Recurrent insertion of 5'-terminal nucleotides and loss of the branchpoint motif in lineages of group II introns inserted in mitochondrial preribosomal RNAs

CHENG-FANG LI,^{1,2} MARIA COSTA,¹ GURMINDER BASSI,^{1,3} YIU-KAY LAI,² and FRANÇOIS MICHEL^{1,4}

¹Centre de Génétique Moléculaire du C.N.R.S., 91190 Gif-sur-Yvette, France

²Department of Life Science and Institute of Biotechnology, National Tsing Hua University, Hsinchu, Taiwan 30013

ABSTRACT

A survey of sequence databases revealed 10 instances of subgroup IIB1 mitochondrial ribosomal introns with 1 to 33 additional nucleotides inserted between the 5' exon and the consensus sequence at the intron 5' end. These 10 introns depart further from the IIB1 consensus in their predicted domain VI structure: In contrast to its basal helix and distal GNRA terminal loop, the middle part of domain VI is highly variable and lacks the bulging A that serves as the branchpoint in lariat formation. In vitro experiments using two closely related IIB1 members inserted at the same ribosomal RNA site in the basidiomycete fungi *Grifola frondosa* and *Pycnoporellus fulgens* revealed that both ribozymes are capable of efficient self-splicing. However, whereas the *Grifola* intron was excised predominantly as a lariat, the *Pycnoporellus* intron, which possesses six additional nucleotides at the 5' end, yielded only linear products, consistent with its predicted domain VI structure. Strikingly, all of the introns with 5' terminal insertions lack the EBS2 exon-binding site. Moreover, several of them are part of the small subset of group II introns that encode potentially functional homing endonucleases of the LAGLIDADG family rather than reverse transcriptases. Such coincidences suggest causal relationships between the shift to DNA-based mobility, the loss of one of the two ribozyme sites for binding the 5' exon, and the exclusive use of hydrolysis to initiate splicing.

Keywords: mitochondrial group II introns; linear intron; lariat intron branchpoint; homing endonucleases

INTRODUCTION

Bacterial group II introns result from the association of a reverse transcriptase (RT) gene with a large ribozyme: the latter catalyzes the branching and ligation reactions that result in an excised intron lariat and spliced exons (for review, see Lambowitz and Zimmerly 2004; Beauregard et al. 2008). Several lineages of these widely distributed prokaryotic retrotransposons found their way into the genomes of organelles and proliferated in diverse eukaryote clades. A majority of present-day group II members from organelles subsequently lost their RT component, but a number of individual introns have retained the potential to code for a protein and still behave as mobile elements (Kennell et al. 1993; Lazowska et al. 1994).

Even in the absence of an RT gene, identifying a group II intron in sequence data remains reasonably straightforward. Of the six secondary structure domains of the ribozyme, the

small domain V tends to be sufficiently conserved in terms of structure and sequence to lend itself to the design of an efficient search engine, with relatively few false negatives and positives (e.g., Griffiths-Jones et al. 2005; Lang et al. 2007). And once a candidate domain V has been spotted, it is generally feasible to use comparative sequence analysis and start building step-by-step the potential secondary structure of the rest of the ribozyme, all the way to the 3' and 5' splicing junctions (Michel et al. 1989). Although there exist a few noteworthy exceptions (Michel et al. 1989; Vogel and Börner 2002; Stabell et al. 2007), the 3' terminus normally lies 2–3 nt downstream from domain VI, while the latter carries on its 3' side a bulging adenine that serves as branchpoint for lariat formation (Fig. 1). On the other side of the intron, the last six or so nucleotides of the 5' exon take part in a long-range, intron–exon pairing, EBS1–IBS1 (Fig. 2A), whose ultimate base pair precedes the splice site. Moreover, the first five nucleotides of the intron tend to obey a characteristic consensus sequence, GUGYG, which is conserved as such in some 85% of known group II members (to the exclusion of the “degenerate” group II introns in the chloroplast genomes of euglenoids) (see Hallick et al. 1993); the actual extent of conservation of individual nucleotides

³Deceased.

⁴Corresponding author.

E-mail michel@cgm.cnrs-gif.fr.

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.2655911>.

FIGURE 1. Predicted structures of domain VI in 10 introns with a 5' terminal insertion. (dV) Domain V; (arrow) the 3' splice site; (boxed) well-conserved G:C pairs in the basal helix of domain VI, as well as its terminal loop, when it obeys the GNRA consensus. The structures are compared with the strongly conserved consensus structure and sequence of domain VI in 34 introns devoid of additional nucleotides at the 5' end (*inset*; the asterisk indicates the branchpoint; nucleotides and base pairs shown are at least 90% conserved; M: A or C; the highly divergent sequences of the *cox1* introns of *Paracoccidioides brasiliensis* and *Candida parapsilosis* [see Table 1] were not taken into account).

varies from 92% for position three to virtually 100% for position five (Michel et al. 2009).

While taking a census of established and candidate group II introns in organelle DNA sequences, our attention was brought to a small subset of introns that appeared to diverge somewhat from these rules. In these 10 group II members, the end of the IBS1 sequence and the GUGYG consensus sequence are separated from one another by one to as many as 33 intervening nucleotides (Fig. 2B; Table 1). Moreover, at the other intron end, the potential secondary structure of domain VI lacks a bulging A at the expected location for the branchpoint (Fig. 1). These introns, which happen to belong to the same ribozyme structural subgroup (IIB1) (Michel et al. 1989) and are inserted in ribosomal RNA precursor transcripts, exhibit additional remarkable features (Table 1): They all lack the EBS2–IBS2 pairing between the ribozyme and the 5' exon, which is potentially present in most group II introns, and several of them code for a homing endonuclease, rather than a reverse transcriptase (see also Michel and Ferat 1995; Toor and Zimmerly 2002; Monteiro-Vitorello et al. 2009).

We have cloned one of the members of this peculiar subset of intron sequences, as well as a closely related, but

apparently “normal” intron inserted at the same genomic site in another host, and we now show that the self-splicing reaction of the former (but not latter) molecule is initiated by hydrolysis, resulting in excision of the intron in linear form, rather than by transesterification, which generates a lariat structure (as is normal for group II introns) (for review, see Michel and Ferat 1995). More generally, we propose that the loss of the ability to form a branched structure should be regarded as an ultimate consequence of the recently documented (Mullineux et al. 2010) evolutionary conversion of some mitochondrial group II introns into DNA transposons (the class II mobile elements of Wicker et al. [2007] that move at the DNA level, contrary to retrotransposons that change location as RNA).

RESULTS

Distribution of 5'-terminal inserts in mitochondrial subgroup IIB1 introns

Subgroup IIB1 is widespread both in bacteria and organelles and includes two members whose ribozyme is used as a model system (*Saccharomyces cerevisiae* *cox1/5γ* and

FIGURE 2. Secondary structure models of (A) the *G. frondosa* SSU788 intron, (B) the *P. fulgens* SSU788 intron. (Boldface) Nucleotides in common between the two introns. (Arrowheads) Point to splice junctions; (asterisk) points to the branchpoint of the *Grifola* intron. Labeling of secondary structure components and tertiary interactions is as in Michel et al. (2009). In domain IV, nucleotides not shown were replaced by the sequence CTCGAG in “ORF-less” constructs.

Pylaiella littoralis LSU1787, also known as Pl.LSU/2). A list of published sequences of mitochondrial subgroup IIB1 members, which contains the 10 organelle group II introns we found to possess additional nucleotides at their very 5' extremity, is provided in Table 1 (several subgroup IIB1 members from land plants other than *Marchantia* are missing from this list; they were excluded from our analyses because of the likelihood of [partly] undocumented editing of intron nucleotides prior to splicing) (see Bonen 2008).

To ascertain that a candidate intron contained additional nucleotides at its 5' end, the location of splice junctions was first inferred by comparison with intron-less copies of the host gene in related organisms. We then looked for entries in which the 5'-terminal consensus sequence (GUGCGAC in the case of subgroup IIB1 introns) was separated from the predicted 5' splice site by one or more nucleotides. Finally, a complete secondary structure model was generated for each candidate ribozyme (Fig. 2B; Supplemental Data Set), and the EBS1 terminal loop in domain I was verified to base-pair with the last nucleotides of the inferred 5' exon, rather than with the sequence preceding the 5'-terminal consensus.

An alternative interpretation to the existence of an insert at the intron 5' end could be that the additional nucleotides are not removed during splicing. However, all 10 candidate introns happen to be in ribosomal RNA genes, and their inferred sites of insertion (SSU788, LSU2059, LSU2449, and LSU2586) (see Table 1) lie within segments of sequence that are extremely conserved and most unlikely to tolerate insertions in mature, functional molecules (the three LSU segments are either part of, or lie next to, the catalytic site for peptide synthesis, while the SSU site is part of the loop that separates the tRNA P and E sites) (Nissen et al. 2000; Schuwirth et al. 2005). As for the possibility that the sequence under scrutiny was that of a pseudogene, it can be ruled out for entries that correspond to completely sequenced genomes (the three placozoan sequences and those of *Amoebidium*, *Rhizophyidium*, and *Trametes*) and have a single, intron-containing copy of the ribosomal RNA gene of interest.

