1238–1250 *Nucleic Acids Research, 2001, Vol. 29, No. 5*

Phylogenetic relationships among group II intron ORFs

Steven Zimmerly*, Georg Hausner and Xu-chu Wu

Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada

Received September 7, 2000; Revised November 30, 2000; Accepted January 17, 2001 DDBJ/EMBL/GenBank accession no. ALIGN_000044

ABSTRACT

Group II introns are widely believed to have been ancestors of spliceosomal introns, yet little is known about their own evolutionary history. In order to address the evolution of mobile group II introns, we have compiled 71 open reading frames (ORFs) related to group II intron reverse transcriptases and subjected their derived amino acid sequences to phylogenetic analysis. The phylogenetic tree was rooted with reverse transcriptases (RTs) of non-long terminal repeat retroelements, and the inferred phylogeny reveals two major clusters which we term the mitochondrial and chloroplast-like lineages. Bacterial ORFs are mainly positioned at the bases of the two lineages but with weak bootstrap support. The data give an overview of an apparently high degree of horizontal transfer of group II intron ORFs, mostly among related organisms but also between organelles and bacteria. The Zn domain (nuclease) and YADD motif (RT active site) were lost multiple times during evolution. Differences in domain structures suggest that the oldest ORFs were concise, while the ORF in the mitochondrial lineage subsequently expanded in three locations. The data are consistent with a bacterial origin for mobile group II introns.

INTRODUCTION

Group II introns are self-splicing RNAs that are widely believed to have been ancestors of nuclear pre-mRNA introns (1). Some group II introns are also active retroelements due to reverse transcriptases (RTs) encoded within the introns (2). These mobile group II introns are linked mechanistically and phylogenetically to non-long terminal repeat (non-LTR) elements, an abundant class of retroelements in eukaryotes (3). Based on these relationships, it has been proposed that group II introns might be ancestors of non-LTR retroelements as well as spliceosomal introns (3,4). Yet despite the potential importance of group II introns to the evolution of eukaryotic genomes, little information is available about the evolutionary history of group II introns themselves.

Mobile group II introns consist of an ~600 nt self-splicing RNA structure surrounding an ~2 kb open reading frame (ORF; Fig. 1). The conserved secondary structure of the intron comprises six domains (5), and the ORF is invariably located in domain IV (Fig. 1A and B). The ORF itself is divided into conserved domains RT, X and Zn. The RT domain forms the bulk of the ORF and consists of subdomains 1–7 [palm and finger regions in the crystal structure of HIV-RT (6)]. Subdomain 0 can be considered an N-terminal extension of the RT domain and is conserved among non-LTR RTs. [Domain 0 was formerly called domain Z but was more recently renamed domain 0 in non-LTR RTs (7)]. Domain X (analogous to the thumb structure of HIV-RT) is implicated in the splicing, or maturase, function of the RT protein (8), while the Zn domain contains a potential zinc finger, and contributes a nuclease activity. The Zn domain is related to a family of bacterial colicin and pyocin nucleases, as well as to some group I intron ORFs (9,10).

Splicing and mobility of group II introns require catalytic activities of both the intron and intron-encoded protein (see 2 for complete description). In brief, intron splicing occurs *in vivo* when the RT protein binds to unspliced intron transcript and stimulates the inherently RNA-catalyzed splicing reaction. Mobility of the intron is initiated when the RNP product of splicing (RT bound to lariat intron) encounters a homing site DNA (fused exons lacking intron). The intron reverse splices either partially or completely into the exon sequences of DNA. Then the Zn domain of the RT nicks the antisense strand of the DNA target 9 or 10 bp downstream of the exon junction, and the RT reverse transcribes the intron using the cleaved antisense strand as a primer. Intron insertion is completed by host repair enzymes.

Our understanding of the evolutionary history of group II introns is based primarily on speculation and general observations. It has been proposed that mobile group II introns were created when an RT was inserted into a pre-existing group II intron (11,12), although other possibilities have been considered (13). Such a formative event may have occurred in bacteria, after which introns migrated to mitochondria and chloroplasts. The theory of a bacterial origin was prompted by the discovery of ORF-containing group II introns in bacterial species related to the ancestors of organelles [*Calothrix*,

*To whom correspondence should be addressed. Tel: +1 403 220 7933; Fax: +1 403 289 9311; Email: zimmerly@ucalgary.ca
Present addresses:
Georg Hausner, Department of Botany, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada
Xu-chu Wu, Department of Cell Biology and Anatomy, University of Calgary School of Medicine, Calgary, Alberta T2N 4N1, Canada

