support improvements in gene function. We note that the four most efficiently spliced protointrons we observed (*S. cerevisiae MTR2*, *USV1*, and *MCR1*, *S. bayanus YTA12*) reside in 5' UTRs. We realized that genes with long distances

between their transcription start sites and their start codons (i. e. with large 5' UTRs) are at risk for mutations that create a uORF in the 5' UTR, which often negatively influences translation [36, 65]. Removal of a large region of the UTR by splicing would buffer this genetic risk. Secondary structures or other detrimental sequences that might arise in long 5'UTRs [73] might also be safely removed by splicing. To test the plausibility of this idea, we examined the frequency of uORFs in 5' UTRs of *S. cerevisiae* genes that have or do not have 5' UTR introns and found that uORFs are significantly more prevalent in 5' UTRs that have introns as compared to other 5' UTRs (see Results). This suggests that the presence of a 5' UTR intron may help buffer an mRNA against any detrimental effects of uORFs or RNA secondary structure, and provides evidence that intronization in particular in 5' UTRs may be adaptive in *Saccharomyces* species.

A second way that protointrons may become functional is by producing in frame splicing events within ORFs to create mRNAs encoding shorter protein isoforms with new functions. The frequency of splicing signals is lower than expected in *S. cerevisiae* ORFs (Fig 3), supporting the expectation that most introns that arise within ORFs would be detrimental to fitness. Despite this, we found 20 protointrons contained within ORFs that do not interrupt the reading frame, and that may lead to the translation of alternative protein products (Table S8). If such shorter proteins contribute to fitness, mutations that increase the splicing of the protointron (without disrupting the function of the full-length protein) may lead to the establishment and conservation of a standard intron that allows production of both protein forms. This may be the mechanism by which the conserved in-frame introns of *PTC7* [27, 56] and *MRM2* ([42], this work) have evolved. In these cases, both the intron and the protein sequence encoded by the intron are conserved in *sensu stricto* yeasts, suggesting both contribute to fitness across the genus *Saccharomyces*. Many protointrons span the boundaries between conserved and nonconserved sequences (Fig 2B), increasing the chances that a new splicing event will alter one or the other end of an existing protein. Studies of protein evolution indicate that proteins evolve at their edges [74], suggesting

that protointrons may contribute to this as well. Although there is as yet no evidence for new function, the 5' UTR protointron in the *S. cerevisiae MTR2* gene has arisen sufficiently close to the start codon that different alternative splicing events add different peptides to the amino terminus of the annotated protein sequence [61, 62]. Thus, protointrons that appear in frame within existing genes, or that span the edges of existing genes, create protein expression variation that may provide fitness advantages, in particular under stress conditions.

A third way that protointrons may prove advantageous is through controlled downregulation through splicing and NMD. A standard intron in the *PRP5* gene is conserved in the *Saccharomyces* genus and destroys the *PRP5* mRNA by removing the start codon and creating a transcript that is subject to NMD [55]. The intron must have appeared since the divergence of the *Saccharomyces* species from their common ancestor with *Lachancia kluveri*, since this more distant relative has a different intron in its *PRP5* gene. This situation may evolve where overexpression of a particular protein may be detrimental. *PRP5* encodes a splicing factor, and increase (or decrease) in Prp5 protein activity may increase splicing and reduce (or decrease splicing and accumulate) *PRP5* mRNA levels by using this conserved out of frame intron to create a homeostatic regulatory loop. A more difficult to recognize but no less important mechanism is illustrated by the *BDF2* gene in which abortive splicing downregulates expression through spliceosome-mediated decay [40]. It is unclear whether to annotate this location and others like it [69] as an intron, since it does not appear that 3' splice site selection is required or important for its activity. Thus, even protointrons that are out of frame within coding regions, or pseudo-intron locations at which abortive splicing takes place may provide opportunities for adaptive regulatory controls to evolve.

Conclusions

Protointrons are a class of splicing events that that represent the action of the spliceosome on RNA without a necessary connection to the expression of a mature gene. In mammalian cells the spliceosome is no less constrained, and a very large number of alternative splicing events that appear unrelated to "correct" gene expression support this [75]. In particular, newly evolved lncRNAs have introns that are inefficiently spliced and have multiple alternative splice sites, unlike older, more conserved lncRNA and mRNA encoding genes [70, 71]. These observations indicate that a general feature of the evolution of introns is that any transcribed sequence has a chance of being spliced by the spliceosome, should that sequence evolve recognizable splicing signals. Additionally, any sequence that suddenly becomes transcribed can be expected to contain sequences by chance that are immediately recognized as introns. Since the sequences required for splicing are ubiquitous and have low information, many such newly appearing sequences will immediately produce diverse RNA transcripts. If these confer some advantage, or if mutations that improve splicing become fixed by neutral genetic drift, then a standard intron may evolve. This general pathway may be a source of new introns whose splicing contributes to diversification of the transcriptome, and to the appearance of new genes and new products from existing genes, as genomes evolve.

Materials and Methods

Strains and culture conditions

Two independent cultures of *S. cerevisiae* strain BY4741 *upf1Δ* (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 upf1Δ::KANMX*) were grown in YEPD medium at 30°C to an optical density at 600 nm (OD_{600}) ≈ 0.5. The cultures were split and rapamycin was added to one half at a final concentration of 200 ng/ml for 1 hour. *S. bayanus* strain JRY9195 (*MATa hoD::loxP his3 lys2 ura3*) and *S. mikatae* strain JRY9184 (*MATa hoD::NatMX trp1D::HygMX ura3D::HygMX*) were grown in YEPD medium at 26 °C, and were

treated with rapamycin as for *S. cerevisiae* except at 26°C. These strains were a kind gift of Chris Hettinger [60].

RNA isolation

RNA was extracted from yeast cells using Procedure 1 as described [76]. Prior to RNAseq library construction (see below), RNA was DNased using Turbo DNase (Life Technologies) and RNA quality was evaluated using the 2100 Bioanalyzer (Agilent).

RNAseq library preparation

5-10 ug of total *S. cerevisiae* RNA was depleted of ribosomal RNA using the RiboZero Gold rRNA Removal Kit (Illumina) according to the manufacturer's instructions. Strand-specific cDNA libraries were prepared using the Kapa Stranded RNA-Seq Library Preparation Kit for Illumina Platforms (Kapa Biosciences) following the manufacturer's instructions with the following modifications. Sequencing adapters and oligonucleotides used for PCR barcoding were from the NEBNext Multiplex Oligos for Illumina Kit (New England Biolabs, NEB). Prior to PCR amplification of the library, adapter-ligated cDNA was treated with USER enzyme (NEB). Adapter-ligated libraries were then PCR amplified for 10 cycles using NEB index primers compatible with Illumina sequencing. After amplification, size selection of the libraries was performed using an E-gel Safe Imager and 2% E-gel size select gels (Invitrogen). Indexed libraries were pooled and 100 bp paired-end sequenced on the same flow cell of an Illumina HiSeq4000 instrument at the Berkeley sequencing facility.

RNA extracted from *S. bayanus and S. mikatae* was depleted of ribosomal RNA as described above. Strand-specific cDNA sequencing libraries were prepared as described [77] and 50 bp paired-end sequenced on the HiSeq2000 platform (Illumina).

Mapping and Analysis of RNAseq Data

RNAseq data is deposited in GEO under the accession number GSE102615. For *S. cerevisiae* libraries, all mappings were done using 100x100 bp reads to the SacCer3 Apr. 2011 genome assembly (Saccharomyces Genome Database, SGD, [78]). For *S. bayanus* and *S. mikatae* libraries, mappings were done using 51x51 bp reads to the SacBay2 and SacMik2 genome assemblies (Saccharomyces Sensu Stricto Database, [60]), respectively. Reads mapping by BowTie2 [79] to *S. cerevisiae* tRNA and rRNA defined by Ensemble or to Ty elements defined by SGD were discarded, however mappings to Y' elements were recovered. For each library, reads were remapped to their respective genomes using STAR with two-pass mode [80]. PCR duplicate reads (reads with identical positions at both ends) were discarded and reduced down to one read. Changes in gene expression upon treatment of cells with rapamycin were determined using DESeq2 [81], comparing untreated and treated cells. Splice junctions were identified by STAR mapping [80].

Splicing Indexes (ratios of splicing measurements) were calculated by comparing reads that cross the intron, reads that cross the splice junction, and reads in exon 2 in different ways. Splice junction coverage is taken as the number of reads that cross the splice junction. Intron coverage was taken as the average per nucleotide coverage across the whole intron. When introns overlapped, a minimal length intron was used such that start of the intron was the most downstream start of the overlapping introns and the end was the most upstream start of the overlapping introns. Exon2 coverage was the average coverage for 100 bases of the following exon, using the most downstream 3' ss to define the exon. Log2-transformed ratios (Indexes) were calculated for the three comparisons: intron/exon2, splice junction/exon2, splice junction/intron. Figure 1A shows how the splice junction/intron index changes with rapamycin treatment by plotting the value of log2[splice junction-60/intron-60] – log2[splice junction-0/intron-0] for each intron in each replicate. The splicing events plotted here are for locations

whose overall transcript level changes less than 2 fold, and whose junctions are supported by at least 35 reads for standard introns or at least 50 reads for protointrons. The general shift of the points to the upper right quadrant indicates increased splicing efficiency (increased junction relative to intron reads) after rapamycin treatment.

Intron splice sites and candidate branch sits were extracted for analysis using the mapped splice junctions and by choosing a best branch point using the following heuristics. The likely branch points (underlined) were identified by searching introns for the following sequences in order until a branchpoint was identified: 1. ACTA<u>A</u>, 2. RYTR<u>A</u>YR, 3. YTR<u>A</u>Y (where R = A or G, Y = C or T) constrained to be 45 or more bases away from the 5' ss and no closer than 7 nucleotides upstream from the 3' ss. Candidate introns not matching YTRAY were considered to have no good match to a branchpoint consensus. If multiple equally good branchpoints are identified the one closer to the 3' ss was recorded. Details and scripts are at: <https://github.com/donoyoyo/intron_bp_generator>. To evaluate conservation, phastCons conservation scores were extracted from a window surrounding the splice site or branchpoint using data from the UCSC Human Genome Browser for *S. cerevisiae* at <http://genome.ucsc.edu>. Weblogos [82] were created using the site at https://weblogo.berkeley.edu.

Reverse Transcription and PCR

RNA was reverse transcribed using SuperScript III (Life Technologies) according to the manufacturer's instructions using a mixture of anchored oligo-dT (T24VN) and random hexamers as primer. Primers to validate and sequence products of splicing from protointrons by RT-PCR were designed using Primer3 [83]. PCR was performed using oligonucleotides listed in Table S5. PCR products were resolved by electrophoresis on agarose gels and visualized with ethidium bromide staining.