The 10 introns with a 5'-terminal insert also stand out in that the sequence and predicted secondary structure of ribozyme domain VI is strikingly variable, even among closely related introns, and departs in multiple ways from the consensus domain VI structure shared by all other mitochondrial members of the IIB1 subgroup (Fig. 1). Not only is the branchpoint adenine missing at its expected location, but the well-conserved 3-bp helix and (GAA:CUA) internal loop immediately distal of it are unrecognizable. This is all the more striking since the

TABLE 1. A list of mitochondrial subgroup IIB1 introns

Organism, gene, and intron	Accession number	Intron coordinates ^b	5' insert ^c	EBS2	ORF product ^d and location ^j
Fungi					
<i>Paracoccidioides brasiliensis</i> cox1/2 ^e	AY955840	41,071–43,890			RT (IV)
<i>Candida parapsilosis</i> cox1 ^e	X74411	12,690–15,605			RT (IV)
<i>Saccharomyces cerevisiae</i> cox1/5γ	V00694	8746–9632*			
<i>Saccharomyces pastorianus</i> cox1/5	EU852811	53,565–54,476			
<i>S. cerevisiae</i> cytb/1	EU004203	38,472–39,239			
<i>Candida zemplinina</i> LSU2584 ^f	AY445918	2516–324*			RT (IV)
<i>Candida ipomoeae</i> SSU531	AY393889	176–801			
<i>Glomus intraradices</i> C8.3b_18 LSU1787	AM950206	1449–2257			
<i>G. intraradices</i> C16g1_2 LSU1787	AM950209	2541–3896			LAGLIDADG (IV)
Uncultured <i>Glomus</i> W9/1 LSU1787	FN377588	1827–3215			LAGLIDADG (IV)
<i>Allomyces macrogynus</i> LSU2059	U41288	2416–3192			
<i>Rhizophydium</i> sp. 136 LSU2059	NC_003053	3880–4564	+6 GUGCGAC	No	
<i>Pichia angusta</i> LSU2059	AL432964	625–1; 469–1; 294–879*		No	LAGLIDADG (IV)
	AL434946				
	AL433470				
<i>Ustilago maydis</i> SRX2 LSU2059	EU921807	3413–5372			LAGLIDADG (IV)
<i>Agroclype aegerita</i> LSU2059	AF087656	9088–10,871*	+23 UUGCGAC	No	LAGLIDADG ^k (IV)
<i>Suillus luteus</i> LSU2059	L47586	2675–3341	+25 UAGCGAC	No	
<i>Cryphonectria parasitica</i> SSU952	AF029891	7168–9235			LAGLIDADG (III)
<i>Leptographium truncatum</i> 1435 SSU952	GU949593	800–2639		No	LAGLIDADG (III)
<i>Cordyceps konnoana</i> SSU952	AB031194	897–2724*			LAGLIDADG (III)
<i>Paracoccidioides brasiliensis</i> SSU952	AY955840	25,574–27,362			Unidentified (III)
<i>Aleurodiscus botryosus</i> SSU788	FR773980 ^g			No	LAGLIDADG (IV)
<i>Ceriporiopsis subvermisporea</i> SSU788	EU546103	345–907*		No	
<i>Grifola frondosa</i> SSU788	FR773978 ^g			No	LAGLIDADG (IV)
<i>Pycnoporellus fulgens</i> SSU788	FR773979 ^g		+6 UUGCGAC	No	LAGLIDADG (IV)
<i>Ganoderma lucidum</i> SSU788	AF214475 ^h	1056–2562	+6 UUGCGAC	No	LAGLIDADG (IV)
<i>Trametes cingulata</i> SSU788	GU723273	39,037–40,442*	+6 AUGCGAC	No	LAGLIDADG ^k (IV)
<i>Usnea antarctica</i> SSU788	DQ990920	397–1473*			
<i>C. parasitica</i> SSU788	AF029891	2415–4596*		No	LAGLIDADG (IV)
Ichthyosporea					
<i>Amoebidium parasiticum</i> SSU788	AF538044	855–2198		No	GIY-YIG (IV)
<i>A. parasiticum</i> LSU2449	AF538042	5337–5909	+33 GAGCGAC	No	
Plants^a					
<i>Chaetosphaeridium globosum</i> trnG(UCC)	AF494279	16,893–16,066*		No	
<i>Chlorokybus atmophyticus</i> trnS(GCU)	NC_009630	134,254–136,307			
<i>Marchantia polymorpha</i> trnS(GCU)	NC_001660	48,902–49,892			
<i>M. polymorpha</i> LSU787	NC_001660	149,541–148,824			
<i>M. polymorpha</i> cox3/2	NC_001660	87,980–88,911			
<i>M. polymorpha</i> cox2/1	M68929	81,191–82,268			
<i>Scenedesmus obliquus</i> LSU2455	NC_002254	28,831–29,438			
<i>S. obliquus</i> SSU968	NC_002254	40,780–41,627*			
<i>Pedinomonas minor</i> LSU1787	AF116775	1198–2027			
Stramenopiles					
<i>Pylaiella littoralis</i> LSU575	Z48620	543–2952			RT (IV)
<i>P. littoralis</i> LSU1787	Z48620	4052–6489			RT (IV)
<i>P. littoralis</i> LSU2451	Z48620 ⁱ	7134–8304			

(continued)

TABLE 1. Continued

Organism, gene, and intron	Accession number	Intron coordinates ^b	5' insert ^c	EBS2	ORF product ^d and location ^j
Animals					
<i>Trichoplax adhaerens</i> LSU2586	NC_008151	19,460–20,691*	+1 GUGCGAC	No	
Placozoan sp. BZ2423 LSU2586	NC_008834	16,339–17,068*	+1 GUGCGAC	No	
Placozoan sp. BZ10101 LSU2586	NC_008832	17,186–17,990*	+1 GUGCGAC	No	
Placozoan sp. BZ10101 LSU1787	NC_008832	15,476–16,215			

^aIntrons from land plants other than *Marchantia* were excluded because of the likelihood of RNA editing.

^bIntron coordinates were inferred by comparison with closely related uninterrupted gene sequences; (asterisks) introns the coordinates of which differ from those indicated in the GenBank entry or were missing.

^cFor introns with insertions at the 5' end, the sequence that best matches the IIB1 5' terminal consensus is provided, together with the number of nucleotides inserted in front of it.

^d(RT) Reverse transcriptase; (LAGLIDADG and GIY . . YIG) putative homing endonucleases belonging, respectively, to these families.

^eThese introns are inserted at homologous sites in the *cox1* gene, their sequences are only distantly related to the rest of the IIB1 set, and they were used as the outgroup in Figure 5.

^fFollowing Johansen and Haugen (2001), introns in ribosomal RNA genes are designated by the name of the ribosomal subunit followed by the coordinate of the nucleotide preceding the insertion site (according to the *Escherichia coli* numbering scheme).

^gThis study.

^hThe reference for this entry is Hong et al. (2002).

ⁱThe sequence of domain VI was from entry AB281597.

^jRibozyme secondary structure domain within which the ORF is located.

^kPresence of in-frame stop codons and frameshifts.

base of the domain VI stem tends to be well-conserved—it begins with a G:C pair in nine out of 10 introns—and seven out of 10 sequences share a GNRA loop at the tip of the domain; in related introns, that loop participates in the η - η' tertiary interaction (Chanfreau and Jacquier 1996; Costa et al. 1997a) between domains II and VI (Fig. 2).

Cloning and sequence analysis of the *Grifola frondosa* and *Pycnoporellus fulgens* SSU788 introns

Of the 10 intron sequences with 5' terminal insertions in Table 1, that of the *P. fulgens* SSU788 intron (GenBank entry AF518690) was incomplete. We chose to clone and sequence this intron and its flanking exons, as well as two partially sequenced, insert-lacking, related SSU788 introns in the basidiomycete fungi *G. frondosa* and *Aleurodiscus botryosus* (accession numbers AF334880 and AF026646).

As shown in Figure 2, the predicted secondary structure models of the *Grifola* and *Pycnoporellus* ribozymes are very similar, and the same is true of the *Aleurodiscus* ribozyme (Supplemental Fig. S1). As expected, the identity of nucleotides at sites known to participate in intra- or inter-domain, long-range tertiary interactions (Toor et al. 2008a; Michel et al. 2009; Pyle 2010) is especially well conserved. The only exception is the δ - δ' Watson-Crick base pair, which contributes to the stability of the EBS1-IBS1 intron-exon pairing (Costa et al. 2000): the U:A δ - δ' pair of the *Grifola* intron is replaced by G:A in the *Pycnoporellus* molecule (the closely related *Ganoderma lucidum* and *Trametes cingulata* introns have A:A at these sites, whereas other 5'-insert-

bearing introns, and also the *A. botryosus* molecule, have diverse Watson-Crick base pairs) (data not shown). Also very well-conserved is domain III, which contributes to the efficiency of catalysis (Fedorova and Pyle 2005).