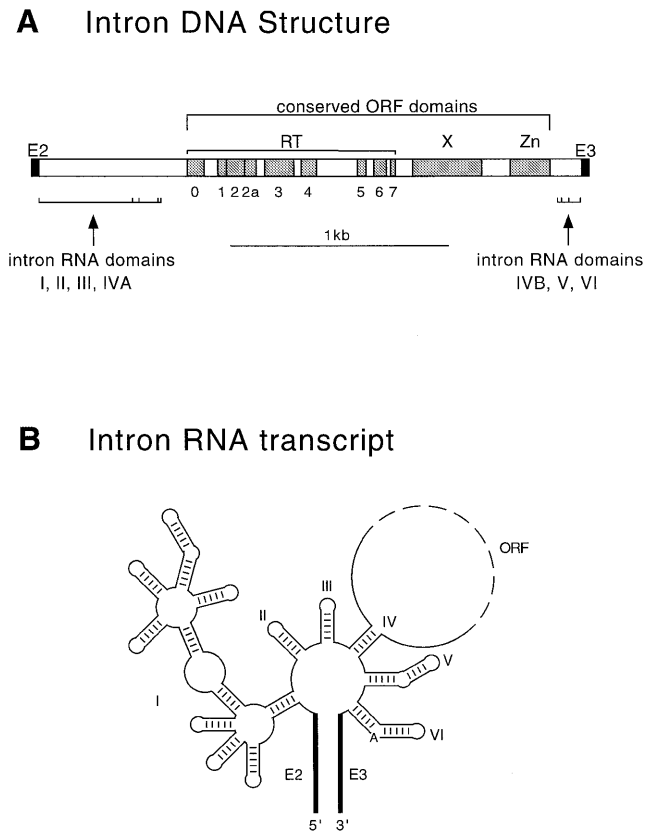


Figure 1. Structure of a typical mobile group II intron based on *S.c.cox1*I2 of yeast mitochondria. (**A**) DNA structure showing the upstream and downstream exons (E2, E3), intron domains I–VI, and the conserved domains of the ORF [RT(0–7), X, Zn] (drawn to scale). (**B**) An unspliced intron transcript of *S.c.cox1*I2 showing a simplified secondary structure of the six conserved intron structural domains, with the ORF looped out of domain IV (not drawn to scale).

Azotobacter and *Escherichia coli* (14,15)]. Group II introns have since been reported in diverse bacterial species including *Lactococcus* (16), *Clostridium* (17), *Pseudomonas* (18), *Sinorhizobium* (19), *Bacillus* (20) and *Sphingomonas* (21), indicating a widespread presence of group II introns in bacteria.

Deciphering the history of group II introns is complicated by the fact that group II introns are inherited horizontally as well as vertically. Horizontal transfer is suggested by an idiosyncratic distribution of introns among species and strains, and also by the observation that related introns are sometimes found in seemingly less related host genes or host organisms (11). For example, *Kluyveromyces lactis cox1*I1 and *Saccharomyces cerevisiae cox1*I2 are 96% identical in DNA sequence (intron and ORF) and are located at the same site within the *cox1* gene, yet the *cox1* genes are 88% identical (22). Another example is the group II intron ORF *ltrA* of *Lactococcus lactis*, which is more closely related to ORFs of mitochondrial introns than to other bacterial intron ORFs (16).

Vertical inheritance of group II introns is best exemplified by introns with degenerate ORFs. The most extreme example is the *mat*K family of proteins found in the chloroplast *trnK* genes of higher plants. *Mat*K proteins do not resemble other group II intron-encoded proteins except for some conservation in RT subdomains 5–7 and domain X. It is believed that *mat*K ORFs lost mobility functions but retained their maturase function which is associated with the domain X motif still present in these ORFs (8,23).

To help elucidate the evolutionary past of mobile group II introns, we have undertaken phylogenetic analysis of the intron-encoded proteins. Evolution of these proteins is obviously distinct from the evolution of the intron RNA structure itself, although one report indicated coevolution for a limited subset of RT-encoding introns (24). We have compiled group II intron ORFs from the databases, including a number of bacterial ORFs not explicitly reported as group II introns in the literature. Alignment of the ORF sequences and construction of a phylogenetic tree suggest a model for the history of mobile group II introns. This model gives an overview of horizontal versus vertical inheritance and predicts how the ORF structure, and possibly activities, evolved.

