Molecular Biology and Evolution

Large-Scale tRNA Intron Transposition in the Archaeal Order Thermoproteales Represents a Novel Mechanism of Intron Gain

Kosuke Fujishima,[1,2] Junichi Sugahara,[1,3] Masaru Tomita,[1,2,3] and Akio Kanai*[,1,2,3]

[1]Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan

[2]Department of Environmental Information, Keio University, Fujisawa, Japan

[3]Systems Biology Program, Graduate School of Media and Governance, Keio University, Fujisawa, Japan

***Corresponding author**: E-mail: akio@sfc.keio.ac.jp.

Associate editor: Martin Embley

Abstract

Recently, diverse arrangements of transfer RNA (tRNA) genes have been found in the domain Archaea, in which the tRNA is interrupted by a maximum of three introns or is even fragmented into two or three genes. Whereas most of the eukaryotic tRNA introns are inserted strictly at the canonical nucleotide position (37/38), archaeal intron-containing tRNAs have a wide diversity of small tRNA introns, which differ in their numbers and locations. This feature is especially pronounced in the archaeal order Thermoproteales. In this study, we performed a comprehensive sequence comparison of 286 tRNA introns and their genes in seven Thermoproteales species to clarify how these introns have emerged and diversified during tRNA gene evolution. We identified 46 intron groups containing sets of highly similar sequences (>70%) and showed that 16 of them contain sequences from evolutionarily distinct tRNA genes. The phylogeny of these 16 intron groups indicates that transposition events have occurred at least seven times throughout the evolution of Thermoproteales. These findings suggest that frequent intron transposition occurs among the tRNA genes of Thermoproteales. Further computational analysis revealed limited insertion positions and corresponding amino acid types of tRNA genes. This has arisen because the bulge–helix–bulge splicing motif is required at the newly transposed position if the pre-tRNA is to be correctly processed. These results clearly demonstrate a newly identified mechanism that facilitates the late gain of short introns at various noncanonical positions in archaeal tRNAs.

Key words: Archaea, tRNA intron, transposition, Thermoproteales.

Introduction

Transfer RNA (tRNA) is a universally conserved essential molecule that decodes genetic information during protein biosynthesis. tRNA introns are found in all three domains of life (bacteria, Archaea, and Eukarya), representing the most common disruption of the host tRNA genes. However, their splicing mechanisms and locations in the tRNA genes of the three biological domains differ. Bacterial tRNA introns are known to be self-splicing group I introns and are located in the anticodon loop of the tRNA genes of the α-, β-, and γ-Proteobacteria (Reinhold-Hurek and Shub 1992). In contrast, the tRNA introns of the Archaea and Eukarya are processed by the specific enzymatic reactions of tRNA splicing endonucleases. Eukaryotic tRNA introns are predominantly inserted at canonical position 37/38, a universally conserved tRNA intron insertion position in Archaea and Eukarya, and are removed by the heterotetrameric eukaryotic splicing endonucleases of the Sen family, based on a measuring system that recognizes the distance between the mature domain and the splice sites (Li et al. 1998). A few exceptions occur in the nuclear genomes of photosynthetic eukaryotes and the nucleomorph, the endosymbiont-derived eukaryotic nucleus found in certain plastids, in which 1–3 introns are located at various locations within a single tRNA gene (Kawach et al. 2005; Soma et al. 2007; Maruyama et al. 2009).

Similarly, archaeal species have one of three types of homologous tRNA splicing endonucleases (homodimeric, homotetrameric, or heterotetrameric), which recognize an exon–intron junction characterized by the bulge–helix–bulge (BHB) motif (Diener and Moore 1998; Xue, Calvin, and Li 2006). Archaeal species expressing the heterotetrameric splicing endonuclease tend to have increased proportions of tRNA introns inserted at noncanonical positions (other than 37/38) compared with the archaeal species that express the other types of endonucleases. Species belonging to the thermoacidophilic order Thermoproteales display a particularly elevated proportion of tRNA introns at noncanonical sites, which are found at over 30 nucleotide positions within various tRNA genes (Sugahara et al. 2008). Unusual tRNA genes that contain three introns are currently known in only three species, *Thermofilum pendens*, *Pyrobaculum calidifontis*, and *P. islandicum*, all members of the Thermoproteales. However, the numbers of intron-containing tRNA genes do not follow the phylogenetic trends of the Thermoproteales species, which suggests an active gain and loss of tRNA introns within this order (Sugahara et al. 2009).

Several models have been proposed for the gain and loss of spliceosomal introns in eukaryotes. For example, reverse transcription and recombination of the spliced RNA could lead to intron loss, whereas reverse splicing, the insertion of transposable elements, and gene duplication could result in intron gain (Roy and Gilbert 2006). In contrast, the origin and evolution of tRNA introns are poorly understood. Therefore, whether the models proposed for the spliceosomal introns can also be applied to the enzymatically cleaved tRNA introns remains a major question. A unique type of disrupted tRNA, the so-called "split tRNA" (tRNA encoded on two or three separate minigenes), has recently been discovered in two hyperthermophilic Archaea, *Nanoarchaeum equitans* and *Caldivirga maquilingensis* (Randau et al. 2005; Fujishima et al. 2009). Our recent studies have shown that the precursor sequences of these split tRNAs have high sequence similarity to the tRNA intron sequences in related archaeal species (Fujishima et al. 2008, 2009). This finding clearly suggests an evolutionary relationship between the two types of tRNA genes, although the transition state between the two is unknown. One explanation is that the tRNA intron is a remnant of the residual leader sequence produced during the *trans* splicing of the split tRNA gene ("split-early" hypothesis; Di Giulio 2006, 2008). An alternative explanation is that the insertion of the intron occurred first and the gene subsequently underwent fragmentation at the site of insertion ("split-late" hypothesis; Randau and Soll 2008; Heinemann et al. 2009).

In this study, we performed a comprehensive sequence analysis of all the tRNA intron sequences in seven members of the Thermoproteales and characterized the sets of "intron groups" with similar sequences. Approximately one-third of the groups contained introns from evolutionarily distinct tRNA genes, which we defined as the "transposable intron" group. Phylogenetic analysis revealed that these groups originated at various times during the evolution of the Thermoproteales and have contributed to the late gain of introns in the tRNA genes. This new phenomenon is inconsistent with any of the known theories of intron gain. It also has important implications for the origin and evolution of the enzymatically cleaved short introns that are present in the domains Archaea and Eukarya.

Materials and Methods

tRNA Sequences

In total, 2662 tRNA genes from 58 archaeal species were downloaded from the archaeal tRNA database SPLITSdb (http://splits.iab.keio.ac.jp/splitsdb/) (Sugahara et al. 2008). Intron sequences and their RNA secondary structures were also obtained from SPLITSdb.

Classification of tRNA Introns

A total of 286 tRNA intron sequences from seven Thermoproteales species were aligned using ClustalW 2.0 (Larkin et al. 2007) with the default parameters. In this study, we defined an intron group as more than two intron sequences that cluster with a sequence similarity above 70%. Transposable introns were further extracted by manually investigating each tRNA intron groups to see whether it satisfied either of the following criteria: 1) the introns belong to more than two evolutionarily distinct tRNA genes (e.g., $tRNA^{Thr}$ and $tRNA^{Val}$), or 2) the introns are located at more than two different nucleotide positions. The same criteria were also applied to 217 tRNA introns from 12 crenarchaeal species: *Aeropyrum pernix*, *Cenarchaeum symbiosum*, *Desulfurococcus kamchatkensis*, *Hyperthermus butyricus*, *Ignicoccus hospitalis*, *Metallosphaera sedula*, *Staphylothermus marinus*, *Sulfolobus acidocaldarius*, *Su. islandicus* L.S.2.15, *Su. islandicus* M.14.25, *Su. solfataricus*, and *Su. tokodaii*.