A striking feature that the two secondary structure models have in common is the lack of EBS2-IBS2, an extended canonical pairing that involves nucleotides upstream of IBS1, on the one hand, and a single-stranded loop in the distal section of subdomain ID, on the other. The EBS2-IBS2 pairing is present in a majority of group II introns, with the exception of members of subgroup IIC, whose 5' exon displays a hairpin structure at the expected location for the IBS2 sequence (Granlund et al. 2001; Quiroga et al. 2008). What has been lost, in fact, is not only the EBS2 loop, but an entire subdomain that, in subgroup IIB, branches off the 5' strand of the stem connecting the internal loops that contain the EBS3 and α' nucleotides. This subdomain carries, in addition, a sequence that, in many introns, potentially participates in the β - β' long-range interaction with subdomain IC2 (Michel et al. 1989). Interestingly (Table 1), the EBS2 loop and associated subdomain are missing from all 10 introns with 5'-terminal inserts and also all other known SSU788 introns with the exception of *Usnea antarctica* (data not shown).

Subdomains that are known (Toor et al. 2008a) or suspected (Pyle 2010) to lie at the surface of the ribozyme three-dimensional structure tend to be the most variable ones. This is especially true of domain IV, only the first three base pairs of which are conserved between the *Grifola* and *Pycnoporellus* sequences. Still, the contents of domain IV are similar in the two introns (and in the *A. botryosus*

SSU788 intron), consisting primarily of open reading frames (ORFs; 260 codons in *Grifola* and 266 in *Pycnoporellus*) that potentially encode related (38% identical at the amino acid level) members of the LAGLIDADG family of DNA double-stranded homing endonucleases (Stoddard 2005). As seen in fact in Table 1, four out of the other seven published sequences of SSU788 introns contain coding sequences for additional LAGLIDADG homing endonucleases (the gene appears defective in *T. cingulata*), while a fifth one (in *Amoebidium parasiticum*) potentially encodes a GIY-YIG protein, the second most common family of homing endonucleases in mitochondrial genomes.

Contrasting self-splicing products of the *Grifola* and *Pycnoporellus* SSU788 introns

The lack of a group II branchpoint structure in domain VI of the *Pycnoporellus* SSU788 intron suggested that splicing was

initiated by hydrolysis at the 5' splice junction, rather than by transesterification (Jarrell et al. 1988; Jacquier and Jacquesson-Breuleux 1991; Daniels et al. 1996; Podar et al. 1998; Vogel and Börner 2002). This was confirmed by incubating precursor transcripts containing the *Grifola* and *Pycnoporellus* SSU788 introns under conditions that allow in vitro self-splicing.

In vitro self-splicing of the *Grifola* SSU788 intron (Fig. 3) is reasonably efficient at 42°C in 1 M NH₄Cl and at a moderately high magnesium concentration (20 mM). As reported for other group II introns (Daniels et al. 1996; Costa et al. 1997a,b), reaction of precursor molecules is a kinetically complex process, converting only about half of the material to products in ~2 min and the rest much more slowly if at all (Fig. 3B). The distribution of splicing products is also typical of most group II introns, being dominated by the lariat intron and ligated exons (Fig. 3A), the identity of which was verified by gel extraction followed by reverse transcription (Fig. 4A,C). Only small amounts of

FIGURE 3. Self-splicing of the *Grifola* and *Pycnoporellus* SSU788 introns. (A) Time course of self-splicing reactions at 42°C in 1 M NH₄Cl, 20 mM MgCl₂, 40 mM Na-MES (pH 6.2). Products were identified based on (1) reverse transcription of gel-extracted molecules (see Fig. 4) and (2) their electrophoretic mobility, compared to that of known splicing products of a *P. littoralis* LSU1787 (Table 1; Costa et al. 1997b) precursor transcript (MW lane: band 1, 640 nt, lariat; band 2, 872 nt, precursor; band 3, 640 nt, linear intron; band 4, 232 nt, ligated exons). (B) Time course of self-splicing reactions of a *Grifola* SSU788 precursor RNA at 42°C in 40 mM Tris-Cl (pH 7.5 at 25°C), 20 mM MgCl₂, and 1 M NH₄Cl (circles and solid curve, generated by a biphasic exponential fit with $k_1 = 0.9 \pm 0.2 \text{ min}^{-1}$ and $k_2 = 0.03 \text{ min}^{-1}$) (see Materials and Methods) or 1 M KCl (squares and dashed curve, lariat intron; lozenges and dotted curve, linear intron; both from single exponential fits). (C) Time course of self-splicing reactions of a *Pycnoporellus* SSU788 precursor RNA in 40 mM Tris-Cl (pH 7.5 at 25°C), 1 M NH₄Cl, and 10 mM MgCl₂ (empty squares), 20 mM MgCl₂ (empty circles), 50 mM MgCl₂ (empty lozenges), or in 40 mM Na-MES (pH 6.2) and 20 mM MgCl₂ (filled circles and dashed curve). Reactions at 10 and 20 mM Mg (pH 7.5) were fitted to a biphasic process ($k_1 = 0.32 \pm 0.03 \text{ min}^{-1}$, $k_2 = 0.030 \pm 0.016 \text{ min}^{-1}$), the other ones to simple exponentials.

FIGURE 4. Mapping of intron–exon junctions and the branch site. Sequencing lanes are labeled by the base complementary to the dideoxynucleotide added. (A) Sequencing by reverse transcription of gel-extracted ligated exons; (left panel) *Pycnoporellus*; (right panel) *Grifola*. (Arrows) Splicing junctions. (B) Mapping of the 5' extremity of gel-extracted linear intron molecules generated by in vitro self-splicing of a *Pycnoporellus* precursor transcript; the latter was used as a template to generate the sequencing lanes at right with a primer located downstream from the intron 5' extremity. Elongation from the same primer using the excised intron molecules as template generated the strong stop in the lane at left; (arrow) the 5' splice site. (C) Mapping of the branchpoint and 5' extremity of gel-extracted lariat intron molecules generated by in vitro self-splicing of a *Grifola* precursor transcript. (Left panel) Elongation from a primer located downstream from the intron 5' extremity; the stop (marked by an asterisk) corresponds to the first intron nucleotide; sequencing lanes (at right) were generated by the same primer on a precursor RNA template. (Right panel) elongation from a primer located in the 3' exon (intron–3' exon branched molecules were used as template); (asterisk) the branch site (elongation stops on the nucleotide immediately 3' of the branch site); sequencing lanes were generated by the same primer on a precursor RNA template.

a linear intron form were observed, unless ammonium ions were replaced by potassium ions (Jarrell et al. 1988). Even then, the final molar fraction of linear intron molecules—presumably generated by hydrolysis at the 5' splice site—did not exceed 15% of intron-containing products (Fig. 3B).

Under the conditions used for the *Grifola* intron, self-splicing of the *Pycnoporellus* SSU788 intron is also a rather rapid process (Fig. 3C, solid curve), with ~80% of precursor molecules converted to products in 10 min. However, the lariat intron is absent from reaction products, which consist primarily of the ligated exons and a linear intron form. The 5' extremity of the latter was verified by reverse transcription (Fig. 4B) to coincide with the 5' splice site, as determined by alignment with uninterrupted versions of the host gene, on the one hand, and sequencing of the ligated exons (Fig. 4A), on the other. Additional products include small amounts of molecules with the expected electrophoretic mobility of the linear intron–3' exon splicing intermediate and a molecule of ~550 nt, which could have been generated by ribozyme-catalyzed, hydrolytic cleavage of the linear intron at position 110, 3' of the sequence AGGAC. The latter offers a better match to EBS1 (GUCCU) than the IBS1 sequence (AGGAU) at the 3' end of the 5' exon (see Fig. 2).

Varying the concentration of magnesium (Fig. 3C) did not make it possible to observe lariat molecules among self-splicing products of the *Pycnoporellus* intron but confirmed that the optimal magnesium concentration in terms of

reaction rate and final extent of reaction is ~10–20 mM (Fig. 3C). On the other hand, lowering the pH from 7.0 to 6.2 did lead to a substantial reduction of the reaction rate (Fig. 3C), suggesting that catalysis is at least partly rate-limiting when splicing is initiated by hydrolysis.

Phylogenetic relationships of mitochondrial subgroup IIB1 introns with and without a 5' terminal insert

To generate a phylogenetic tree of mitochondrial subgroup IIB1 introns, their ribozyme sequences were aligned over shared components of the subgroup IIB1 secondary structure (see Materials and Methods; Supplemental Data Set). The number of sites that can be unambiguously aligned (526) is too small to resolve the complete phylogenetic relationships of all members of this subgroup (Fig. 5). Nevertheless, bootstrap analysis indicates that introns inserted at the same ribosomal site

tend to form well-supported clades, consistent with a common origin. The only exception comes from introns inserted at position 1787 of the large ribosomal RNA: neither the four available sequences, from *P. littoralis*, *Pedinomonas minor*, *Glomus intraradices*, and *Placozoon* sp. BZ10101 (Fig. 5; Table 1), nor the corresponding secondary structure models (data not shown) reveal any particularly close similarity.

