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 (MA)
(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 we call protointrons that do not overlap known introns. 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 most are not under selection. Some are found in non-coding RNAs (e. g. CUTs and SUTs), where they may contribute to the creation of new genes. Others are found across boundaries between noncoding and coding sequences, or within coding sequences, where they offer pathways to the creation of new protein variants, or new regulatory controls for existing genes. We define protointrons as (1) nonconserved intron-like sequences that are (2) infrequently spliced, and importantly (3) are not currently understood to contribute to gene expression or regulation in the way that standard introns function. A very few protointrons in *S. cerevisiae* challenge this classification by their increased splicing frequency and potential function, consistent with the proposed evolutionary process of "intronization", whereby new standard introns are created. This snapshot of intron evolution highlights the important role of the spliceosome in the expansion of transcribed genomic sequence space, providing a pathway for the rare events that may lead to the birth of new eukaryotic genes and the refinement of existing gene function.

Author Summary

The protein coding information in eukaryotic genes is broken by intervening sequences called introns that are removed from RNA during transcription by a large protein-RNA complex called the spliceosome. Where introns come from and how the spliceosome contributes to genome evolution are open questions. In this study, we find more than 150 new places in the yeast genome that are recognized by the spliceosome and spliced out as introns. Since they appear to have arisen very recently in evolution by sequence drift and do not appear to contribute to gene expression or its regulation, we call these protointrons. Protointrons are found in both protein-coding and non-coding RNAs and are not efficiently removed by the splicing machinery. Although most protointrons are not conserved, a few are spliced more efficiently, and are located where they might begin to play functional roles in gene expression, as predicted by the proposed process of intronization. The challenge now is to understand how spontaneously appearing splicing events like protointrons might contribute to the creation of new genes, new genetic controls, and new protein isoforms as genomes evolve.

57 ## Introduction

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

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

75 Until recently, gain or loss of introns has been detected by comparing closely related
76 genomes. Many such "presence/absence" variations are inferred to be intron loss, in which reverse
77 transcription of a spliced RNA, followed by homologous recombination of the intronless cDNA back
78 into the gene of origin, erases the intron [6, 7]. Several mechanisms for the gain of new introns have
79 been proposed (for review see [8, 9]). For example, single nucleotide changes that create new splice

sites (and thus new introns) can lead to "intron sliding" or new alternative splicing events [10]. "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 [11]. These intron gain mechanisms rely on pre-existing local splicing events, and represent intron diversification, rather than de novo intron creation at sites where no intron previously existed.

De novo intron creation appears to occur by two main pathways, "intron transposition" whereby an intron at one location is copied and inserted at a new location, and "intronization". First described in the marine alga *Micromonas* [12], transposition of introns called "Introner Elements" appears to have expanded an intron repeat family in some lineages [13-17] perhaps through an "armed spliceosome" 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, 9]). More recently, intron transposition through an RNA intermediate has been documented in *S. cerevisiae*, supporting the idea that reverse splicing may operate to spread introns [18]. Other models suggest that introns transposition may arise by the action of DNA damage repair [19] or non-autonomous DNA transposons [20].

A distinct pathway for *de novo* intron creation is called "intronization" whereby mutations arise either through drift [4, 8, 21], or other sequence changes [22] to create sequences recognized as introns by the splicing machinery. As a genome sequence distant from other introns drifts, it may accumulate mutations that by chance allow its transcripts to be recognized by the splicing machinery and spliced. This process is thought to occur gradually over evolutionary time, generating sequences that exhibit properties of both exons and introns that are often alternatively spliced through several, different, weak splicing signals [4, 23]. Whether these sequences evolve to become bona fide introns depends on whether their removal through splicing provides a fitness advantage.

103 Splicing in the *S. cerevisiae* genome appears to have been streamlined by evolution, such that
104 about 5% of genes have introns, and most that do have only one. Despite their scarcity in genes,
105 introns appear in about 25% of transcripts when cells are growing in rich medium [24, 25]. More
106 than a third of annotated introns are found in genes for ribosome biogenesis (ribosomal protein
107 genes, RPGs), and their mRNAs account for 90% of the splicing performed in rapidly growing cells
108 [25, 26]. This unusual distribution of introns in a highly expressed class of genes with shared
109 function presents both challenges and opportunities for studying integration of splicing into core
110 cellular regulation. For example, repressing transcription of the RPGs increases the efficiency of
111 splicing for the majority of non-RPG introns and can suppress temperature-sensitive spliceosomal
112 protein mutations [27, 28]. Based on this we proposed that relieving pre-mRNA competition by
113 reducing RPG expression frees the spliceosome to process less competitive, splicing substrates it
114 normally ignores. This phenomenon has been shown to contribute to the efficiency of meiotic
115 splicing [28, 29], as well as regulation of Coenzyme Q_6 synthesis [30], whereby regulated repression
116 of RPGs potentiates splicing control mechanisms at other genes.
117 In this study, we use rapamycin to repress RPGs and create hungry spliceosome conditions in
118 three related yeast species that have diverged over ~10-20 million years [31], and find 163 locations
119 in the *S. cerevisiae* genome that are substrates of the spliceosome, but distinct from the current set of
120 annotated introns. Other studies have also found undocumented splicing events in yeast [32-39],
121 including alternative splicing of known introns. Here we focus strictly on splicing events not
122 associated with a known intron, which we call protointrons. To better understand intron creation, we
123 distinguish, standard introns from protointrons as follows. Standard introns are efficiently spliced
124 under normal growth conditions (median efficiency ~93% spliced), highly conserved in related
125 *Saccharomyces* species, and have known functions in gene expression (93% reside in protein coding

126 regions). Protointrons on the other hand generally splice with very low efficiency (median efficiency
127 ~1% spliced), are often specific to different *Saccharomyces* species, and most importantly do not
128 have a clear role in correct gene expression. These protointrons are found in mRNAs as well as
129 noncoding RNAs such as promoter-associated transcripts, CUTs, SUTs, XUTs and other noncoding
130 RNAs [40-48]. Related yeasts *S. bayanus* and *S. mikatae* also have protointrons, but most are
131 species-specific, indicating that protointrons appear and disappear during evolution. This suggests
132 that protointrons arise initially through genetic drift and provide the raw material for intronization in
133 order to enable intron creation through a mechanism that is distinct from gene duplication or intron
134 transposition [4]. While only a very tiny fraction of protointrons might ever evolve into new
135 standard introns, we observe several, more efficiently spliced protointrons that appear to be more
136 advanced along the intronization pathway. This intermediate class of introns tend to occur in
137 5'UTRs where they might buffer the negative effects of RNA secondary structure or micro-ORFs
138 (uORFs) on translation, however there is currently no evidence that these have any adaptive value.
139 This work reveals the extent to which the spliceosome recognizes and splices intron-like sequences,
140 thus expanding the information contained in the genome. This contribution of the spliceosome to the
141 information content of the transcriptome may enhance the rate at which new genes and regulatory
142 mechanisms appear in eukaryotes.
143

144 ## Results

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

147 In experiments where we observed increases in splicing efficiency of standard introns after
148 repression of RPG expression [28], we also observed unannotated splicing events, distinct from

149 known introns, whose splicing efficiency improved. In addition, inventive new RNAseq methods
150 designed to capture branchpoints [32, 34] have provided evidence for splicing events elsewhere in
151 the transcriptome. Some of these events appear to be activated in response to stress [32], which
152 down-regulates RPG expression and creates hungry spliceosome conditions. Still others are more
153 readily detected during meiosis [32, 39], or when cells are deleted for RNA decay pathway
154 components that degrade unstable transcripts [32, 33, 38]. Our interest in understanding the
155 evolution of splicing prompted us to focus on these distinct new introns. To capture more of them,
156 we obtained additional RNAseq data that included non-polyadenylated RNAs and RNAs sensitive to
157 nonsense-mediated decay. We made four libraries, one each from rRNA-depleted RNA from
158 untreated (0 min) and rapamycin treated (blocks nutrient signaling and represses RPGs, 60 min)
159 replicate cultures of a yeast strain deficient in NMD (*upf1Δ*, [49]). We obtained more than 300
160 million reads that show excellent between-replicate coherence in gene expression changes (Fig.
161 S1A). These data confirm our previous observation [28] that splicing of a majority of standard
162 introns in non-RPGs increases after rapamycin treatment. A splicing index relating change in the
163 ratio of junction/intron reads over time (SJ index = log2[junctions-t60/intron-t60] – log2[junctions-
164 t0/intron-t0]) increases for most introns in transcripts whose total transcript levels change less than
165 two-fold during the experiment (Fig 1A, blue circles, Table S1, NB: RPGs are repressed >2 fold and
166 are excluded). In addition to the standard introns, we observe more than 600 splicing events that
167 result from use of alternative 3' or 5' splice sites overlapping the standard introns (see also the study
168 by Douglass et al. [38]). Some of these splicing events have been previously characterized [33, 36,
169 37], and produce out of frame mRNAs that are more easily detected in the *upf1Δ* strain due to the
170 loss of NMD (Table S2). Because our interest here is in introns appearing at novel locations, we
171 have not studied the splicing events that overlap the standard introns any further. At this sequencing

172 depth, a set of splicing events that do not overlap any standard intron is evident (Table 1 and Table
173 S3). These are supported by reads that span sequences with known characteristics of introns but
174 occur at much lower frequencies than those for standard introns. As is the case for standard introns,
175 splicing of the majority of these nonstandard introns increases upon rapamycin treatment (Fig 1A,
176 orange circles, positive values indicate increased splicing), suggesting they are also in competition
177 with RPG pre-mRNAs for the spliceosome.
178

179 **Table 1. Easily detectable protointrons in *Saccharomyces cerevisiae upf1Δ* mutant cells**

Gene	Overlapping Junctions	5′SS	Consv 5'SS	BP (predicted)	Consv BP	3'SS	Consv 3'SS	% Spliced 0 min.	% Spliced 60 min.
MTR2	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*	2	GTATGA GTTTGT	0.79 0.00	TACTAACA	0.77	AAG	0.16 0.14	22	35
MCR1	5	GTACGT GTACTC	0.74 0.01	TACTAAC AACTAACA TACTAACG	0.00 0.25 0.01	AAG TAG CAG	0.38 0.03 0.04	36	46
YEL023C	1	GTATGG	0.00	CACTAACA	0.27	TAG	0.41 0.18	23	17
ZTA1	2	GTATGA	0.95	TACTAATC	0.57	CAG AAG	0.98 0.99	1.6	5.3
PDX3	0	GTATCA	1.00	TACTGACG	0.50	AAG	1.00	1.6	1.9
TVP15	1	GTATGC	0.98	TACTAAGT	0.16	CAG TAG	0.62 0.62	2.2	4.7
PUS7	3	GTAAGG	0.99	TACTAACA	0.89	AAG TAG AAG AAG	0.97 0.62 0.62 0.96	6.5	4.4
YDR336W	1	GTACGT	0.06	CACTAAAA	0.12	TAG AAG	0.02 0.08	18	44
NTH1	1	GTATGG	0.99	TGCTAACA	1.00	CAG TAG	1.00 1.00	1.4	4.0
SPO1	0	GTATGT	0.52	TATTAACC	0.96	CAG	0.66	6.8	14
MCA1	0	GTACAG	0.02	GACTAATG	0.97	AAG	0.68	2.7	2.5
Opposing *IPT1*	1	GTAAGT	0.37	TACTAACT	0.25	AAG CAG	0.44 0.29	0.11	18
PDC1	1	GTATGT	1.00	CACTGACA	1.00	CAG	1.00 1.00	0.03	0.05
Downstream of *PDR12*	1	GTGTGT	0.01	GACTAACA	0.04	CAG AAG	0.00 0.02	36	31

SYN8	2	GTATGG	0.99	AACTAATA	0.23	TAG CAG CAG	0.96 0.95 0.96	2.1	3.4

180

181 **Identification and validation of protointrons**

182 To extract and validate nonstandard splicing locations from the RNAseq data, we inspected
183 reads that span and are missing genomic sequences bounded by GT or GC on the 5' side and AG on
184 the 3' side, aggressively filtering out those that were unlikely to have been generated by the
185 spliceosome (Table S4, see also Methods). For example, we ignored reads that are abundant in only
186 one library due to spurious PCR-derived "jackpot" amplification, or that are incorrectly mapped as
187 spliced over naturally repeated sequences. We also filtered out reads that appeared fewer than three
188 times, spanned sequences with no discernable match to a relaxed and appropriately positioned
189 branchpoint consensus sequence (RYURAY, >45 from a 5'ss, >7 from a 3'ss), or that use a GAG 3'
190 ss (although some of these may be true). Finally, in order to focus on new intron locations, we
191 separated out reads that overlap known standard introns (these are found in Table S2). Finally, we
192 merged overlapping alternative splice sites observed at each new intron location into one. From this
193 sequencing experiment, we identify 226 splicing events at 163 intron locations in the *Saccharomyces*
194 *cerevisiae* genome that do not overlap with standard introns. We call these protointrons (Table 1 and
195 Table S3, NB: we have reclassified some previously annotated introns as protointrons based on their
196 less efficient splicing and lack of conservation in close relatives, see below).

197 To characterize protointrons as a distinct class of splicing events, we inspected the
198 alignments of many individual protointrons (for example Fig 1B) and validated more than a dozen of
199 them by RT-PCR, cloning, and Sanger sequencing (Fig 1B and C, Table S5). These events map to a
200 diversity of transcribed locations, including mRNAs, a variety of non-coding RNAs (CUTs, SUTs,
201 XUTs, etc.), and within telomeric Y' repeats. 39% of protointrons can be found entirely within a

202 coding region (e. g. *TAF13* Fig 1B, C, Fig S1B and E), whereas others reside in noncoding regions
203 or within RNA antisense to a standard gene (29%, e. g. *antiASH1*, Fig 1B, C, Fig S1B and C). Often
204 alternative splice sites are observed, for example in the noncoding RNA *XUT12R-370* antisense to
205 *TUS1* (Fig 1B, note only the product derived from use of the downstream 3' ss is visible on the gel
206 in Fig 1C, Fig S1D). An internally initiating RNA from the meiotic *SPO1* gene expressed only
207 during vegetative growth has a protointron (Fig 1C, Fig S1F). Protointrons can also be found
208 crossing boundaries from the coding region to either UTR in mRNAs (17% in 5'UTR^coding; 1% in
209 coding^3'UTR), and still others can be found completely within a UTR (9% in 5'UTR; 4% in
210 3'UTR). In many cases, excision, cloning and sequencing the faint band near the size predicted by
211 the RNAseq reads identifies additional alternatively spliced forms not observed by RNAseq. For the
212 four events shown in Fig 1C, we successfully confirmed splice junctions indicated by RNAseq
213 (Table S3).

214 Several of the protointrons predicted by the RNAseq data appear to use unusual 5' splice
215 sites (5' ss) not anticipated by examination of standard introns. Standard intron 5' ss show strong
216 conservation of the G residue at position 5 (G5, underlined here: GUAY<u>G</u>U), which contributes to
217 intron recognition through interactions first with U1 snRNA and later with U6 snRNA [50, 51]. The
218 5' ss consensus in mammals has a less strongly conserved G5, and a subset of mammalian introns
219 use G at position 6 instead [52]. Validation tests of several protointrons whose 5' ss lack G5 show
220 that they are authentic products of splicing (Fig 1D, Fig S1G). During this effort we also detected
221 splicing at an AT-AC junction in the noncoding RNA *SUT635* (Fig 1E). Although yeast does not
222 have a minor (U12) spliceosome, some major (U2) spliceosomal introns use AT-AC junctions [53],
223 and mutation of a standard yeast intron shows that AT-AC is the most efficient splice junction
224 dinucleotide combination after GT-AG [54]. Alternative AT-AC junction use has been reported for

225 the standard intron (a normal GT-AG intron) in *RPL30* [32], however no standard yeast intron uses
226 AT-AC junctions normally. The appearance of an AT-AC protointron in *SUT635* suggests that AT-
227 AC introns may represent an alternative path to evolution of standard introns.
228
229 **Distinct features of protointrons: sequence, conservation, size, and splicing efficiency**
230 To begin contrasting the features of protointrons and standard introns, we evaluated
231 evolutionary conservation, the most obvious difference. To analyze conservation around splicing
232 signals of both intron classes, we extracted a sequence window surrounding the 5'ss, predicted
233 branchpoint, and 3'ss from each standard intron and protointron and compared them. Position-
234 specific weight matrix-based logos of the splicing signals (Fig 2A) reveal that protointrons use a
235 more divergent collection of splicing signals than do standard introns. As noted above, G5 of the
236 5'ss is a prominent feature of the standard yeast intron 5'ss but is less well represented among
237 protointron 5'ss. Similarly, the predicted branchpoint sequences of protointrons lack strong
238 representation of bases to either side of the core CUAA of the UACUAAC consensus for standard
239 introns. Branchpoints of standard introns are enriched for Us upstream of the consensus and for A
240 just downstream, and neither of these context elements are prominent in the predicted branchpoints
241 of protointrons. The 3'ss are more similar, but as with the branchpoint, the U-rich context detectable
242 upstream of the standard intron 3'ss is not observed in protointrons. The more diverse collection of
243 splicing signals used by protointrons is similar to that of the overlapping alternative splice sites of
244 standard introns [33], and a mixed set of overlapping and distinct potential introns detected using a
245 branchpoint sequencing approach [32]. Furthermore, we note that protointrons use a less constrained
246 set of 5' splice sites than standard introns as compared to the branchpoint and 3'ss sequences. This
247 may reflect important interactions between the pre-mRNA cap binding complex and the U1 snRNP

248 during earliest steps of spliceosome assembly [55-58] and suggests these robust interactions allow
249 greater drift of the 5'ss than other splicing signals. This is also consistent with our observation that
250 26% of protointrons overlap 5' UTRs as compared with 5% for 3' UTRs and agrees with a
251 prediction of the intronization model developed for metazoans [8, 59]. We conclude that
252 protointrons use a wider variety of branchpoints and splice sites than do standard introns in *S.*
253 *cerevisiae* and hypothesize that protointrons may evolve toward standard intron status by acquiring
254 mutations that enhance the context and match to the consensus of the core splicing signals, in part to
255 increase their ability to compete with RPG pre-mRNAs [28].