Cloning and Sanger sequencing of PCR products

PCR products generated by *T. aquaticus* DNA polymerase (Taq) were cut from low melting point agarose gels and purified using Machernary-Nagel gel extraction kits, then cloned using TOPO-cloning (Invitrogen). Inserts were sequenced by Sanger sequencing at the U. C. Berkeley sequencing center. Splice junctions were identified using BLAT [84] running behind a home copy of the UCSC Genome Browser [85] publicly available at http://intron.ucsc.edu/.

Estimation of background frequency of splicing signals in codon-permuted yeast genes

To test the hypothesis that "ACTAAC" (proxy for the branchpoint sequence), GTATGT (proxy for the 5' ss), or any other 6-mer nucleotide sequence within extant yeast ORFs might be enriched or depleted, we created 10,000 codon-permuted versions of the *S. cerevisiae* ORF set and counted the number of each of the 4096 possible 6-mers in each, computing a Z-score for each that compares representation of each in the extant ORF set to the mean representation of each in the 10,000 permuted ORF sets. To create permuted ORF sets in a way that preserves the GC content and codon usage of the extant set, we permuted the codons within each ORF (except for the start and stop codons) in the complete set of ORFs. Scripts for creating permuted ORF sets and analysis related to this question can be found under this github link: https://github.com/rshelans/genePermuter

CRISPR/Cas9 mediated intron transplantation

Yeast CRISPR editing was done essentially as described by DiCarlo et al [86], except that we rearranged the elements from different plasmids into a simplified single plasmid system by Gibson assembly. We obtained p426-crRNA-CAN1.Y and p414-TEF1p-Cas9-CYC1t [86] from Addgene. To create a BaeI cleavable cassette for easy guide cloning, we annealed oligos newguide1 and newguide2 together, and separately newguide3 and newguide4, and filled to make two fragments which were mixed

and then PCR amplified using newguide1 and newguide4 as primers (Table S5). This duplex was purified and assembled using Gibson mix (NEB) with p426-crRNA-CAN1.Y that had been cut with NheI and Acc65I to replace the CAN1.Y guide target region with a stuffer fragment that could be released by BaeI (NEB) and allow any guide to be inserted easily (p426-crRNA-BaeI). We then used p426-crRNA-BaeI as a template to amplify a fragment containing the new cassette with the *SNR52* promoter and the *URA3* gene using oligos trp1-S-ura3 and Cyc-K-SNR52 (Table S5). This fragment was combined with p414-TEF1p-Cas9-CYC1t that had been cut with SnaBI and Acc65I and assembled using Gibson mix to create p416-TEF1p-Cas9-NLS-crRNA-BaeI. The net effect of these manipulations is to (1) combine the guide RNA and Cas9 genes on a single centromeric (low copy) plasmid, (2) create a flexible entry site for any guide sequence, and (3) replace the *TRP1* marker with *URA3*. To target the *S. cerevisiae YTA12* 5'UTR, we cleaved p416-TEF1p-Cas9-NLS-crRNA-BaeI with BaeI, annealed the YTA12_top YTA12_bot 25-mers together (Table S5) and ligated them to the BaeI cleaved plasmid to produce p416-TEF1-Cas9-NLS-CYC1t-crRNA-YTA12. The advantage of this single plasmid system is that guide sequences are more easily inserted, only one plasmid is needed, and cells lacking the plasmid can be selected after editing on 5-fluororotic acid (5-FOA) plates.

Rescue fragments were created by annealing combinations of synthetic oligonucleotides (Table S5) and filling them in with DNA polymerase. These fragments contained the sequences needed to edit the *S. cerevisiae YTA12* 5' UTR so that it contained the *S. bayanus* intron, or was deleted of the intron-syntenic sequences, or contained different 5' ss mutations of the *S. bayanus* intron. Candidate edited yeast clones were grown, and DNA was isolated and analyzed by PCR using primers on either side of the edited site. PCR products were purified and sequenced at the U. C. Berkeley sequencing center to confirm correct editing. Yeast strains determined to contain the correct sequence were streaked on 5-FOA to select clones that have lost the p416-TEF1-Cas9-NLS-CYC1t-crRNA-YTA12 plasmid.

Figure Legends

Figure 1. Identification and Validation of Splicing at Unannotated Genomic Locations. RNA sequencing reads corresponding to spliced RNAs defined as described in the text were used to identify and measure splicing. Reads (~300M) were obtained from untreated cells and cells treated for 60 minutes with rapamycin from two biological replicate experiments. **(A)** Splicing efficiency of many introns improves after rapamycin treatment as judged by change in log2 fold change in the ratio of splice junction reads to intron reads (splicing index) for standard introns (blue dots, ≥35 total junction reads) and for unannotated and non-overlapping locations where splicing takes place (orange dots, ≥50 total junction reads, see text). **(B)** Genomic alignment and lack of conservation for three example protointrons. Splicing signals are boxed. Protointrons in a divergent upstream transcript *antiASH1*, a XUT *XUT12R-370*, and an mRNA for *TAF13* are shown. The vertical bars indicate where additional intron sequences are not shown. **(C)** RT-PCR validation of protointron splicing and increased splicing after rapamycin is shown for the protointrons in **(B)**, and for a non-coding vegetative cell transcript of a meiotic gene *ncSPO1*. PCR products corresponding in size to spliced (predicted based on RNAseq read structure) and unspliced RNAs are labeled. Splice junctions were confirmed by sequencing cloned PCR products. **(D)** Validation of protointron splicing through 5' ss not observed in standard introns. Junctions were validated by sequencing cloned PCR products. **(E)** A protointron in *SUT635* uses both GT-AG and AT-AC splice sites. The sequences of two cloned PCR products from *SUT635* are aligned to the genome (above) and the sequencing trace from the clone representing the use of AT-AC junctions is shown (below).

Figure 2. Differences between Standard Introns and Protointrons. (A) Splice site and branchpoint sequences of protointrons are less constrained in sequence than the standard introns. Weblogos representing position-specific weight matrices of the 5' ss, branch points, and 3' ss of the standard

introns (top) and the protointrons (bottom) are shown. **(B)** The pattern of sequence conservation in the context of protointron splicing signals (orange bars in each panel) is bimodal as compared to the standard introns (blue bars in each panel). Histograms of the standard introns (blue bars) and the protointrons (orange bars) showing the distributions of PhastCons scores in windows containing the 5' ss, branchpoints, and 3' ss of the standard introns and protointrons. See text. **(C)** The size distribution of protointrons is distinct from that of the standard introns. Histograms show the distribution of intron sizes for the standard introns (blue bars) and the protointrons (orange bars). A Kolmogorov-Smirnoff test indicates the two distributions are different ($D = 0.22$, p value $\leq 10^{-4}$). **(D)** Protointrons are much less efficiently spliced than standard introns. The scatter plot shows the relationship between splicing efficiency in untreated (time 0) cells and the change in splicing efficiency after one hour in rapamycin. Standard introns are shown in blue, protointrons in orange.

Figure 3. Hexamers representing splicing signals are depleted from annotated *S. cerevisiae* ORFs. (**A)** Histogram of counts of two hexamers (6-mers) serving as proxies for the branch point (ACTAAC, yellow) and the 5' ss (GTATGT, blue) in the extant ORF set of *S. cerevisiae* (vertical lines) as compared with the distribution of counts in each of 10,000 codon-permuted ORF sets. (**B)** Histogram of Z-scores computed for each of 4096 6-mers in the extant *S. cerevisiae* ORF set relative to their corresponding mean representation in 10,000 codon-permuted *S. cerevisiae* ORF sets. The number of 6-mers (y-axis) with the given Z-score (x-axis) is represented as a histogram in grey. Similar distributions are shown for two subclasses: those containing stop codons (blue histogram) and those containing start codons (maroon histogram). The Z-scores for the branchpoint proxy hexamer ACTAAC and the 5' splice proxy hexamer GTATGT are marked in the plot. The 6-mer "ACTAAC" had a Z-score of -12.25 and ranked 153rd lowest among all 4096 6-mers, and 91st lowest of 759 6-mers carrying stop codons. The 6-mer

GTATGT had a Z-score of -6.98 and ranked 671st lowest among all 4096 6-mers, and 7th lowest of 255 6-mers carrying start codons.

Figure 4. Protointrons are found in other *Saccharomyces* species but are not conserved. RT-PCR products from RNA of *S. mikatae* (left) and *S. bayanus* (right) at different protointrons identified by RNAseq. Splice junctions were validated by cloning and sequencing the PCR products indicated by a white dot. Below the gel image are shown alignments of the RT-PCR product sequences from *S. mikatae antiYCR060W* and *S. bayanus YOL122C* to their corresponding genomes to show lack of conservation of splicing signals (boxed).

Figure 5. Unusually efficient protointrons that may be evolving toward standard introns. Positions of 5 efficiently spliced protointrons that share similarity with standard introns on the unrooted tree of sensu stricto *Saccharomyces* species are shown. Grey arrows indicate separation points that delineate boundaries between species having or lacking the indicated protointron sequence. Bars in the alignments indicate that sequences between these blocks are not shown. Splicing signals are boxed. Although these protointrons are restricted to one or two closely related species, their splicing efficiency approaches that of standard introns. Most protointrons are unique to a species and are very inefficiently spliced.

Figure 6. Introns in the Y' element repeat family. Two different introns are found in the transcribed Y' repeat elements. **(A)** Y' intron 1. Top: expanded view of the protein encoded by *YRF1-6* located in the Y' element at the left end of chromosome XIV with the Sir1 and DECD helicase homology regions indicated. An expanded segment from the upstream part of the gene shows the alignment of the detected intron relative to the annotated predicted intron at SGD. At the bottom is shown the alignment of the seven different versions of this intron from the seventeen Y' elements in the *S. cerevisiae* genome that

possess it. Splicing signals are highlighted. **(B)** Y' intron 2. Top: expanded view of the protein encoded by *YLR464W* located in the Y' element at the right end of chromosome XII. The alignment shows the detected intron relative to the annotated predicted intron at SGD. At the bottom is shown the alignment of the nine different versions of this intron from the nine Y' elements in the *S. cerevisiae* genome that possess it. Splicing signals are highlighted.

Figure 7. The *S. bayanus* YTA12 5'UTR intron is efficiently spliced in *S. cerevisiae* (A) Alignment of the *YTA12* promoter and 5'UTR from *S.cerevisiae* (sacCer, no intron), *S. bayanus* (sacBay, very efficient intron), and the *S. cerevisiae* strain carrying the *S. bayanus* intron (Sc-SbI), showing the major transcription start sites and the cryptic start sites (>), the splice sites (underlined), and the aligned base pairs (*). **(B)** Strategy for CRISPR/Cas9 editing-based transplantation of the *S. bayanus* intron into *S. cerevisiae*. A guide sequence was designed to recognize a sequence present in the *S. cerevisiae YTA12* 5'UTR but not present in the *S. bayanus* intron. A plasmid derived from those provided by DiCarlo et al. [86] expressing this guide along with Cas9 was co-transformed with a synthetic rescue fragment that contained the *S. bayanus* intron sequence between "exons" from *S. cerevisiae*. Repair of the double-stranded break using this rescue fragment results in transplantation of the *S. bayanus* intron into *S. cerevisiae*. **(C)** Reverse transcriptase primer extension analysis of RNA from the YTA12 locus of S. cerevisiae strains with the transplanted *S. bayanus* intron and mutant derivatives. The cDNAs representing unspliced (native) start sites, spliced (or deleted) RNAs initiating from the normal start site, and unspliced RNAs arising from cryptic start sites are indicated at left. Lane 1, wild type; lane 2, deletion of the region that aligns with the *S. bayanus* intron; lane 3, transplantation of the wild type S. bayanus intron; lane 4, GaAUGU mutation of the *S. bayanus* intron 5' ss; lane 5, GaAcGU mutant eliminating both the 5' ss and the start codon of a uORF; lane 6, GUAcGU mutant that creates a common functional ss while removing the start codon of the uORF; lane m, 100 bp ladder markers.