Provided it is assumed that the insertion of nucleotides at the 5' intron extremity and the accompanying loss of the branchpoint structure are irreversible events, the minimal number of occurrences that gave birth to lineages of introns with 5'-terminal inserts may be estimated from the phylogeny proposed in Figure 5. The most parsimonious interpretation of the data implies at least four founding insertion events, and a fifth one would become necessary should the hypothetical relationship of the *A. parasiticum* LSU2449 intron with the *Placozoon* LSU2586 introns prove nonsignificant (in that case, two events would need to be postulated at LSU2059 and one at each of the other three ribosomal RNA sites occupied by introns with 5'-terminal inserts).

Interestingly, the presence in mitochondrial members of subgroup IIB1 of intron-contained homing endonuclease ORFs results as well from multiple, independent acquisition events (Supplemental Fig. S2). As already pointed out by Monteiro-Vitorello et al. (2009), the proteins potentially encoded by the SSU788 and SSU952 introns of *Cryptosporidium parvum* are not closely related. More generally, whereas

FIGURE 5. Phylogenetic relationships of mitochondrial subgroup IIB1 introns based on aligned ribozyme sequences (see Materials and Methods and Supplemental Data Set). Introns are designated as in Table 1; the *cox1* introns from *P. brasiliensis* and *C. parapsilosis*, the sequences of which are markedly divergent from the rest, were used as outgroups. Numbers next to nodes are bootstrap proportions (200 replicates) $\geq 75\%$ (corresponding branches are thickened). The roots of well-supported, major clades of ribosomal introns are indicated. The length of the 5' terminal insertion, when present, is provided to the right of an intron name (boxed numbers preceded by + sign). RT, LAGLIDADG, GIY . . YIG, and "Unknown" designate proteins potentially encoded by the introns (see Table 1; note that only some versions of the *G. intraradices* LSU1787 intron include an ORF).

LAGLIDADG proteins encoded by introns inserted at the same ribosomal site tend to be rather similar—they form monophyletic groups—and may have coevolved with their intron host, introns located at different sites encode proteins that belong to separate lineages within the LAGLIDADG phylogenetic tree (Supplemental Fig. S2). Note also that in contrast to the ORFs located in introns inserted at positions LSU2059, SSU788, and SSU952, which contain two LAGLIDADG motifs, the much shorter *Glomus* LSU1787 intron ORFs (200 and 208 codons; accession numbers AM950209 and FN377588, respectively) contain a single LAGLIDADG element, so that the corresponding homing endonuclease must be a homodimer, rather than a monomer (see Stoddard 2005).

The following facts provide further evidence that mitochondrial subgroup IIB1 introns acquired ORFs for proteins other than reverse transcriptases through independent insertion events: (1) in introns inserted at the SSU952 site, the ORF is inserted in ribozyme domain III, rather than in

domain IV (Table 1; Mullineux et al. 2010); (2) the ORF in the *A. parasiticum* SSU788 intron encodes a protein belonging to the GIY-YIG family of homing endonucleases, rather than a member of the LAGLIDADG family; (3) the proteins possibly encoded by the LSU2059 introns are closely related (Supplemental Fig. S2) to the protein specified by a group I intron inserted at position 2066 of the mitochondrial LSU gene of *Tuber melanosporum*, just 7 nt 3' of position 2059.

The latter observation is obviously in keeping with a model (Bonocora and Shub 2009) in which mobilization of an intron by a homing endonuclease precedes the acquisition of the endonuclease gene by the intron. Should the proteins encoded by the LSU2059 and LSU2066 introns eventually be found to share the same site of cleavage, as appears likely, this would constitute an additional instance (after those reported by Zeng et al. 2009 and Bonocora and Shub 2009) of an endonuclease-coding sequence being translocated without the endonuclease changing its cleavage specificity. Note that even though available phylogenetic data provide no indication as to whether translocation was from the group I to the group II intron subclade or vice versa in this particular case, the much greater abundance of endonuclease-encoding group I introns in fungal mitochondrial genomes makes it far

more likely that they act as donors. The occasional transposition of an entire group I intron into domain IV (or the periphery of domain III), followed by the rapid degeneration of the group I ribozyme sections, constitutes an obvious way for an endonuclease-coding gene to invade a group II intron: The resulting genomic arrangement should be readily selected whenever the recipient intron already happens to lie within the recognition sequence of the endonuclease.

DISCUSSION

Additional nucleotides at the intron 5' end and the inability to initiate splicing by transesterification

We have shown that under in vitro self-splicing conditions, the SSU788 intron of *P. fulgens* generates only linear intron forms, in contrast to its close relative in *G. frondosa*, the excision of which yields the expected lariat (branched)

intron. Either the absence of a bulging A at the expected location for the branchpoint or the presence of an insert at the intron 5' end could be invoked to account for the inability of the *Pycnoporellus* intron to perform the branching reaction. Deletion or base-pairing of the branchpoint adenosine has long been known to inhibit branching of the *S. cerevisiae* *cox1/5γ* intron (van der Veen et al. 1987; Chu et al. 1998), although splicing remains possible via hydrolysis at the 5' splice site. Similarly, the insertion of additional nucleotides at the intron 5' end was reported by Jacquier and Jacquesson-Breuleux (1991) to result in the loss of the *cox1/5γ* branching reaction in vitro; splicing could be initiated only by hydrolysis, and they showed the 5' splice site to coincide with the 3' end of the IBS1 sequence, rather than with the 5' end of the GUGCG intron consensus sequence, just as we now report for the *Pycnoporellus* intron.

The existence of natural group II introns that lack a bulging A on the 3' side of domain VI was noted long ago (Michel et al. 1989), and one of these introns, in the gene encoding the tRNA^{Val} (UAC) of plant chloroplasts, was later shown to be excised without forming lariats (Vogel and Börner 2002). On the other hand, this is the first time that the existence of group II introns with additional nucleotides at the 5' end is explicitly reported (the presence of a 5' terminal insert in the *Agroclype aegerita* LSU2059 intron was apparent in the secondary structure model in Figure 3 of Gonzalez et al. [1999] but was not discussed in the text). 5'-Terminal inserts can be surprisingly long: In *Paxillus atrotomentosus* isolate TDB-782 (Bruns et al. 1998), an intron closely related to the LSU2059 intron of *Suillus luteus* has no fewer than 48 additional nucleotides inserted between the presumed 5' splice site and the UAGCGAC sequence motif that these two introns share on the 5' side of domain I (this sequence [accession number AD001614] was not listed in Table 1 because it stops 73 nt within the intron). At the other end of the length spectrum, only 1 nt separates the inferred 5' splice site from the canonical GUGCG sequence motif in the three placozoan LSU2586 sequences. Whether a single-nucleotide insert is sufficient to abolish branching is questionable: Insertion of just 1 nt at the 5' end of *S. cerevisiae* intron *cytb1* does not prevent branching, even though it results in a marked shift toward initiation of splicing by hydrolysis (Wallasch et al. 1991). Still, the noncanonical secondary structure of domain VI in the placozoan LSU2586 introns (Fig. 1) makes it unlikely that these ribozymes would succeed in catalyzing branch formation.

As inferred from experiments in which phosphodiester bonds were replaced by phosphorothioates (Steitz and Steitz 1993; also, for review, see Michel and Ferat 1995; Jacquier 1996), the geometry of the reactive bond in the branching step must differ from the one that prevails during reversal of exon ligation, and also in 5' hydrolysis. Introns in which the end of the IBS1 sequence is not directly connected to the GUGCG consensus sequence are unable to catalyze branch-

ing, probably because interactions between the ribozyme and nucleotides bordering the 5' splice site on both its 5' and 3' sides are necessary to drive the phosphodiester bond between the intron and 5' exon into the appropriate, presumably highly constrained conformation required for first-step transesterification. In contrast, the two exons are believed to be maintained in helical continuity by the EBS1-IBS1 and EBS3-IBS3 interactions in the ligation step (Costa et al. 2000; Toor et al. 2008b). Now, since 5' hydrolysis has the same phosphorothioate requirements as the reversal of exon ligation, one would expect EBS3 to base-pair not only with the first nucleotide of the 3' exon, but also with the first nucleotide of a 5'-terminal insert, when present. Jacquier and Jacquesson-Breuleux (1991) did observe that for *S. cerevisiae* *cox1/5γ* constructs with a 5' insert, hydrolysis was facilitated when the nucleotide following the 5' splice site was an A (which could base-pair with the U at the site that would come to be known as EBS3). In nature, however, while the EBS3-IBS3 interaction is maintained in all 10 introns with 5'-terminal inserts, the first intron nucleotide forms U:U mismatches with EBS3 in the placozoan introns and an A:A mismatch in the *Pycnoporellus* intron. Thus, the identity of the nucleotide following (or to be linked to) the IBS1 sequence may be less important for 5' splice site hydrolysis (see also Su et al. 2001) than it is for exon ligation (Costa et al. 2000) or retrotransposition (Jimenez-Zurdo et al. 2003).