256 With the exception of the protointrons found entirely within protein coding sequences,
257 typically at least one (and often all three) of the splicing signals for a given protointron in *S.*
258 *cerevisiae* is imbedded in sequence that is not conserved in closely related *Saccharomyces* species
259 (Fig 1B, E, Fig S1C, F). To analyze whether this is a distinguishing characteristic of protointrons,
260 we recorded the average phastCons score (range between 0, evolving as not conserved, and 1,
261 evolving as highly conserved, [60]) of the nucleotides within sequence windows containing the 5'ss,
262 the branchpoint, and the 3'ss of the standard introns and protointrons and plotted them (blue bars,
263 standard introns; orange bars, protointrons, Fig 2B). Many standard introns are embedded in
264 conserved protein coding sequences, and thus the splice sites and their immediate exon context are
265 also conserved, such that the average phastCons scores for both the 5' and 3'ss windows rise above
266 0.7 (Fig 2B). The branchpoints of standard introns are also conserved but have a broader and lower
267 score distribution, because constraining protein coding exon sequences are not usually found near
268 the branchpoints of standard introns. In contrast, the distributions of phastCons scores for each of
269 the protointron splicing signals is clearly bimodal, meaning that protointron splicing signals are
270 either highly conserved, falling within protein coding sequence, or are poorly conserved, falling in

271 UTRs or intergenic regions (Fig 2B). This distribution illuminates the sequence landscape within
272 which most protointrons arise. Since the strongly transcribed regions of the genome code mostly for
273 protein, and transcription is a prerequisite for splicing, protointrons tend to appear in or span less
274 well conserved noncoding RNA sequences such as UTRs and ncRNAs (Supplemental Figure S1B).
275 This distinguishes protointrons from standard introns and suggests that most of the protointrons we
276 detect have appeared only recently in the *S. cerevisiae* genome, when transcribed non-protein coding
277 regions acquire intron-like features by mutation.

278 Standard introns show a bimodal length distribution, with peaks at about 100 nt and 400 nt
279 ([61], Fig 2C). In contrast, the distribution of protointrons is on average shorter, with a single main
280 peak at around 100 nt in length, and few larger than 300 nt. These distributions are significantly
281 different (Kolmogorov-Smirnoff test, $p \leq 10^{-4}$), suggesting that if protointrons evolve into standard
282 introns, they may become longer by acquiring additional sequence features that enhance their
283 recognition by the spliceosome. Many of the larger standard introns are found in RPGs [61], where
284 secondary structures and other long distance RNA-RNA interactions promote efficient and accurate
285 splicing [62, 63]. The elaboration of such structures during evolution of increased splicing efficiency
286 may explain the increased intron length that characterizes the large intron class in yeast.

287 Protointrons are less efficiently spliced than standard introns (Fig 2D, Tables 1, S1 and S3).
288 The vast majority of standard introns are spliced at greater than 80% efficiency (median ~ 93%) by
289 comparison of splice junction reads to intron base coverage. Exceptions include meiotic introns
290 whose efficient splicing may require repression of RPGs or the expression of a meiosis-specific
291 splicing factor like Mer1 [27-29]. In contrast, most protointrons have splicing efficiencies below
292 20% at best (median ~ 1%). A few protointrons, such as the introns in the *S. cerevisiae MTR2*,
293 *USV1*, and *MCR1* genes or the *S. bayanus YTA12* gene, are uncharacteristically well spliced,

294 suggesting that they may be transitional intron forms or species-specific standard introns (see
295 below). Splicing improves for most protointrons and standard introns after rapamycin treatment,
296 however some standard introns appear to show reduced splicing, suggesting splicing repression in
297 response to rapamycin at those introns.
298
299 **Coding regions are depleted of sequences required for splicing**
300 Protointrons that emerge within coding regions (39%, Supplemental Figure S1B) might
301 disrupt gene expression by reducing mRNA levels or creating toxic proteins. Since protointrons
302 emerge readily from nonconserved sequence (Fig 2B), we wondered whether the appearance of
303 protointrons within ORFs most often reduces fitness, and thus whether the frequency of splice site
304 and branchpoint sequences might be lower than would be expected by chance in protein coding
305 regions. In rare cases such introns might allow advantageous mRNA regulation through NMD [5,
306 64] to emerge (as appears to have been the case with a recently evolved standard intron in *PRP5*,
307 [65]) or make mRNA for beneficial alternative proteins (as may be the case for *PTC7*, [30, 66, 67];
308 and *MRM2* [37], see below), which have in-frame conserved coding sequences through their introns.
309 Analysis of several diverged genomes by Farlow et al. revealed that the consensus 5' ss sequence is
310 significantly underrepresented in the coding strand of genes compared to the noncoding strand [68].
311 To test whether *S. cerevisiae* coding sequences might be depleted of splicing signals, we compared
312 their frequency in the ORF set of the extant *S. cerevisiae* genome with that in 10,000 synthetic ORF
313 sets derived by randomizing the order of codons for each ORF. This approach maintains ORF
314 length, integrity, GC content, and codon usage in the permuted ORF sets while generating partially
315 randomized nucleotide sequences that can be used as a background sequence set for comparison, and
316 has been used to evaluate co-evolution of RNA processing signals within coding sequence [69].

317 To assess the representation of 5' ss and branch point sequences, we chose two 6-mers as
318 proxies – one for the 5' ss (GTATGT), and one for the branch point (ACTAAC). We counted the
319 number of times each appeared in the extant S. cerevisiae ORF set, and in each of the 10,000
320 permuted ORF sets, and plotted them. For both 6-mers, their counts in the extant ORF set are less
321 than 3 standard deviations below the mean of their respective counts from the 10,000 permuted ORF
322 sets (vertical lines), indicating that these 6-mers are significantly underrepresented in the natural *S.*
323 *cerevisiae* coding sequences as compared to the randomized coding sequences (Fig 3A).
324 To determine whether these proxy 6-mers were unusually depleted as compared to other 6-
325 mers, we calculated a Z-score for each of the 4096 6-mer sequences in turn, comparing the counts of
326 each in the extant *S. cerevisiae* ORF set with its mean counts in the 10,000 permuted ORF sets, and
327 plotted the distribution (Fig 3B, grey bars). The intron branchpoint 6-mer "ACTAAC" is found
328 comparatively much less frequently in the extant genome than are other 6-mers (Fig 3B), and much
329 less frequently than the average stop codon-containing 6-mer (blue bars). The 5' ss 6-mer
330 "GTATGT" is also more depleted than the average 6-mer, especially when compared to the subset of
331 ATG containing 6-mers (maroon bars, Fig 3B). Because there is a large amount of information in
332 coding sequences, we cannot be certain the depletion of the splicing signal 6-mers is due to splicing.
333 Numerous other features are being randomized by the process of codon permutation, for example 6-
334 mer representation may be influenced by di-codon frequencies that affect translation [70]. Even so,
335 these observations are consistent with the idea that splicing signals within ORFs carry a risk of
336 reduced fitness. This suggests that a robust level of spliceosome activity may be sufficient to lead to
337 loss of correct mRNA should genetic drift create splice sites within ORFs, providing a rationale for
338 tight regulation of a splicing activity limited to pre-mRNAs that can compete [28].
339

340 **Protointrons are idiosyncratic to closely related species**

341 *S. cerevisiae* protointrons are not conserved in closely related yeasts, suggesting that they
342 have appeared or disappeared in *S. cerevisiae* in the time since these lineages diverged. To explore
343 the hypothesis that protointrons arise and disappear differently in the other *Saccharomyces* lineages,
344 we treated cultures of *S. mikatae* and *S. bayanus* with rapamycin for 0 or 60 minutes, isolated RNA,
345 depleted rRNA, and made cDNA libraries for sequencing. These strains are NMD competent
346 (*UPF1*), so our ability to detect transcripts subject to NMD is limited. Regardless, nearly all
347 annotated introns in *S. cerevisiae* are present in the *S. mikatae* and *S. bayanus* genomes, and *S.*
348 *bayanus* has an additional standard intron in a *CHA4*-like gene that has no ortholog in *S. cerevisiae*
349 (Tables S6 and S7). Some annotated *S. cerevisiae* introns are missing in these close relatives and
350 based in part on their locations and splicing efficiencies, we propose reclassifying them as
351 protointrons (see below). In contrast to the high conservation of standard intron locations, we detect
352 distinct sets of protointrons in each species (Tables S6 and S7). We validated a subset of these (Fig
353 4) and find that most all of the *S. mikatae* protointrons are not present in either *S. cerevisiae* or *S.*
354 *bayanus*, and that the *S. bayanus* protointrons are not found in *S. cerevisiae* or *S. mikatae* (Tables S3,
355 S6, and S7). An exception to this is *YIL048W/NEO1*, in which the same protointron is observed in
356 both *S. cerevisiae* (Table S3) and *S. mikatae* (Fig 4, Table S6). High sequence conservation in the
357 coding region of *YIL048W/NEO1*, most likely due to functional constraints on the protein coding
358 function of the sequence, has fortuitously preserved the splicing signals in all of the *Saccharomyces*
359 yeasts. The intron is in frame with the coding sequence (Table S8), thus although the splicing of this
360 protointron is not efficient, it remains possible this intron could generate a functional alternative
361 protein. We conclude that protointrons are idiosyncratic in closely related yeast species. This is
362 evidence for rapid evolutionary appearance and disappearance of sequences that can be functionally

363 recognized by the spliceosome. The dynamics of creation of protointrons thus appears consistent
364 with genetic drift, primarily in the rapidly evolving nonconserved sequences of recently diverged
365 genomes.
366
367 **Introns with features of both protointrons and standard introns may be intermediates in de**
368 **novo intron creation**
369 Based on the above analysis and those described elsewhere that note "novel" splicing [32-34,
370 37-39, 67], we propose defining standard introns as (1) conserved in related organisms or clades, (2)
371 efficiently spliced under appropriate physiological conditions, and (3) established in the pathway for
372 production or regulation of a functional gene product. Furthermore, we propose defining
373 protointrons as (1) not conserved in closely related species, (2) inefficiently spliced, and (3) not
374 clearly understood to contribute to correct expression or regulation of a gene. This simple definition
375 allows classification of most all of the observed splicing events in the yeast genome as either
376 protointron or standard intron, with only a few exceptions. The vast majority of protointrons arise
377 by neutral drift and likely provide no fitness advantage, and most probably disappear. A few introns
378 do not neatly fall into one or the other category and may be transitioning from protointron to
379 standard intron status, as predicted by the intronization model. Like protointrons these intermediate
380 class introns not conserved, but unlike typical protointrons they have increased splicing efficiency,
381 and may appear positioned to influence expression of the gene that carries them.
382 Examples of protointrons that may be on an evolutionary path toward standard intron status
383 include introns in the 5' UTRs of *S. cerevisiae MTR2, USV1, YEL023C*, and *MCR1*, and in the 5'
384 UTR of *S. bayanus YTA12* (Fig 5). Using the unrooted tree describing relationships between the
385 genomes of the sensu stricto yeasts [71], we map the appearance of these high efficiency

386 protointrons as predicted by their presence in extant genomes that have diverged over 10-20 million
387 years. The *MTR2* intron contains essential sequences [72] and can diversify the N-terminal sequence
388 of the mRNA export protein Mtr2 [73]. When sequences of related yeasts became available [31], it
389 became clear that the *MTR2* intron is unique to *S. cerevisiae* (Fig 5). High efficiency protointrons are
390 found in the 5'UTRs of *USV1* and *MCR1*. The *USV1* intron is efficiently spliced after rapamycin
391 treatment in *S. cerevisiae* and is also functional in *S. mikatae* (Fig 5, Table S6). However, *S.*
392 *kudiravzevii* and *S. bayanus* have different sets of nucleotide changes that eliminate splice sites and
393 branchpoints required for this intron. The *MCR1* intron is spliced at about 30%, and appears to be
394 shared by *S. paradoxus*, but is absent in *S. mikatae*, *S. bayanus*, and *S. kudriavzevii*. None of these
395 splicing events alter the N-terminus of Usv1 (a stress induced transcription factor) or Mcr1 (a
396 mitochondrial NADH-cytochrome b5 reductase), however both introns remove uORFs from the 5'
397 UTRs of these genes, suggesting that splicing could affect 5'UTR function in mRNA translation or
398 stability for both genes. Finally, a very efficiently spliced (>95%) intron is found in the 5'UTR of
399 the *S. bayanus YTA12* gene, as well as in *S. kudriavzevii* (Fig 5). Removal of this intron does not
400 alter the N-terminus of Yta12, a mitochondrial protein complex assembly factor [74], but does lead
401 to removal of a uORF. Interestingly the *S. cerevisiae YTA12* gene matches at 88 of 117 (75%)
402 positions in the *S. bayanus* intron and has neither an intron nor any uORF (see below).
403 Ten percent of the protointrons identified in *S. cerevisiae* are located exclusively in 5'UTRs
404 (Supplemental Figure S1B). The apparent relationship between introns and uORFs leads to the idea
405 that 5'UTR introns may be adaptive by protecting mRNAs with long 5'UTRs from the general
406 negative effect of uORFs [48, 75]. To test this idea, we asked whether uORFs are present more
407 frequently in 5'UTRs that contain standard introns, as compared to similarly sized 5'UTRs that do
408 not. There are 22 yeast genes (7%) with standard introns in their 5'UTRs, (this number does not

409 count the annotated introns in *MCR1*, *MTR2*, or *USV1*, which are not conserved across the *sensu*
410 *stricto* group). The size range of the (unspliced) 5'UTRs for these 22 is from ~240 to 950
411 nucleotides, and there are 91 intronless genes with 5'UTRs in this size range. We counted uORFs
412 longer than 4 codons (including the AUG, but not the stop codon) within 5'UTRs in this size range.
413 Among the 91 genes without 5' UTR introns, 53 lack any uORFs, whereas 38 have at least one
414 uORF. All 22 genes with 5'UTR introns have at least one uORF, and for 20 of these all the uORFs
415 in the 5' UTR are removed by splicing. This distribution of uORFs in 5'UTRs with introns is
416 unlikely to have been generated by chance (Fisher's exact test, $p < 10^{-5}$). One hypothesis to explain
417 this is that by removing much of the 5' UTR RNA, an intron may protect a gene that has a long
418 5'UTR from genetic drift that creates uORFs, or other translational inhibitory features like RNA
419 secondary structure [48]. Additional experiments will be needed to determine which if any of these
420 splicing events promotes gene expression and whether or not the effect contributes to fitness.
421 A second evolutionary scenario whereby protointrons may be adaptive concerns in frame
422 splicing, which would produce an alternative polypeptide. This appears to be the case for the
423 standard intron in *PTC7* [30, 66]. A similar intron is found in *MRM2*, where as many as three
424 different proteins may be produced (Tables S1 and S2, see also [37], NB: not annotated in SGD, but
425 this fits the standard intron definition). Despite forces that seem to deplete splicing signals within
426 ORFs (Fig 3), we find 20 (out of 63) 3n protointrons within ORFs that do not interrupt the reading
427 frame and thus could produce functional proteins, particularly under stress or other conditions where
428 RPG transcription is reduced (Table S8). We suggest that such protointrons provide evolutionary
429 opportunities to create new protein isoforms from existing genes.
430
431 **Telomeric Y' repeats**

432 Intron-like sequences have been noted in the telomeric Y' family of repeat sequences for
433 more than 25 years and continue to be annotated in the Saccharomyces Genome Database (SGD). So
434 far, molecular tests for splicing of these annotated introns have been negative [61, 76]. Intron
435 predictions allow some Y' repeat element copies to encode a large (1838 amino acid) protein (e. g.
436 YNL339C, Fig 6A), that carries an N-terminal Sir1 domain and a central DExD helicase domain.
437 Other Y' elements differ in sequence and can only encode fragments of the open reading frame that
438 may nonetheless produce smaller functional proteins, for example the helicase overexpressed in
439 telomerase-deficient "escaper" colonies [76, 77]. The function(s) of any of the Y' element predicted
440 proteins in normal cells are not known.
441 To determine if Y' element transcripts are spliced, we allowed RNAseq reads to map to the
442 repetitive Y' elements (i. e. without masking). Although the mapped locations may not be the precise
443 origin of the RNA that created the read, this allows us to identify spliced reads and assign them to
444 possible members of the Y' repeat family. We find two introns within the Y' repeat family (Fig 6),
445 one of which lies on the far left of the repeat and is required to create the open reading frame for the
446 longest predicted protein (exemplified by YNL339C near TEL14L, Fig 6A). The other is in the
447 center of the Y' repeat (exemplified by YLR464W near TEL12R, Fig 6B). Neither of these introns
448 matches the annotations at SGD, and instead in both cases, downstream 3' ss are used (see also
449 [37]). It is not uncommon for the yeast spliceosome to skip proximal 3' ss in favor of a distal 3' ss,
450 in some cases due to secondary structure of the pre-mRNA [78]. The Y' elements in *S. cerevisiae*
451 differ from each other; not all can express a protein as large as YNL399C after removal of intron 1
452 using the distal 3' ss. Intron 2 splicing does not greatly extend the open reading frame of YLR464W.
453 To confirm that the reads arise from splicing rather than from a deleted copy of the Y' element
454 precisely lacking the intron, we searched the genome using the "spliced" sequence produced for