Supplemental Figure 1. (A) Coherence of gene expression changes after 60 minute rapamycin treatment between the two replicate experiments. Log2ratio of treatment to control read coverage over genes was plotted giving an R^2 value of ~0.99. Supplemental to Fig 1A. **(B)** Coverage tracks showing transcription through the genomic locus upstream of *ASH1* where the *antiASH1* protointron is located. Supplemental to Figs 1 B and C. **(C)** Coverage tracks showing transcription through the genomic locus upstream of *TUS1* where the *XUT12R-370* protointron is located. Supplemental to Figs 1 B and C. **(D)** Coverage tracks showing transcription through the genomic locus of *TAF13* where the *TAF13* protointron is located. Supplemental to Figs 1 B and C. **(E)** Coverage tracks showing transcription through the genomic locus of *SPO1* where the *ncSPO1* protointron is located. Supplemental to Figs 1 B and C. **(F)** Alignment of sequenced RT-PCR products showing the location of protointrons with unusual 5' ss. Supplemental to Fig 1D.

Acknowledgements

Thanks very much to Joshua Arribere and Russ Corbett-Detig for insightful critical analysis of an earlier draft of this work. We appreciate Sam Fagg, Jen Quick-Cleveland, Stephanie Nystrom, and Santiago Sanchez for their constructive comments on this work. Thanks to Kevin Karplus for suggesting that codon-permutation could be used to create randomized ORF sets that preserve codon usage and GC content. We thank UCLA colleagues Tracy Johnson and Stephen Douglass for discussions and sharing results prior to publication. This work was supported by grant GM040478 from NIGMS to M.A.

References

1. Fica SM, Nagai K. Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. Nat Struct Mol Biol. 2017;24(10):791-9. Epub 2017/10/06. doi: 10.1038/nsmb.3463. PubMed PMID: 28981077.
2. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu Rev Biochem. 2015;84:291-323. Epub 2015/03/19. doi: 10.1146/annurev-biochem-060614-034316. PubMed PMID: 25784052; PubMed Central PMCID: PMCPMC4526142.
3. Scheres SH, Nagai K. CryoEM structures of spliceosomal complexes reveal the molecular mechanism of pre-mRNA splicing. Curr Opin Struct Biol. 2017;46:130-9. Epub 2017/09/10. doi: 10.1016/j.sbi.2017.08.001. PubMed PMID: 28888105.
4. Catania F, Lynch M. Where do introns come from? PLoS Biol. 2008;6(11):e283. Epub 2008/12/11. doi: 10.1371/journal.pbio.0060283. PubMed PMID: 19067485; PubMed Central PMCID: PMCPMC2586383.
5. Lynch M, Richardson AO. The evolution of spliceosomal introns. Curr Opin Genet Dev. 2002;12(6):701-10. PubMed PMID: 12433585.
6. Carmel L, Wolf YI, Rogozin IB, Koonin EV. Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. 2007;17(7):1034-44. Epub 2007/05/15. doi: 10.1101/gr.6438607. PubMed PMID: 17495008; PubMed Central PMCID: PMCPMC1899114.
7. Fink GR. Pseudogenes in yeast? Cell. 1987;49(1):5-6. Epub 1987/04/10. PubMed PMID: 3549000.
8. Yenerall P, Zhou L. Identifying the mechanisms of intron gain: progress and trends. Biol Direct. 2012;7:29. Epub 2012/09/12. doi: 10.1186/1745-6150-7-29. PubMed PMID: 22963364; PubMed Central PMCID: PMCPMC3443670.
9. Rogozin IB, Lyons-Weiler J, Koonin EV. Intron sliding in conserved gene families. Trends Genet. 2000;16(10):430-2. Epub 2000/10/26. PubMed PMID: 11050324.
10. Sorek R. The birth of new exons: mechanisms and evolutionary consequences. RNA. 2007;13(10):1603-8. Epub 2007/08/22. doi: 10.1261/rna.682507. PubMed PMID: 17709368; PubMed Central PMCID: PMCPMC1986822.
11. Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW. Origin of introns by 'intronization' of exonic sequences. Trends Genet. 2008;24(8):378-81. Epub 2008/07/04. doi: 10.1016/j.tig.2008.05.007. PubMed PMID: 18597887.
12. Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, et al. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science. 2009;324(5924):268-72. Epub 2009/04/11. doi: 10.1126/science.1167222. PubMed PMID: 19359590.
13. Collemare J, Beenen HG, Crous PW, de Wit PJ, van der Burgt A. Novel Introner-Like Elements in fungi Are Involved in Parallel Gains of Spliceosomal Introns. PLoS One. 2015;10(6):e0129302. Epub 2015/06/06. doi: 10.1371/journal.pone.0129302. PubMed PMID: 26046656; PubMed Central PMCID: PMCPMC4457414.
14. Collemare J, van der Burgt A, de Wit PJ. At the origin of spliceosomal introns: Is multiplication of introner-like elements the main mechanism of intron gain in fungi? Commun Integr Biol. 2013;6(2):e23147. Epub 2013/06/12. doi: 10.4161/cib.23147. PubMed PMID: 23750299; PubMed Central PMCID: PMCPMC3609843.
15. Simmons MP, Bachy C, Sudek S, van Baren MJ, Sudek L, Ares M, Jr., et al. Intron Invasions Trace Algal Speciation and Reveal Nearly Identical Arctic and Antarctic Micromonas Populations. Mol Biol Evol. 2015;32(9):2219-35. Epub 2015/05/23. doi: 10.1093/molbev/msv122. PubMed PMID: 25998521; PubMed Central PMCID: PMCPMC4540971.
16. Verhelst B, Van de Peer Y, Rouze P. The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. Genome Biol Evol. 2013;5(12):2393-401.

Epub 2013/11/26. doi: 10.1093/gbe/evt189. PubMed PMID: 24273312; PubMed Central PMCID: PMCPMC3879977.
17. van der Burgt A, Severing E, de Wit PJ, Collemare J. Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. Curr Biol. 2012;22(13):1260-5. Epub 2012/06/05. doi: 10.1016/j.cub.2012.05.011. PubMed PMID: 22658596.
18. Li W, Tucker AE, Sung W, Thomas WK, Lynch M. Extensive, recent intron gains in Daphnia populations. Science. 2009;326(5957):1260-2. Epub 2009/12/08. doi: 10.1126/science.1179302. PubMed PMID: 19965475; PubMed Central PMCID: PMCPMC3878872.
19. Huff JT, Zilberman D, Roy SW. Mechanism for DNA transposons to generate introns on genomic scales. Nature. 2016. doi: 10.1038/nature20110. PubMed PMID: 27760113.
20. Lee S, Stevens SW. Spliceosomal intronogenesis. Proc Natl Acad Sci U S A. 2016;113(23):6514-9. doi: 10.1073/pnas.1605113113. PubMed PMID: 27217561; PubMed Central PMCID: PMCPMC4988565.
21. Lopez PJ, Seraphin B. Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition. RNA. 1999;5(9):1135-7. Epub 1999/09/25. PubMed PMID: 10496213; PubMed Central PMCID: PMCPMC1369835.
22. Ares M, Jr., Grate L, Pauling MH. A handful of intron-containing genes produces the lion's share of yeast mRNA. RNA. 1999;5(9):1138-9. Epub 1999/09/25. PubMed PMID: 10496214; PubMed Central PMCID: PMCPMC1369836.
23. Warner JR. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci. 1999;24(11):437-40. Epub 1999/11/05. PubMed PMID: 10542411.
24. Munding EM, Igel AH, Shiue L, Dorighi KM, Trevino LR, Ares M, Jr. Integration of a splicing regulatory network within the meiotic gene expression program of Saccharomyces cerevisiae. Genes Dev. 2010;24(23):2693-704. Epub 2010/12/03. doi: 10.1101/gad.1977410. PubMed PMID: 21123654; PubMed Central PMCID: PMCPMC2994042.
25. Munding EM, Shiue L, Katzman S, Donohue JP, Ares M, Jr. Competition between pre-mRNAs for the splicing machinery drives global regulation of splicing. Mol Cell. 2013;51(3):338-48. doi: 10.1016/j.molcel.2013.06.012. PubMed PMID: 23891561; PubMed Central PMCID: PMCPMC3771316.
26. Venkataramanan S, Douglass S, Galivanche AR, Johnson TL. The chromatin remodeling complex Swi/Snf regulates splicing of meiotic transcripts in Saccharomyces cerevisiae. Nucleic Acids Res. 2017;45(13):7708-21. Epub 2017/06/24. doi: 10.1093/nar/gkx373. PubMed PMID: 28637241; PubMed Central PMCID: PMCPMC5570110.
27. Awad AM, Venkataramanan S, Nag A, Galivanche AR, Bradley MC, Neves LT, et al. Chromatin-remodeling SWI/SNF complex regulates coenzyme Q6 synthesis and a metabolic shift to respiration in yeast. J Biol Chem. 2017;292(36):14851-66. Epub 2017/07/26. doi: 10.1074/jbc.M117.798397. PubMed PMID: 28739803; PubMed Central PMCID: PMCPMC5592666.
28. Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, et al. Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. Cell Rep. 2014;7(6):1858-66. doi: 10.1016/j.celrep.2014.05.023. PubMed PMID: 24931603; PubMed Central PMCID: PMCPMC4105149.
29. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012;487(7407):370-4. doi: 10.1038/nature11184. PubMed PMID: 22722833; PubMed Central PMCID: PMCPMC3401362.
30. van Dijk EL, Chen CL, d'Aubenton-Carafa Y, Gourvennec S, Kwapisz M, Roche V, et al. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature. 2011;475(7354):114-7. doi: 10.1038/nature10118. PubMed PMID: 21697827.

31. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, Camblong J, et al. Bidirectional promoters generate pervasive transcription in yeast. Nature. 2009;457(7232):1033-7. doi: 10.1038/nature07728. PubMed PMID: 19169243; PubMed Central PMCID: PMCPMC2766638.
32. Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature. 2009;457(7232):1038-42. doi: 10.1038/nature07747. PubMed PMID: 19169244.
33. Ito T, Miura F, Onda M. Unexpected complexity of the budding yeast transcriptome. IUBMB Life. 2008;60(12):775-81. doi: 10.1002/iub.121. PubMed PMID: 18649367.
34. Davis CA, Ares M, Jr. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2006;103(9):3262-7. Epub 2006/02/18. doi: 10.1073/pnas.0507783103. PubMed PMID: 16484372; PubMed Central PMCID: PMCPMC1413877.
35. Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, et al. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell. 2005;121(5):725-37. Epub 2005/06/07. doi: 10.1016/j.cell.2005.04.030. PubMed PMID: 15935759.
36. Arribere JA, Gilbert WV. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. Genome Res. 2013;23(6):977-87. Epub 2013/04/13. doi: 10.1101/gr.150342.112. PubMed PMID: 23580730; PubMed Central PMCID: PMCPMC3668365.
37. Gould GM, Paggi JM, Guo Y, Phizicky DV, Zinshteyn B, Wang ET, et al. Identification of new branch points and unconventional introns in Saccharomyces cerevisiae. RNA. 2016;22(10):1522-34. Epub 2016/07/31. doi: 10.1261/rna.057216.116. PubMed PMID: 27473169; PubMed Central PMCID: PMCPMC5029451.
38. Kawashima T, Douglass S, Gabunilas J, Pellegrini M, Chanfreau GF. Widespread use of non-productive alternative splice sites in Saccharomyces cerevisiae. PLoS Genet. 2014;10(4):e1004249. doi: 10.1371/journal.pgen.1004249. PubMed PMID: 24722551; PubMed Central PMCID: PMCPMC3983031.
39. Qin D, Huang L, Wlodaver A, Andrade J, Staley JP. Sequencing of lariat termini in S. cerevisiae reveals 5' splice sites, branch points, and novel splicing events. RNA. 2016;22(2):237-53. Epub 2015/12/10. doi: 10.1261/rna.052829.115. PubMed PMID: 26647463; PubMed Central PMCID: PMCPMC4712674.
40. Volanakis A, Passoni M, Hector RD, Shah S, Kilchert C, Granneman S, et al. Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast. Genes Dev. 2013;27(18):2025-38. Epub 2013/09/26. doi: 10.1101/gad.221960.113. PubMed PMID: 24065768; PubMed Central PMCID: PMCPMC3792478.
41. Aslanzadeh V, Huang Y, Sanguinetti G, Beggs JD. Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. Genome Res. 2018;28(2):203-13. Epub 2017/12/20. doi: 10.1101/gr.225615.117. PubMed PMID: 29254943; PubMed Central PMCID: PMCPMC5793784.
42. Schreiber K, Csaba G, Haslbeck M, Zimmer R. Alternative Splicing in Next Generation Sequencing Data of Saccharomyces cerevisiae. PLoS One. 2015;10(10):e0140487. Epub 2015/10/16. doi: 10.1371/journal.pone.0140487. PubMed PMID: 26469855; PubMed Central PMCID: PMCPMC4607428.
43. Douglass S, Leung CS, Johnson TJ. Extensive Splicing across the Saccharomyces genome. bioRxiv. 2019. doi: 10.1101/515163.
44. Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, et al. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. Proc Natl Acad Sci U S A. 2006;103(47):17846-51. Epub 2006/11/15. doi: 10.1073/pnas.0605645103. PubMed PMID: 17101987; PubMed Central PMCID: PMCPMC1693835.

45. He F, Jacobson A. Identification of a novel component of the nonsense-mediated mRNA decay pathway by use of an interacting protein screen. Genes Dev. 1995;9(4):437-54. Epub 1995/02/15. PubMed PMID: 7883168.
46. Kandels-Lewis S, Seraphin B. Involvement of U6 snRNA in 5' splice site selection. Science. 1993;262(5142):2035-9. Epub 1993/12/24. PubMed PMID: 8266100.
47. Lesser CF, Guthrie C. Mutations in U6 snRNA that alter splice site specificity: implications for the active site. Science. 1993;262(5142):1982-8. Epub 1993/12/24. PubMed PMID: 8266093.
48. Roca X, Krainer AR. Recognition of atypical 5' splice sites by shifted base-pairing to U1 snRNA. Nat Struct Mol Biol. 2009;16(2):176-82. Epub 2009/01/27. doi: 10.1038/nsmb.1546. PubMed PMID: 19169258; PubMed Central PMCID: PMCPMC2719486.
49. Wu Q, Krainer AR. Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. RNA. 1997;3(6):586-601. Epub 1997/06/01. PubMed PMID: 9174094; PubMed Central PMCID: PMCPMC1369508.
50. Parker R, Siliciano PG. Evidence for an essential non-Watson-Crick interaction between the first and last nucleotides of a nuclear pre-mRNA intron. Nature. 1993;361(6413):660-2. Epub 1993/02/18. doi: 10.1038/361660a0. PubMed PMID: 8437627.
51. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-50. Epub 2005/07/19. doi: 10.1101/gr.3715005. PubMed PMID: 16024819; PubMed Central PMCID: PMCPMC1182216.
52. Spingola M, Grate L, Haussler D, Ares M, Jr. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. RNA. 1999;5(2):221-34. Epub 1999/02/19. PubMed PMID: 10024174; PubMed Central PMCID: PMCPMC1369754.
53. Howe KJ, Ares M, Jr. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. Proc Natl Acad Sci U S A. 1997;94(23):12467-72. Epub 1997/11/14. PubMed PMID: 9356473; PubMed Central PMCID: PMCPMC25003.
54. Libri D, Stutz F, McCarthy T, Rosbash M. RNA structural patterns and splicing: molecular basis for an RNA-based enhancer. RNA. 1995;1(4):425-36. Epub 1995/06/01. PubMed PMID: 7493320; PubMed Central PMCID: PMCPMC1482409.
55. Karaduman R, Chanarat S, Pfander B, Jentsch S. Error-Prone Splicing Controlled by the Ubiquitin Relative Hub1. Mol Cell. 2017;67(3):423-32 e4. Epub 2017/07/18. doi: 10.1016/j.molcel.2017.06.021. PubMed PMID: 28712727.
56. Juneau K, Nislow C, Davis RW. Alternative splicing of PTC7 in Saccharomyces cerevisiae determines protein localization. Genetics. 2009;183(1):185-94. Epub 2009/07/01. doi: 10.1534/genetics.109.105155. PubMed PMID: 19564484; PubMed Central PMCID: PMCPMC2746143.
57. Zhang Z, Hesselberth JR, Fields S. Genome-wide identification of spliced introns using a tiling microarray. Genome Res. 2007;17(4):503-9. Epub 2007/03/14. doi: 10.1101/gr.6049107. PubMed PMID: 17351133; PubMed Central PMCID: PMCPMC1832097.
58. Zhou Z, Dang Y, Zhou M, Yuan H, Liu Y. Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. Elife. 2018;7. Epub 2018/03/17. doi: 10.7554/eLife.33569. PubMed PMID: 29547124; PubMed Central PMCID: PMCPMC5869017.
59. Gamble CE, Brule CE, Dean KM, Fields S, Grayhack EJ. Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. Cell. 2016;166(3):679-90. Epub 2016/07/05. doi: 10.1016/j.cell.2016.05.070. PubMed PMID: 27374328; PubMed Central PMCID: PMCPMC4967012.
60. Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, et al. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces

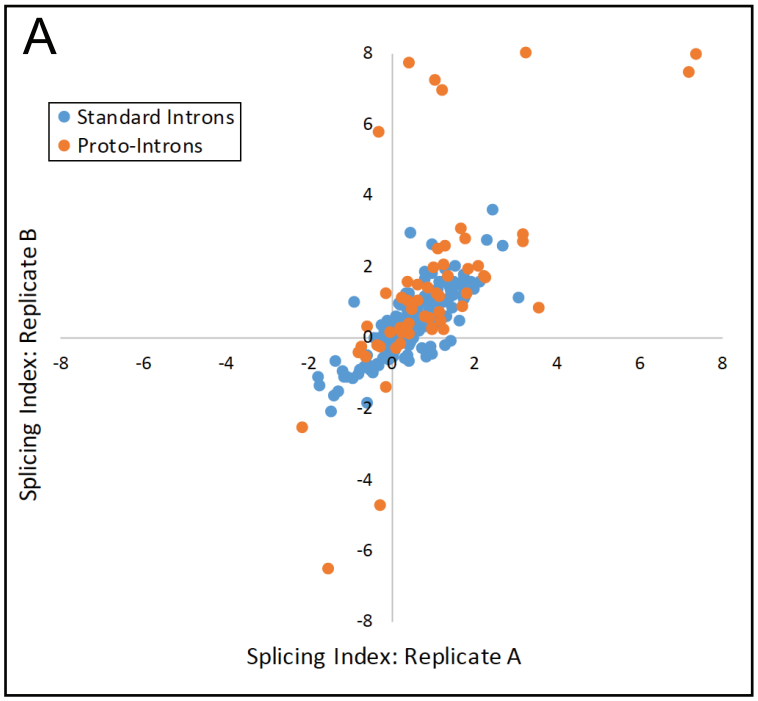
sensu stricto Genus. G3 (Bethesda). 2011;1(1):11-25. Epub 2012/03/03. doi: 10.1534/g3.111.000273. PubMed PMID: 22384314; PubMed Central PMCID: PMCPMC3276118.
61. Parenteau J, Durand M, Veronneau S, Lacombe AA, Morin G, Guerin V, et al. Deletion of many yeast introns reveals a minority of genes that require splicing for function. Mol Biol Cell. 2008;19(5):1932-41. Epub 2008/02/22. doi: 10.1091/mbc.E07-12-1254. PubMed PMID: 18287520; PubMed Central PMCID: PMCPMC2366882.
62. Davis CA, Grate L, Spingola M, Ares M, Jr. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. Nucleic Acids Res. 2000;28(8):1700-6. Epub 2000/03/29. PubMed PMID: 10734188; PubMed Central PMCID: PMCPMC102823.
63. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 2003;423(6937):241-54. doi: 10.1038/nature01644. PubMed PMID: 12748633.
64. Arlt H, Tauer R, Feldmann H, Neupert W, Langer T. The YTA10-12 complex, an AAA protease with chaperone-like activity in the inner membrane of mitochondria. Cell. 1996;85(6):875-85. Epub 1996/06/14. PubMed PMID: 8681382.
65. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science. 2012;335(6068):552-7. Epub 2011/12/24. doi: 10.1126/science.1215110. PubMed PMID: 22194413; PubMed Central PMCID: PMCPMC3414261.
66. Louis EJ, Haber JE. The structure and evolution of subtelomeric Y' repeats in Saccharomyces cerevisiae. Genetics. 1992;131(3):559-74. Epub 1992/07/01. PubMed PMID: 1628806; PubMed Central PMCID: PMCPMC1205030.
67. Yamada M, Hayatsu N, Matsuura A, Ishikawa F. Y'-Help1, a DNA helicase encoded by the yeast subtelomeric Y' element, is induced in survivors defective for telomerase. J Biol Chem. 1998;273(50):33360-6. Epub 1998/12/05. PubMed PMID: 9837911.
68. Meyer M, Plass M, Perez-Valle J, Eyras E, Vilardell J. Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. Mol Cell. 2011;43(6):1033-9. Epub 2011/09/20. doi: 10.1016/j.molcel.2011.07.030. PubMed PMID: 21925391.
69. Kannan R, Hartnett S, Voelker RB, Berglund JA, Staley JP, Baumann P. Intronic sequence elements impede exon ligation and trigger a discard pathway that yields functional telomerase RNA in fission yeast. Genes Dev. 2013;27(6):627-38. Epub 2013/03/08. doi: 10.1101/gad.212738.112. PubMed PMID: 23468430; PubMed Central PMCID: PMCPMC3613610.
70. Mele M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. Genome Res. 2017;27(1):27-37. Epub 2016/12/09. doi: 10.1101/gr.214205.116. PubMed PMID: 27927715; PubMed Central PMCID: PMCPMC5204342.
71. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505(7485):635-40. Epub 2014/01/28. doi: 10.1038/nature12943. PubMed PMID: 24463510.
72. Liu J, Martin-Yken H, Bigey F, Dequin S, Francois JM, Capp JP. Natural yeast promoter variants reveal epistasis in the generation of transcriptional-mediated noise and its potential benefit in stressful conditions. Genome Biol Evol. 2015;7(4):969-84. Epub 2015/03/13. doi: 10.1093/gbe/evv047. PubMed PMID: 25762217; PubMed Central PMCID: PMCPMC4419794.
73. Pickering BM, Willis AE. The implications of structured 5' untranslated regions on translation and disease. Semin Cell Dev Biol. 2005;16(1):39-47. Epub 2005/01/22. doi: 10.1016/j.semcdb.2004.11.006. PubMed PMID: 15659338.