Loss of the ability to initiate splicing by branching entails only limited degeneration of ribozyme domain VI

The diversity of domain VI structures in introns with a 5'-terminal insert (Fig. 1), which stands in striking contrast to the well-conserved structure and sequence of this domain in the rest of the mitochondrial IIB1 subset, is strongly suggestive of rapid, unconstrained divergent evolution. Still, in all but the *Rhizophyidium* intron, apparent degeneration is limited to sections in the middle part of domain VI that have been shown to matter to the efficiency and specificity of the branching reaction. Specifically affected are (1) the branchpoint bulging A, of which the deletion or base-pairing inhibit branching (Schmelzer and Muller 1987; van der Veen et al. 1987; Chu et al. 1998); (2) the two G:U pairs flanking the branchpoint, whose replacement by G:C pairs specifically decreases the rate of branching compared to hydrolysis (Chu et al. 1998); (3) the AAA:CUA internal loop (and its closing base pairs). Replacement of this loop by base pairs has moderate, yet significant, effects on the efficiency of branching relative to hydrolysis under stringent conditions (Chu et al. 1998). Moreover, atomic group substitutions on the 5' side of the loop were found to interfere with branching (Chanfreau and Jacquier 1994; Boudvillain and Pyle 1998), while its deletion was reported to have a marked effect on the accessibility to the solvent of

the branchpoint nucleotide in a magnesium-bound domain VI construct (Schlatterer and Greenbaum 2008).

In contrast, both the base and tip of domain VI remain highly constrained in introns with a 5' terminal insertion. All but the *Rhizophydium* intron retain a 3–4-bp helix at the base of domain VI, which is connected by 3-nt joining segments to domain V on one side and the 3' splice site on the other (Fig. 1). Complete deletion of domain VI has long been known to interfere with the choice of the proper 3' splice site (Jacquier and Jacquesson-Breuleux 1991). Moreover, shortening and, to some extent, lengthening of the segment connecting domains V and VI in *S. cerevisiae* intron *cox1/5γ* (Boulanger et al. 1996) not only interfere with branching, but can lead to mis-splicing, even when the reaction is initiated by hydrolysis: In deletion mutants, only a fraction of molecules used the correct 3' splice site, despite both the 5' and 3' flanking nucleotides of the latter being involved in tertiary interactions (γ – γ' and EBS3–IBS3) (Fig. 2). Interestingly, the data in Figure 1 suggest that the identity of base pairs in the basal helix of domain VI is important as well for efficient and faithful exon ligation: in introns with a 5'-terminal insert, G:C (not C:G) base pairs predominate at positions 1, 3, and 4 of the basal helix, being present in nine, eight, and seven sequences, respectively.

At the other, distal end of domain VI, seven of the 10 intron sequences with additional nucleotides at the 5' extremity have retained a 4-nt terminal loop of the GNRA family, like nearly all mitochondrial and bacterial members of subgroup IIB1 (Fig. 1). The GUAA loop (η) that caps domain VI of the *S. cerevisiae* *cox1/5γ* intron was shown by Chanfreau and Jacquier (1996) to interact with a specific receptor (η') in ribozyme subdomain IIA: Such a receptor potentially exists in all the intron sequences in Figure 1 that share a GNRA loop at the tip of domain VI (Fig. 2; data not shown). Binding of domain VI to domain II after branch formation was proposed to drag the first-step product—i.e., the 2'–5' bonded A-G dinucleotide—out of the catalytic site, so as to make way for the 3' splice site (Chanfreau and Jacquier 1996). However, persistence of the η – η' interaction in introns that have lost the branchpoint structure and, presumably, the ability to carry out the branching reaction implies that formation of this interaction does not merely sequester domain VI (see Pyle 2010), but contributes also to the specific positioning of the 3' splice site for exon ligation. In fact, disruption of η – η' impairs specifically the second step of splicing in vitro (Chanfreau and Jacquier 1996). Even though the strikingly diverse structures of the middle part of domain VI in introns with 5'-terminal inserts (Fig. 1) may not all be capable of positioning precisely the proximal and distal ends of domain VI with respect to one another, formation of η – η' may favor correct exon ligation simply by reducing the complexity of the conformational space to be explored to bring the 3' splice site into the catalytic center of the ribozyme.

Endonuclease-mediated homing and the loss of the lariat structure

Compared to the thousands of group II introns that have been sequenced from hundreds of organisms (Rfam database) (Griffiths-Jones et al. 2005), the number of group II introns that possess 5'-terminal inserts is quite small. Moreover, these introns have a limited distribution, being confined to ribosomal RNA precursor molecules transcribed from mitochondrial genomes, and they belong to a single subgroup of ribozyme structures (IIB1). This makes it all the more striking that the 10 known instances of 5'-terminal inserts should result from no fewer than four to five independent insertion events (Fig. 5).

Just as remarkable, insertion of additional nucleotides at an intron 5' end is not the only process that has been at play specifically in mitochondrial ribosomal genes and that recurrently led to the creation of novel lineages of unusual group II introns. All known examples of group II introns encoding proteins completely unrelated to reverse transcriptases also come from mitochondrial genes encoding ribosomal RNA precursor transcripts (Toor and Zimerly 2002; Monteiro-Vitorello et al. 2009; Mullineux et al. 2010); moreover, these introns belong again to subgroup IIB1, and multiple events of the insertion of an ORF (at least six of them) (Fig. 5; Supplemental Fig. S2) need as well be postulated to account for the phylogenetic distribution of ORF–ribozyme pairs. There exist, in fact, introns—the SSU788 introns of *Pycnoporellus*, *Trametes*, and *Ganoderma* and the LSU2059 intron of *Agrocybe*—that possess a 5'-terminal insert and encode a protein of the LAGLIDADG family at the same time (Fig. 5) (the ORF of the *Agrocybe* intron is defective, but a closely related, apparently intact, ORF exists in the *Ustilago maydis* SRX2 LSU2059 intron, which belongs to the same ribozyme lineage). Such a coincidence inevitably raises suspicion that some causal relationship may exist between the acquisition of a 5'-terminal insert and that of a non-RT ORF, encoding a protein with proven (in the case of *Leptographium truncatum*) (Mullineux et al. 2010) or putative endonuclease activity.

Admittedly, six out of 10 introns with 5' inserts lack any significant protein-coding potential, while a majority of the introns that contain non-RT ORFs are devoid of 5' inserts and have a normal domain VI, which was shown to support efficient branching in the case of the *G. frondosa* SSU788 (this study) and *L. truncatum* SSU952 (Mullineux et al. 2010) introns. However, whereas degeneration of the middle part of domain VI, which closely precedes or follows the insertion of nucleotides at the 5' splice site, must be irreversible, acquisition of the coding sequence of a homing endonuclease is likely temporary. The reason is that, in a panmictic host population, the selective advantage provided by homing decreases rapidly as previously empty insertion sites become filled by a copy of the intron (Goddard and Burt 1999), so that the coding sequence of the endonuclease

should soon begin to accumulate deleterious mutations and degenerate beyond recognition. That is, unless the protein has become essential to its host by acquiring “maturase” activity. Maturase function, by which the intron-encoded protein participates in the splicing process, typically by helping the ribozyme to fold into an active structure, is commonplace in LAGLIDADG proteins encoded by group I introns (Ho et al. 1997; Bassi et al. 2002), but has not been detected so far for their counterparts in group II introns (Mullineux et al. 2010; G Bassi, unpubl., experiments with the *Grifola* SSU788 intron). To summarize, it is not unreasonable to hypothesize that not only the intron clades at the SSU788 and LSU2059 sites, but also those at LSU2586 and LSU2449, experienced invasion by the coding sequences of homing endonucleases and that sequencing of other group II introns inserted at these sites will eventually reveal their presence in some organisms.

Assuming then that all group II introns with 5'-terminal inserts had ancestors that encoded LAGLIDADG or other DNA endonucleases, why should relying on these proteins for homing eventually lead to the loss of branching? On the one hand, the lariat structure appears essential for retrotransposition by inverse splicing; the linear intron molecules that result from hydrolysis at the 5' splice site are unable to perform the second transesterification reaction (reverse of branch formation) and to complete their integration into a DNA target by themselves. On the other hand, the intron RNA, whether branched or linear, does not play any part in the homing process mediated by DNA endonucleases of the LAGLIDADG and GIY-YIG families, which rests on resealing of a double-strand break by general, homologous recombination, using the intact, intron-carrying copy as template. Thus, once a group II retrotransposon has been converted into a DNA transposon (class II mobile element) (Wicker et al. 2007) by the loss of its reverse transcriptase and the acquisition of the coding sequence of a homing endonuclease, a 2'-5' phosphodiester bond should no longer be required for mobility: The ability to generate this bond could become lost through mutations at, or next to, the branchpoint or else, the insertion of nucleotides at the 5' splice site.