Construction of a Phylogenetic Tree of tRNAs

A multiple sequence alignment of the tRNAs was constructed by masking all the tRNA intron sequences to consider the phylogeny of the mature tRNA sequences. All the mature tRNA sequences (input as tDNA sequences) were aligned using ClustalW 2.0 (Larkin et al. 2007). The aligned tRNA sequences (.aln file) were manually improved by matching the consensus nucleotides conserved among the archaeal tRNAs: base 8 (T or C), bases 14–15 (AG), bases 18–19 (GG), base 21 (A or G), base 33 (T), base 48 (T or C), bases 53–58 (GTTC[A/G]A), bases 60–61 (TC), and base 72 (T or C) (Marck and Grosjean 2002) to generate a structural alignment. An unrooted neighbor-joining tree was constructed using ClustalW 2.0 and was visualized with iToL (Letunic and Bork 2007).

In Silico Splicing Analysis

The intron sequences belonging to the tRNAs listed in table 2 were computationally inserted into every nucleotide position (from 1/2 to 72/73) of the original tRNA sequence, and into the same nucleotide position in the mature tRNAs with different isoacceptors derived from the same species, to produce two types of artificial intron-containing tRNAs (supplementary fig. S1, Supplementary Material online). To correctly predict the strict/relaxed BHB motif, an extra 20-nt pre-tRNA 5′ leader sequence was added to the tRNA genes containing Group 4 introns located at position 3/4. The artificial intron sequences were subjected to in silico splicing using SPLITS (Sugahara et al. 2006) with the default parameters: $-h = 3$ (minimum length of the central helix) and $-p = 0.6$ (change in the cutoff value of the position weight matrix). We defined "spliceable intron" based on the criteria: 1) is spliced at the same exon–intron boundary; 2) has the same central helix length (either 3 or 4 nt); and 3) has a bulge of at least 3 nt on either side, compared with the original intron.

Results

Characterization of tRNA Introns in Thermoproteales

According to SPLITSdb, a comprehensive archaeal tRNA database (http://splits.iab.keio.ac.jp/splitsdb/) (Sugahara et al. 2008), 648 intron sequences have been found to date

Table 1. Characteristics of 16 Transposable tRNA Intron Groups in Thermoproteales.

				Number of tRNA genes in each species							
Group	Length (nt)	Amino acids[a]	Position	Pae	Pis	Tne	Par	Pca	Cma	Tpe	Total
1	20–27	Ala, Ile, Met, Thr, Val	29/30	5	5	4	5	2	1	—	22
2	20–24	Arg, Asn, Cys, Lys	30/31	2	8	6	—	—	—	—	16
3	15–16	Ala, Gln, Ile, Pro, Thr, Val	53/54	—	—	—	—	12	—	—	12
4	21	Asp, Glu	3/4	3	3	2	3	—	—	—	11
5	16	Arg, Lys, Thr	56/57	2	—	—	—	7	—	—	9
6	17–20	Asn, Asp, Glu, His, iMet, Trp	59/60	—	—	—	—	7	—	—	7
7	17–20	Arg, Asp, Gly, Trp	45/46	—	—	—	—	—	—	6	6
8	18	Gly, Pro	25/26, 29/30	—	—	—	—	5	—	—	5
9	18	Arg, Pro	56/57	—	2	1	1	—	—	—	4
10	28–30	Ile, Val	51/52	—	2	2	—	—	—	—	4
11	29–32	Ala, Val	37/38	—	—	—	—	—	—	4	4
12	17–18	Asn, Trp	45/46	—	1	2	—	—	—	—	3
13	16	Asn, Cys, Trp	45/46	—	—	—	—	3	—	—	3
14	20–22	Cys, Glu, Thr	58/59, 59/60	3	—	—	—	—	—	—	3
15	24	Asn, Ile	22/23	—	—	—	—	—	—	2	2
16	29	Asn, Ile	43/44	—	—	—	—	—	—	2	2
			Total	15	21	17	9	36	1	14	113

[a] Amino acid corresponding to the anticodon of the tRNA

in 549 intron-containing tRNA genes from 58 archaeal species. Of these introns, 44% belong to seven species assigned to the order Thermoproteales (*C. maquilingensis*, *P. aerophilum*, *P. arsenaticum*, *P. calidifontis*, *P. islandicum*, *Thermofilum pendens*, and *Thermoproteus neutrophilus*), representing an extremely enriched proportion of tRNA introns in this specific order. In an attempt to characterize the overall picture of this phenomenon using a comparative genomics approach, we first collected 286 tRNA intron sequences from 200 intron-containing tRNA genes in seven Thermoproteales species and performed a multiple sequence alignment to identify clusters of introns with similar sequences. We then defined an intron group as a set of introns with a sequence similarity above 70% that are likely to have the same origin. Forty-six intron groups were thus extracted and classified into two categories: 30 orthologous intron groups, which only contain sets of intron sequences found in orthologous tRNA genes, and 16 transposable intron groups, each of which contains a set of intron sequences found in more than two evolutionarily distinct tRNA genes. The features of the 113 introns belonging to the 16 transposable intron groups are summarized in Table 1. Each group is characterized by a specific intron length, insertion position, corresponding amino acid, and degree of conservation among the seven Thermoproteales species.

To examine the tRNA transposition event in terms of the evolution of the tRNA genes, we constructed a phylogenetic tree of 323 tRNA genes from the seven Thermoproteales species based on the structural alignment of their exon sequences, and mapped the 113 transposable tRNA introns onto the tRNA tree (fig. 1). The transposable introns were distributed among the tRNA genes corresponding to 16 different amino acids. The tRNA genes corresponding to the four amino acids Leu, Phe, Ser, and Tyr did not contain any transposable introns. Although almost all the transposed introns were found singly in the tRNA genes, six tRNA genes had multiple transposable tRNA introns. Interestingly, these tRNAs can be classified into three tRNA pairs, having either Group 3 and Group 8 introns (two $tRNA^{Pro}$s in *P. calidifontis*), Group 6 and Group 13 introns ($tRNA^{Asn}$ and $tRNA^{Trp}$ in *P. calidifontis*), or Group 15 and Group 16 introns ($tRNA^{Asn}$ and $tRNA^{Ile}$ in *Thermofilum pendens*). Furthermore, within 14 of the 16 transposable tRNA intron groups, insertion has only occurred at a single nucleotide position (e.g., position 29/30 for Group 1). This feature can be explained by the structural requirement that the inserted introns must form a BHB motif with the adjacent exon sequence to allow the correct processing of the pre-tRNA (discussed below). In contrast, introns in Group 8 and Group 13 are located at two different nucleotide positions, making these introns a very rare exception (e.g., Group 8, positions 25/26 and 29/30).

Characteristics of the Sequences, Structures, and Nucleotide Positions of Transposable tRNA Introns

The classification and distribution of the tRNA introns on the tRNA phylogenetic tree revealed that intron transposition has occurred widely, involving approximately 40% (113/286) of the tRNA introns in Thermoproteales. To clarify the mechanism underlying this transposition, we next focused on the sequences and BHB splicing motifs of each transposable intron group. The BHB motif consists of both intron and exon sequences and can be classified into three types based on its RNA secondary structure (Marck and Grosjean 2003; fig. 2A). All 113 transposable introns comprise either the strict hBHBh′ or the relaxed HBh′ motif. From an enzymatic perspective, this observation is compatible with the fact that Thermoproteales species express a heterotetrameric splicing endonuclease, which cleaves both the strict hBHBh′ and relaxed HBh′/hBH motifs (Tocchini-Valentini et al. 2007; Yoshinari et al. 2009). In the Archaea, the hBH motif is only observed at the canonical position 37/38 (a rare site for transposition), which