455 intron 1 or intron 2 and found that there is no such contiguous genomic sequence. We conclude that
456 Y' element transcripts can carry at least two introns that are distinct from current annotations in SGD
457 (Fig 6).
458 To evaluate the sequence relationships of the Y' repeat element introns we aligned them with
459 each other, after merging identical copies into one. All the predicted intron 1 sequences have the
460 second most common 5' ss in the yeast genome (GUACGU, followed by the preferred A at position
461 7, [61]. The most distal 3' ss of several possible creates the large ORF, and is UAG for all except
462 YPR202W, which has a CAG. Several other potential 3' ss (including the one annotated in SGD)
463 are skipped or used alternatively. Interestingly only YPR202W, YRF1-3, YRF1-6, and YRF1-7 have
464 the canonical UACUAAC branch point sequence, whereas most of the others have UAUUAAC, a
465 variant found in some standard introns (Fig 6A). The remaining group (YEL075C, YRF1-2, and
466 YRF1-4) are deleted for the region containing the branch point, suggesting that they are unable to be
467 spliced. Intron 2 has the most common 5'ss GUAUGU, and the most common branch point
468 UACUAAC, and uses the first AAG (a less commonly used but standard 3' ss) downstream from the
469 branch site (Fig 6B). Most copies also contain an alternative 5' ss which is used less frequently. We
470 have not estimated the efficiency of splicing of these introns because we cannot reliably assign reads
471 to specific repeat elements with confidence. The current genome assemblies of *S. mikatae*, *S.*
472 *paradoxus*, and *S. kudriavzevii*, but not the *S. bayanus* assembly include at least one Y' element
473 related to the *S. cerevisiae* elements [71], but the precise numbers and arrangements of the Y'
474 elements in those genomes await refinement of the genome assemblies for those organisms.
475
476 **The *S. bayanus YTA12* protointron functions in *S. cerevisiae***

477 The finding of a highly efficient protointron in the *YTA12* 5' UTR of *S. bayanus* (and
478 putatively in *S. kudriavzevii,* Fig 5) that is not observed in the alignable syntenic sequence of *S.*
479 *cerevisiae* prompted us to test (1) whether this *S. bayanus*-specific intron can be spliced in *S.*
480 *cerevisiae*, and (2) whether the intron might confer some advantage for growth on glycerol, given
481 the function of Yta12 in assembly of mitochondrial protein complexes [74]. Fig 7A shows an
482 alignment of the region including and upstream of the Yta12 start codon from *S. bayanus* (sacBay),
483 *S. cerevisiae* (sacCer) and *S. cerevisiae* in which the 117 bp of the *S. cerevisiae* genome
484 corresponding to the *S. bayanus* intron have been replaced in *S. cerevisiae* with the *S. bayanus* intron
485 (Sc-SbI). This replacement was made using CRISPR/Cas9 guided cleavage of an *S. cerevisiae*-
486 specific target sequence within the syntenic region and a repair fragment containing the *S. bayanus*
487 intron (Fig 7B). As controls, we created *S. cerevisiae* strains precisely deleted for the syntenic
488 region aligning with the *S. bayanus* intron, as well as versions of the *S. bayanus* intron with 5' ss
489 mutations (Fig 7C). We isolated RNA and evaluated the expression of these modified *YTA12* genes
490 by extension of a labeled primer complementary to *YTA12* mRNA with reverse transcriptase (Fig
491 7C). The major transcription start sites for *YTA12* in *S. cerevisiae* map about 300 nt from the 5' end
492 of the primer (Fig 7A and C, lane 1). These start sites are unaffected by the 117 bp deletion (the
493 same collection of cDNAs are shorter by 117 residues, lane 2). The migration of cDNAs from the
494 deletion strain are useful to mark the expected position of spliced RNAs, and indeed replacement of
495 the 117 bp with the *S. bayanus* intron sequence (lane 3) results in the appearance of the same
496 collection of cDNAs with the disappearance of the signal from pre-mRNA, indicating efficient
497 splicing (lane 3, compare with lane 1).
498 Mutation of the 5' ss from GUAUGU to GaAUGU or GaAcGU results in the reduction of
499 the spliced mRNA cDNAs, and the appearance of cDNAs corresponding to pre-mRNA (lanes 4 and

500 5), indicating that splicing is inhibited by these mutations. Changing the 5' ss from GUAUGU to the
501 less commonly used GUAcGU reduces the efficiency of splicing but does not block it, as judged by
502 the slight accumulation of unspliced RNA (lane 6). Unexpectedly, the *S. bayanus* intron sequence
503 activates a set of cryptic start sites in the *S. cerevisiae* sequences downstream of the major start site
504 and the *S. bayanus* intron (Fig 7C, lanes 2-6). These start sites are inefficiently used in the wild type
505 *S. cerevisiae YTA12* promoter (lane 1). One consequence is that new mRNAs are made that initiate
506 downstream of the intron and thus do not require splicing for expression of Yta12. This
507 interpretation is supported by the observation that all the strains grow on YP glycerol plates as well
508 as wild type BY4741 (not shown). This result highlights the challenge of anticipating the effect of
509 mutations in 5' UTRs where transcription, splicing, and translation operate together on the same
510 sequence. This experiment measures changes in splicing due only to differences in the intron, and
511 not due to any differences in exonic sequences or trans-acting factors between *S. bayanus* and *S.*
512 *cerevisiae*. We conclude that the efficiently spliced protointron from *S. bayanus* is equally at home
513 in *S. cerevisiae*. This intron appears to have formed in *S. bayanus* after *S. bayanus* and *S. cerevisiae*
514 last shared a common ancestor, but before the divergence of *S. bayanus* from *S. kudriavzevii*.
515

516 ## Discussion

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

518 Here we characterize a class of splicing events in yeast we call protointrons. Many previous
519 studies have noted "novel" introns in yeast under a variety of experimental conditions and genetic
520 backgrounds [32-39, 67]. Here we distinguish protointrons by several criteria, most importantly that
521 they reside at locations not overlapping known standard introns. We first recognized protointrons
522 while studying how the abrupt disappearance of RPG pre-mRNA during early nutrient deprivation

523 signaling frees the spliceosome to increase splicing of other pre-mRNAs [28, 29]. RNAseq analysis
524 of NMD-deficient yeast cells treated with rapamycin revealed that protointrons are found throughout
525 the transcriptome in both coding and non-coding regions of pre-mRNAs, ncRNAs, and antisense
526 transcripts, such as CUTs, SUTs, and XUTs (Fig 1, Fig S1). Protointrons contain all of the splicing
527 signals necessary for recognition by the spliceosome (5'SS, BP, and 3'SS), however the sequences
528 of these signals are more variable than those of standard introns (Fig 2A and B). Whereas standard
529 introns are conserved in related organisms, efficiently spliced, and established for production or
530 regulation of a functional gene product, protointrons are present in one or a few closely related
531 species, not efficiently spliced, and do not clearly contribute to correct expression or regulation of a
532 gene. Given this redefinition, we propose a revised intron annotation, including the addition of a
533 standard intron in *MRM2*, and molecular evidence for the correct location of expected but not
534 demonstrated splicing of Y' repeat element transcripts. We provide this and related data on a
535 publicly accessible genome browser with several *Saccharomyces* species genomes at
536 http://intron.ucsc.edu/.

537 Splicing events that occur outside our expectation of what is needed to make a protein or a
538 structural RNA have attracted labels like "splicing noise" or "splicing error" [79]. But viewing the
539 spliceosome as an enzyme able to catalyze a complex series of pre-mRNA binding, refolding, and
540 release operations, including two cleavage-ligation reactions, or even just the first one [35, 80], on a
541 very diverse set of substrates (for review see [2]) suggests that such terms should be more carefully
542 defined. The protointrons described here, as well as for example similar newly evolved splicing
543 events observed in mammalian lncRNAs [81, 82] reveal the outer edges of the substrate repertoire of
544 this enzyme in sequence space, and do not represent either splicing noise or splicing errors. We
545 suggest the term "splicing noise" should refer to fluctuations due to stochastic events affecting

546 particular splicing events, just as the term "transcriptional noise" refers to the stochasticity of
547 transcription events (see [83] and references therein). We also suggest the term "splicing error"
548 should refer to events within the spliceosome that lead to spliceosome assembly or catalysis that is
549 incompatible with successful completion of the two splicing reactions, spliced product release, and
550 recycling. In order for splicing to contribute to rapid evolution of multicellular organisms it seems
551 likely that a variety of sequences besides highly evolved introns would need to be recognized and
552 spliced, including those that appear in genomes by genetic drift. The extent to which these
553 spliceosome-generated spliced RNA sequences contribute to fitness would eventually determine
554 their evolutionary fate. The protointron class of splicing substrates represents opportunity to create
555 new genes, create new proteins from existing genes, or impose new regulatory controls on existing
556 genes.
557
558 **Some protointrons show greater splicing efficiency and may be adaptive**
559 The forces and mechanisms that drive intron evolution in eukaryotic genomes are still largely
560 unknown. If protointrons represent raw material for intron creation by the process of intronization,
561 then perhaps the most efficiently spliced protointrons represent intermediates in standard intron
562 formation that are advancing by selection of improving mutations. Our data provide evidence for
563 rapid and complete intronization in the *YTA12* 5' UTR between now and the time *S. bayanus* and *S.*
564 *cerevisiae* last shared a common ancestor (~ 20 Mya, [71]). Over the 117 bp intron sequence, the *S.*
565 *cerevisiae* 5' UTR differs at 29 positions (Fig 7A). Replacement of this region with the *S. bayanus*
566 sequence produces an efficiently spliced intron in *S. cerevisiae* (Fig 7C). This intron transplantation
567 experiment shows that no species-specific barrier prevents this sequence from serving as an efficient
568 intron in *S. cerevisiae*. Although this intron appears fixed in the *S. bayanus* and *S. kudravzevii*

569 branch of the *Saccharomyces* tree, there is currently no evidence for fitness effects, and thus this
570 intron could be a product of neutral evolution.

571 In some cases, a protointron might provide increased fitness that would explain its
572 evolutionary persistence. We suggest three specific ways that protointrons may support
573 improvements in gene function. Approximately 10% of protointrons reside entirely within 5'UTRs
574 (Supplemental Figure S1B), including the four most efficiently spliced protointrons we observed (*S.*
575 *cerevisiae MTR2*, *USV1*, and *MCR1*, *S. bayanus YTA12*). We realized that genes with long distances
576 between their transcription start sites and their start codons (i. e. with large 5' UTRs) are at risk for
577 mutations that create a uORF in the 5' UTR, which often negatively influences translation [48, 75].
578 Removal of a large region of the UTR by splicing would buffer this genetic risk. Secondary
579 structures or other detrimental sequences that might arise in long 5'UTRs [84] might also be safely
580 removed by splicing. To test the plausibility of this idea, we examined the frequency of uORFs in 5'
581 UTRs of *S. cerevisiae* genes that have or do not have 5' UTR introns and found that uORFs are
582 significantly more prevalent in 5' UTRs that have introns as compared to other 5' UTRs (see
583 Results). This suggests that the presence of a 5' UTR intron may help buffer an mRNA against any
584 detrimental effects of uORFs or RNA secondary structure, and provides evidence that intronization
585 in particular in 5' UTRs may be adaptive in *Saccharomyces* species.

586 A second way that protointrons may become functional is by producing in frame splicing
587 events within ORFs to create mRNAs encoding shorter protein isoforms with new functions. The
588 frequency of splicing signals is lower than expected in *S. cerevisiae* ORFs (Fig 3), supporting the
589 expectation that most introns that arise within ORFs would be detrimental to fitness. Despite this, we
590 found 20 protointrons contained within ORFs that do not interrupt the reading frame, and that may
591 lead to the translation of alternative protein products (Table S8). If such shorter proteins contribute

592 to fitness, mutations that increase the splicing of the protointron (without disrupting the function of
593 the full-length protein) may lead to the establishment and conservation of a standard intron that
594 allows production of both protein forms. This may be the mechanism by which the conserved in-
595 frame introns of *PTC7* [30, 66] and *MRM2* ([37], this work) have evolved. In these cases, both the
596 intron and the protein sequence encoded by the intron are conserved in *sensu stricto* yeasts,
597 suggesting both contribute to fitness across the genus *Saccharomyces*. Many protointrons span the
598 boundaries between conserved and nonconserved sequences (Fig 2B), increasing the chances that a
599 new splicing event will alter one or the other end of an existing protein. Studies of protein evolution
600 indicate that proteins evolve at their edges [85], suggesting that protointrons may contribute to this
601 as well. Although there is as yet no evidence for new function, the 5' UTR protointron in the *S.*
602 *cerevisiae MTR2* gene has arisen sufficiently close to the start codon that different alternative
603 splicing events add different peptides to the amino terminus of the annotated protein sequence [72,
604 73]. Thus, protointrons that appear in frame within existing genes, or that span the edges of existing
605 genes, create protein expression variation that may provide fitness advantages, in particular under
606 stress conditions that have yet to be explored.

607 A third way that protointrons may prove advantageous is through controlled downregulation
608 through splicing and NMD. We find that 16% of protointrons in *S. cerevisiae* span the 5' UTR and
609 coding region of twenty-seven genes and upon being spliced, remove the canonical AUG start codon
610 making these transcripts potential targets of NMD. In ten of these protointrons, the AUG start codon
611 is embedded within the GU**AUG** of the 5' ss (e. g. Ade2), suggesting sequences surrounding start
612 codons are particularly susceptible to drifting toward a 5' ss. A recently studied example of this is
613 the standard intron in the *PRP5* gene that is conserved in the *Saccharomyces* genus and destroys the
614 *PRP5* mRNA by removing the start codon and creating a transcript that is subject to NMD [65]. The

615 intron must have appeared since the divergence of the *Saccharomyces* species from their common
616 ancestor with *Lachancia kluveri*, since this more distant relative has a different intron in its *PRP5*
617 gene. This situation may evolve where overexpression of a particular protein may be detrimental.
618 *PRP5* encodes a splicing factor, and increase (or decrease) in Prp5 protein activity may increase
619 splicing and reduce (or decrease splicing and accumulate) *PRP5* mRNA levels by using this
620 conserved out of frame intron to create a homeostatic regulatory loop. A more difficult to recognize
621 but no less important mechanism is illustrated by the *BDF2* gene in which abortive splicing
622 downregulates expression through spliceosome-mediated decay [35]. It is unclear whether to
623 annotate this location and others like it [80] as an intron, since it does not appear that 3' splice site
624 selection is required or important for its activity. Thus, even protointrons that are out of frame within
625 coding regions, or pseudo-intron locations at which abortive splicing takes place may provide
626 opportunities for adaptive regulatory controls to evolve.
627
628 **Conclusions**
629 Protointrons are a rare class of splicing events that represent the action of the spliceosome on
630 RNA without a necessary connection to the expression of a mature gene. In mammalian cells the
631 spliceosome is no less constrained, and a very large number of alternative splicing events that appear
632 unrelated to "correct" gene expression support this [86]. In particular, newly evolved lncRNAs have
633 introns that are inefficiently spliced and have multiple alternative splice sites, unlike older, more
634 conserved lncRNA and mRNA encoding genes [81, 82]. These observations indicate that a general
635 feature of the evolution of introns is that any transcribed sequence has a chance of being spliced by
636 the spliceosome, should that sequence evolve recognizable splicing signals. Additionally, any
637 sequence that suddenly becomes transcribed can be expected to contain sequences by chance that are

638 immediately recognized as introns. Since the sequences required for splicing are ubiquitous and have
639 low information, many such newly appearing sequences will immediately produce diverse RNA
640 transcripts. If these confer some advantage, or if mutations that improve splicing become fixed by
641 neutral genetic drift, then a standard intron may evolve. This general pathway may be a source of
642 new introns whose splicing contributes to diversification of the transcriptome, and to the appearance
643 of new genes and new products from existing genes, as genomes evolve.
644

645 ## Materials and Methods

646 **Strains and culture conditions**

647 Two independent cultures of *S. cerevisiae* strain BY4741 *upf1Δ* (*MATa his3Δ1 leu2Δ0 met15Δ0*
648 *ura3Δ0 upf1Δ::KANMX*) were grown in YEPD medium at 30°C to an optical density at 600 nm
649 (OD_{600}) ≈ 0.5. The cultures were split and rapamycin was added to one half at a final concentration
650 of 200 ng/ml for 1 hour. *S. bayanus* strain JRY9195 (*MATa hoD::loxP his3 lys2 ura3*) and *S.*
651 *mikatae* strain JRY9184 (*MATa hoD::NatMX trp1D::HygMX ura3D::HygMX*) were grown in
652 YEPD medium at 26 °C, and were treated with rapamycin as for *S. cerevisiae* except at 26°C. These
653 strains were a kind gift of Chris Hettinger [71].
654

655 **RNA isolation**

656 RNA was extracted from yeast cells using Procedure 1 as described [87]. Prior to RNAseq library
657 construction (see below), RNA was DNased using Turbo DNase (Life Technologies) and RNA
658 quality was evaluated using the 2100 Bioanalyzer (Agilent).
659

660 **RNAseq library preparation**

661 5-10 ug of total *S. cerevisiae* RNA was depleted of ribosomal RNA using the RiboZero Gold
662 rRNA Removal Kit (Illumina) according to the manufacturer's instructions. Strand-specific cDNA
663 libraries were prepared using the Kapa Stranded RNA-Seq Library Preparation Kit for Illumina
664 Platforms (Kapa Biosciences) following the manufacturer's instructions with the following
665 modifications. Sequencing adapters and oligonucleotides used for PCR barcoding were from the
666 NEBNext Multiplex Oligos for Illumina Kit (New England Biolabs, NEB). Prior to PCR
667 amplification of the library, adapter-ligated cDNA was treated with USER enzyme (NEB). Adapter-
668 ligated libraries were then PCR amplified for 10 cycles using NEB index primers compatible with
669 Illumina sequencing. After amplification, size selection of the libraries was performed using an E-
670 gel Safe Imager and 2% E-gel size select gels (Invitrogen). Indexed libraries were pooled and 100
671 bp paired-end sequenced on the same flow cell of an Illumina HiSeq4000 instrument at the Berkeley
672 sequencing facility. RNA extracted from *S. bayanus and S. mikatae* was depleted of ribosomal RNA
673 as described above. Strand-specific cDNA sequencing libraries were prepared as described [88] and
674 50 bp paired-end sequenced on the HiSeq2000 platform (Illumina).
675
676 **Mapping and Analysis of RNAseq Data**
677 RNAseq data is deposited in GEO under the accession number GSE102615. For *S. cerevisiae*
678 libraries, all mappings were done using 100x100 bp reads to the SacCer3 Apr. 2011 genome
679 assembly (Saccharomyces Genome Database, SGD, [89]). For *S. bayanus* and *S. mikatae* libraries,
680 mappings were done using 51x51 bp reads to the SacBay2 and SacMik2 genome assemblies
681 (Saccharomyces Sensu Stricto Database, [71]), respectively. Reads mapping by BowTie2 [90] to *S.*
682 *cerevisiae* tRNA and rRNA defined by Ensemble or to Ty elements defined by SGD were discarded,
683 however mappings to Y' elements were recovered. For each library, reads were remapped to their

684 respective genomes using STAR with two-pass mode [91]. PCR duplicate reads (reads with identical
685 positions at both ends) were discarded and reduced down to one read. Changes in gene expression
686 upon treatment of cells with rapamycin were determined using DESeq2 [92], comparing untreated
687 and treated cells. Splice junctions were identified by STAR mapping [91].