Why a branched intron structure in the absence of retrotransposition?

While the branching reaction would no longer appear necessary in introns that have lost retrotransposition, the data in Table 1 point to a much more complex reality. The mere fact that a majority of mitochondrial subgroup IIB1 ribozymes have retained a canonical branchpoint means that a branched structure remains somehow important for introns that do not encode an RT gene. Furthermore, since the branchpoint and, presumably, branching have survived the acquisition of a homing endonuclease gene in more than half of the introns expected to propagate (or to have propagated) as DNA, initiation of splicing by transesterification may remain advantageous even in this subset.

At the same time, one might question the need for a branched structure even in retrotransposition. In fact, correct integration of a linear intron that has undergone partial reverse splicing followed by reverse transcription should still be possible, by recombination with the intron-carrying DNA copy: This is how the unidirectional conversion of upstream exon sequences that accompanies insertion of *S. cerevisiae* intron *cox1/2* into its intron-less target has been accounted for (Lazowska et al. 1994; Eskes et al. 2000). However, the 5' exons of the intron-carrying donor and recipient molecules must be homologous, as is the case, indeed, when, but only when, homing—as opposed to ectopic transposition—is involved. Such situations in which an intron is transmitted partly by retrotransposition and partly by homologous recombination may actually reflect transition from one mode of propagation to the other.

Admittedly, retrotransposition even of exon-less, linear intron molecules was recently reported in heterologous systems (see Zhuang et al. 2009). However, that process, which involves nonhomologous end-joining at the 5' intron extremity, is orders of magnitude less efficient than lariat retrohoming. In fact, imprecise recombination at an intron 5' end can generate 5'-terminal inserts, whose presence, and the resulting loss of branching, would trigger rapid degeneration of the branchpoint structure. Alternatively, since even linear intron molecules may retain the ability to attack suitable targets with their 3' extremity and generate partially reverse spliced molecules, it may be argued that the loss of branching should be followed by insertion events at the 5' splice site. Whatever the actual mechanism, degeneration of the branchpoint structure and the acquisition of a 5'-terminal insert must be closely coupled in subgroup IIB1, for evolutionary intermediates have not been found so far.

Coming back to the possible significance of branching for introns devoid of an RT gene, the overall coevolution of the ribozyme and protein components of group II introns (Toor et al. 2001) makes it unlikely that the intimate molecular interactions at its root could form back once they have been lost. Still, RT-less, lariat-forming introns may manage to transpose by diverting, whether on an occasional or more lasting basis, a group II-encoded reverse transcriptase that happens to be synthesized in the cellular compartment in which they reside. (Mitochondrial members of subgroup IIA, another subclass of group II introns that is widely distributed in organelles [Michel et al. 1989; Toor et al. 2001] generally encode reverse transcriptases and at least some of them are indeed mobile [Lazowska et al. 1994].)

A more subtle justification for retaining the ability to form lariats takes its roots in experimental evidence pointing to (some levels of) indiscriminate reverse transcription by group II-encoded reverse transcriptases. In yeast mitochondria, the presence of RT-encoding group II introns has been shown to promote genomic deletion of both group II

and group I introns (Gargouri et al. 1983), presumably via cDNA synthesis from spliced mRNAs. Such occasional reverse transcription could lead as well to the genomic gain of an intron that had happened to reverse-splice into an ectopic RNA site.

The branching reaction may also remain advantageous because it is liable to be more efficient than hydrolysis. This is certainly the case in vitro (Jacquier and Jacquesson-Breuleux 1991; see also Fig. 3), at physiological pH values, and could also be true in vivo, unless folding of precursor molecules were to remain rate-limiting even when compared to hydrolysis. Yet another potential advantage of making lariats is that it provides resistance to digestion by exonucleases. Stabilization of the intron would, in turn, stabilize the mRNA for the intron-encoded protein, as was argued, for instance, to account for the production of mini-, 5'-terminal lariats by a group I-derived ribozyme (Nielsen et al. 2005). However, some alternative mechanisms must exist to allow efficient production of homing endonucleases from linear intron molecules: The extensive, long-range RNA–RNA pairings that flank the ORFs of the *Pycnoporellus* molecule and other introns with 5'-terminal inserts (Fig. 2B; data not shown) may substitute for the 2'–5' phosphodiester bond of the lariat and slow down the progression of exonucleases.

Possible implication of the EBS2–IBS2 pairing in branch formation

Another feature that shows correlation with the loss of branching is the absence of EBS2 (Table 1). As already emphasized (see Results), all 10 introns with a 5' insert actually lack the entire subdomain that the EBS2 segment is normally part of (Michel et al. 1989; Dai et al. 2003). The EBS2–IBS2 pairing is known to be important for insertion of group II introns by reverse-splicing into double-stranded DNA, presumably because it helps stabilize interactions between the intron and its target relative to DNA:DNA base-pairing, but it does not appear to be required for transposition into single-stranded nucleic acids (Coros et al. 2005). Still, that this interaction should persist not only in introns that have lost the coding sequence for a reverse transcriptase, but in several of those that encode a LAGLIDADG homing endonuclease (Table 1), implies that the EBS2–IBS2 pairing has some significant function in splicing as well. Partial disruption of that interaction in *S. cerevisiae* intron *cox1/5γ* decreases the stability of the complex between the 5' exon and intron, resulting in accumulation of the intron–3'exon reaction intermediate in vitro (Jacquier and Michel 1987). Somewhat more unexpectedly, it also appears to affect the chemical step of the reaction by which oligonucleotides that mimic the intron target site are cleaved (Xiang et al. 1998).

Close examination of the data in Table 1 and, in particular, the absence of EBS2 from all but one of the SSU788

introns, suggests that loss of the EBS2–IBS2 pairing precedes, and might even be a necessary step for, the loss of branching. One possibility is that the deletion of EBS2 somehow facilitates hydrolysis at the 5' splice site (although for the *Grifola* ribozyme, hydrolysis-initiated self-splicing was found to constitute but a minor reaction pathway even in the presence of potassium) (Fig. 3B). This would not be without precedent, for the only intron subclass—subgroup IIC—in which these components are systematically missing is noteworthy for (most of) its members initiating self-splicing in vitro by hydrolysis (Granlund et al. 2001; Toor et al. 2006).

Unfortunately, possible ways in which EBS2 and the structures that surround it might affect the balance between transesterification and hydrolysis remain difficult to think of at present. The only currently available group II crystal structure (Toor et al. 2008a,b) happens to be that of a subgroup IIC intron, and it lacks not only domain VI, but the EBS2–IBS2 interaction and a number of additional RNA subdomains and devices that a majority of other lineages of group II ribozymes have opted to conserve (for review, see Pyle 2010). Additional group II structures, in which domain VI and the branchpoint can be visualized in interaction with the rest of the ribozyme, are a prerequisite if we are eventually to reach a complete understanding of why the branching reaction has been so stubbornly, although not universally, retained during the diversifying evolution of group II introns.

MATERIALS AND METHODS

Sequence analyses of mitochondrial subgroup IIB1 ribozymes

Published sequences of mitochondrial introns that possessed characteristic sequence and secondary structure features of subgroup IIB1 (Michel et al. 1989) were collected (Table 1), and their ribozyme sections were manually aligned (Supplemental Data Set) based on conservation of both sequence and potential secondary structure (the distal sections of stems IC2, ID2, IIA, IIIB, and IV [see Fig. 2; Michel et al. 2009] could not be reliably aligned and were discarded). Starting from this alignment, a phylogenetic tree was generated by PAUP* 4.0b10 (Swofford 2002) using the Neighbor-Joining algorithm and a matrix that had been obtained by using the LogDet measure of distance, which is insensitive to differences in base composition (Lockhart et al. 1994). Note that (1) to avoid biasing the tree-building procedure in favor of subsets constituted by introns that share homologous insertion sites, the EBS1, EBS2, and EBS3 sites were removed, leaving 526 sites in the final alignment; (2) of the three closely related *G. intraradices* LSU1787 intron sequences in Table 1, only the first one, which does not include an ORF, was retained for the tree-building process.

Sequence analyses of LAGLIDADG proteins

To investigate the phylogenetic relationships of LAGLIDADG proteins potentially encoded by subgroup IIB1 introns, apparently

intact or defective sequences generated from the intron nucleotide sequences (Table 1) were compared to the NCBI nonredundant protein data set, and for each comparison, the 10 target sequences with the highest BLASTP scores were retained. The resulting sequence data set was then aligned together with the 91 sequences in the Pfam LAGLIDADG 1 (PF00961) “seed” set. After manual refinement, the final alignment (available from the authors) consisted of 146 sites and 174 sequences, 26 of which were of presumably dimeric proteins (Stoddard 2005), with a single LAGLIDADG motif, while the rest corresponded to monomeric proteins (in which case, each of the two sections following a LAGLIDADG motif was aligned separately).