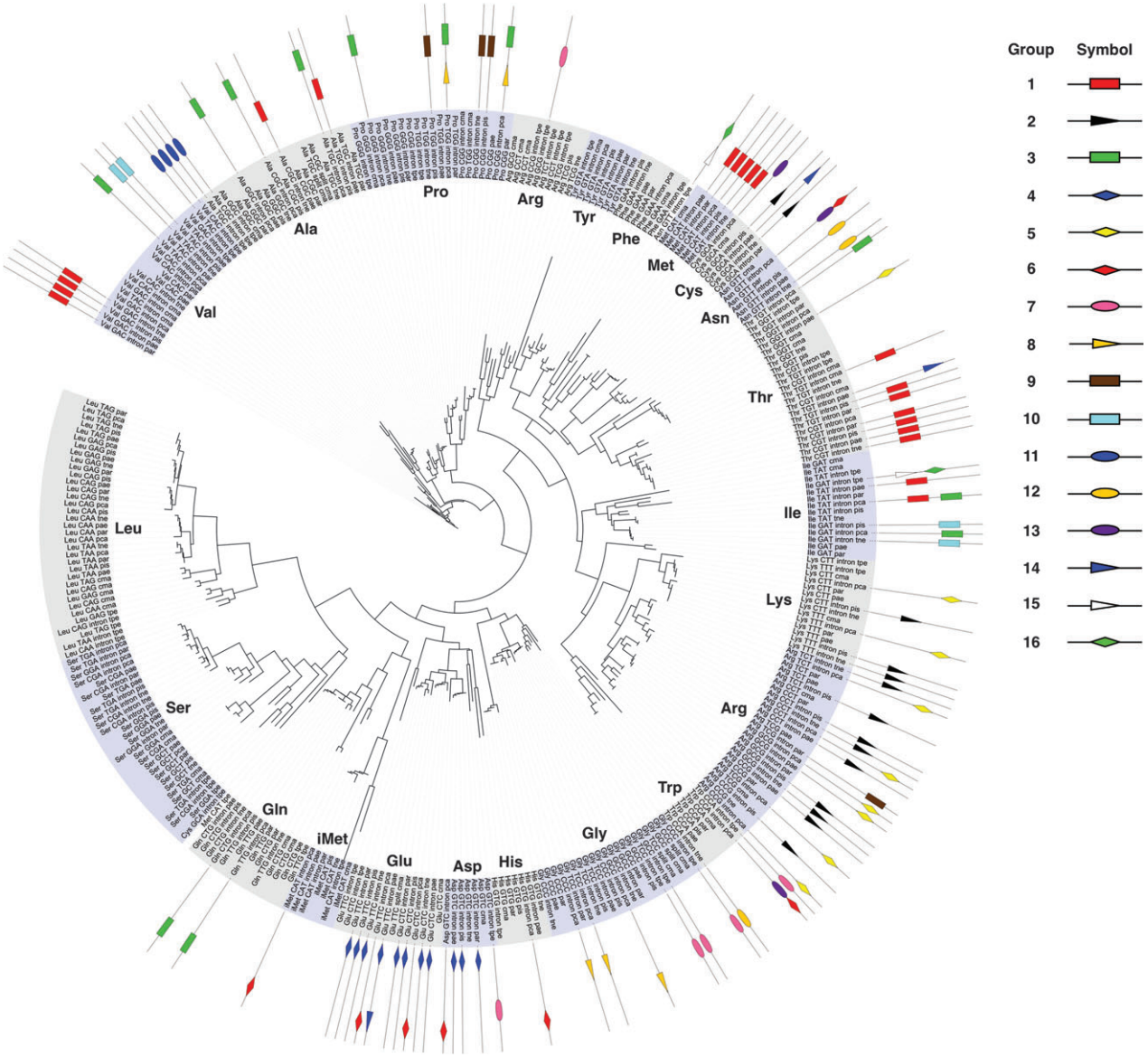

FIG. 1. Distribution of transposable introns on the phylogenetic tree of tRNA genes. A phylogenetic tree of the 323 tRNA genes in the seven Thermoproteales species was constructed based on the neighbor-joining method using ClustalW 2.0 (Larkin et al. 2007). In total, 113 transposable tRNA introns, classified into 16 individual groups (Group 1–Group 16) were mapped onto the tree, according to the symbols represented by different shapes and colors.

explains why the hBH structure was not associated with the transposable introns. We especially focused on three unique intron groups that have different sequence features (fig. 2*B*). Group 1 represents the largest population, with 22 intron sequences from six different Thermoproteales species. These introns are located specifically at position 29/30 within tRNA genes that correspond to five different amino acids (Ala, Ile, Met, Thr, and Val). In the $tRNA^{Ile}$(TAT) intron of *P. arsenaticum*, an extra seven nucleotides (CAGCCCC) was observed. These extra nucleotides probably result from the duplication of the adjacent sequence. Therefore, this intron may originally have been the same length as the other introns. Group 4 contains introns of exactly identical lengths. A multiple sequence alignment of the 11 introns of Group 4 shows identical lengths of 21 nt, and these introns occur only in the sequences of $tRNA^{Asp}$ and $tRNA^{Glu}$ of the four *Pyrobaculum* species. These introns are also unique in their insertion position at 3/4, which is the only position that requires a pre-tRNA 5′ leader sequence for the BHB motif. Group 3 is a species-specific intron group found in *P. calidifontis*, in which all the intron sequences are located at nucleotide position 53/54 in 12 tRNA genes corresponding to six different amino acids (Ala, Gln, Ile, Pro, Thr, and Val). Eight other groups are also species specific (Groups 6–8, 11, 13–16; see supplementary fig. S2, Supplementary Material online), indicating the late gain of these tRNA introns.

If these introns were acquired recently, do their insertion positions have any specific characteristics? To answer this question, we compared the proportion of transposable introns at each tRNA nucleotide position with the proportion of other archaeal tRNA introns. As shown in fig. 3,