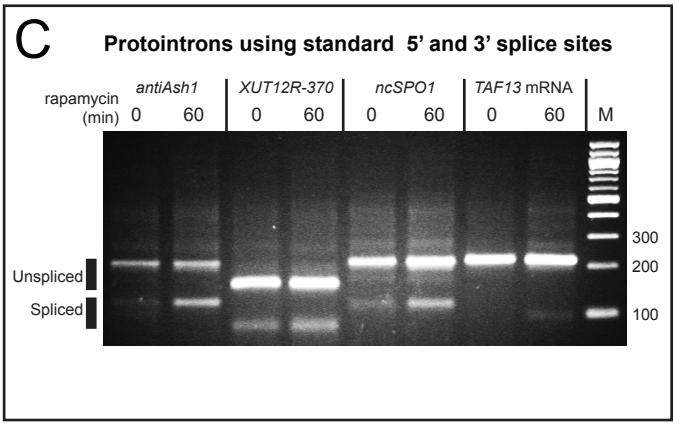
74. Andreatta ME, Levine JA, Foy SG, Guzman LD, Kosinski LJ, Cordes MH, et al. The Recent De Novo Origin of Protein C-Termini. Genome Biol Evol. 2015;7(6):1686-701. doi: 10.1093/gbe/evv098. PubMed PMID: 26002864; PubMed Central PMCID: PMCPMC4494051.
75. Tress ML, Abascal F, Valencia A. Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem Sci. 2017;42(2):98-110. Epub 2016/10/08. doi: 10.1016/j.tibs.2016.08.008. PubMed PMID: 27712956.
76. Ares M. Isolation of total RNA from yeast cell cultures. Cold Spring Harb Protoc. 2012;2012(10):1082-6. Epub 2012/10/03. doi: 10.1101/pdb.prot071456. PubMed PMID: 23028070.
77. Yassour M, Pfiffner J, Levin JZ, Adiconis X, Gnirke A, Nusbaum C, et al. Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. Genome Biol. 2010;11(8):R87. Epub 2010/08/28. doi: 10.1186/gb-2010-11-8-r87. PubMed PMID: 20796282; PubMed Central PMCID: PMCPMC2945789.
78. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 2012;40(Database issue):D700-5. Epub 2011/11/24. doi: 10.1093/nar/gkr1029. PubMed PMID: 22110037; PubMed Central PMCID: PMCPMC3245034.
79. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9. Epub 2012/03/06. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PubMed Central PMCID: PMCPMC3322381.
80. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. Epub 2012/10/30. doi: 10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PubMed Central PMCID: PMCPMC3530905.
81. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. Epub 2014/12/18. doi: 10.1186/s13059-014-0550-8. PubMed PMID: 25516281; PubMed Central PMCID: PMCPMC4302049.
82. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188-90. Epub 2004/06/03. doi: 10.1101/gr.849004. PubMed PMID: 15173120; PubMed Central PMCID: PMCPMC419797.
83. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):e115. Epub 2012/06/26. doi: 10.1093/nar/gks596. PubMed PMID: 22730293; PubMed Central PMCID: PMCPMC3424584.
84. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64. Epub 2002/04/05. doi: 10.1101/gr.229202. PubMed PMID: 11932250; PubMed Central PMCID: PMCPMC187518.
85. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42(Database issue):D764-70. Epub 2013/11/26. doi: 10.1093/nar/gkt1168. PubMed PMID: 24270787; PubMed Central PMCID: PMCPMC3964947.
86. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. 2013;41(7):4336-43. Epub 2013/03/06. doi: 10.1093/nar/gkt135. PubMed PMID: 23460208; PubMed Central PMCID: PMCPMC3627607.

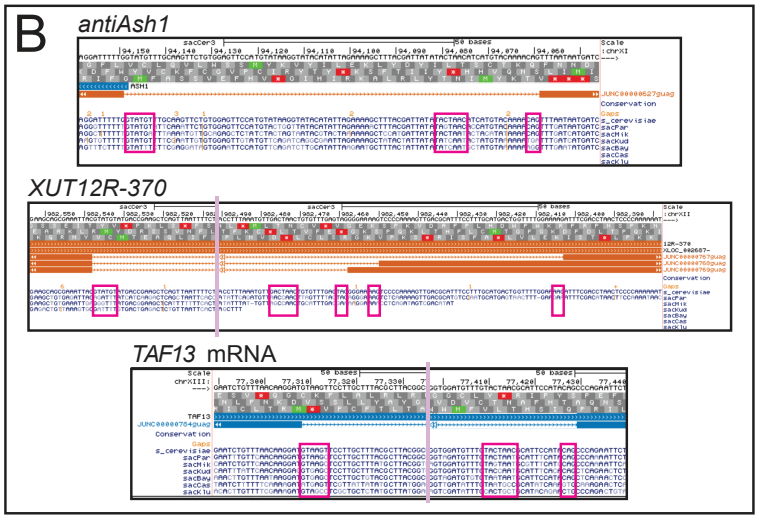
Table 1. Easily detectable protointrons in *Saccharomyces cerevisiae*

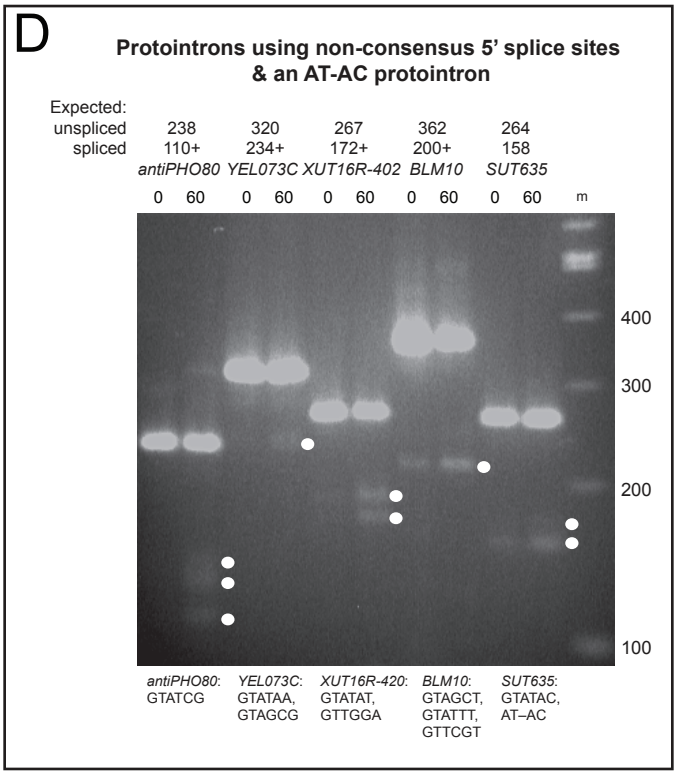
Gene	Chr	Overlapping Junctions	5'SS	Consv 5'SS	BP (predicted)	Consv BP	3'SS	Consv 3'SS	% Spliced 0 min.	% Spliced 60 min.
MTR2	XI	5	GTACGT GTATGT	0.04 0.00	AACTAATA AACTAACC	0.02 0.00	CAG CAG TAG	0.42 0.03 0.00	57	62
Opposing YPL216W	XVI	2	GTATGA GTTTGT	0.79 0.00	TACTAACA	0.77	AAG	0.16 0.14	22	35
YEL023C	V	1	GTATGG	0.00	CACTAACA	0.27	TAG	0.41 0.18	23	17
ZTA1	II	2	GTATGA	0.95	TACTAATC	0.57	CAG AAG	0.98 0.99	1.6	5.3
PDX3	II	0	GTATCA	1.00	TACTGACG	0.50	AAG	1.00	1.6	1.9
TVP15	IV	1	GTATGC	0.98	TACTAAGT	0.16	CAG TAG	0.62 0.62	2.2	4.7
PUS7	XV	3	GTAAGG	0.99	TACTAACA	0.89	AAG TAG AAG AAG	0.97 0.62 0.62 0.96	6.5	4.4
YDR336W	IV	1	GTACGT	0.06	CACTAAAA	0.12	TAG AAG	0.02 0.08	18	44
NTH1	IV	1	GTATGG	0.99	TGCTAACA	1.00	CAG TAG	1.00 1.00	1.4	4.0
SPO1	XIV	0	GTATGT	0.52	TATTAACC	0.96	CAG	0.66	6.8	14
MCA1	XV	0	GTACAG	0.02	GACTAATG	0.97	AAG	0.68	2.7	2.5
Opposing IPT1	IV	1	GTAAGT	0.37	TACTAACT	0.25	AAG CAG	0.44 0.29	0.11	18
PDC1	XII	1	GTATGT	1.00	CACTGACA	1.00	CAG	1.00 1.00	0.03	0.05
Downstream of PDR12	XVI	1	GTGTGT	0.01	GACTAACA	0.04	CAG AAG	0.00 0.02	36	31
SYN8	I	2	GTATGG	0.99	AACTAATA	0.23	TAG CAG CAG	0.96 0.95 0.96	2.1	3.4