688 Splicing Indexes (ratios of splicing measurements) were calculated by comparing reads that
689 cross the intron, reads that cross the splice junction, and reads in exon 2 in different ways. Splice
690 junction coverage is taken as the number of reads that cross the splice junction. Intron coverage was
691 taken as the average per nucleotide coverage across the whole intron. When introns overlapped, a
692 minimal length intron was used such that start of the intron was the most downstream start of the
693 overlapping introns and the end was the most upstream start of the overlapping introns. Exon2
694 coverage was the average coverage for 100 bases of the following exon, using the most downstream
695 3' ss to define the exon. Log2-transformed ratios (Indexes) were calculated for the three
696 comparisons: intron/exon2, splice junction/exon2, splice junction/intron. Figure 1A shows how the
697 splice junction/intron index changes with rapamycin treatment by plotting the value of log2[splice
698 junction-60/intron-60] – log2[splice junction-0/intron-0] for each intron in each replicate. The
699 splicing events plotted here are for locations whose overall transcript level changes less than 2 fold,
700 and whose junctions are supported by at least 35 reads for standard introns or at least 50 reads for
701 protointrons. The general shift of the points to the upper right quadrant indicates increased splicing
702 efficiency (increased junction relative to intron reads) after rapamycin treatment.

703 Intron splice sites and candidate branch sits were extracted for analysis using the mapped
704 splice junctions and by choosing a best branch point using the following heuristics. The likely
705 branch points (underlined) were identified by searching introns for the following sequences in order
706 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

707 or T) constrained to be 45 or more bases away from the 5' ss and no closer than 7 nucleotides
708 upstream from the 3' ss. Candidate introns not matching YTRAY were considered to have no good
709 match to a branchpoint consensus. If multiple equally good branchpoints are identified the one closer
710 to the 3' ss was recorded. Details and scripts are at:
711 <https://github.com/donoyoyo/intron_bp_generator>. To evaluate conservation, phastCons
712 conservation scores were extracted from a window surrounding the splice site or branchpoint using
713 data from the UCSC Human Genome Browser for *S. cerevisiae* at <http://genome.ucsc.edu>.
714 Weblogos [93] were created using the site at https://weblogo.berkeley.edu.
715

716 **Reverse Transcription and PCR**

717 RNA was reverse transcribed using SuperScript III (Life Technologies) according to the
718 manufacturer's instructions using a mixture of anchored oligo-dT (T24VN) and random hexamers as
719 primer. Primers to validate and sequence products of splicing from protointrons by RT-PCR were
720 designed using Primer3 [94]. PCR was performed using oligonucleotides listed in Table S5. PCR
721 products were resolved by electrophoresis on agarose gels and visualized with ethidium bromide
722 staining.
723

724 **Cloning and Sanger sequencing of PCR products**

725 PCR products generated by *T. aquaticus* DNA polymerase (Taq) were cut from low melting
726 point agarose gels and purified using Machernary-Nagel gel extraction kits, then cloned using
727 TOPO-cloning (Invitrogen). Inserts were sequenced by Sanger sequencing at the U. C. Berkeley
728 sequencing center. Splice junctions were identified using BLAT [95] running behind a home copy of
729 the UCSC Genome Browser [96] publicly available at http://intron.ucsc.edu/.

730

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

732 To test the hypothesis that "ACTAAC" (proxy for the branchpoint sequence), GTATGT
733 (proxy for the 5' ss), or any other 6-mer nucleotide sequence within extant yeast ORFs might be
734 enriched or depleted, we created 10,000 codon-permuted versions of the *S. cerevisiae* ORF set and
735 counted the number of each of the 4096 possible 6-mers in each, computing a Z-score for each that
736 compares representation of each in the extant ORF set to the mean representation of each in the
737 10,000 permuted ORF sets. To create permuted ORF sets in a way that preserves the GC content
738 and codon usage of the extant set, we permuted the codons within each ORF (except for the start and
739 stop codons) in the complete set of ORFs. Scripts for creating permuted ORF sets and analysis
740 related to this question can be found under this github link:
741 https://github.com/rshelans/genePermuter

742

743 **CRISPR/Cas9 mediated intron transplantation**

744 Yeast CRISPR editing was done essentially as described by DiCarlo et al [97], except that we
745 rearranged the elements from different plasmids into a simplified single plasmid system by Gibson
746 assembly. We obtained p426-crRNA-CAN1.Y and p414-TEF1p-Cas9-CYC1t [97] from Addgene.
747 To create a BaeI cleavable cassette for easy guide cloning, we annealed oligos newguide1 and
748 newguide2 together, and separately newguide3 and newguide4, and filled to make two fragments
749 which were mixed and then PCR amplified using newguide1 and newguide4 as primers (Table S5).
750 This duplex was purified and assembled using Gibson mix (NEB) with p426-crRNA-CAN1.Y that
751 had been cut with NheI and Acc65I to replace the CAN1.Y guide target region with a stuffer
752 fragment that could be released by BaeI (NEB) and allow any guide to be inserted easily (p426-

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

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

774

775 **Acknowledgements**

776 Thanks very much to Joshua Arribere and Russ Corbett-Detig for insightful critical analysis of an
777 earlier draft of this work. We appreciate Sam Fagg, Jen Quick-Cleveland, Stephanie Nystrom, and
778 Santiago Sanchez for their constructive comments on this work. Thanks to Kevin Karplus for
779 suggesting that codon-permutation could be used to create randomized ORF sets that preserve codon
780 usage and GC content. We thank UCLA colleagues Tracy Johnson and Stephen Douglass for
781 discussions and sharing results prior to publication. Thank you to Shana McDevitt and the Vincent J.
782 Coates Genomics Sequencing Laboratory for advice on sequencing library preparation and for
783 sequencing services.

784

785 **References**

786 1. Fica SM, Nagai K. Cryo-electron microscopy snapshots of the spliceosome: structural
787 insights into a dynamic ribonucleoprotein machine. Nat Struct Mol Biol. 2017;24(10):791-9. Epub
788 2017/10/06. doi: 10.1038/nsmb.3463. PubMed PMID: 28981077.
789 2. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu Rev
790 Biochem. 2015;84:291-323. Epub 2015/03/19. doi: 10.1146/annurev-biochem-060614-034316.
791 PubMed PMID: 25784052; PubMed Central PMCID: PMCPMC4526142.
792 3. Scheres SH, Nagai K. CryoEM structures of spliceosomal complexes reveal the molecular
793 mechanism of pre-mRNA splicing. Curr Opin Struct Biol. 2017;46:130-9. Epub 2017/09/10. doi:
794 10.1016/j.sbi.2017.08.001. PubMed PMID: 28888105.
795 4. Catania F, Lynch M. Where do introns come from? PLoS Biol. 2008;6(11):e283. Epub
796 2008/12/11. doi: 10.1371/journal.pbio.0060283. PubMed PMID: 19067485; PubMed Central
797 PMCID: PMCPMC2586383.
798 5. Lynch M, Richardson AO. The evolution of spliceosomal introns. Curr Opin Genet Dev.
799 2002;12(6):701-10. PubMed PMID: 12433585.
800 6. Carmel L, Wolf YI, Rogozin IB, Koonin EV. Three distinct modes of intron dynamics in the
801 evolution of eukaryotes. Genome Res. 2007;17(7):1034-44. Epub 2007/05/15. doi:
802 10.1101/gr.6438607. PubMed PMID: 17495008; PubMed Central PMCID: PMCPMC1899114.
803 7. Fink GR. Pseudogenes in yeast? Cell. 1987;49(1):5-6. Epub 1987/04/10. PubMed PMID:
804 3549000.
805 8. Catania F. From intronization to intron loss: How the interplay between mRNA-associated
806 processes can shape the architecture and the expression of eukaryotic genes. Int J Biochem Cell Biol.
807 2017;91(Pt B):136-44. doi: 10.1016/j.biocel.2017.06.017. PubMed PMID: 28673893.
808 9. Yenerall P, Zhou L. Identifying the mechanisms of intron gain: progress and trends. Biol
809 Direct. 2012;7:29. Epub 2012/09/12. doi: 10.1186/1745-6150-7-29. PubMed PMID: 22963364;
810 PubMed Central PMCID: PMCPMC3443670.

811 10. Rogozin IB, Lyons-Weiler J, Koonin EV. Intron sliding in conserved gene families. Trends
812 Genet. 2000;16(10):430-2. Epub 2000/10/26. PubMed PMID: 11050324.
813 11. Sorek R. The birth of new exons: mechanisms and evolutionary consequences. RNA.
814 2007;13(10):1603-8. Epub 2007/08/22. doi: 10.1261/rna.682507. PubMed PMID: 17709368;
815 PubMed Central PMCID: PMCPMC1986822.
816 12. Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, et al. Green evolution and
817 dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science.
818 2009;324(5924):268-72. Epub 2009/04/11. doi: 10.1126/science.1167222. PubMed PMID:
819 19359590.
820 13. Collemare J, Beenen HG, Crous PW, de Wit PJ, van der Burgt A. Novel Introner-Like
821 Elements in fungi Are Involved in Parallel Gains of Spliceosomal Introns. PLoS One.
822 2015;10(6):e0129302. Epub 2015/06/06. doi: 10.1371/journal.pone.0129302. PubMed PMID:
823 26046656; PubMed Central PMCID: PMCPMC4457414.
824 14. Collemare J, van der Burgt A, de Wit PJ. At the origin of spliceosomal introns: Is
825 multiplication of introner-like elements the main mechanism of intron gain in fungi? Commun Integr
826 Biol. 2013;6(2):e23147. Epub 2013/06/12. doi: 10.4161/cib.23147. PubMed PMID: 23750299;
827 PubMed Central PMCID: PMCPMC3609843.
828 15. Simmons MP, Bachy C, Sudek S, van Baren MJ, Sudek L, Ares M, Jr., et al. Intron Invasions
829 Trace Algal Speciation and Reveal Nearly Identical Arctic and Antarctic Micromonas Populations.
830 Mol Biol Evol. 2015;32(9):2219-35. Epub 2015/05/23. doi: 10.1093/molbev/msv122. PubMed
831 PMID: 25998521; PubMed Central PMCID: PMCPMC4540971.
832 16. Verhelst B, Van de Peer Y, Rouze P. The complex intron landscape and massive intron
833 invasion in a picoeukaryote provides insights into intron evolution. Genome Biol Evol.
834 2013;5(12):2393-401. Epub 2013/11/26. doi: 10.1093/gbe/evt189. PubMed PMID: 24273312;
835 PubMed Central PMCID: PMCPMC3879977.
836 17. van der Burgt A, Severing E, de Wit PJ, Collemare J. Birth of new spliceosomal introns in
837 fungi by multiplication of introner-like elements. Curr Biol. 2012;22(13):1260-5. Epub 2012/06/05.
838 doi: 10.1016/j.cub.2012.05.011. PubMed PMID: 22658596.
839 18. Lee S, Stevens SW. Spliceosomal intronogenesis. Proc Natl Acad Sci U S A.
840 2016;113(23):6514-9. doi: 10.1073/pnas.1605113113. PubMed PMID: 27217561; PubMed Central
841 PMCID: PMCPMC4988565.
842 19. Li W, Tucker AE, Sung W, Thomas WK, Lynch M. Extensive, recent intron gains in
843 Daphnia populations. Science. 2009;326(5957):1260-2. Epub 2009/12/08. doi:
844 10.1126/science.1179302. PubMed PMID: 19965475; PubMed Central PMCID: PMCPMC3878872.
845 20. Huff JT, Zilberman D, Roy SW. Mechanism for DNA transposons to generate introns on
846 genomic scales. Nature. 2016. doi: 10.1038/nature20110. PubMed PMID: 27760113.
847 21. Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW. Origin of introns by
848 'intronization' of exonic sequences. Trends Genet. 2008;24(8):378-81. Epub 2008/07/04. doi:
849 10.1016/j.tig.2008.05.007. PubMed PMID: 18597887.
850 22. Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. Comparative
851 analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role
852 in shaping the human transcriptome. Genome Biol. 2007;8(6):R127. doi: 10.1186/gb-2007-8-6-r127.
853 PubMed PMID: 17594509; PubMed Central PMCID: PMCPMC2394776.
854 23. Catania F, Schmitz J. On the path to genetic novelties: insights from programmed DNA
855 elimination and RNA splicing. Wiley Interdiscip Rev RNA. 2015;6(5):547-61. doi:
856 10.1002/wrna.1293. PubMed PMID: 26140477.

24. 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.
25. 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.
26. Warner JR. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci. 1999;24(11):437-40. Epub 1999/11/05. PubMed PMID: 10542411.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. 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

903 2017/12/20. doi: 10.1101/gr.225615.117. PubMed PMID: 29254943; PubMed Central PMCID:
904 PMCPMC5793784.
905 37. Schreiber K, Csaba G, Haslbeck M, Zimmer R. Alternative Splicing in Next Generation
906 Sequencing Data of Saccharomyces cerevisiae. PLoS One. 2015;10(10):e0140487. Epub
907 2015/10/16. doi: 10.1371/journal.pone.0140487. PubMed PMID: 26469855; PubMed Central
908 PMCID: PMCPMC4607428.
909 38. Douglass S, Leung CS, Johnson TJ. Extensive Splicing across the Saccharomyces genome.
910 bioRxiv. 2019. doi: 10.1101/515163.
911 39. Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, et al. A large-scale full-
912 length cDNA analysis to explore the budding yeast transcriptome. Proc Natl Acad Sci U S A.
913 2006;103(47):17846-51. Epub 2006/11/15. doi: 10.1073/pnas.0605645103. PubMed PMID:
914 17101987; PubMed Central PMCID: PMCPMC1693835.
915 40. Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, et al. Translation of
916 small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. Cell
917 Rep. 2014;7(6):1858-66. doi: 10.1016/j.celrep.2014.05.023. PubMed PMID: 24931603; PubMed
918 Central PMCID: PMCPMC4105149.
919 41. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al.
920 Proto-genes and de novo gene birth. Nature. 2012;487(7407):370-4. doi: 10.1038/nature11184.
921 PubMed PMID: 22722833; PubMed Central PMCID: PMCPMC3401362.
922 42. van Dijk EL, Chen CL, d'Aubenton-Carafa Y, Gourvennec S, Kwapisz M, Roche V, et al.
923 XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature.
924 2011;475(7354):114-7. doi: 10.1038/nature10118. PubMed PMID: 21697827.
925 43. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, Camblong J, et al. Bidirectional
926 promoters generate pervasive transcription in yeast. Nature. 2009;457(7232):1033-7. doi:
927 10.1038/nature07728. PubMed PMID: 19169243; PubMed Central PMCID: PMCPMC2766638.
928 44. Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. Widespread
929 bidirectional promoters are the major source of cryptic transcripts in yeast. Nature.
930 2009;457(7232):1038-42. doi: 10.1038/nature07747. PubMed PMID: 19169244.
931 45. Ito T, Miura F, Onda M. Unexpected complexity of the budding yeast transcriptome. IUBMB
932 Life. 2008;60(12):775-81. doi: 10.1002/iub.121. PubMed PMID: 18649367.
933 46. Davis CA, Ares M, Jr. Accumulation of unstable promoter-associated transcripts upon loss of
934 the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A.
935 2006;103(9):3262-7. Epub 2006/02/18. doi: 10.1073/pnas.0507783103. PubMed PMID: 16484372;
936 PubMed Central PMCID: PMCPMC1413877.
937 47. Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, et al. Cryptic pol II
938 transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase.
939 Cell. 2005;121(5):725-37. Epub 2005/06/07. doi: 10.1016/j.cell.2005.04.030. PubMed PMID:
940 15935759.
941 48. Arribere JA, Gilbert WV. Roles for transcript leaders in translation and mRNA decay
942 revealed by transcript leader sequencing. Genome Res. 2013;23(6):977-87. Epub 2013/04/13. doi:
943 10.1101/gr.150342.112. PubMed PMID: 23580730; PubMed Central PMCID: PMCPMC3668365.
944 49. He F, Jacobson A. Identification of a novel component of the nonsense-mediated mRNA
945 decay pathway by use of an interacting protein screen. Genes Dev. 1995;9(4):437-54. Epub
946 1995/02/15. PubMed PMID: 7883168.
947 50. Kandels-Lewis S, Seraphin B. Involvement of U6 snRNA in 5' splice site selection. Science.
948 1993;262(5142):2035-9. Epub 1993/12/24. PubMed PMID: 8266100.