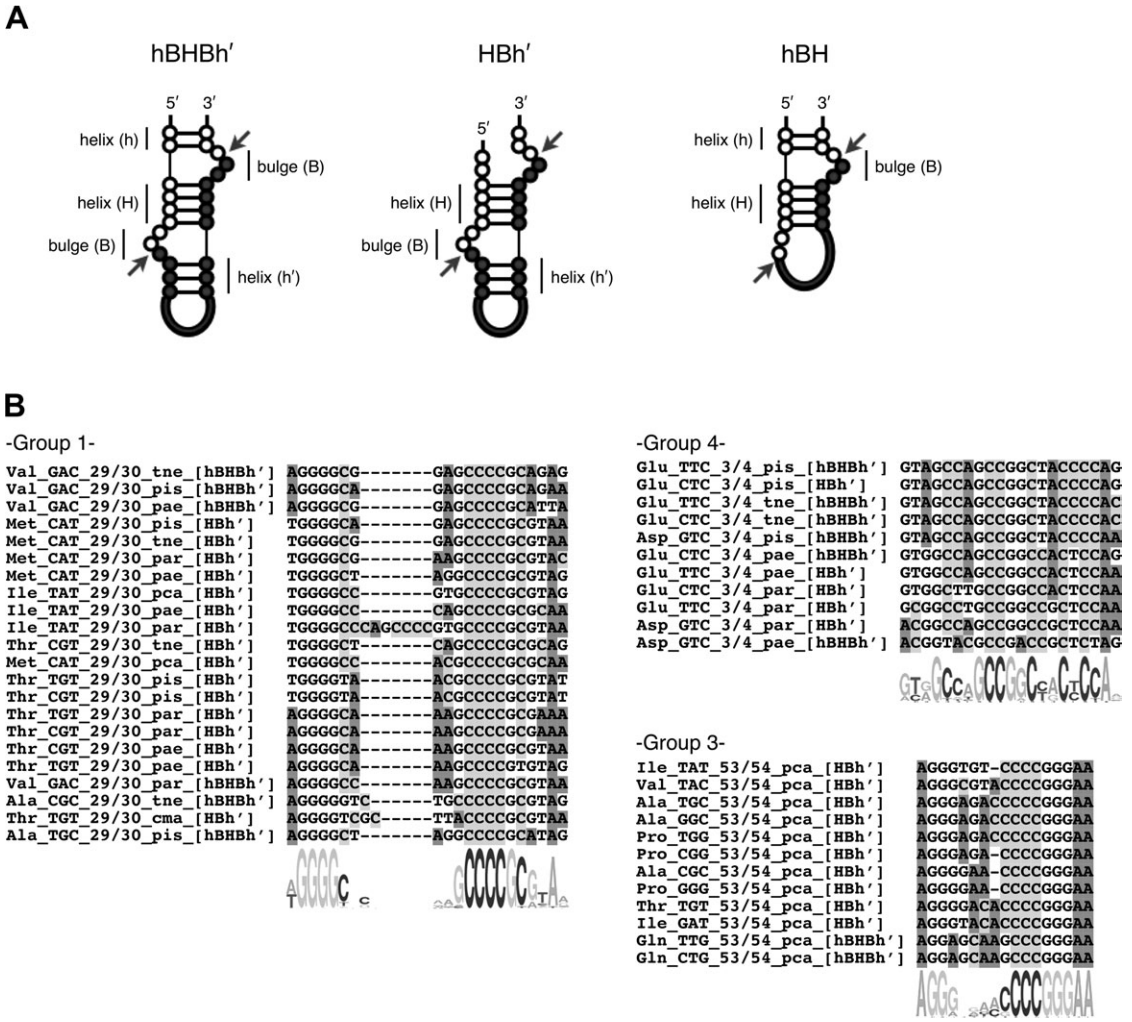

Fig. 2. Structural and sequence features of transposable tRNA introns. (*A*) Three types of splicing motifs are defined based on a previous study (Marck and Grosjean 2003): the strict hBHBh′ motif (left), the relaxed HBh′ motif (middle), and the relaxed hBH motif (right). The hBH motif is also known as the BHL motif (Tocchini-Valentini et al. 2005). The two arrows indicate the exon (white)–intron (gray) boundaries cleaved by the splicing endonuclease. (*B*) The tRNA intron sequence alignment is shown for three intron groups, each characterized by different features: the greatest number of members (Group 1); exactly equal lengths (Group 4); and specific to *P. calidifontis* (Group 3). The corresponding amino acid, anticodon, insertion position, species name, and type of splicing motif are given for each intron sequence. Consensus sequences were created with WebLogo (Crooks et al. 2004).

Note.—cma, *Caldivirga maquilingensis*; pae, *P. aerophilum*; par, *P. arsenaticum*; pca, *P. calidifontis*; pis, *P. islandicum*; and tne, *Thermoproteus neutrophilus*.

most of the transposable tRNA introns are located at noncanonical positions (other than 37/38). In other words, introns located at the canonical position are not amenable to transposition. Overall, approximately 80% (210 of 260) of the noncanonical tRNA introns currently found in the Archaea belong to the seven Thermoproteales species, and we have shown that about half of these noncanonical introns (109 of 210) are transposable. Furthermore, 10 nucleotide positions are occupied by more than 50% of the transposable introns, representing hot spots for intron transposition: 3/4, 29/30, 30/31, 43/44, 45/46, 51/52, 53/54, 56/57, 58/59, and 59/60. These positions include the only insertion site found in the acceptor stem (3/4) and the two major insertion sites in the anticodon stem (29/30 and 30/31), whereas the majority of these hot spots are distributed in the T-arm region of the tRNA, indicating that introns located in the T-arm region have been specifically acquired through intron transposition.

Possible Evolutionary Timing of tRNA Intron Transposition

To estimate the origin and evolution of the transposable introns, we mapped the possible times of the acquisition of the ancestral intron sequences of the 16 groups to the phylogenetic tree of Thermoproteales based on the conservation of the tRNA intron groups among the species (fig. 4*A*). We used the phylogenetic tree from the UCSC archaeal genome browser (http://archaea.ucsc.edu/; Schneider et al. 2006), which was constructed from a multiple genome alignment. The mapping results clearly demonstrate that the transposable tRNA introns emerged on at

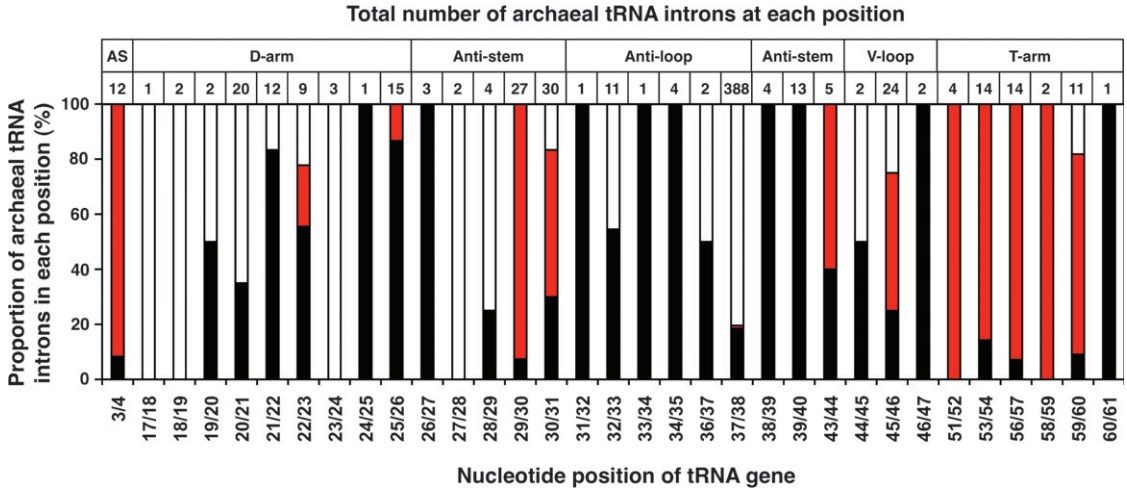

FIG. 3. Proportion of transposable introns at each nucleotide position in archaeal tRNAs. The boxed numbers represent the distribution of the total 647 tRNA introns found in 58 archaeal species. The tRNA introns are classified into three categories: transposable introns in the seven Thermoproteales species (red), other introns in the seven Thermoproteales species (black), and tRNA introns found in the remaining 51 archaeal species (white). Their proportions are given for each tRNA nucleotide position. NOTE.—AS, acceptor stem; D-arm, dihydrouridine arm; Anti-stem, anticodon stem; Anti-loop, anticodon loop; V-loop, variable loop; T-arm, TΨC arm.

least seven different occasions during the speciation of the Thermoproteales, which implies an early origin of the transposition mechanism. Because other crenarchaeal species also have relatively high proportions (20–50%) of intron-containing tRNAs (Sugahara et al. 2008), we performed the same multiple sequence alignment procedure on 217 intron sequences from 12 crenarchaeal species (see Materials and Methods for details) to identify transposable introns in crenarchaeal species. However, not a single transposable intron was identified in these species, suggesting that this mechanism arose in the early stage of Thermoproteales evolution and has been maintained since then. It is also clear that the number of intron groups correlates with the number of tRNA introns within a species. For instance, *P. calidifontis* and *Thermofilum pendens* each have four species-specific intron groups and a large number of tRNA