Talkish et al Fig 1

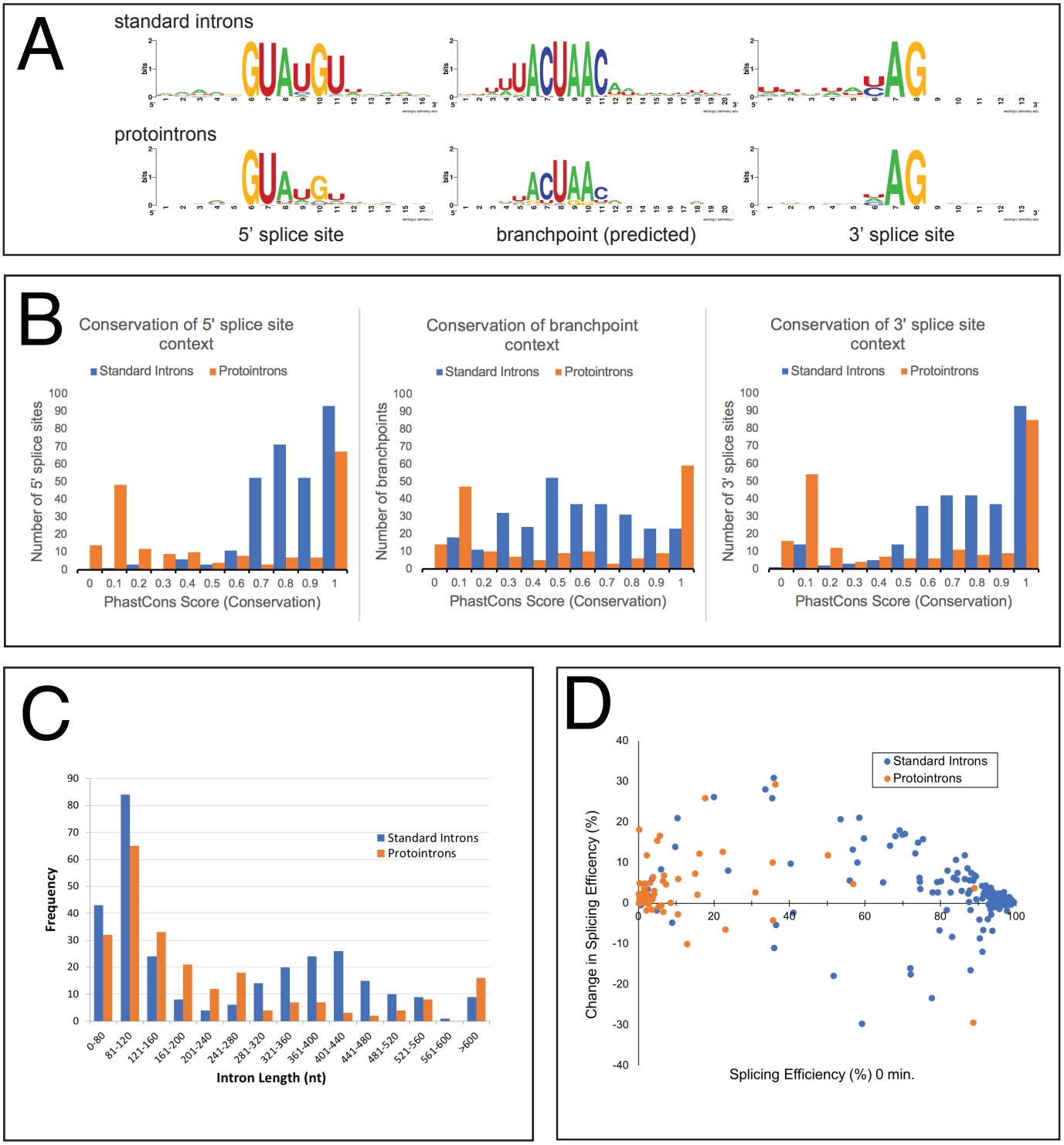




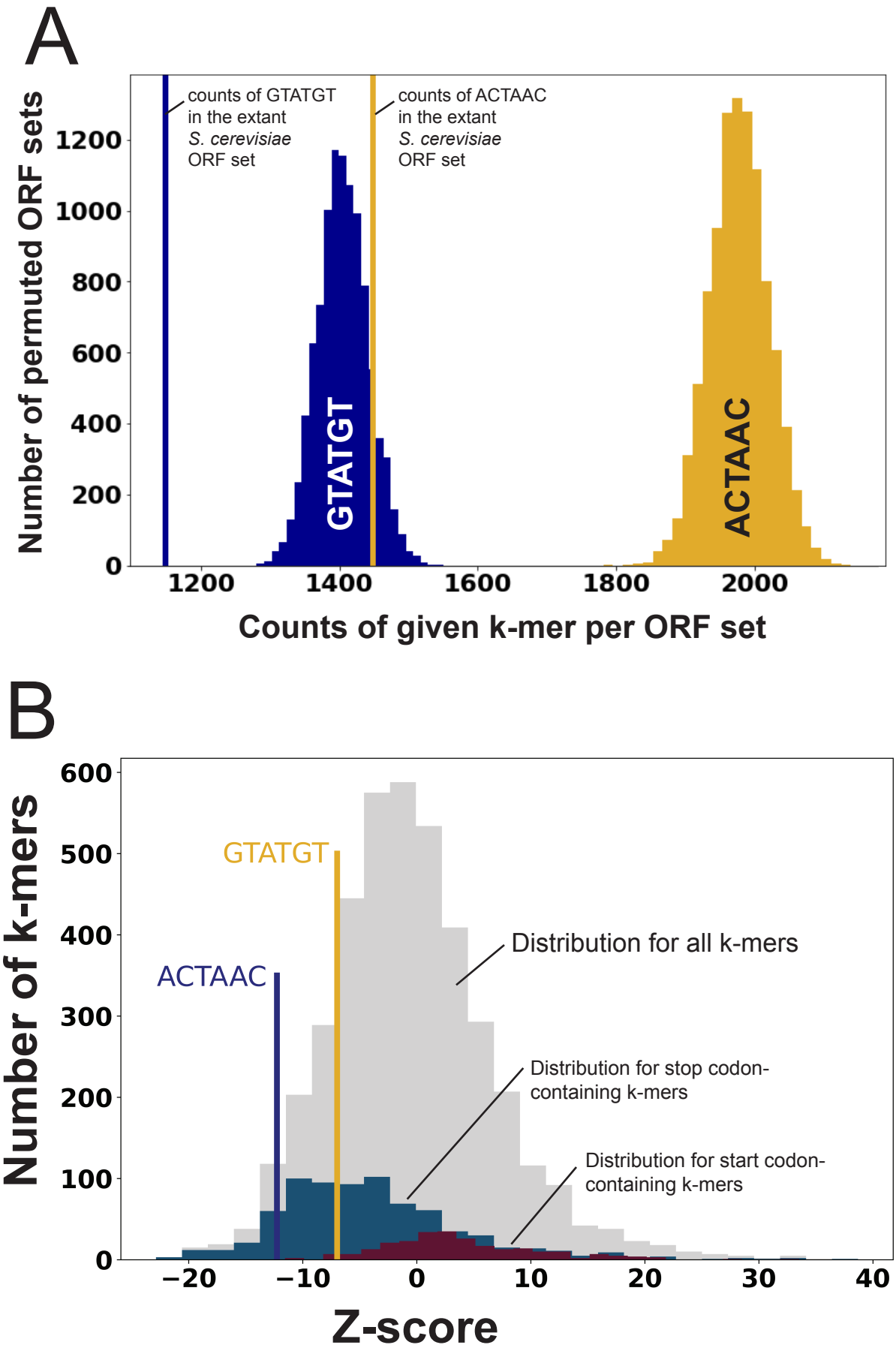




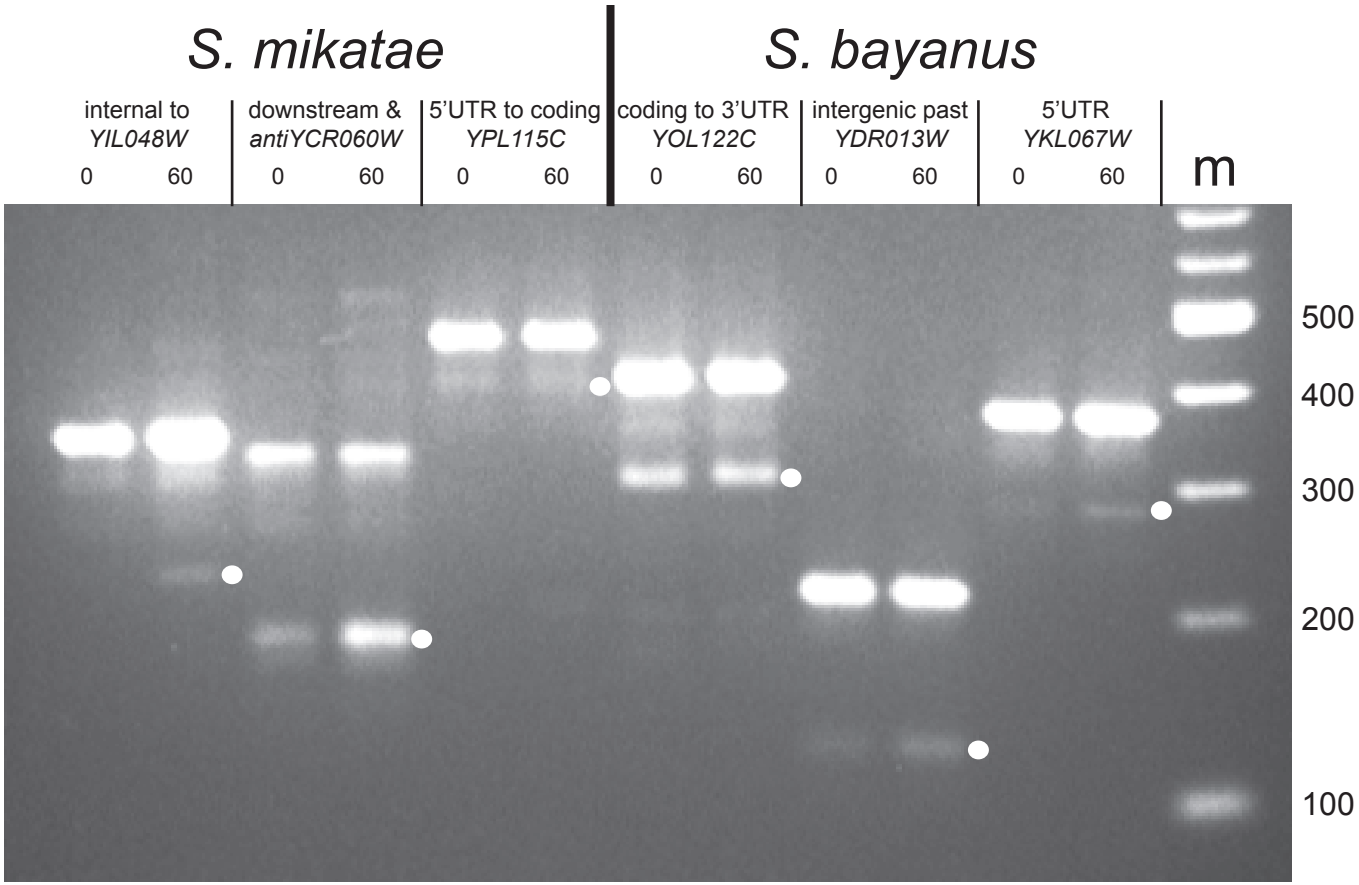
Talkish et al Fig 2



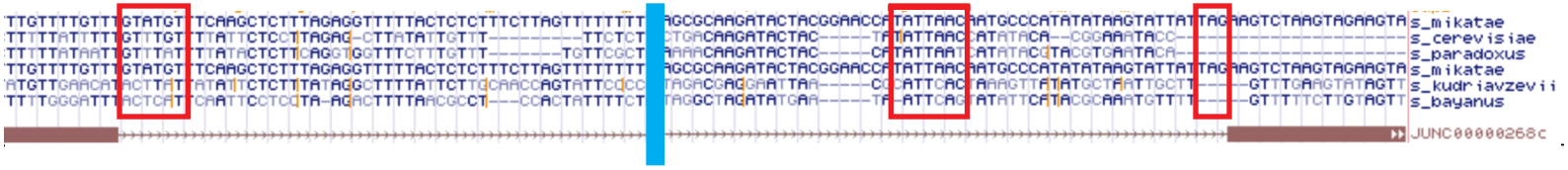
Talkish et al. Fig 3



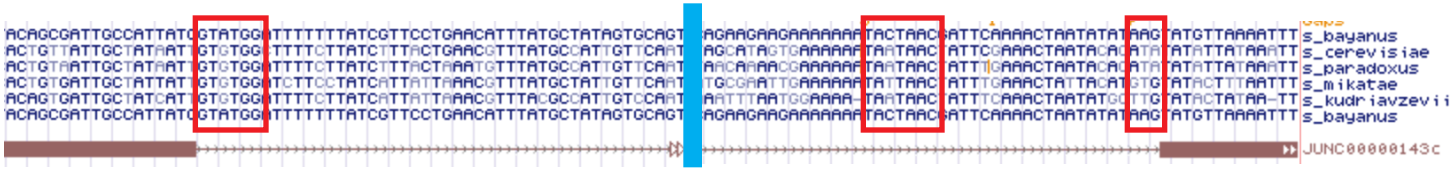
Talkish et al Fig 4



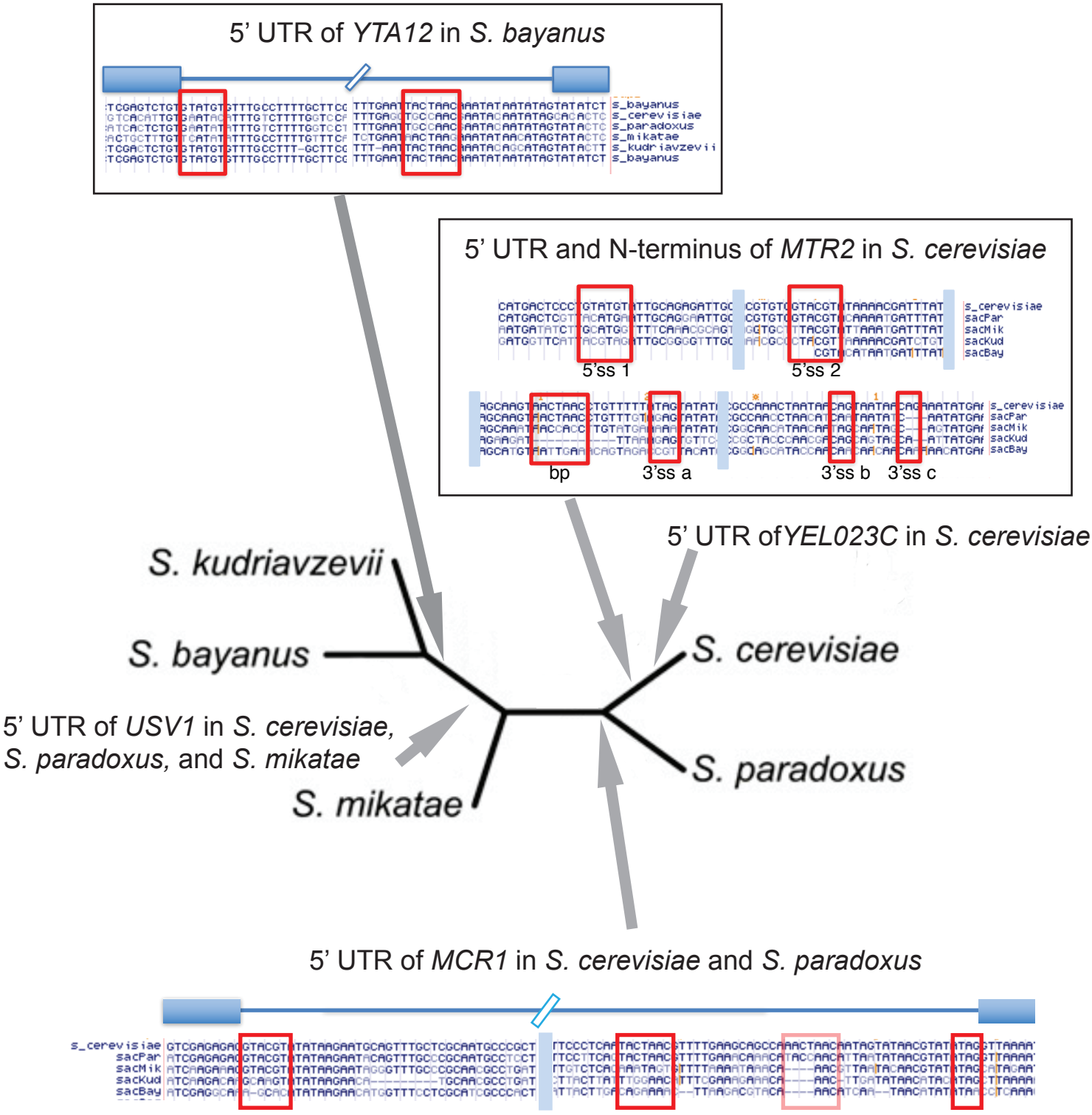
S. mikatae downstream & *antiYCR060W*



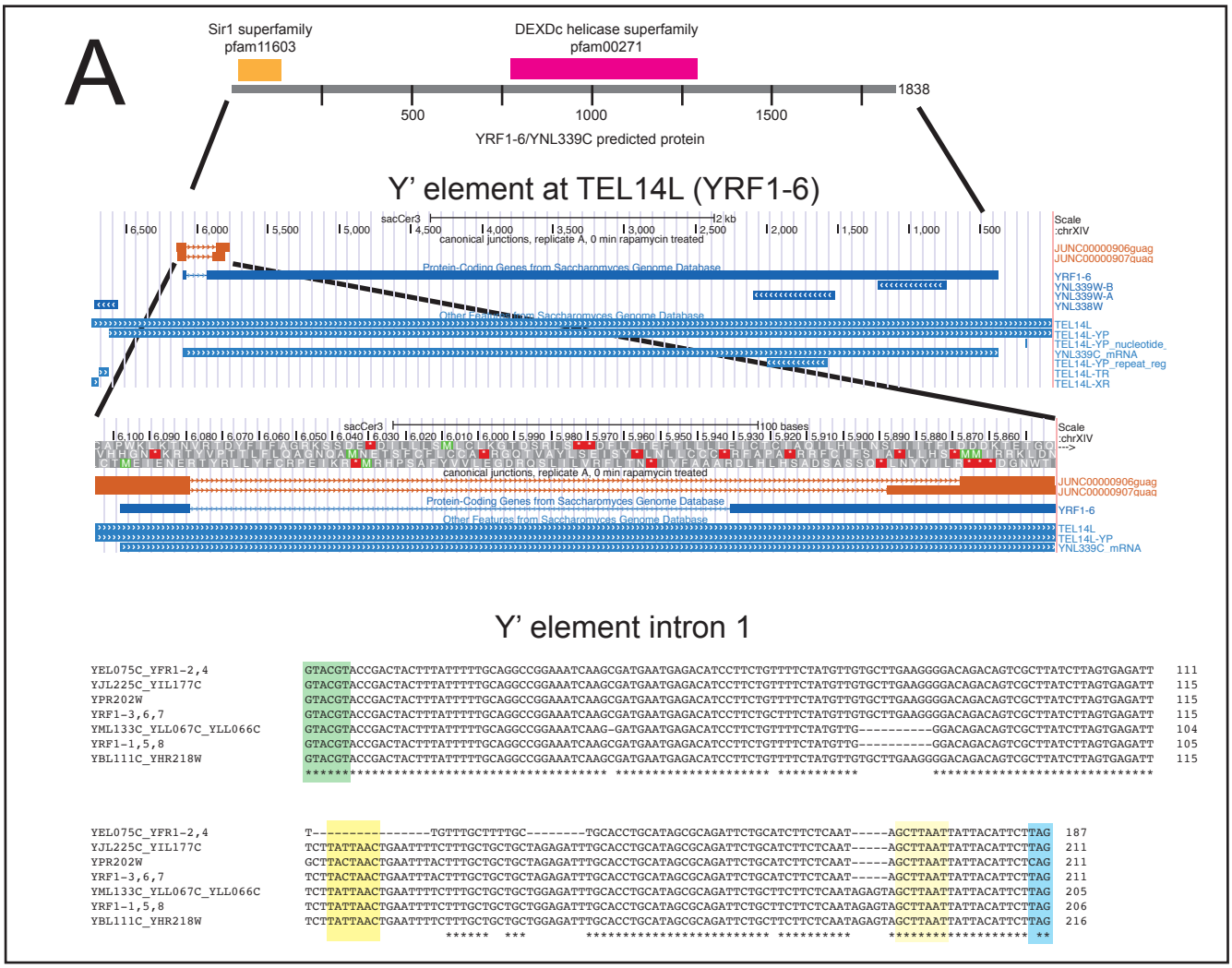
S. bayanus coding to 3'UTR *YOL122C*

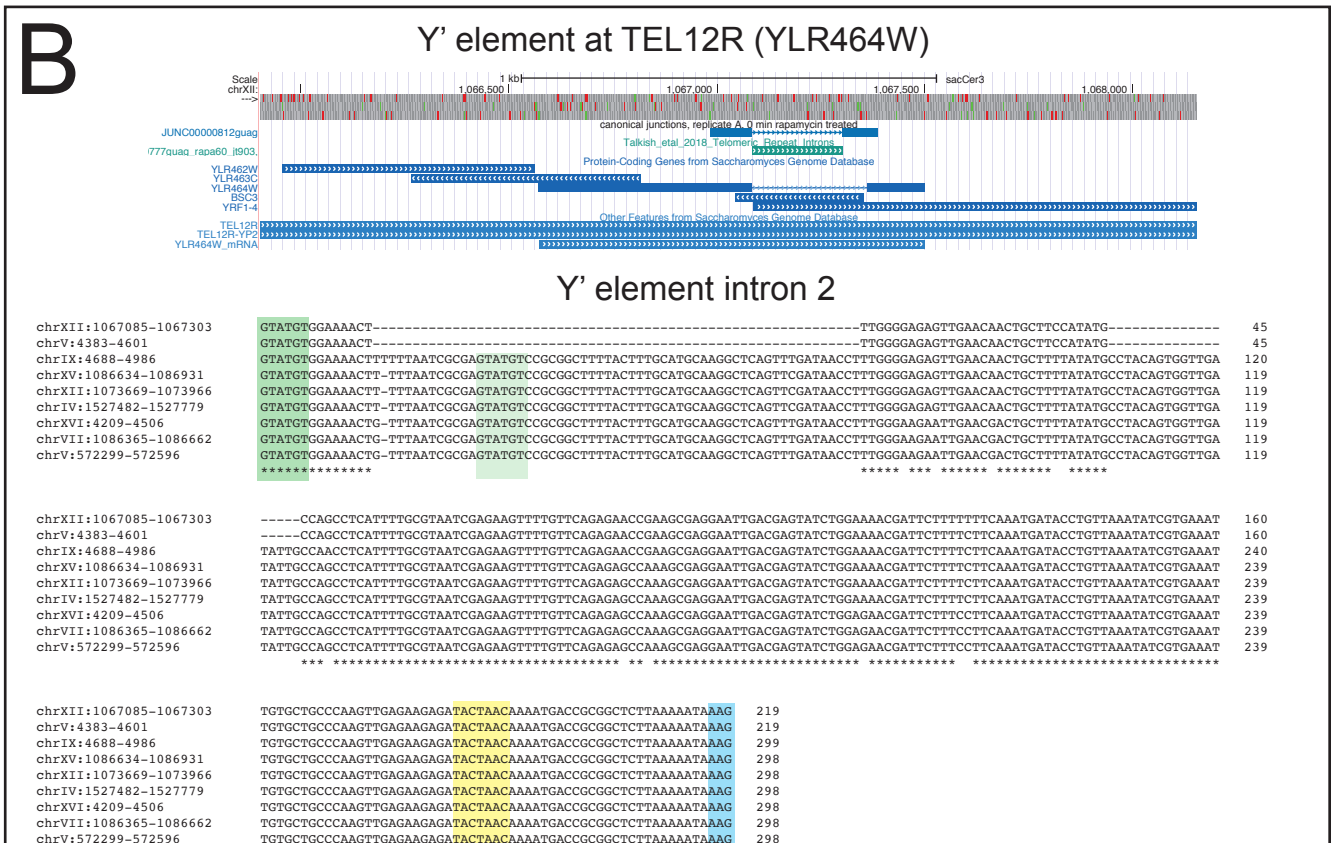


Talkish et al Fig 5



Talkish et al Fig 6



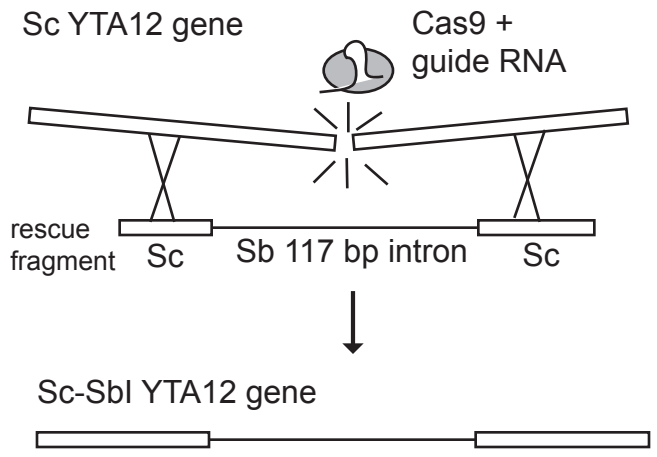
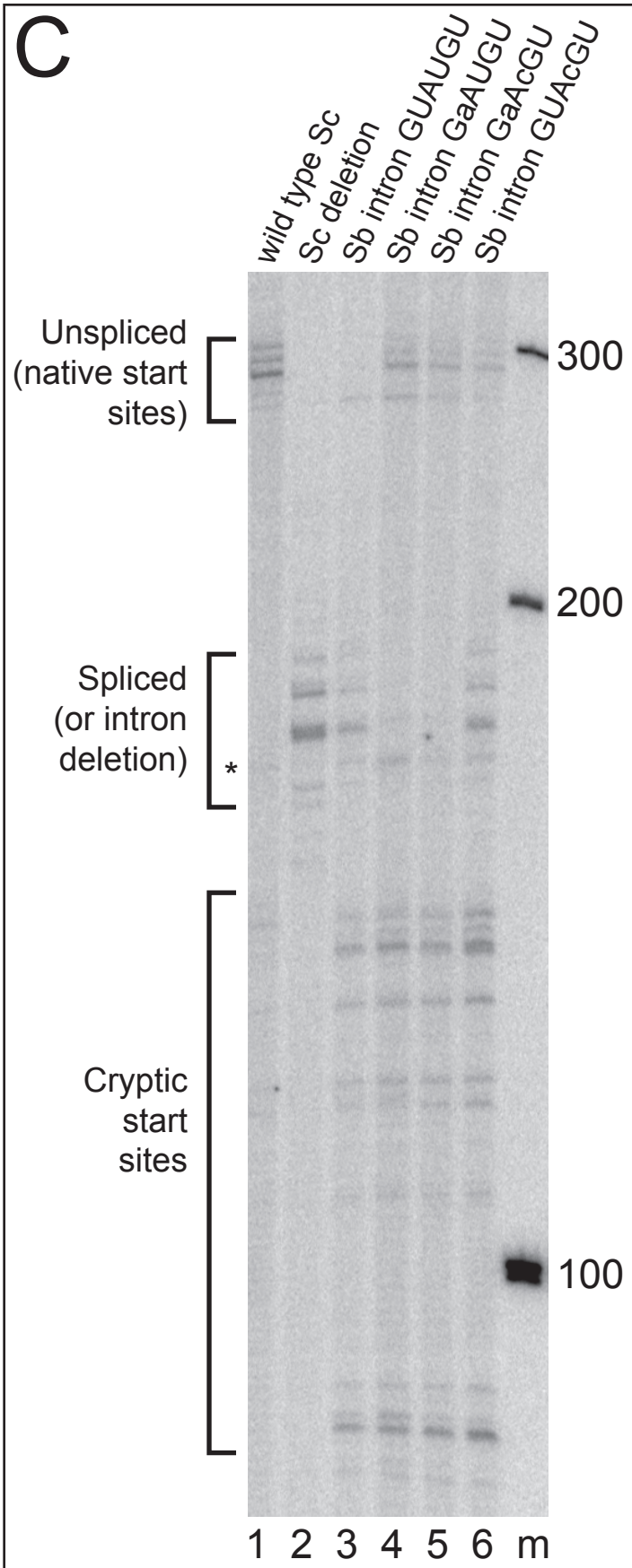


Talkish et al Fig 7

A

```
Sc-SbI  ACGTAACATCCTCCATATGTGTGAATTATTTTGTCCGGCGTCTT-GTTTGAATATCCCAT
sacCer  ACGTAACATCCTCCATATGTGTGAATTATTTTGTCCGGCGTCTT-GTTTGAATATCCCAT
sacBay  ACGTATGATTCTCTATAGGTGTAGATTATTTTGTCCGGGGTCTTTGTTTCAATATCTCAT
        *****  ** *** *** ****  ************** ***** **** ****** ***

                   "TATA box"
Sc-SbI  CCGGACGTATTTCTTTATAGGATAACGCATGCACATATCGGTTCGTTCAATAAGAAAGTC