949 51. Lesser CF, Guthrie C. Mutations in U6 snRNA that alter splice site specificity: implications
950 for the active site. Science. 1993;262(5142):1982-8. Epub 1993/12/24. PubMed PMID: 8266093.
951 52. Roca X, Krainer AR. Recognition of atypical 5' splice sites by shifted base-pairing to U1
952 snRNA. Nat Struct Mol Biol. 2009;16(2):176-82. Epub 2009/01/27. doi: 10.1038/nsmb.1546.
953 PubMed PMID: 19169258; PubMed Central PMCID: PMCPMC2719486.
954 53. Wu Q, Krainer AR. Splicing of a divergent subclass of AT-AC introns requires the major
955 spliceosomal snRNAs. RNA. 1997;3(6):586-601. Epub 1997/06/01. PubMed PMID: 9174094;
956 PubMed Central PMCID: PMCPMC1369508.
957 54. Parker R, Siliciano PG. Evidence for an essential non-Watson-Crick interaction between the
958 first and last nucleotides of a nuclear pre-mRNA intron. Nature. 1993;361(6413):660-2. Epub
959 1993/02/18. doi: 10.1038/361660a0. PubMed PMID: 8437627.
960 55. Colot HV, Stutz F, Rosbash M. The yeast splicing factor Mud13p is a commitment complex
961 component and corresponds to CBP20, the small subunit of the nuclear cap-binding complex. Genes
962 Dev. 1996;10(13):1699-708. doi: 10.1101/gad.10.13.1699. PubMed PMID: 8682299.
963 56. Lewis JD, Izaurralde E, Jarmolowski A, McGuigan C, Mattaj IW. A nuclear cap-binding
964 complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. Genes Dev.
965 1996;10(13):1683-98. doi: 10.1101/gad.10.13.1683. PubMed PMID: 8682298.
966 57. Lewis JD, Gorlich D, Mattaj IW. A yeast cap binding protein complex (yCBC) acts at an
967 early step in pre-mRNA splicing. Nucleic Acids Res. 1996;24(17):3332-6. doi:
968 10.1093/nar/24.17.3332. PubMed PMID: 8811086; PubMed Central PMCID: PMCPMC146107.
969 58. Schwer B, Shuman S. Conditional inactivation of mRNA capping enzyme affects yeast pre-
970 mRNA splicing in vivo. RNA. 1996;2(6):574-83. PubMed PMID: 8718686; PubMed Central
971 PMCID: PMCPMC1369396.
972 59. Lepennetier G, Catania F. mRNA-Associated Processes and Their Influence on Exon-Intron
973 Structure in Drosophila melanogaster. G3 (Bethesda). 2016;6(6):1617-26. doi:
974 10.1534/g3.116.029231. PubMed PMID: 27172210; PubMed Central PMCID: PMCPMC4889658.
975 60. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al.
976 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res.
977 2005;15(8):1034-50. Epub 2005/07/19. doi: 10.1101/gr.3715005. PubMed PMID: 16024819;
978 PubMed Central PMCID: PMCPMC1182216.
979 61. Spingola M, Grate L, Haussler D, Ares M, Jr. Genome-wide bioinformatic and molecular
980 analysis of introns in Saccharomyces cerevisiae. RNA. 1999;5(2):221-34. Epub 1999/02/19.
981 PubMed PMID: 10024174; PubMed Central PMCID: PMCPMC1369754.
982 62. Howe KJ, Ares M, Jr. Intron self-complementarity enforces exon inclusion in a yeast pre-
983 mRNA. Proc Natl Acad Sci U S A. 1997;94(23):12467-72. Epub 1997/11/14. PubMed PMID:
984 9356473; PubMed Central PMCID: PMCPMC25003.
985 63. Libri D, Stutz F, McCarthy T, Rosbash M. RNA structural patterns and splicing: molecular
986 basis for an RNA-based enhancer. RNA. 1995;1(4):425-36. Epub 1995/06/01. PubMed PMID:
987 7493320; PubMed Central PMCID: PMCPMC1482409.
988 64. Farlow A, Meduri E, Dolezal M, Hua L, Schlotterer C. Nonsense-mediated decay enables
989 intron gain in Drosophila. PLoS Genet. 2010;6(1):e1000819. doi: 10.1371/journal.pgen.1000819.
990 PubMed PMID: 20107520; PubMed Central PMCID: PMCPMC2809761.
991 65. Karaduman R, Chanarat S, Pfander B, Jentsch S. Error-Prone Splicing Controlled by the
992 Ubiquitin Relative Hub1. Mol Cell. 2017;67(3):423-32 e4. Epub 2017/07/18. doi:
993 10.1016/j.molcel.2017.06.021. PubMed PMID: 28712727.

994 66. Juneau K, Nislow C, Davis RW. Alternative splicing of PTC7 in Saccharomyces cerevisiae
995 determines protein localization. Genetics. 2009;183(1):185-94. Epub 2009/07/01. doi:
996 10.1534/genetics.109.105155. PubMed PMID: 19564484; PubMed Central PMCID:
997 PMCPMC2746143.
998 67. Zhang Z, Hesselberth JR, Fields S. Genome-wide identification of spliced introns using a
999 tiling microarray. Genome Res. 2007;17(4):503-9. Epub 2007/03/14. doi: 10.1101/gr.6049107.
1000 PubMed PMID: 17351133; PubMed Central PMCID: PMCPMC1832097.
1001 68. Farlow A, Dolezal M, Hua L, Schlotterer C. The genomic signature of splicing-coupled
1002 selection differs between long and short introns. Mol Biol Evol. 2012;29(1):21-4. doi:
1003 10.1093/molbev/msr201. PubMed PMID: 21878685; PubMed Central PMCID: PMCPMC3245539.
1004 69. Zhou Z, Dang Y, Zhou M, Yuan H, Liu Y. Codon usage biases co-evolve with transcription
1005 termination machinery to suppress premature cleavage and polyadenylation. Elife. 2018;7. Epub
1006 2018/03/17. doi: 10.7554/eLife.33569. PubMed PMID: 29547124; PubMed Central PMCID:
1007 PMCPMC5869017.
1008 70. Gamble CE, Brule CE, Dean KM, Fields S, Grayhack EJ. Adjacent Codons Act in Concert to
1009 Modulate Translation Efficiency in Yeast. Cell. 2016;166(3):679-90. Epub 2016/07/05. doi:
1010 10.1016/j.cell.2016.05.070. PubMed PMID: 27374328; PubMed Central PMCID:
1011 PMCPMC4967012.
1012 71. Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, et al. The Awesome
1013 Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the
1014 Saccharomyces sensu stricto Genus. G3 (Bethesda). 2011;1(1):11-25. Epub 2012/03/03. doi:
1015 10.1534/g3.111.000273. PubMed PMID: 22384314; PubMed Central PMCID: PMCPMC3276118.
1016 72. Parenteau J, Durand M, Veronneau S, Lacombe AA, Morin G, Guerin V, et al. Deletion of
1017 many yeast introns reveals a minority of genes that require splicing for function. Mol Biol Cell.
1018 2008;19(5):1932-41. Epub 2008/02/22. doi: 10.1091/mbc.E07-12-1254. PubMed PMID: 18287520;
1019 PubMed Central PMCID: PMCPMC2366882.
1020 73. Davis CA, Grate L, Spingola M, Ares M, Jr. Test of intron predictions reveals novel splice
1021 sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. Nucleic
1022 Acids Res. 2000;28(8):1700-6. Epub 2000/03/29. PubMed PMID: 10734188; PubMed Central
1023 PMCID: PMCPMC102823.
1024 74. Arlt H, Tauer R, Feldmann H, Neupert W, Langer T. The YTA10-12 complex, an AAA
1025 protease with chaperone-like activity in the inner membrane of mitochondria. Cell. 1996;85(6):875-
1026 85. Epub 1996/06/14. PubMed PMID: 8681382.
1027 75. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. High-resolution
1028 view of the yeast meiotic program revealed by ribosome profiling. Science. 2012;335(6068):552-7.
1029 Epub 2011/12/24. doi: 10.1126/science.1215110. PubMed PMID: 22194413; PubMed Central
1030 PMCID: PMCPMC3414261.
1031 76. Louis EJ, Haber JE. The structure and evolution of subtelomeric Y' repeats in Saccharomyces
1032 cerevisiae. Genetics. 1992;131(3):559-74. Epub 1992/07/01. PubMed PMID: 1628806; PubMed
1033 Central PMCID: PMCPMC1205030.
1034 77. Yamada M, Hayatsu N, Matsuura A, Ishikawa F. Y'-Help1, a DNA helicase encoded by the
1035 yeast subtelomeric Y' element, is induced in survivors defective for telomerase. J Biol Chem.
1036 1998;273(50):33360-6. Epub 1998/12/05. PubMed PMID: 9837911.
1037 78. Meyer M, Plass M, Perez-Valle J, Eyras E, Vilardell J. Deciphering 3'ss selection in the yeast
1038 genome reveals an RNA thermosensor that mediates alternative splicing. Mol Cell.

1039 2011;43(6):1033-9. Epub 2011/09/20. doi: 10.1016/j.molcel.2011.07.030. PubMed PMID:
1040 21925391.
1041 79. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy splicing drives mRNA isoform diversity in
1042 human cells. PLoS Genet. 2010;6(12):e1001236. doi: 10.1371/journal.pgen.1001236. PubMed
1043 PMID: 21151575; PubMed Central PMCID: PMCPMC3000347.
1044 80. Kannan R, Hartnett S, Voelker RB, Berglund JA, Staley JP, Baumann P. Intronic sequence
1045 elements impede exon ligation and trigger a discard pathway that yields functional telomerase RNA
1046 in fission yeast. Genes Dev. 2013;27(6):627-38. Epub 2013/03/08. doi: 10.1101/gad.212738.112.
1047 PubMed PMID: 23468430; PubMed Central PMCID: PMCPMC3613610.
1048 81. Mele M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. Chromatin
1049 environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. Genome
1050 Res. 2017;27(1):27-37. Epub 2016/12/09. doi: 10.1101/gr.214205.116. PubMed PMID: 27927715;
1051 PubMed Central PMCID: PMCPMC5204342.
1052 82. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution
1053 of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505(7485):635-40. Epub
1054 2014/01/28. doi: 10.1038/nature12943. PubMed PMID: 24463510.
1055 83. Liu J, Martin-Yken H, Bigey F, Dequin S, Francois JM, Capp JP. Natural yeast promoter
1056 variants reveal epistasis in the generation of transcriptional-mediated noise and its potential benefit
1057 in stressful conditions. Genome Biol Evol. 2015;7(4):969-84. Epub 2015/03/13. doi:
1058 10.1093/gbe/evv047. PubMed PMID: 25762217; PubMed Central PMCID: PMCPMC4419794.
1059 84. Pickering BM, Willis AE. The implications of structured 5' untranslated regions on
1060 translation and disease. Semin Cell Dev Biol. 2005;16(1):39-47. Epub 2005/01/22. doi:
1061 10.1016/j.semcdb.2004.11.006. PubMed PMID: 15659338.
1062 85. Andreatta ME, Levine JA, Foy SG, Guzman LD, Kosinski LJ, Cordes MH, et al. The Recent
1063 De Novo Origin of Protein C-Termini. Genome Biol Evol. 2015;7(6):1686-701. doi:
1064 10.1093/gbe/evv098. PubMed PMID: 26002864; PubMed Central PMCID: PMCPMC4494051.
1065 86. Tress ML, Abascal F, Valencia A. Alternative Splicing May Not Be the Key to Proteome
1066 Complexity. Trends Biochem Sci. 2017;42(2):98-110. Epub 2016/10/08. doi:
1067 10.1016/j.tibs.2016.08.008. PubMed PMID: 27712956.
1068 87. Ares M. Isolation of total RNA from yeast cell cultures. Cold Spring Harb Protoc.
1069 2012;2012(10):1082-6. Epub 2012/10/03. doi: 10.1101/pdb.prot071456. PubMed PMID: 23028070.
1070 88. Yassour M, Pfiffner J, Levin JZ, Adiconis X, Gnirke A, Nusbaum C, et al. Strand-specific
1071 RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across
1072 yeast species. Genome Biol. 2010;11(8):R87. Epub 2010/08/28. doi: 10.1186/gb-2010-11-8-r87.
1073 PubMed PMID: 20796282; PubMed Central PMCID: PMCPMC2945789.
1074 89. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al.
1075 Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res.
1076 2012;40(Database issue):D700-5. Epub 2011/11/24. doi: 10.1093/nar/gkr1029. PubMed PMID:
1077 22110037; PubMed Central PMCID: PMCPMC3245034.
1078 90. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods.
1079 2012;9(4):357-9. Epub 2012/03/06. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PubMed
1080 Central PMCID: PMCPMC3322381.
1081 91. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
1082 universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. Epub 2012/10/30. doi:
1083 10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PubMed Central PMCID:
1084 PMCPMC3530905.

1085 92. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-
1086 seq data with DESeq2. Genome Biol. 2014;15(12):550. Epub 2014/12/18. doi: 10.1186/s13059-014-
1087 0550-8. PubMed PMID: 25516281; PubMed Central PMCID: PMCPMC4302049.
1088 93. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator.
1089 Genome Res. 2004;14(6):1188-90. Epub 2004/06/03. doi: 10.1101/gr.849004. PubMed PMID:
1090 15173120; PubMed Central PMCID: PMCPMC419797.
1091 94. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3--new
1092 capabilities and interfaces. Nucleic Acids Res. 2012;40(15):e115. Epub 2012/06/26. doi:
1093 10.1093/nar/gks596. PubMed PMID: 22730293; PubMed Central PMCID: PMCPMC3424584.
1094 95. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64. Epub
1095 2002/04/05. doi: 10.1101/gr.229202. PubMed PMID: 11932250; PubMed Central PMCID:
1096 PMCPMC187518.
1097 96. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC
1098 Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42(Database issue):D764-70.
1099 Epub 2013/11/26. doi: 10.1093/nar/gkt1168. PubMed PMID: 24270787; PubMed Central PMCID:
1100 PMCPMC3964947.
1101 97. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in
1102 Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. 2013;41(7):4336-43.
1103 Epub 2013/03/06. doi: 10.1093/nar/gkt135. PubMed PMID: 23460208; PubMed Central PMCID:
1104 PMCPMC3627607.
1105

1106 **Figure Legends**

1107 **Figure 1. Identification and Validation of Splicing at Unannotated Genomic Locations.** RNA
1108 sequencing reads corresponding to spliced RNAs defined as described in the text were used to
1109 identify and measure splicing. Reads (~300M) were obtained from untreated cells and cells treated
1110 for 60 minutes with rapamycin from two biological replicate experiments. **(A)** Splicing efficiency of
1111 many introns improves after rapamycin treatment as judged by the log2 fold change in the ratio of
1112 splice junction reads to intron reads (splicing index; see Methods) for replicate experiments.
1113 Standard introns with ≥35 total junction reads are shown as blue dots. Unannotated and non-
1114 overlapping splicing events (protointrons) with ≥50 total junction reads are shown as orange dots.
1115 Data points in quadrant I indicate introns in which splicing improved after treatment with rapamycin
1116 in both replicate experiments. **(B)** Genomic alignment and lack of conservation for three example
1117 protointrons. Protointrons in a divergent upstream transcript *antiASH1*, a XUT *XUT12R-370*, and an

1118 mRNA for *TAF13* are shown. The 5' ss is green, the branchpoint sequence is yellow, and the 3' ss is
1119 blue. The vertical bars indicate where additional intron sequences are not shown. **(C)** RT-PCR
1120 validation of protointron splicing and increased splicing after rapamycin is shown for the
1121 protointrons in (B), and for a non-coding vegetative cell transcript of a meiotic gene *ncSPO1*. PCR
1122 products corresponding in size to spliced (predicted based on RNAseq read structure) and unspliced
1123 RNAs are labeled. Splice junctions were confirmed by sequencing cloned PCR products. **(D)**
1124 Validation of protointron splicing through 5' ss not observed in standard introns. Junctions were
1125 validated by sequencing cloned PCR products. **(E)** A protointron in *SUT635* uses both GT-AG and
1126 AT-AC splice sites. The sequences of two cloned PCR products from *SUT635* are aligned to the
1127 genome (above) and the sequencing trace from the clone representing the use of AT-AC junctions is
1128 shown (below).
1129
1130 **Figure 2. Differences between Standard Introns and Protointrons. (A)** Splice site and
1131 branchpoint sequences of protointrons are less constrained in sequence than the standard introns.
1132 Weblogos representing position-specific weight matrices of the 5' ss, branch points, and 3' ss of the
1133 standard introns (top) and the protointrons (bottom) are shown. **(B)** The pattern of sequence
1134 conservation in the context of protointron splicing signals (orange bars in each panel) is bimodal as
1135 compared to the standard introns (blue bars in each panel). Histograms of the standard introns (blue
1136 bars) and the protointrons (orange bars) showing the distributions of PhastCons scores in windows
1137 containing the 5' ss, branchpoints, and 3' ss of the standard introns and protointrons. See text. **(C)**
1138 The size distribution of protointrons is distinct from that of the standard introns. Histograms show
1139 the distribution of intron sizes for the standard introns (blue bars) and the protointrons (orange bars).
1140 A Kolmogorov-Smirnoff test indicates the two distributions are different ($D = 0.22$, p value $\leq 10^{-4}$).