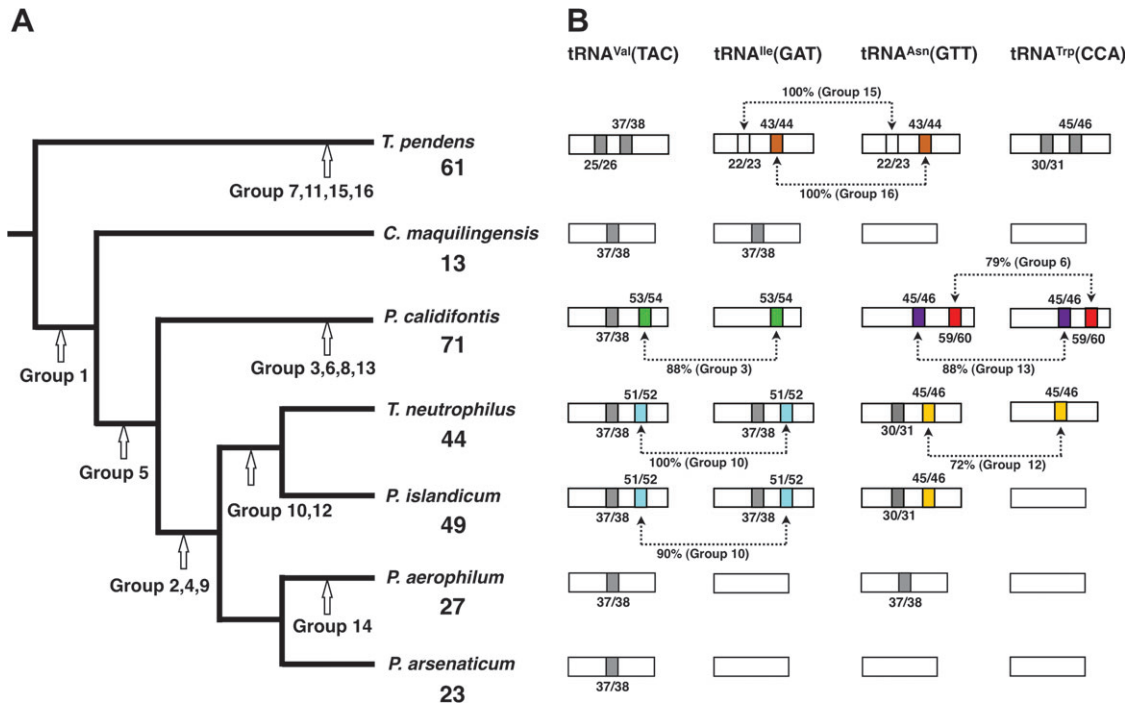

FIG. 4. Earliest insertion of the ancestral sequences of the 16 intron groups. (*A*) The estimated evolutionary timing of the insertion of the ancestral intron sequences is shown on the phylogenetic tree of seven Thermoproteales species (white arrow). The number of tRNA introns is given below the name of each species. (*B*) An example of tRNA transposition is shown for the four orthologous tRNA genes: $tRNA^{Val}$(TAC), $tRNA^{Ile}$(GAT), $tRNA^{Asn}$(GTT), and $tRNA^{Trp}$(CCA). Transposed introns (colored boxes) and other introns (gray boxes) are mapped onto the tRNA genes (rectangles). Group names and the sequence similarities of the transposed intron pairs are represented by dotted lines.

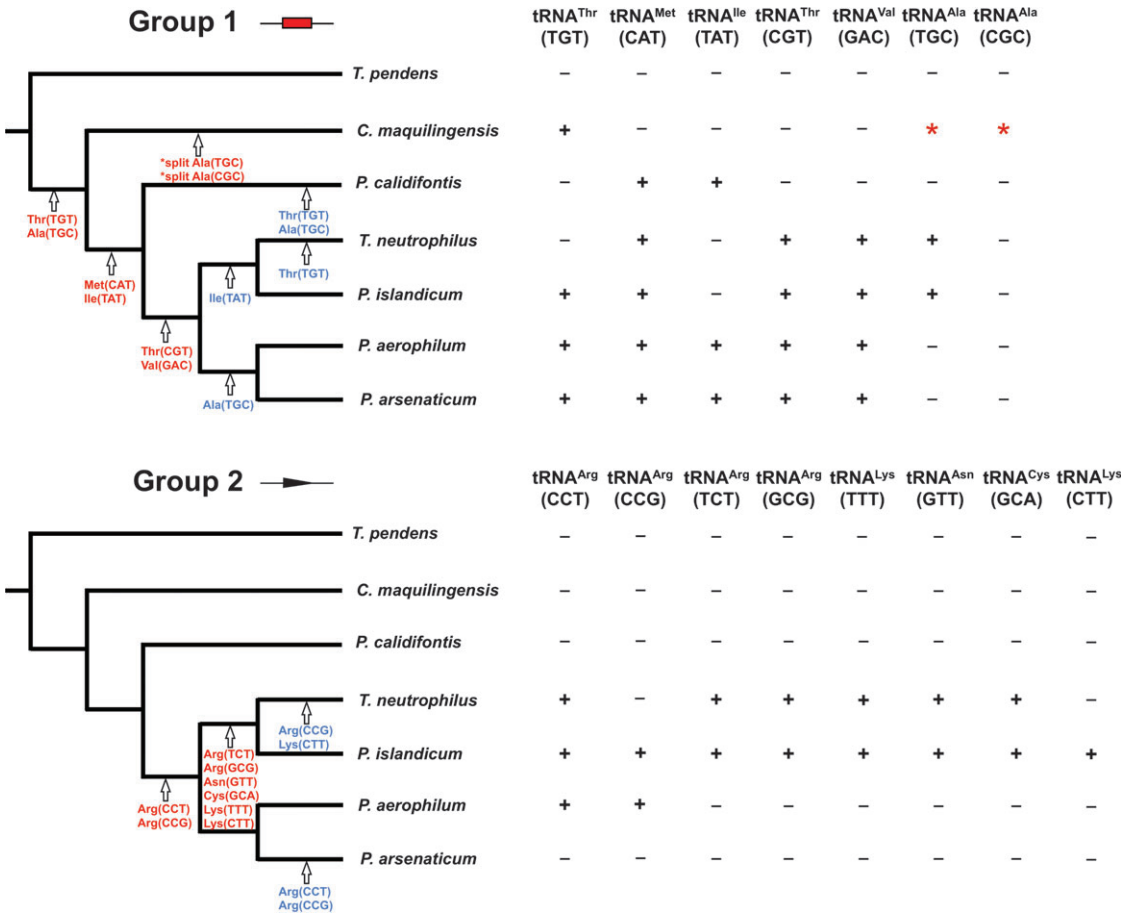

FIG. 5. Examples of tRNA intron gains and losses during the evolution of Thermoproteales. Possible timing of the gains and losses of transposable tRNA introns is shown for (*A*) Group 1 and (*B*) Group 2. The conservation patterns of the introns among species and tRNA genes are illustrated by the symbols '+' and '−', and the asterisk indicates the split tRNA. Arrows represent the evolutionary times of intron gain (red) and loss (blue) in the corresponding tRNA genes.

introns (71 and 61, respectively). In contrast, *C. maquilingensis* completely lacks species-specific introns and contains only 13 tRNA introns, which is the least observed in the seven Thermoproteales species (fig. 4*A*).

Concrete examples of transposable introns are illustrated by four orthologous tRNA genes that contain seven intron groups (fig. 4*B*). In *Thermoproteus neutrophilus* and *P. islandicum*, the tRNA^{Val}(TAC) and tRNA^{Ile}(GAT) genes both have two tRNA introns, at positions 37/38 and 51/52. A sequence comparison revealed that the introns located at canonical position 37/38 share no similarity in length or sequence, whereas in contrast, the 30-nt intron sequences located at position 51/52, which belong to Group 10, share 100% and 90% sequence identity in each species, respectively. Because the orthologous tRNA^{Val}(TAC) and tRNA^{Ile}(GAT) genes in other species do not contain similar introns, the insertion and transposition of this 30-nt intron sequence probably took place in the common ancestor of *Thermoproteus neutrophilus* and *P. islandicum*. Figure 4*B* also shows two examples of unique tRNA gene pairs containing two transposable tRNA introns simultaneously. Intriguingly, the two introns located at 22/23 (Group 15) and 43/44 (Group 16) in the tRNA^{Ile}(GAT) and tRNA^{Asn}(GTT) genes of *Thermofilum pendens* both share 100% identity, suggesting that the transposition of the two introns took place recently, within a short time frame.

To explore the origin of these newly acquired introns, we performed a nucleotide Blast search (http://blast.ncbi.nlm.nih.gov/) using transposable introns longer than 24 nt as queries against the GenBank nucleotide collection (nr/nt). However, no sequence other than introns belonging to the same group was detected below the threshold *E*-value of 0.1, revealing that these transposed intron sequences are not the derivatives of host genome fragments or any known transposable elements. It is possible that novel tRNA introns were acquired from the extracellular environment and later expanded within the species via a transposition mechanism.