Because of the rather large number of sequences in this data set, we resorted to the efficient Neighbor-Joining algorithm, using distances generated by the program PROTDIST (Felsenstein 2004), to generate a phylogenetic tree and calculate bootstrap percentages (Supplemental Fig. S2). Even though the resulting phylogeny is far from being completely resolved, our main conclusions regarding phylogenetic relationships of the proteins encoded by group II introns (see above) are supported by relatively high bootstrap percentages and/or the fact that groupings were found to be the same whether the first or second pseudo-repeat of monomeric proteins was used for comparisons (Supplemental Fig. S2).

Amplification, sequencing, and cloning of fungal introns

DNA extracts from *A. botryosus* CBS195.91, *G. frondosa* CBS 480.63, and *P. fulgens* T-325 were obtained from David Hibbett (Clark Fungal Database at Clark University). PCR amplifications of the SSU788 intron and surrounding exons were performed in 50 μ L with 1 μ M primers BMS65MOD and BMS103E (Supplemental Table S1) using 1 unit of high-fidelity Phusion polymerase in HF buffer (Finnzymes) and 33 cycles (10 sec at 98°C, 45 sec at 60°C, 90 sec at 72°C). Sequencing of amplification products was carried out on both strands by GATC Biotech using the same primers as well as species-specific primers listed in Supplemental Table S1. Accession numbers for the assembled sequences are FR773978, FR773979, and FR773980.

For cloning into *Escherichia coli*, amplification products were reamplified with primers BMS65MODT7 and BMS103EZ, digested with BamHI and XmaI, and ligated into the pUC19 vector plasmid. For deletion of ORF sequences from ribozyme domain IV of the *G. frondosa* and *P. fulgens* introns, primers GRXHOREV (or PYXHOREV) and GRXHOFWD (or PYXHOFWD) (see Supplemental Table S1) were used in combination with vector-specific primers ANT7 and 24mer, respectively, to generate PCR products. These products were digested with XhoI and either BamHI or XmaI, and cloned back into pUC19. The resulting constructs, pUC19-GR1 Δ ORF and pUC19-PY1 Δ ORF, in which most of domain IV has been replaced by an XhoI site (Fig. 2, legend), were verified by sequencing.

In vitro transcription and purification of precursor RNA

Templates for synthesis of the *Grifola* and *Pycnoporellus* precursor RNAs were obtained by digestion of plasmids pUC19-GR1 Δ ORF and pUC19-PY1 Δ ORF with SmaI. RNA synthesis and purification were carried out as described in Costa et al. (1997b), except that the

transcription mixture contained 10% DMSO so as to avoid premature transcription stops and a 1.55 molar concentration ratio of magnesium over nucleotides was used to prevent premature intron splicing.

Self-splicing reactions

Precursor transcripts internally labeled with ^{32} P-UTP were denatured in water at 90°C prior to cooling to reaction temperature. Reactions were started by addition of an equal volume of 2 $\times$ -concentrated splicing buffer. The final concentration of precursor molecules was 20 nM. Reactions were stopped by addition of an equal volume of a solution of formamide containing EDTA at a concentration appropriate to complex all of the magnesium, and products were separated on a denaturing polyacrylamide gel (50% [w/v] urea, 4% total acrylamide, 0.2% bis-acrylamide). Radioactivity was quantitated on fixed, dried gels using a PhosphorImager (MolecularDynamics), and the molar fraction of each product was calculated. Reaction time courses were fitted to either single $\{m_1[1 - \exp(-kt)]\}$ or double $\{m_1[1 - \exp(-k_1t)] + m_2[1 - \exp(-k_2t)]\}$ exponentials (m_1 or $m_1 + m_2$ are final fractions of reacted product).

Reverse transcription of splicing products

Preparative self-splicing reactions were carried out in 40 mM Tris-Cl (pH 7.5), 20 mM MgCl₂, and 1 M NH₄Cl at 42°C (the *G. frondosa* intron–3' exon lariat molecule was isolated from a splicing reaction that included 20 mM CaCl₂). Purification of splicing products from preparative denaturing polyacrylamide gels and their reverse transcription with ^{32} P-labeled, gel-purified oligonucleotides were performed essentially as described by Costa et al. (1997b). The following oligonucleotides (see Supplemental Table S1) were used for reverse transcription: BMS103B, to sequence ligated exons and determine the branchpoint of the *G. frondosa* intron–3' exon lariat; Gr-R2, to determine the 5' splice of the *G. frondosa* intron lariat; Py-R2, to determine the 5' extremity of *P. fulgens* linear intron molecules.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We are especially grateful to David Hibbett for sending us fungal DNA extracts, and to Franz Lang and Taylor Mullineux for attempting to improve our text. This work was funded by the Centre National de la Recherche Scientifique. C.-F.L. was supported by a Joseph Fourier fellowship from the French Government and the National Science Council of Taiwan.

Received February 1, 2011; accepted April 1, 2011.

REFERENCES

Bassi GS, de Oliveira DM, White MF, Weeks KM. 2002. Recruitment of intron-encoded and co-opted proteins in splicing of the b13 group I intron RNA. *Proc Natl Acad Sci* **99**: 128–133.