sacCer  CCGGACGTATTTCTTTATAGGATAACGCATGCACATATCGGTTCGTTCAATAAGAAAGTC
sacBay  CCGGGCACATTTGTTTATAGGATAACGCAATGCACATATCGGCCCACCTCAAAGAAAGTA
        **** *  **** ****************          *  *   *   ********
                                     ^350

                        Sc major tsc start sites
                                 > >    >
Sc-SbI  CCTTCTGGTAT--TATTCGGTGTCCTGTAAGAT-CTATATACTAATCACCAAGCTATCTG
sacCer  CCTTCTGGTAT--TATTCGGTGTCCTGTAAGAT-CTATATACTAATCACCAAGCTATCTG
sacBay  CCCGTTCTGTCTATAGCGAGTGTCATACAAGATCACGTATACCTAATACTAGACCAACTA
        **   *       **    ***** *  *****    *****  *  ** *  * * **
                              ^300

Sc-SbI  GTGTCACATTGTGTATGTGTTTGCCTTTTGCTTCGTTTTAACCCTTATTTTTCTCGTGAA
sacCer  GTGTCACATTGTGAATACATTTGTCTTTTGGTCCATTTTTCCCCATCATTTTCTCGTGAA
sacBay  ACTCGAGTCTGTGTATGTGTTTGCCTTTTGCTTCGTTTTAACCCTTATTTTTCTCGTGAA
             *   **** **   **** ****** * * ****  *** *  ************
                   ^250

Sc-SbI  AACGCTTTTTAACTTTTGCCACTATAGTTCTCCAAGGAAGGATTTGAATTACTAACAAAT