We estimated the evolutionary timing of both intron gains and losses for the two major groups of introns (Group 1 and Group 2) and investigated the expansion of the tRNA introns throughout Thermoproteales evolution (fig. 5). The conservation profiles of the intron sequences in the tRNA genes showed that the ancestral sequence of Group 1 was first inserted into position 29/30 of either the tRNA^{Thr}(TGT) or tRNA^{Ala}(TGC) gene, and then transposed

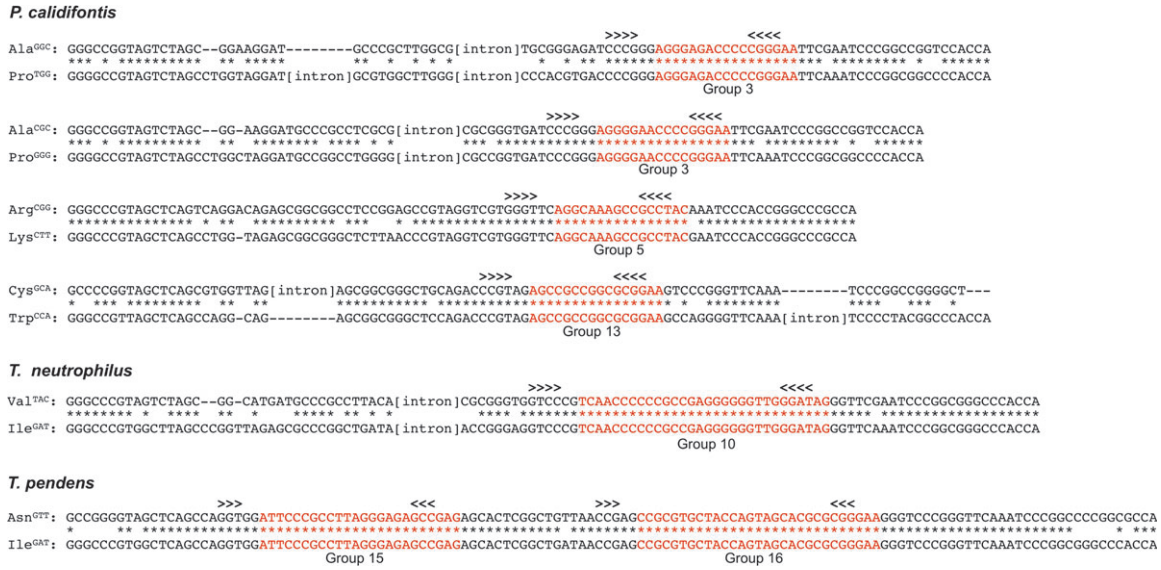

FIG. 6. Nucleotide sequence alignment of unrelated tRNA genes with identical intron sequences. Sequence alignment of six tRNA gene pairs with 100% identical introns (four pairs from *P. calidifontis* and one pair each from *Thermoproteus neutrophilus* and *Thermofilum pendens*) is shown. Identical intron regions are shown in red and other introns are represented by the character '[intron]'. Symbols ">" and "<" represent the hybridization of the central helix of the BHB motif.

into other tRNA genes in the later stages of evolution. Because the tRNA^{Ala}(TGC) gene in *C. maquilingensis* is the only split tRNA^{Ala}(TGC) gene found in the seven species, the fragmentation of tRNA^{Ala}(TGC) seems to have occurred after the insertion of the intron. Furthermore, at least in this case, the highly conserved promoters (93%) and leader sequences (100%) of the two tRNA^{Ala}s (Fujishima et al. 2009) support a later duplication of the split tRNA^{Ala}(TGC) gene, producing a synonymous split tRNA^{Ala}(CGC) gene. Similarly, an ancestral Group 2 intron is estimated to have first emerged at position 30/31 of the tRNA^{Arg}(CCT) or tRNA^{Arg}(CCG) gene after the divergence of *P. calidifontis*, and was later transposed into six different tRNA genes in the common ancestor of *Thermoproteus neutrophilus* and *P. islandicum*. It should be noted that these results also show that there is a trend toward intron loss during the transposition event, although its frequency is relatively low compared with that of intron gain.

BHB Splicing Motif is One of the Limiting Factors Required for tRNA Intron Transposition

Transposable intron sequences are relatively short compared with those of the self-splicing and spliceosomal introns (Hong et al. 2006), ranging from 15 to 32 nt. To clarify the rules behind the mechanism that facilitates the mobility of such short introns, we focused on the evolutionarily distinct tRNA genes with 100% identical introns, a good example of the most recently transposed introns. We identified six such tRNA pairs in three different species (*P. calidifontis*, *Thermoproteus neutrophilus*, and *Thermofilum pendens*), and their pre-tRNA sequences, including both the exons and identical introns, were aligned (fig. 6). Whereas the overall exon sequence identities varied from 75% to 87%, the nucleotides forming the central helix of the BHB motif were perfectly conserved in all six tRNA pairs. Based on these results, we assume that intron transposition takes place in a limited manner, only when the newly inserted intron can form the BHB motif with the adjacent exon sequence to facilitate the splicing reaction in the recipient tRNA gene.

Using a computational analysis, we tested this hypothesis against a variety of intron groups. The intron sequences from 13 tRNA gene pairs representing 11 intron groups were computationally inserted into various tRNA genes and positions to produce artificial intron-containing tRNAs (see Materials and Methods). The splicing potential of each tRNA was assessed with the tRNA prediction software SPLITS (Sugahara et al. 2006). We examined the effect of the BHB splicing motif on two features: the insertion position and the tRNA variation. First, the spliceability of the artificial introns at each nucleotide position in its own tRNA was strictly regulated by the structural requirement for a strict/relaxed BHB motif, and the proportions of spliceable positions ranged from 1% to 9% (Table 2). This result is consistent with the actual observation that almost no transposition has occurred among different positions in the same tRNA intron group. Second, when an intron was inserted at the same position in other tRNA genes, the number of spliceable tRNA genes varied greatly among the different intron groups. The introns of Group 4 and Group 16 showed a particularly significant reduction in the number of spliceable tRNA genes, to 7% and 11%, respectively (table 2). We successfully reconstructed the tRNA gene variations of the Group 4 introns and found that they only inserted at position 3/4 of tRNA^{Asp} and tRNA^{Glu}. This strong selection bias was maintained by the requirement for the tetranucleotide GGGG in the 5′ leader sequence to form either a strict or relaxed BHB

Table 2. In Silico Splicing Analysis of tRNA Gene Pairs with Highly Similar Introns.