- Beauregard A, Curcio MJ, Belfort M. 2008. The take and give between retrotransposable elements and their hosts. *Annu Rev Genet* **42**: 587–617.
- Bonen L. 2008. *Cis*- and *trans*-splicing of group II introns in plant mitochondria. *Mitochondrion* **8**: 26–34.
- Bonocora RP, Shub DA. 2009. A likely pathway for formation of mobile group I introns. *Curr Biol* **19**: 223–228.
- Boudvillain M, Pyle AM. 1998. Defining functional groups, core structural features and inter-domain tertiary contacts essential for group II intron self-splicing: a NAIM analysis. *EMBO J* **17**: 7091–7104.
- Boulanger SC, Faix PH, Yang H, Zhuo J, Franzen JS, Perlman PS. 1996. Length changes in the joining segment between domains 5 and 6 of a group II intron inhibit self-splicing and alter 3' splice site selection. *Mol Cell Biol* **16**: 5896–5904.
- Bruns TD, Szaro TM, Gardes M, Cullings KW, Pan JJ, Taylor DL, Horton TR, Kretzer A, Garbelotto M, Li Y. 1998. A sequence database for the identification of ectomycorrhizal basidiomycetes by phylogenetic analysis. *Mol Ecol* **7**: 257–272.
- Chanfreau G, Jacquier A. 1994. Catalytic site components common to both splicing steps of a group II intron. *Science* **266**: 1383–1387.
- Chanfreau G, Jacquier A. 1996. An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J* **15**: 3466–3476.
- Chu VT, Liu Q, Podar M, Perlman PS, Pyle AM. 1998. More than one way to splice an RNA: Branching without a bulge and splicing without branching in group II introns. *RNA* **4**: 1186–1202.
- Coros CJ, Landthaler M, Piazza CL, Beauregard A, Esposito D, Perutka J, Lambowitz AM, Belfort M. 2005. Retrotransposition strategies of the *Lactococcus lactis* Ll.LtrB group II intron are dictated by host identity and cellular environment. *Mol Microbiol* **56**: 509–524.
- Costa M, Deme E, Jacquier A, Michel F. 1997a. Multiple tertiary interactions involving domain II of group II self-splicing introns. *J Mol Biol* **267**: 520–536.
- Costa M, Fontaine JM, Loiseaux-de Goër S, Michel F. 1997b. A group II self-splicing intron from the brown alga *Pylaiella littoralis* is active at unusually low magnesium concentrations and forms populations of molecules with a uniform conformation. *J Mol Biol* **274**: 353–364.
- Costa M, Michel F, Westhof E. 2000. A three-dimensional perspective on exon binding by a group II self-splicing intron. *EMBO J* **19**: 5007–5018.
- Dai L, Toor N, Olson R, Keeping A, Zimmerly S. 2003. Database for mobile group II introns. *Nucleic Acids Res* **31**: 424–426.
- Daniels DL, Michels WJ Jr, Pyle AM. 1996. Two competing pathways for self-splicing by group II introns: A quantitative analysis of in vitro reaction rates and products. *J Mol Biol* **256**: 31–49.
- Eskes R, Liu L, Ma H, Chao MY, Dickson L, Lambowitz AM, Perlman PS. 2000. Multiple homing pathways used by yeast mitochondrial group II introns. *Mol Cell Biol* **20**: 8432–8446.
- Fedorova O, Pyle AM. 2005. Linking the group II intron catalytic domains: tertiary contacts and structural features of domain 3. *EMBO J* **24**: 3906–3916.
- Felsenstein J. 2004. *PHYMLIP (Phylogeny Inference Package) version 3.6*. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
- Gargouri A, Lazowska J, Slonimski PP. 1983. DNA-splicing of introns in the gene: A general way of reverting intron mutations. In *Mitochondria 1983: Nucleo-mitochondrial interactions* (ed. RJ Schweyen et al.), pp. 259–268. de Gruyter, Berlin.
- Goddard MR, Burt A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc Natl Acad Sci* **96**: 13880–13885.
- Gonzalez P, Barroso G, Labarère J. 1999. Molecular gene organisation and secondary structure of the mitochondrial large subunit ribosomal RNA from the cultivated Basidiomycota *Agrocybe aegerita*: a 13 kb gene possessing six unusual nucleotide extensions and eight introns. *Nucleic Acids Res* **27**: 1754–1761.
- Granlund M, Michel F, Norgren M. 2001. Mutually exclusive distribution of IS1548 and GBSi1, an active group II intron identified in human isolates of group B streptococci. *J Bacteriol* **183**: 2560–2569.
- Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**: D121–D124.
- Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res* **21**: 3537–3544.
- Ho Y, Kim SJ, Waring RB. 1997. A protein encoded by a group I intron in *Aspergillus nidulans* directly assists RNA splicing and is a DNA endonuclease. *Proc Natl Acad Sci* **94**: 8994–8999.
- Hong SG, Jeong W, Jung HS. 2002. Amplification of mitochondrial small subunit ribosomal DNA of polypores and its potential for phylogenetic analysis. *Mycologia* **94**: 823–833.
- Jacquier A. 1996. Group II introns: Elaborate ribozymes. *Biochimie* **78**: 474–487.
- Jacquier A, Jacquesson-Breuleux N. 1991. Splice site selection and role of the lariat in a group II intron. *J Mol Biol* **219**: 415–428.
- Jacquier A, Michel F. 1987. Multiple exon-binding sites in class II self-splicing introns. *Cell* **50**: 17–29.
- Jarrell KA, Peebles CL, Dietrich RC, Romiti SL, Perlman PS. 1988. Group II intron self-splicing. Alternative reaction conditions yield novel products. *J Biol Chem* **263**: 3432–3439.
- Jimenez-Zurdo JJ, Garcia-Rodriguez FM, Barrientos-Duran A, Toro N. 2003. DNA target site requirements for homing in vivo of a bacterial group II intron encoding a protein lacking the DNA endonuclease domain. *J Mol Biol* **326**: 413–423.
- Johansen S, Haugen P. 2001. A new nomenclature of group I introns in ribosomal DNA. *RNA* **7**: 935–936.
- Kennell JC, Moran JV, Perlman PS, Butow RA, Lambowitz AM. 1993. Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. *Cell* **73**: 133–146.
- Lambowitz AM, Zimmerly S. 2004. Mobile Group II introns. *Annu Rev Genet* **38**: 1–35.
- Lang BF, Laforest MJ, Burger G. 2007. Mitochondrial introns: a critical view. *Trends Genet* **23**: 119–125.
- Lazowska J, Meunier B, Macadre C. 1994. Homing of a group II intron in yeast mitochondrial DNA is accompanied by unidirectional co-conversion of upstream-located markers. *EMBO J* **13**: 4963–4972.
- Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol* **11**: 605–612.
- Michel F, Ferat JL. 1995. Structure and activities of group II introns. *Annu Rev Biochem* **64**: 435–461.
- Michel F, Umeson K, Ozeki H. 1989. Comparative and functional anatomy of group II catalytic introns—a review. *Gene* **82**: 5–30.
- Michel F, Costa M, Westhof E. 2009. The ribozyme core of group II introns: a structure in want of partners. *Trends Biochem Sci* **34**: 189–199.
- Monteiro-Vitorello CB, Hausner G, Searles DB, Gibb EA, Fulbright DW, Bertrand H. 2009. The *Cryptosporidia parasitica* mitochondrial *rns* gene: Plasmid-like elements, introns and homing endonucleases. *Fungal Genet Biol* **46**: 837–848.
- Mullineux ST, Costa M, Bassi GS, Michel F, Hausner G. 2010. A group II intron encodes a functional LAGLIDADG homing endonuclease and self-splices under moderate temperature and ionic conditions. *RNA* **16**: 1818–1831.
- Nielsen H, Westhof E, Johansen S. 2005. An mRNA is capped by a 2', 5' lariat catalyzed by a group I-like ribozyme. *Science* **309**: 1584–1587.
- Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. 2000. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**: 920–930.
- Podar M, Chu VT, Pyle AM, Perlman PS. 1998. Group II intron splicing in vivo by first-step hydrolysis. *Nature* **391**: 915–918.
- Pyle AM. 2010. The tertiary structure of group II introns: Implications for biological function and evolution. *Crit Rev Biochem Mol Biol* **45**: 215–232.

- Quiroga C, Roy PH, Centron D. 2008. The S.ma.I2 class C group II intron inserts at integron attC sites. *Microbiology* **154**: 1341–1353.
- Schlatterer JC, Greenbaum NL. 2008. Specificity of Mg²⁺ binding at the Group II intron branch site. *Biophys Chem* **136**: 96–100.
- Schmelzer C, Muller MW. 1987. Self-splicing of group II introns in vitro: Lariat formation and 3' splice site selection in mutant RNAs. *Cell* **51**: 753–762.
- Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, Holton JM, Cate JH. 2005. Structures of the bacterial ribosome at 3.5 Å resolution. *Science* **310**: 827–834.
- Stabell FB, Tourasse NJ, Ravnum S, Kolsto AB. 2007. Group II intron in *Bacillus cereus* has an unusual 3' extension and splices 56 nucleotides downstream of the predicted site. *Nucleic Acids Res* **35**: 1612–1623.
- Steitz TA, Steitz JA. 1993. A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci* **90**: 6498–6502.
- Stoddard BL. 2005. Homing endonuclease structure and function. *Q Rev Biophys* **38**: 49–95.
- Su LJ, Qin PZ, Michels WJ, Pyle AM. 2001. Guiding ribozyme cleavage through motif recognition: The mechanism of cleavage site selection by a group II intron ribozyme. *J Mol Biol* **306**: 655–668.
- Swofford DL. 2002. *PAUP*: Phylogenetic analysis using parsimony [and other methods]*. 4.0 Beta. Sinauer Associates, Sunderland, MA.
- Toor N, Zimmerly S. 2002. Identification of a family of group II introns encoding LAGLIDADG ORFs typical of group I introns. *RNA* **8**: 1373–1377.
- Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* **7**: 1142–1152.
- Toor N, Robart AR, Christianson J, Zimmerly S. 2006. Self-splicing of a group IIC intron: 5' exon recognition and alternative 5' splicing events implicate the stem loop motif of a transcriptional terminator. *Nucleic Acids Res* **34**: 6461–6471.
- Toor N, Keating KS, Taylor SD, Pyle AM. 2008a. Crystal structure of a self-spliced group II intron. *Science* **320**: 77–82.
- Toor N, Rajashankar K, Keating KS, Pyle AM. 2008b. Structural basis for exon recognition by a group II intron. *Nat Struct Mol Biol* **15**: 1221–1222.
- van der Veen R, Kwakman JH, Grivell LA. 1987. Mutations at the lariat acceptor site allow self-splicing of a group II intron without lariat formation. *EMBO J* **6**: 3827–3831.
- Vogel J, Börner T. 2002. Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. *EMBO J* **21**: 3794–3803.
- Wallasch C, Morl M, Niemer I, Schmelzer C. 1991. Structural requirements for selection of 5'- and 3' splice sites of group II introns. *Nucleic Acids Res* **19**: 3307–3314.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973–982.
- Xiang Q, Qin PZ, Michels WJ, Freeland K, Pyle AM. 1998. Sequence specificity of a group II intron ribozyme: Multiple mechanisms for promoting unusually high discrimination against mismatched targets. *Biochemistry* **37**: 3839–3849.
- Zeng Q, Bonocora RP, Shub DA. 2009. A free-standing homing endonuclease targets an intron insertion site in the psbA gene of cyanophages. *Curr Biol* **19**: 218–222.
- Zhuang F, Mastroianni M, White TB, Lambowitz AM. 2009. Linear group II intron RNAs can retrohome in eukaryotes and may use nonhomologous end-joining for cDNA ligation. *Proc Natl Acad Sci* **106**: 18189–18194.

RNA

A PUBLICATION OF THE RNA SOCIETY

Recurrent insertion of 5'-terminal nucleotides and loss of the branchpoint motif in lineages of group II introns inserted in mitochondrial preribosomal RNAs

Cheng-Fang Li, Maria Costa, Gurminder Bassi, et al.

RNA 2011 17: 1321-1335 originally published online May 25, 2011

Access the most recent version at doi:[10.1261/rna.2655911](https://doi.org/10.1261/rna.2655911)

Supplemental Material <http://rnajournal.cshlp.org/content/suppl/2011/05/05/rna.2655911.DC1>

References This article cites 65 articles, 22 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/17/7/1321.full.html#ref-list-1>

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:
<http://rnajournal.cshlp.org/subscriptions>