sacCer  AACGATCTCTAACTTTTGCCACTATAGTTTTAAAGGAAAGATTTTGAGCTGCCAACGAAT
sacBay  AACGCTTTTTAACTTTTGCCACTATAGTTCTCCAAGGAAGGATTTGAATTACTAACAAAT
        **** * * ******************* *  * * ***  *****  * * *** ***
             ^200                                              ^150

                Sb intron activated cryptic start sites
                     > >      >          > >     >          > >
Sc-SbI  ATAATATAGCACACTCAAGAGTCTTGGCAAGCAACTTTAAAGTTAGAAAGAGTACTAATG
sacCer  ACAATATAGCACACTCAAGAGTCTTGGCAAGCAACTTTAAAGTTAGAAAGAGTACTAATG
sacBay  ATAATATAGTATATCTAAGAGTCGTGACAACAGGAGAAATAACAAGAAAGGATGTCAATG
        * ******* * *   ******* ** ***        * *   ******  *   ****
                                                   ^100

Sc-SbI  TTGCTACTTTCTTGGTCAAGAATTGCTACTAAGGTAGTT
sacCer  TTGCTACTTTCTTGGTCAAGAATTGCTACTAAGGTAGTT
sacBay  CTAATACATTCTTGGTCGAGAATTATGGTCAAGGTGGCC
         *  *** ********* ******      ***** *
                                     ^50 from primer
```

B



C