Observed Introns						In Silico Splicing Analysis			
Species	Identity (%)	tRNA Gene Pairs	Group	Position	Length (nt)	Spliceable Introns in Different Positions[a]	Spliceable Introns in Different tRNAs[b]	Amino Acid Variations in the Spliceable tRNAs	Free Energy of the BHB Motif[c]
Pyrobaculum arsenaticum	90	Thr(CGT) Val(GAC)	1	29/30	20	2 (3%)	26 (57%)	Ala, Asp, Gln, Glu, Ile, Met, Phe, Pro, Ser, Thr, Tyr, Val	−5.00/−19.50
P. calidifontis	100	Ala(GGC) Pro(TGG)	3	53/54	17	4 (5%)	28 (61%)	Ala, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Met, Phe, Pro, Thr, Tvr, Val	−8.70/−10.50
	100	Ala(CGC) Pro(GGG)		53/54	16	3 (4%)	28 (61%)	Ala, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Met, Phe, Pro, Thr, Tvr, Val	−6.40/−8.90
	100	Arg(CCG) Lys(CTT)	5	56/57	16	3 (4%)	33 (72%)	Arg, Asn, Asp, Cys, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tvr, Val	−5.60/−5.80
	100	Cys(GCA) Trp(CCA)	13	45/46	16	1 (1%)	14 (30%)	Arg, Asn, Cys, Gly, His, Lys, Tip	−6.80/−6.90
	95	Asp(GTC) Glu(GTC)	6	59/60	19	1 (1%)	44 (96%)	Ala, Arg, Asn, Asp, Cys, Gin, Glu, Gly, His, Ile, Leu, Lvs, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val	−13.80/−23.60
	94	Ile(GAT) Thr(TGT)	3	53/54	17	5 (7%)	28 (61%)	Ala, Asn, Asp, Cys, Gin, Glu, Gly, His, Lle, Met, Phe, Pro, Thr, Tvr, Val	−8.90/−11.40
P. islandicum	95	Asp(GTC) Glu(TTC)	4	3/4	21	2 (3%)	3 (07%)	Asp, Glu	−10.80/−13.80
Thermoproteus neutrophlius	100	Ile(GAT) Val(TAC)	10	51/52	30	3 (4%)	19 (42%)	Ala, Asn, Cys, Gly, iMet, Ile, Pro, Thr, Tyr, Val	−22.50/−22.80
	90	Arg(TCT) Cys(GCA)	2	30/31	20	2 (3%)	14 (31%)	Arg, Asn, Cys, Gly, His, Lys, Trp	−10.60/−14.60
T. pendens	100	Asn(GTT) Ile(GAT)	15	2223	24	7 (9%)	18 (39%)	Asn, Gly, iMet, Ile, Lys, Pro, Thr, Trp, Tyr, Val	−2.60/−4.50
	100	Asn(GTT) Ile(GAT)	16	43/44	29	5 (7%)	5 (11%)	Asn, Ile, Phe, Tyr	−19.50
	94	Asp(GTC) Gly(CCC)	7	45/46	17,18	2 (3%)	18 (39%)	Arg, Asp, Gln, Gly, Leu, Lys	−8.50/−10.20

[a] Number of tRNA nucleotide positions that can splice introns inserted at each nucleotide positions (from 1/2 to 73/74). This analysis is performed in either one of the tRNA gene pairs (indicated in bold font at "tRNA gene pairs" column). The original position is excluded from the count.
[b] Number of tRNA genes with different isoacceptors that can splice introns inserted at the original position. Percentage is calculated taking the total number of tRNA genes in each species to be 100%. In all cases, the results of the two highly similar introns were combined due to the same output.
[c] Maximum and minimum free energy of the BHB motif found in the artificial intron-containing tRNAs predicted by SPLITS.

motif. On the contrary, Group 6 introns require the ubiquitously conserved nucleotides TTC(A/G) in the T-loop region, allowing 96% of all tRNA genes to be their recipients. Finally, a 2D correlation analysis revealed a positive correlation between the type of actually transposed tRNA genes and the artificially spliceable tRNA genes (fig. 7A). The correlation coefficient was $r = 0.67$, suggesting that the BHB motif is indeed a key limiting factor for the transposition event. However, the actual transposition was further restricted to a limited combination of tRNA gene pairs (fig. 7B) and no transposable introns were found in the four types of tRNA genes $tRNA^{Leu}$, $tRNA^{Phe}$, $tRNA^{Tyr}$, and $tRNA^{Ser}$ (although they are predicted to be the recipients of some intron groups by *in silico* splicing analysis). These two facts imply that yet another unknown limiting factor affects the transposition mechanism.

Discussion

This is the first clear example of the unique mobility of the enzymatically cleaved tRNA introns, characterized by a comprehensive sequence analysis of archaeal tRNAs and their introns. Enzymatically cleaved tRNA introns (recognized by a splicing endonuclease) are universally present in the domains Archaea and Eukarya (Chan and Lowe 2009) and some introns even encode C/D box RNA, which is required for tRNA modification (Singh et al. 2004; Clouet-d'Orval et al. 2005). Compared with the progress in the study of spliceosomal introns (Roy and Gilbert 2006), the origin and evolution of tRNA introns remain unclear. In this study, we have shown that transposition events have occurred on an unexpected scale, including approximately 40% of the total tRNA introns in the seven Thermoproteales species. However, because we excluded the intron groups that are conserved only in synonymous tRNA genes from our data set, the actual proportion of transposable introns may be even greater.

The observed transposable introns are widely distributed among various tRNA genes (fig. 1). However, the transposition phenomenon was not observed in the tRNAs corresponding to four amino acids, Leu, Phe, Ser and Tyr.

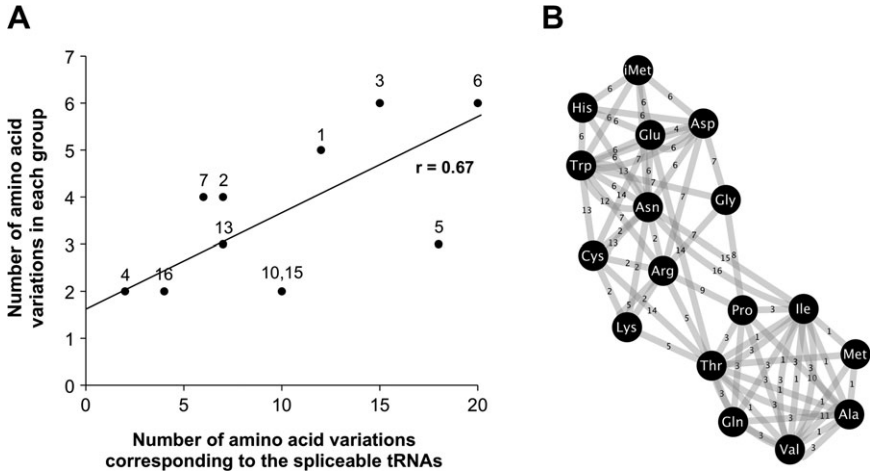

FIG. 7. Comparison of actual and predicted intron transposition. (*A*) Two-dimensional correlation analysis of 11 transposable intron groups (black spots). The *x* axis represents the amino acid variation of spliceable tRNAs estimated by in silico splicing analysis. The *y* axis represents the actual variation observed in each intron group (table 1). The correlation coefficient (*r*) is indicated next to the regression bar. The intron group names are given on the spots. (*B*) Network visualization of the actual intron transposition that has occurred among the tRNA genes. The tRNA genes are classified based on their corresponding amino acids (circles) and each connection (line) represents the intron transposition of one of the 16 intron groups.

Considering that $tRNA^{Leu}$ and $tRNA^{Ser}$ constitute approximately 20% of the whole tRNA gene population, a random occurrence of intron transposition seems unlikely. Because $tRNA^{Leu\ and\ tRNA}Ser$ are the only tRNAs that have a long variable loop, certain structural or sequence requirements for intron transposition were expected. One of the challenges in this study was to determine the underlying limiting factor for intron transposition. In this regard, our in silico splicing analysis showed that the BHB motif is strictly required for transposition. In the Archaea, each tRNA isoacceptor is generally encoded by a single gene, so an archaeon will only be viable when the intron is inserted at a position in the tRNA at which it can form the BHB motif. Accordingly, the requirement for nonredundant tRNA genes further defines the positions and types of tRNA genes amenable to transposition (table 2), which is consistent with the actual trends in insertion positions (table 1) and receptive tRNA genes (fig. 7*A*). It is noteworthy that the BHB motif is a specifically RNA secondary structure, so transposition may take place at the RNA level rather than at the DNA level. Furthermore, introns that contain the BHB motif are also found in archaeal rRNA (Tang et al. 2002) and archaeal mRNA sequences (Yoshinari et al. 2006), indicating the wide use of the BHB motif as a landmark for recombination at the transcript level. These issues should be addressed in future studies when the molecular mechanisms underlying RNA processing (splicing and ligation) have been determined.