1141 **(D)** Protointrons are much less efficiently spliced than standard introns. The scatter plot shows the
1142 relationship between splicing efficiency in untreated (time 0) cells and the change in splicing
1143 efficiency after one hour in rapamycin. Standard introns are shown in blue, protointrons in orange.
1144

1145 **Figure 3. Hexamers representing splicing signals are depleted from annotated *S. cerevisiae***
1146 **ORFs**. (**A)** Histogram of counts of two hexamers (6-mers) serving as proxies for the branch point
1147 (ACTAAC, blue) and the 5' ss (GTATGT, yellow) in the extant ORF set of *S. cerevisiae* (vertical
1148 lines) as compared with the distribution of counts in each of 10,000 codon-permuted ORF sets. (**B)**
1149 Histogram of Z-scores computed for each of 4096 6-mers in the extant *S. cerevisiae* ORF set relative
1150 to their corresponding mean representation in 10,000 codon-permuted *S. cerevisiae* ORF sets. The
1151 number of 6-mers (y-axis) with the given Z-score (x-axis) is represented as a histogram in grey.
1152 Similar distributions are shown for two subclasses: those containing stop codons (blue histogram)
1153 and those containing start codons (maroon histogram). The Z-scores for the branchpoint proxy
1154 hexamer ACTAAC and the 5' splice proxy hexamer GTATGT are marked in the plot. The 6-mer
1155 "ACTAAC" had a Z-score of -12.25 and ranked 153rd lowest among all 4096 6-mers, and 91st lowest
1156 of 759 6-mers carrying stop codons. The 6-mer GTATGT had a Z-score of -6.98 and ranked 671st
1157 lowest among all 4096 6-mers, and 7th lowest of 255 6-mers carrying start codons.
1158

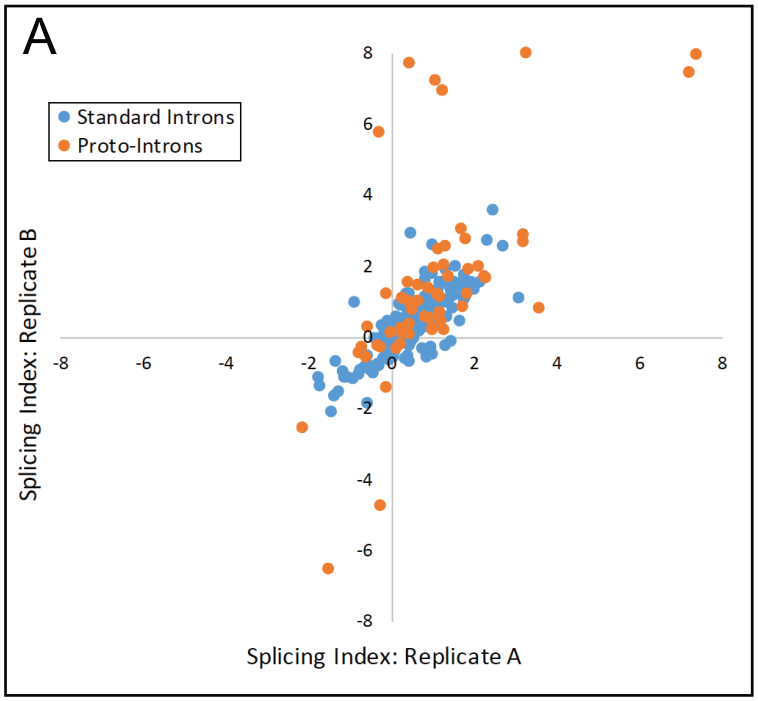
1159 **Figure 4. Protointrons are found in other *Saccharomyces* species but are not conserved.** RT-
1160 PCR products from RNA of *S. mikatae* (left) and *S. bayanus* (right) at different protointrons
1161 identified by RNAseq. Splice junctions were validated by cloning and sequencing the PCR products
1162 indicated by a white dot. Below the gel image are shown alignments of the RT-PCR product

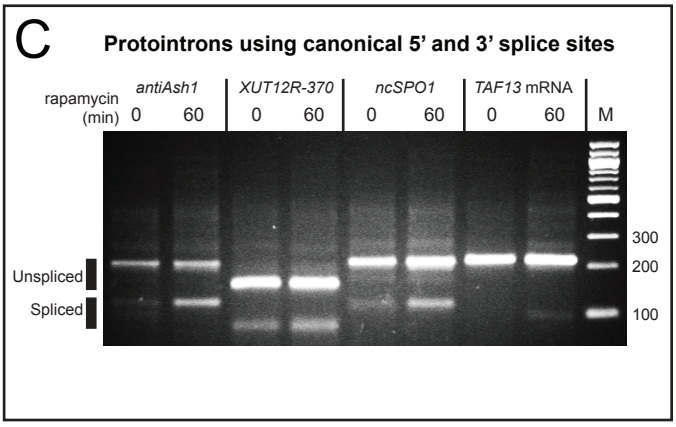
1163 sequences from *S. mikatae antiYCR060W* and *S. bayanus YOL122C* to their corresponding genomes
1164 to show lack of conservation of splicing signals (boxed).
1165
1166 **Figure 5. Unusually efficient protointrons that may be evolving toward standard introns.**
1167 Positions of 5 efficiently spliced protointrons that share similarity with standard introns on the
1168 unrooted tree of sensu stricto *Saccharomyces* species are shown. Grey arrows indicate separation
1169 points that delineate boundaries between species having or lacking the indicated protointron
1170 sequence. Bars in the alignments indicate that sequences between these blocks are not shown. 5' ss
1171 are green, branchpoint sequences are yellow, and 3' ss are blue. Although these protointrons are
1172 restricted to one or two closely related species, their splicing efficiency approaches that of standard
1173 introns. Most protointrons are unique to a species and are very inefficiently spliced.
1174
1175 **Figure 6. Introns in the Y' element repeat family.** Two different introns are found in the
1176 transcribed Y' repeat elements. **(A)** Y' intron 1. Top: expanded view of the protein encoded by
1177 *YRF1-6* located in the Y' element at the left end of chromosome XIV with the Sir1 and DECD
1178 helicase homology regions indicated. An expanded segment from the upstream part of the gene
1179 shows the alignment of the detected intron relative to the annotated predicted intron at SGD. At the
1180 bottom is shown the alignment of the seven different versions of this intron from the seventeen Y'
1181 elements in the *S. cerevisiae* genome that possess it. Sequence names are based on standard and
1182 systematic annotations from the *Saccharomyces* Genome Database (SGD). 5' ss are green,
1183 branchpoint sequences are yellow, and 3' ss are blue. **(B)** Y' intron 2. Top: expanded view of the
1184 protein encoded by *YLR464W* located in the Y' element at the right end of chromosome XII. The
1185 alignment shows the detected intron relative to the annotated predicted intron at SGD. At the bottom

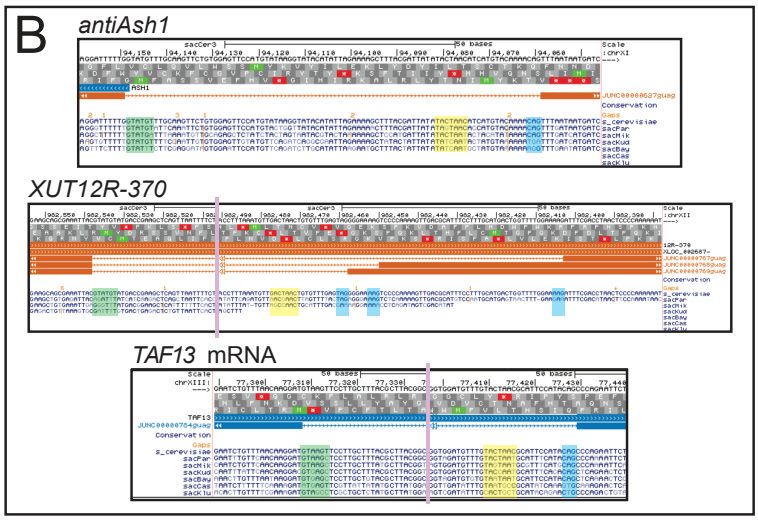
1186 is shown the alignment of the nine different versions of this intron from the nine Y' elements in the
1187 *S. cerevisiae* genome that possess it. Splicing signals are highlighted as in (A).
1188
1189 **Figure 7. The *S. bayanus* YTA12 5'UTR intron is efficiently spliced in *S. cerevisiae* (A)**
1190 Alignment of the *YTA12* promoter and 5'UTR from *S.cerevisiae* (sacCer, no intron), *S. bayanus*
1191 (sacBay, very efficient intron), and the *S. cerevisiae* strain carrying the *S. bayanus* intron (Sc-SbI),
1192 showing the major transcription start sites and the cryptic start sites (>), the splice sites (underlined),
1193 and the aligned base pairs (*). **(B)** Strategy for CRISPR/Cas9 editing-based transplantation of the *S.*
1194 *bayanus* intron into *S. cerevisiae*. A guide sequence was designed to recognize a sequence present in
1195 the *S. cerevisiae YTA12* 5'UTR but not present in the *S. bayanus* intron. A plasmid derived from
1196 those provided by DiCarlo et al. [97] expressing this guide along with Cas9 was co-transformed with
1197 a synthetic rescue fragment that contained the *S. bayanus* intron sequence between "exons" from *S.*
1198 *cerevisiae*. Repair of the double-stranded break using this rescue fragment results in transplantation
1199 of the *S. bayanus* intron into *S. cerevisiae*. **(C)** Reverse transcriptase primer extension analysis of
1200 RNA from the YTA12 locus of S. cerevisiae strains with the transplanted *S. bayanus* intron and
1201 mutant derivatives. The cDNAs representing unspliced (native) start sites, spliced (or deleted) RNAs
1202 initiating from the normal start site, and unspliced RNAs arising from cryptic start sites are indicated
1203 at left. Lane 1, wild type; lane 2, deletion of the region that aligns with the *S. bayanus* intron; lane 3,
1204 transplantation of the wild type S. bayanus intron; lane 4, GaAUGU mutation of the *S. bayanus*
1205 intron 5' ss; lane 5, GaAcGU mutant eliminating both the 5' ss and the start codon of a uORF; lane
1206 6, GUAcGU mutant that creates a common functional ss while removing the start codon of the
1207 uORF; lane m, 100 bp ladder markers.
1208

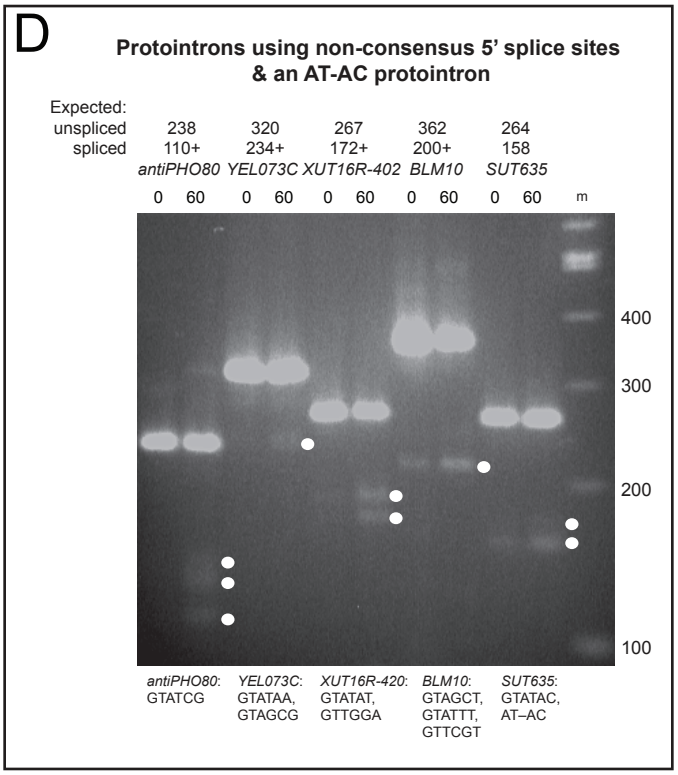
1209 **Supplemental Figure 1. (A)** Coherence of gene expression changes after 60 minute rapamycin
1210 treatment between the two replicate experiments. Log2ratio of treatment to control read coverage
1211 over genes was plotted giving an R^2 value of ~0.99. Supplemental to Fig 1A. **(B)** Percentage of
1212 standard introns and protointrons that are located in non-coding, 5'UTR, 5'UTR^coding, coding,
1213 coding^3'UTR and 3'UTR regions. **(C)** Coverage tracks showing transcription through the genomic
1214 locus upstream of *ASH1* where the *antiASH1* protointron is located. Supplemental to Figs 1 B and C.
1215 **(D)** Coverage tracks showing transcription through the genomic locus upstream of *TUS1* where the
1216 *XUT12R-370* protointron is located. Supplemental to Figs 1 B and C. **(E)** Coverage tracks showing
1217 transcription through the genomic locus of *TAF13* where the *TAF13* protointron is located.
1218 Supplemental to Figs 1 B and C. **(F)** Coverage tracks showing transcription through the genomic
1219 locus of *SPO1* where the *ncSPO1* protointron is located. Supplemental to Figs 1 B and C. **(G)**
1220 Alignment of sequenced RT-PCR products showing the location of protointrons with unusual 5' ss.
1221 Supplemental to Fig 1D.
1222
1223 **Table S1:** *Saccharomyces cerevisiae* standard introns
1224 **Table S2:** *Saccharomyces cerevisiae* overlapping standard introns
1225 **Table S3:** *Saccharomyces cerevisiae* protointrons
1226 **Table S4:** *Saccharomyces cerevisiae* filtered reads
1227 **Table S5:** Oligonucleotides
1228 **Table S6:** *Saccharomyces mikatae* introns
1229 **Table S7:** *Saccharomyces bayanus* introns
1230 **Table S8:** *Saccharomyces cerevisiae* in-frame protointrons

Talkish et al Fig 1

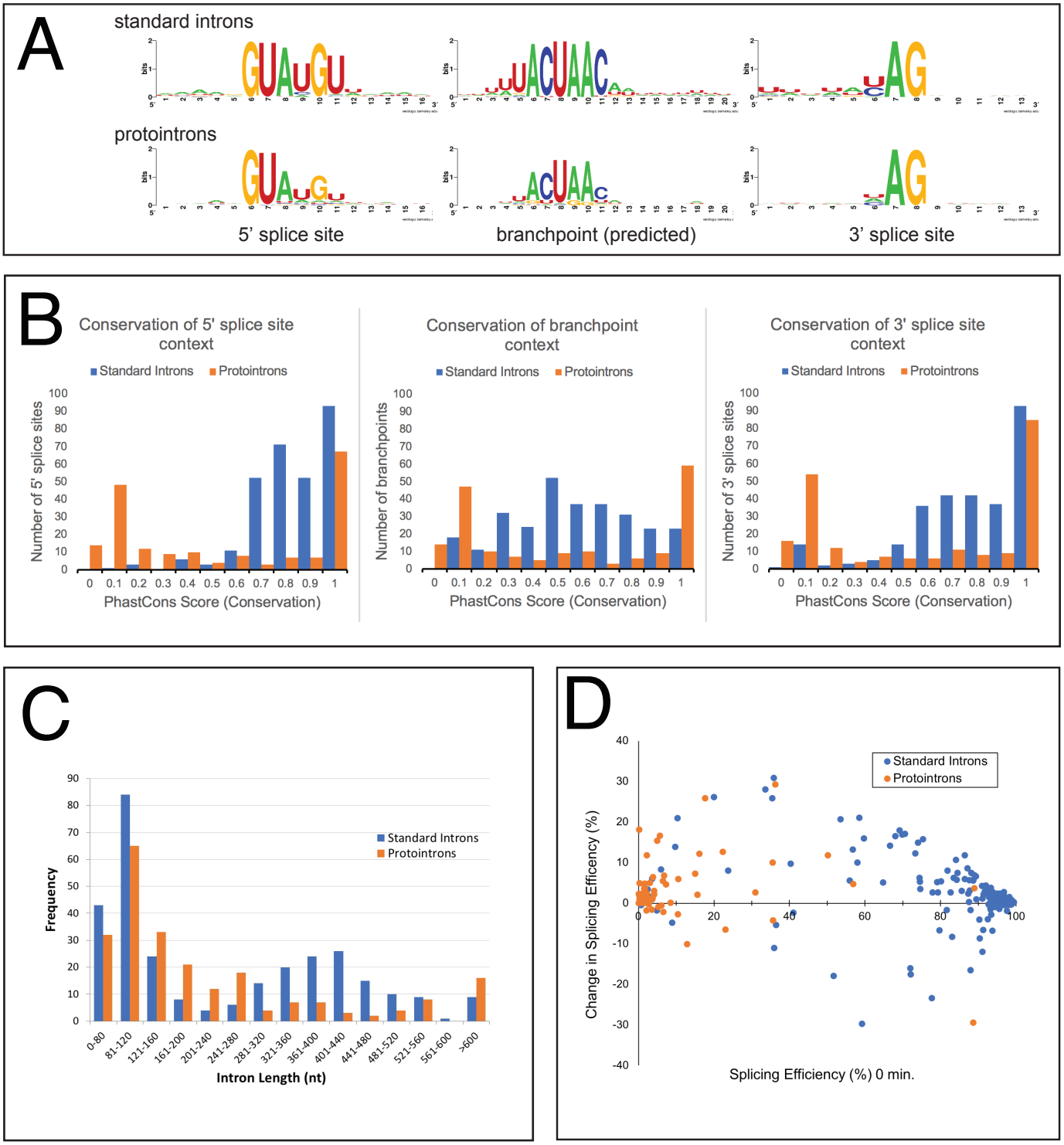




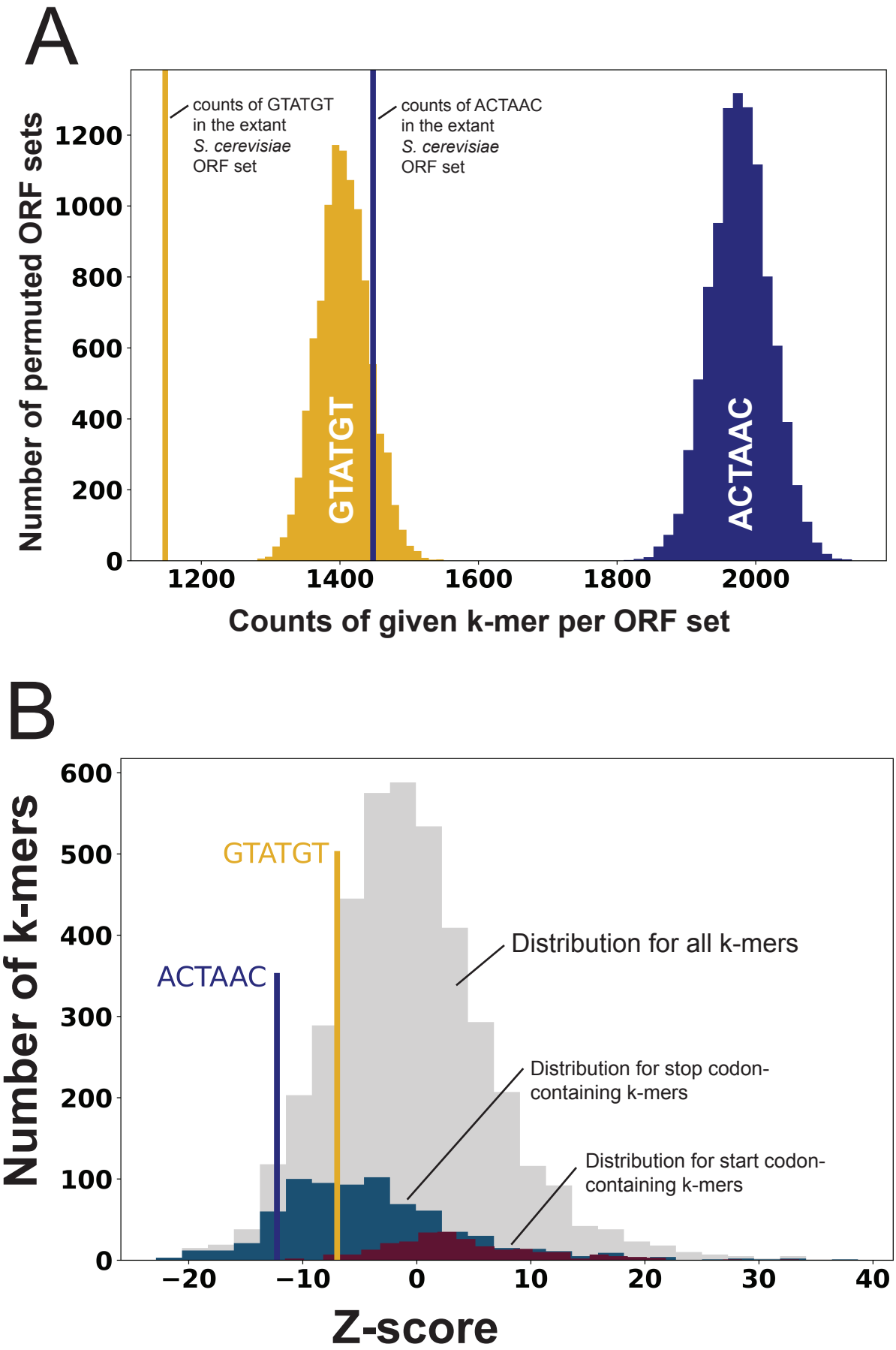




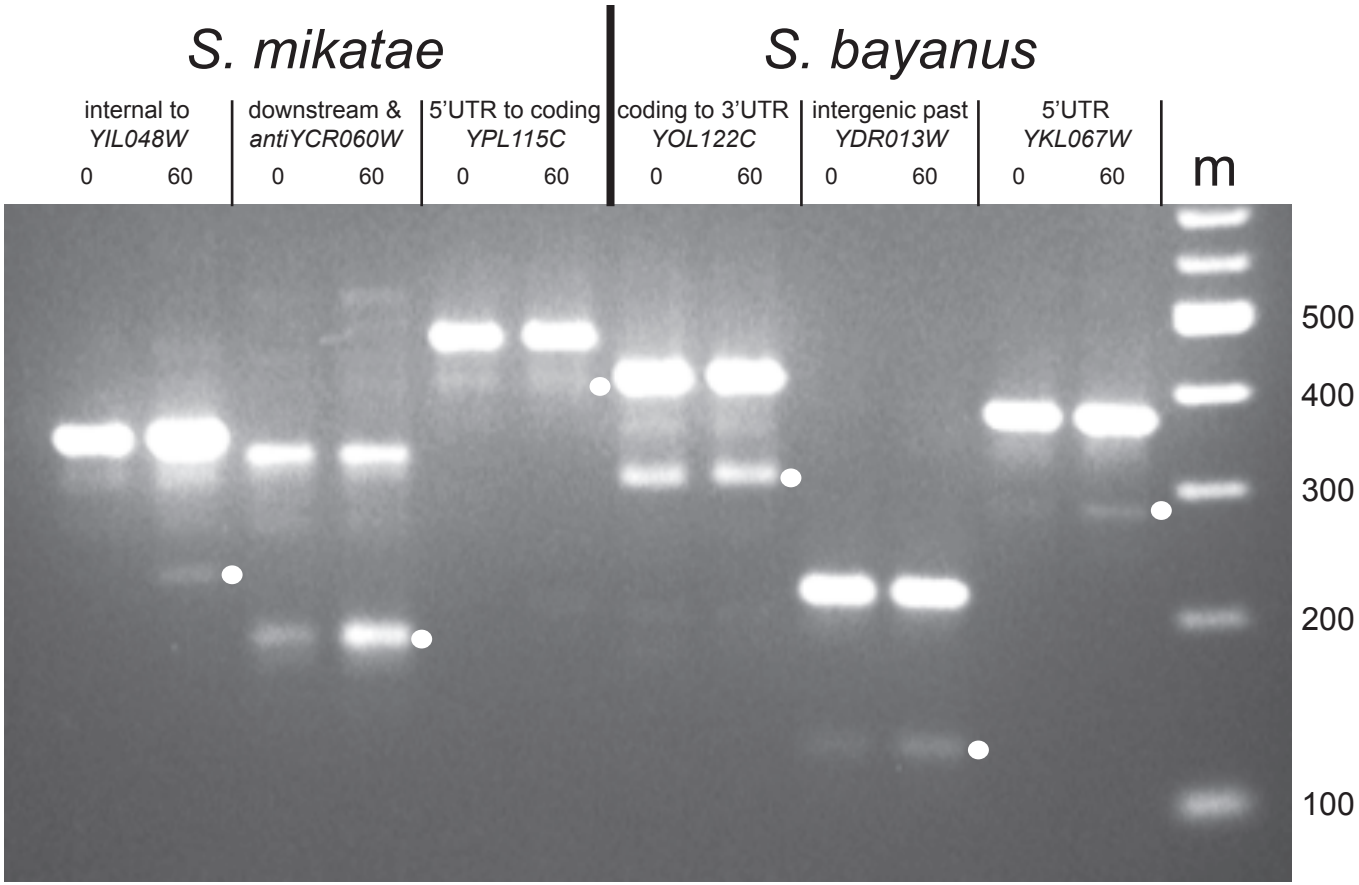
Talkish et al Fig 2



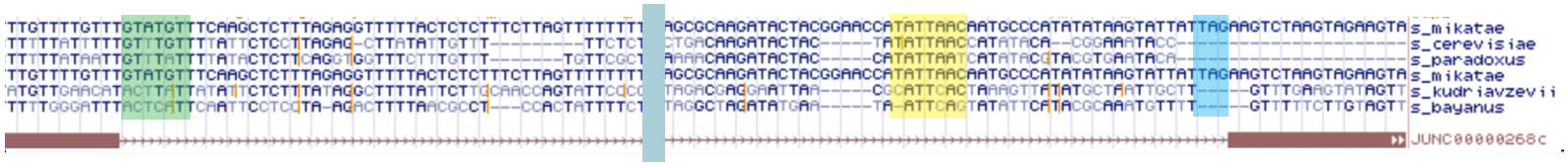
Talkish et al. Fig 3



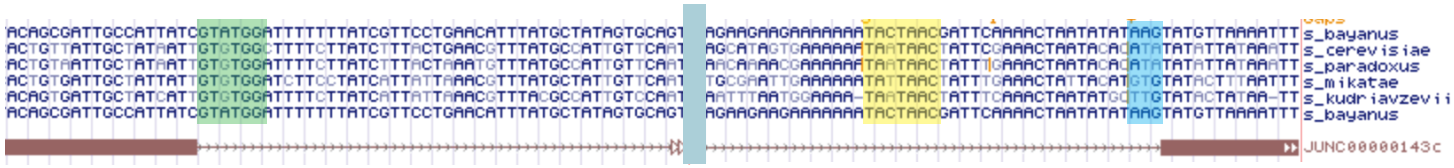
Talkish et al Fig 4



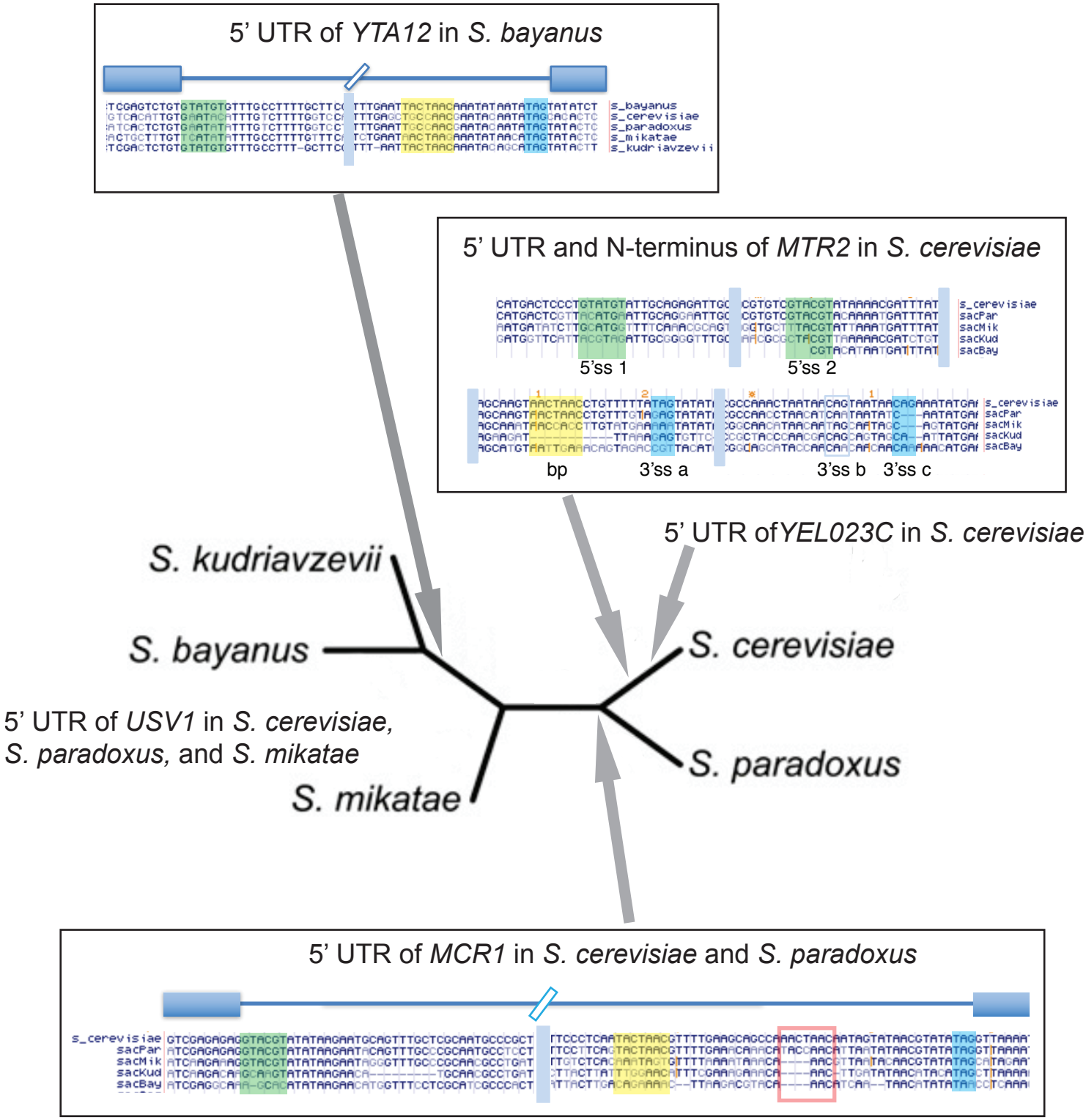
S. mikatae downstream & *antiYCR060W*



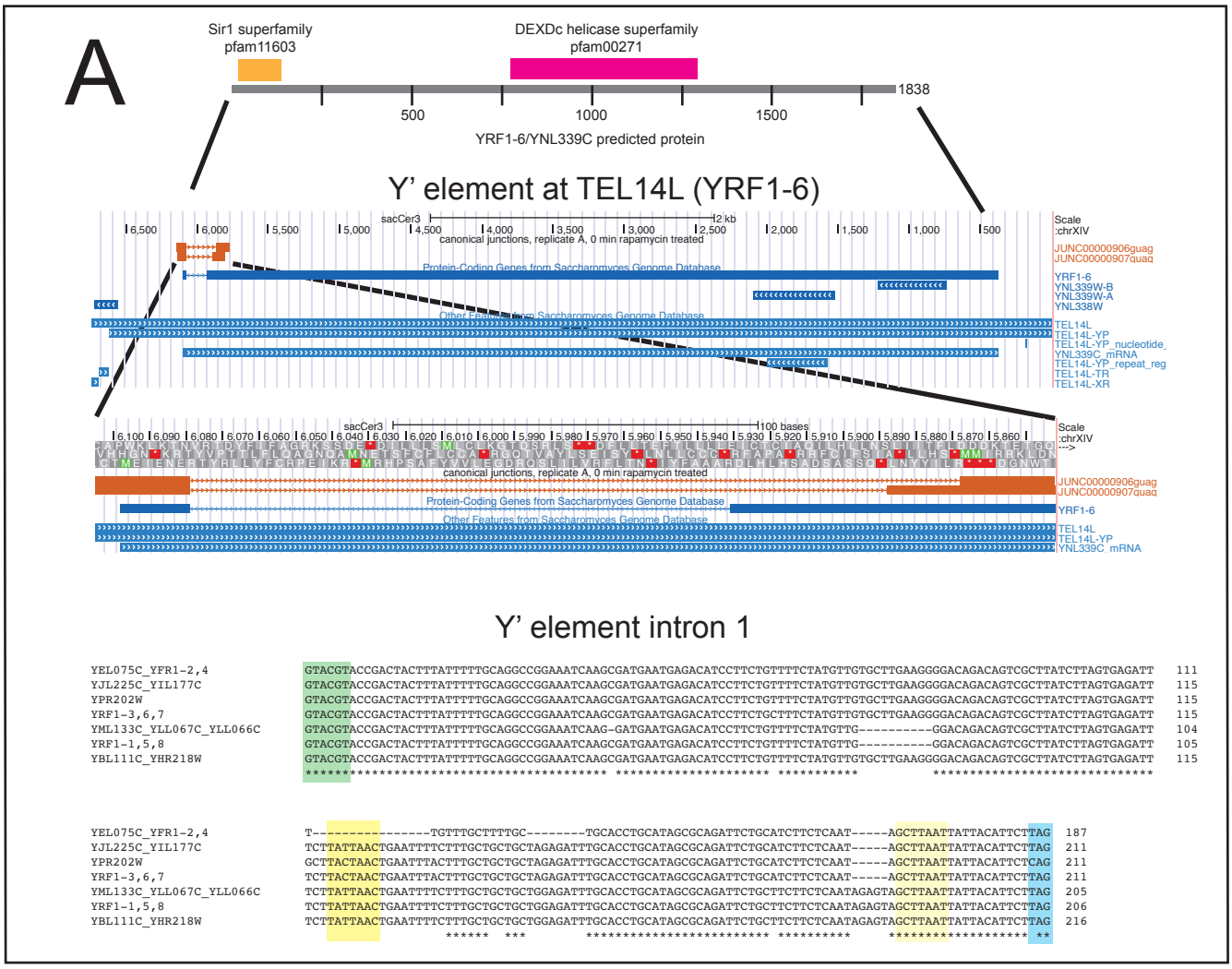
S. bayanus coding to 3'UTR *YOL122C*

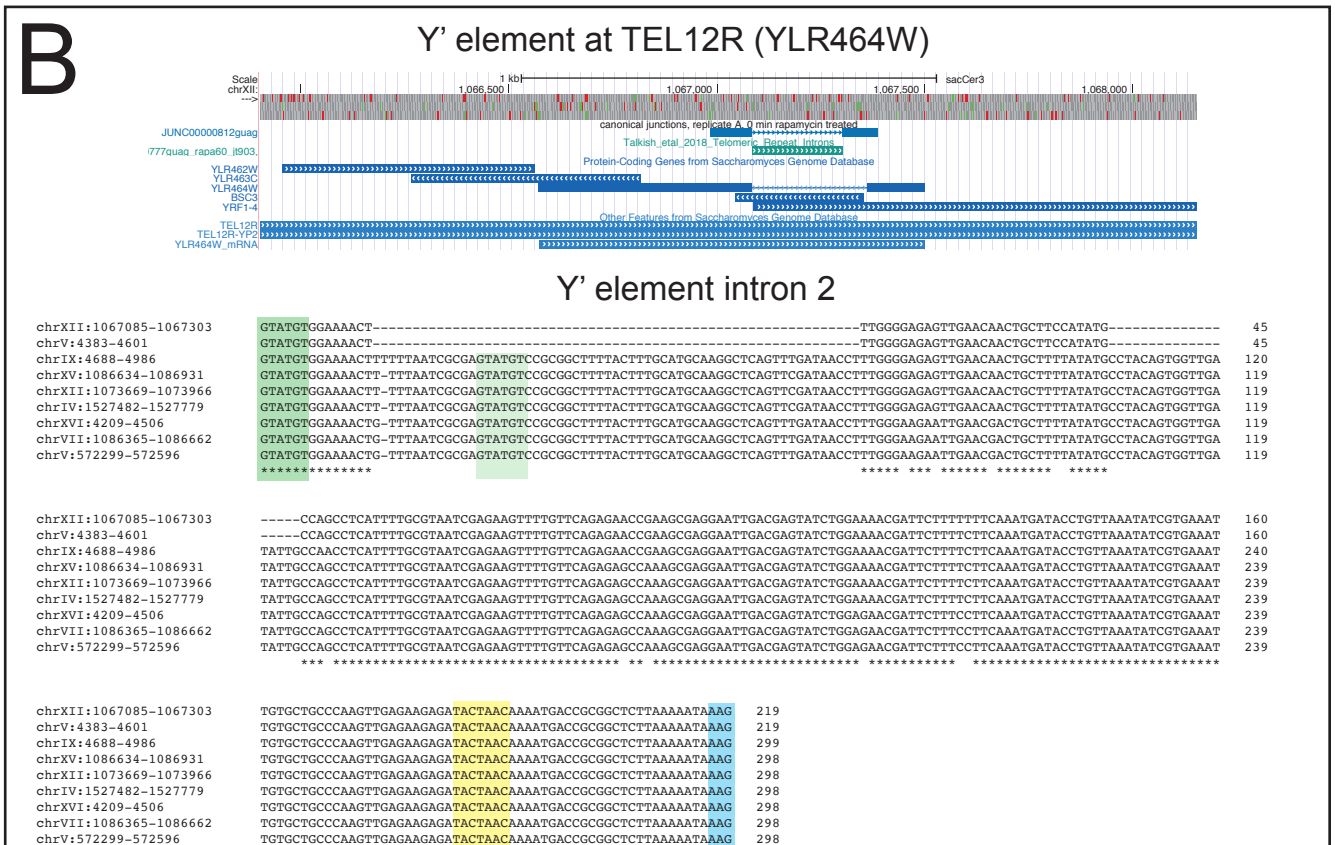


Talkish et al Fig 5

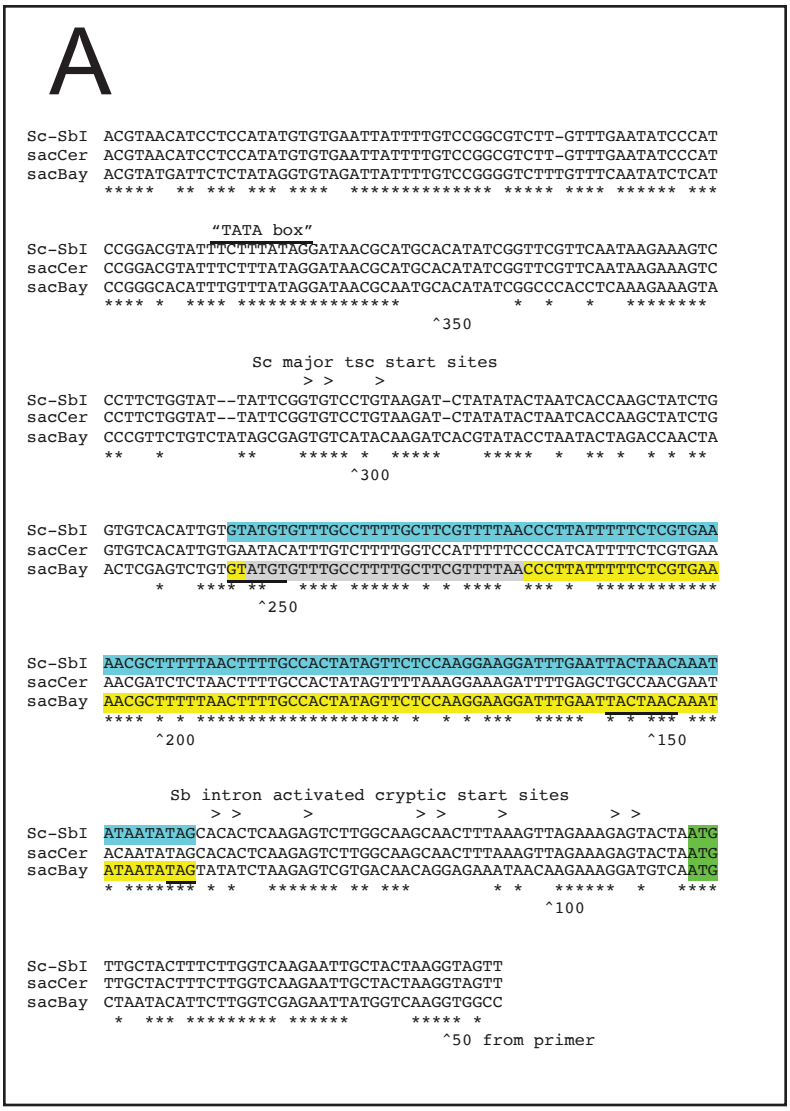


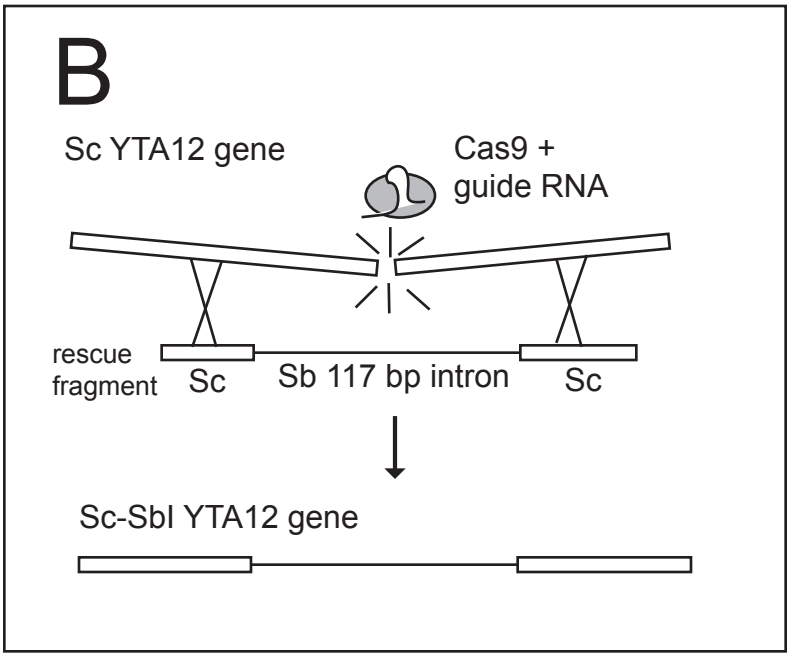
Talkish et al Fig 6

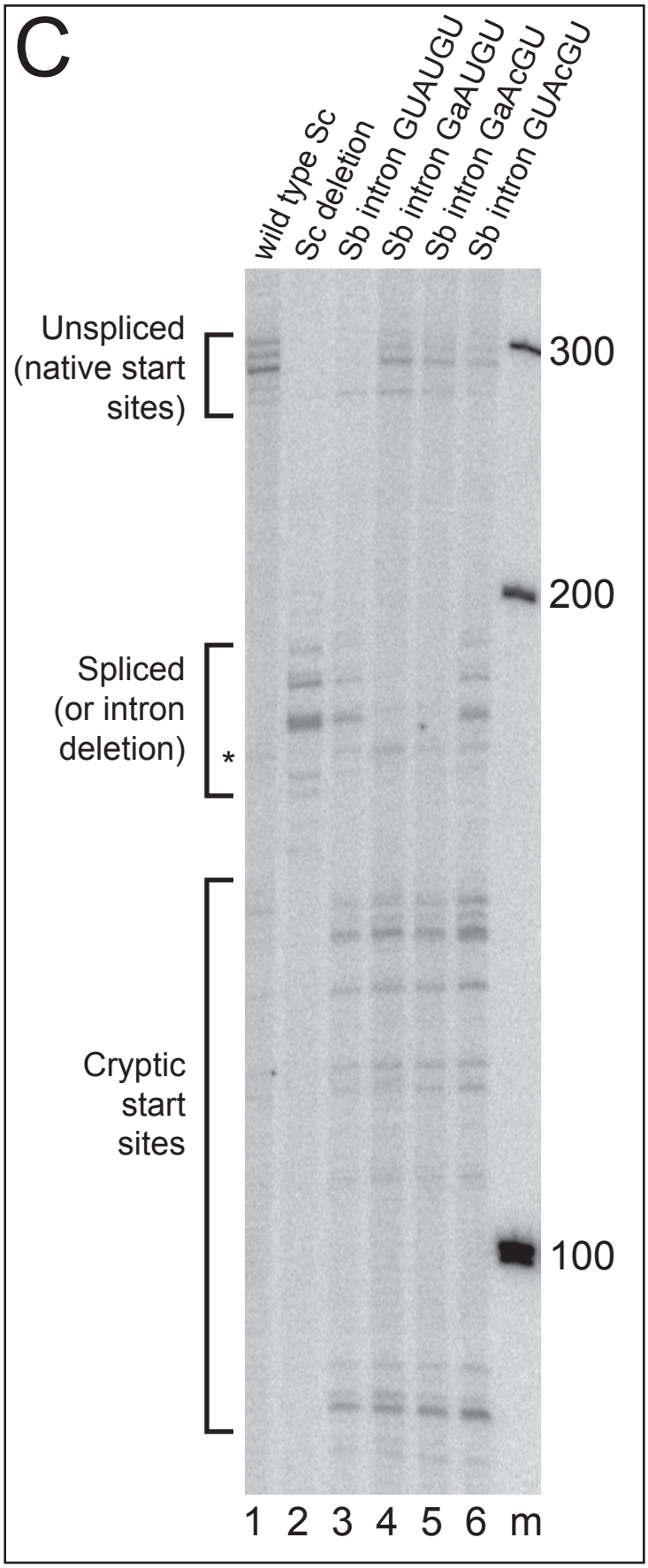


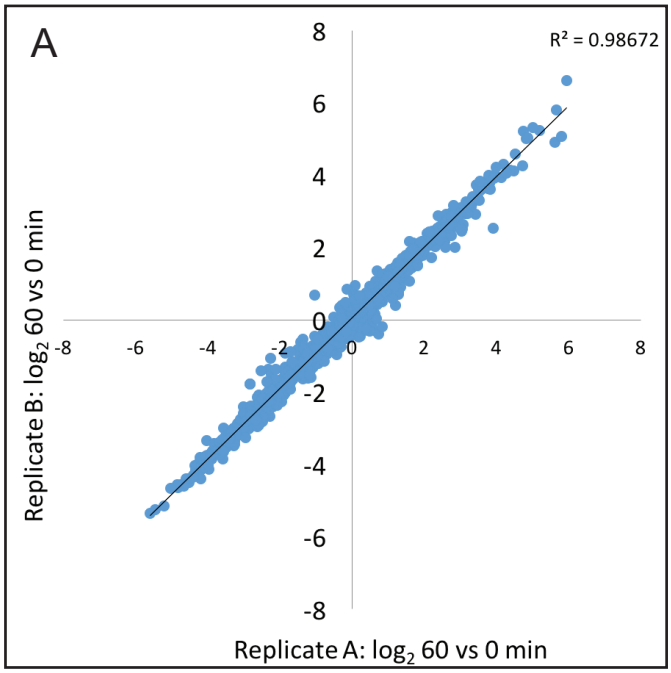


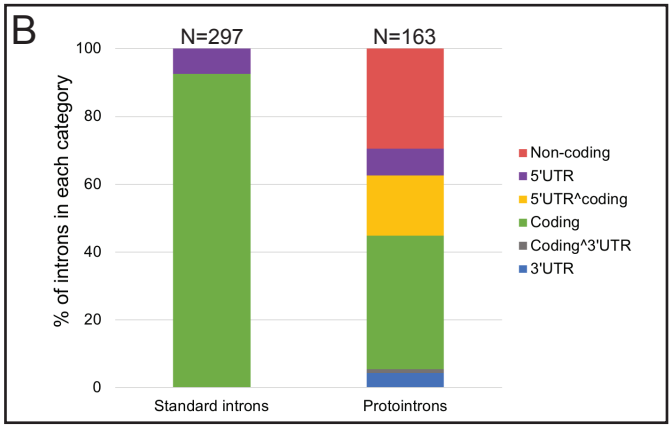
Talkish et al Fig 7

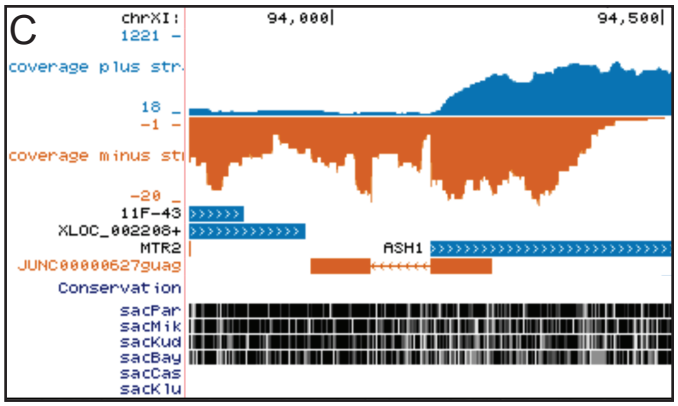


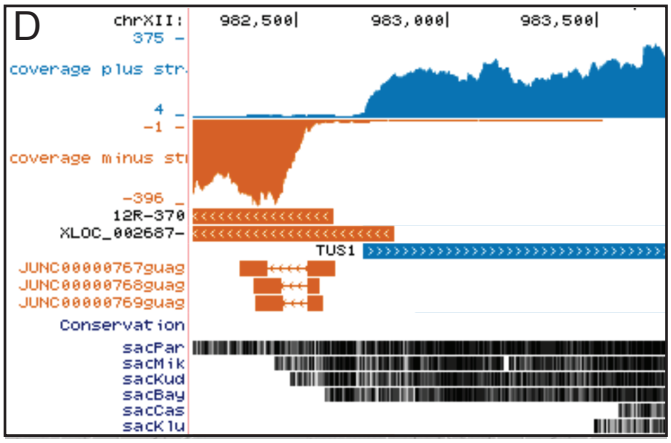


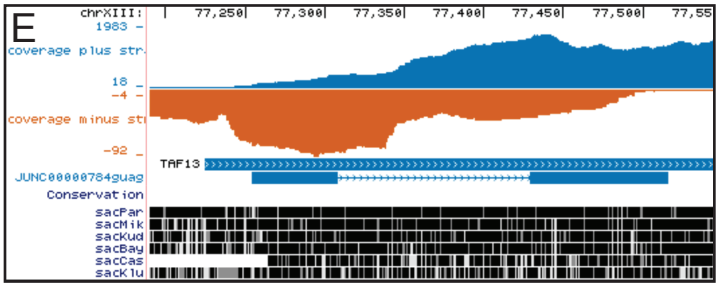


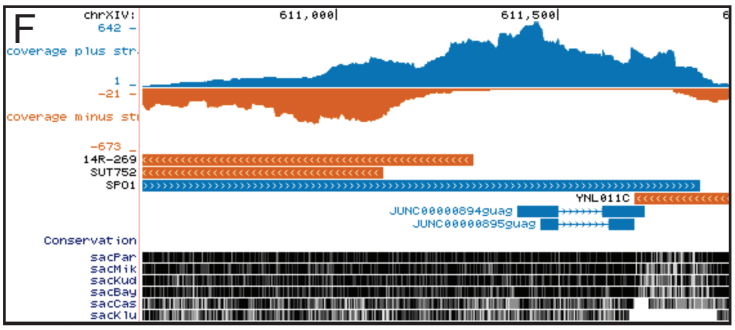

A
R² = 0.98672