We also confirmed that this unique mechanism contributes to the late gain of introns in the deep-branching archaeal order Thermoproteales. The rapid increase in tRNA introns is especially pronounced in *P. calidifontis* and *T. pendens*, in which about 90% of the tRNA genes contain introns (Sugahara et al. 2008). These two species each contains four species-specific intron groups, indicating that intron transposition is highly active in these species. In contrast, the archaeon *C. maquilingensis*, rooted between the two extremely intron-rich archaeal species, lacks a species-specific intron group. Therefore, we assume that transposition has been inactivated in this archaeon or it has acquired a specific mechanism during its evolution to prevent transposition. Interestingly, ten split tRNA genes have recently been identified in the genome of *C. maquilingensis*, of which the two split $tRNA^{Ala}$s are separated at position 29/30, known as the insertion site of introns belonging to Group 1. The leader sequences of these split $tRNA^{Ala}$s share 90% sequence identity with the Group 1 intron inserted in the orthologous $tRNA^{Ala}$ gene in *P. islandicum*, indicating a clear evolutionary relationship between the intron-containing and split tRNAs (Fujishima et al. 2009). Much less is known about the origin of split tRNAs, although we could roughly estimate the origin of the split tRNA genes according to the trends in the gain and loss of Group 1 introns. Therefore, at least for the case of the split $tRNA^{Ala}$s, the fragmentation of the $tRNA^{Ala}$ gene may have occurred in addition to the insertion of the intron sequence at position 29/30 (fig. 5), supporting the split-late hypothesis (Randau and Soll 2008).

Finally, the question remains regarding the origin of the transposable tRNA introns. No related sequences were found with a homology search of the NCBI nucleotide database, except for tRNA introns belonging to the same group. This observation clearly indicates that this transposition is strongly restricted to tRNA genes and therefore, the newly acquired intron sequences could have derived from unregistered microbial or viral genomes. Many crenarchaeal viruses are known to live in the same hot acidic environments as Thermoproteales (Ortmann et al. 2006) and some of them do integrate into the distal regions of tRNA genes (Krupovic and Bamford 2008). Therefore, it is possible that these short introns are fragments of infecting archaeal viruses.

Supplementary Material

Supplementary Figs. S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank all members of the RNA group (Institute for Advanced Biosciences, Keio University) for their insightful discussions. This work was supported in part by the Japan Society for the Promotion of Science (JSPS), the Yamagata Prefectural Government, the city of Tsuruoka, Japan and the Institute for Fermentation, Osaka, Japan. These sources of funding had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* 37:D93–D97.

Clouet-d'Orval B, Gaspin C, Mougin A. 2005. Two different mechanisms for tRNA ribose methylation in Archaea: a short survey. *Biochimie* 87:889–895.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–1190.

Di Giulio M. 2006. The non-monophyletic origin of the tRNA molecule and the origin of genes only after the evolutionary stage of the last universal common ancestor (LUCA). *J Theor Biol.* 240:343–352.

Di Giulio M. 2008. Transfer RNA genes in pieces are an ancestral character. *EMBO Rep.* 9:820; author reply. 820–821.

Diener JL, Moore PB. 1998. Solution structure of a substrate for the archaeal pre-tRNA splicing endonucleases: the bulge-helix-bulge motif. *Mol Cell.* 1:883–894.

Fujishima K, Sugahara J, Kikuta K, Hirano R, Sato A, Tomita M, Kanai A. 2009. Tri-split tRNA is a transfer RNA made from 3 transcripts that provides insight into the evolution of fragmented tRNAs in archaea. *Proc Natl Acad Sci U S A.* 106:2683–2687.

Fujishima K, Sugahara J, Tomita M, Kanai A. 2008. Sequence evidence in the archaeal genomes that tRNAs emerged through the combination of ancestral genes as 5' and 3' tRNA halves. *PLoS ONE.* 3:e1622.

Heinemann IU, Soll D, Randau L. 2009. Transfer RNA processing in archaea: unusual pathways and enzymes. *FEBS Lett.* 584: 303–309.

Hong X, Scofield DG, Lynch M. 2006. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol.* 23:2392–2404.

Kawach O, Voss C, Wolff J, Hadfi K, Maier UG, Zauner S. 2005. Unique tRNA introns of an enslaved algal cell. *Mol Biol Evol.* 22:1694–1701.

Krupovic M, Bamford DH. 2008. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. *Virology* 375:292–300.

Larkin MA, Blackshields G, Brown NP, et al. (13 co-authors). 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127–128.

Li H, Trotta CR, Abelson J. 1998. Crystal structure and evolution of a transfer RNA splicing enzyme. *Science* 280:279–284.

Marck C, Grosjean H. 2002. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA* 8:1189–1232.

Marck C, Grosjean H. 2003. Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. *RNA* 9:1516–1531.

Maruyama S, Sugahara J, Kanai A, Nozaki H. 2010. Permuted tRNA genes in the nuclear and nucleomorph genomes of photosynthetic eukaryotes. *Mol Biol Evol.* 27:1070–1076.

Ortmann AC, Wiedenheft B, Douglas T, Young M. 2006. Hot crenarchaeal viruses reveal deep evolutionary connections. *Nat Rev Microbiol.* 4:520–528.

Randau L, Munch R, Hohn MJ, Jahn D, Soll D. 2005. Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves. *Nature* 433:537–541.

Randau L, Soll D. 2008. Transfer RNA genes in pieces. *EMBO Rep.* 9:623–628.

Reinhold-Hurek B, Shub DA. 1992. Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature* 357:173–176.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet.* 7:211–221.

Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM. 2006. The UCSC Archaeal Genome Browser. *Nucleic Acids Res.* 34:D407–D410.

Singh SK, Gurha P, Tran EJ, Maxwell ES, Gupta R. 2004. Sequential 2'-O-methylation of archaeal pre-tRNATrp nucleotides is guided by the intron-encoded but trans-acting box C/D ribonucleoprotein of pre-tRNA. *J Biol Chem.* 279:47661–47671.

Soma A, Onodera A, Sugahara J, Kanai A, Yachie N, Tomita M, Kawamura F, Sekine Y. 2007. Permuted tRNA genes expressed via a circular RNA intermediate in Cyanidioschyzon merolae. *Science* 318:450–453.

Sugahara J, Fujishima K, Morita K, Tomita M, Kanai A. 2009. Disrupted tRNA gene diversity and possible evolutionary scenarios. *J Mol Evol.* 69:497–504.

Sugahara J, Kikuta K, Fujishima K, Yachie N, Tomita M, Kanai A. 2008. Comprehensive analysis of archaeal tRNA genes reveals rapid increase of tRNA introns in the order thermoproteales. *Mol Biol Evol.* 25:2709–2716.

Sugahara J, Yachie N, Sekine Y, Soma A, Matsui M, Tomita M, Kanai A. 2006. SPLITS: a new program for predicting split and intron-containing tRNA genes at the genome level. *In Silico Biol.* 6:411–418.

Tang TH, Rozhdestvensky TS, d'Orval BC, Bortolin ML, Huber H, Charpentier B, Branlant C, Bachellerie JP, Brosius J, Huttenhofer A. 2002. RNomics in Archaea reveals a further link between splicing of archaeal introns and rRNA processing. *Nucleic Acids Res.* 30:921–930.

Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP. 2005. Coevolution of tRNA intron motifs and tRNA endonuclease architecture in Archaea. *Proc Natl Acad Sci U S A.* 102:15418–15422.

Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP. 2007. The dawn of dominance by the mature domain in tRNA splicing. *Proc Natl Acad Sci U S A.* 104:12300–12305.

Xue S, Calvin K, Li H. 2006. RNA recognition and cleavage by a splicing endonuclease. *Science* 312:906–910.

Yoshinari S, Itoh T, Hallam SJ, DeLong EF, Yokobori S, Yamagishi A, Oshima T, Kita K, Watanabe Y. 2006. Archaeal pre-mRNA splicing: a connection to hetero-oligomeric splicing endonuclease. *Biochem Biophys Res Commun.* 346:1024–1032.

Yoshinari S, Shiba T, Inaoka DK, Itoh T, Kurisu G, Harada S, Kita K, Watanabe Y. 2009. Functional importance of crenarchaea-specific extra-loop revealed by an X-ray structure of a hetero-tetrameric crenarchaeal splicing endonuclease. *Nucleic Acids Res.* 37:4787–4798.