Replicate B: log2 60 vs 0 min
Replicate A: log2 60 vs 0 min

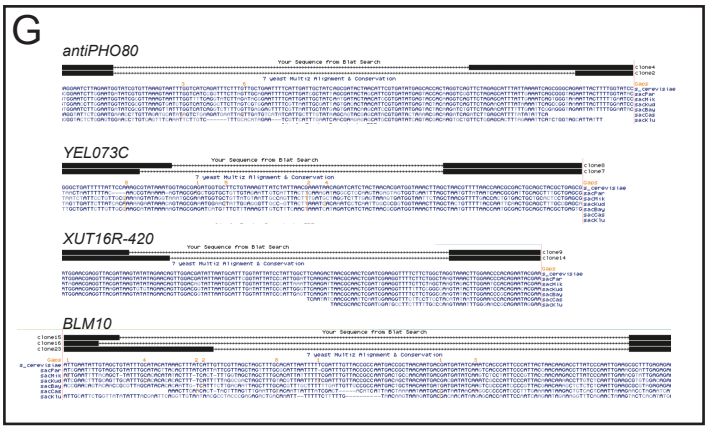

B
N=297
N=163
% of introns in each category
Non-coding
5'UTR
5'UTR^coding
Coding
Coding^3'UTR
3'UTR
Standard introns
Protointrons


C
chrXI:
94,000
94,500
coverage plus str
coverage minus str
11F-43
XLOC_002208+
MTR2
ASH1
JUNC00000627guag
Conservation
sacPar
sacMik
sackud
sacBay
saccas
sackLu


D
chrXII:
982,500
983,000
983,500
coverage plus str
coverage minus str
12R-370
XLOC_002687-
TUS1
JUNC00000767guag
JUNC00000768guag
JUNC00000769guag
Conservation
sacPar
sacMik
sackud
sacBay
saccas
sackLu


E
chrXIII:
77,250
77,300
77,350
77,400
77,450
77,500
coverage plus str
coverage minus str
TAF13
JUNC00000784guag
Conservation
sacPar
sacMik
sackud
sacBay
saccas
sackLu


F
chrXIV:
611,000
611,500
coverage plus str
coverage minus str
14R-269
SUT752
SPO1
YNL011C
JUNC00000894guag
JUNC00000895guag
Conservation
sacPar
sacMik
sackud
sacBay
saccas
sackLu


G
antiPHO80
YEL073C
XUT16R-420
BLM10