MATERIALS AND METHODS

Compilation of sequences

Putative group II intron ORFs were identified by BLAST searches (25) at the National Center for Biotechnology Information (NCBI) web site http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html using a selection of known group II intron ORFs as query sequences. After the phylogenetic tree was constructed, further searches were done based on representatives from each of the major branches to ensure all ORFs related to each subfamily were identified. Sequences were aligned by PILEUP (Wisconsin Package v.8; Genetics Computer Group, Madison, WI) on the ANGIS (Australian National Genomic Information Service) supercomputing server (http://morgan.angis.su.oz.au/) and by CLUSTAL X (26), followed by manual refinement using the editing programs GeneDoc (K.B.Nicholas and H.B.Nicholas Jr, for Windows; http://www.psc.edu/biomed/genedoc/) and SeqApp (D.G.Gilbert, for Macintosh; http://iubio.bio.indiana.edu/soft/molbio/seqpup/). Consensus sequences were calculated using the GCG program PRETTY on the ANGIS server, taking into account conservative substitutions.

Phylogenetic estimates were generated by the programs contained within the PHYLIP package (Version 3.573c; http://evolution.genetics.washington.edu/phylip/getme.html). Neighbor-joining analysis utilized PROTDIST (setting: Dayhoff PAM250 substitution matrix) and NEIGHBOR programs. Parsimony analysis was with PROTPARS (protein parsimony algorithm, version 3.55c). SEQBOOT and CONSENSE programs were used for bootstrap analysis and generation of the majority rule consensus trees (27). Maximum-likelihood trees were inferred with PUZZLE 4.02 (ftp://ebi.ac.uk/pub/software; setting: JTT correction matrix and frequency of amino acid usage estimated from data), with 100 quartet puzzling steps (28).

Sequences omitted from phylogenetic analysis were: C.p.*psbC*I4, *P.p.rtl*, *A.s.* (missing domains); *P.co.cox1*I1 and *P.cu.ND5*I1 (redundance); and *E.g.psbD*I8 (divergence). *S.ma.* was omitted because it lacked a group II intron structure and its inclusion reduced statistical support for the tree; we conclude that the *S.ma.* ORF lost its intron structure and is degenerated. The approximate positions of these ORFs on the phylogenetic tree based on BLAST search similarities are: *A.s.*, *E.g.psbD*I8,

chloroplast-like group; *M.l., P.p.rtl,* algal group; *C.p.psbC*I4, euglenoid group; *S.ma.*, unclear.

RESULTS

Compilation of ORF sequences

Group II intron ORF sequences were identified by BLAST searches of GenBank to find relatives of known group II intron ORFs (Materials and Methods). The compiled 71 ORFs and 14 ORF fragments are displayed in Tables 1–3 along with their features and accession numbers. The tables include 40 mitochondrial, 11 chloroplast and 20 bacterial ORFs. The 14 bacterial ORF fragments are listed in the footnote to Table 3. In comparison with the previous compilation (8), additions are: 20 mitochondrial ORFs (five related *nad1*I4 ORFs from higher plants, eight ORFs of green, red and brown algae, seven fungal ORFs), five chloroplast ORFs (four related euglenoid *psbC*I4 ORFs; one cryptomonad ORF) and 19 bacterial ORFs. In cases where nearly identical intron ORFs were reported in the same species (e.g. *E.coli* intron ORFs of accession numbers AE000133 and D37918), only one entry was included in the table. Only five *mat*R ORFs (*nad1*I4 ORFs in higher plants) are included although over 100 are reported (29). We have omitted members of the *mat*K family (*trnK*I1 ORFs in land plants) because they are too divergent for phylogenetic comparisons, and also because over 1000 *mat*K sequences have been reported due to their use in molecular systematics of higher plants (30). The *mat2* ORFs of euglenoids are also too divergent for phylogenetic analysis (31).

The most important ORFs compiled in our search are the bacterial ORFs, only one of which was included in the previous compilation. Because many of the bacterial ORFs were discovered in sequencing projects, their database entries are often poorly annotated and fail to define correct intron and exon boundaries. Only ~50% of the bacterial ORFs have been reported as group II introns in the literature [*B.m.* (20); *C.s.* (14); *C.d.* (17); *E.c.B*, *E.c.D* (15); *L.l.* (16); *P.a.* (18); *S.a.1*, *S.a.2* (21); *S.me.* (19); see Table 3 for intron abbreviations]. For the newly identified bacterial ORFs, we have folded the flanking sequences to verify that all except *S.ma.* are located in group II intron RNA structures (Table 3).

Sequence alignment

ORF sequences were aligned by standard methods, and the alignment has been submitted to the EMBL database (accession number ALIGN_000044). Figure 2 shows a representative alignment of four ORF sequences, each representing a phylogenetic grouping. For convenience we refer to these clusters as the mitochondrial, algal, bacterial and euglenoid groupings, although the bacterial grouping is not a discrete clade, and the mitochondrial and algal groups have mixed compositions (Fig. 3). A 75% consensus sequence is presented for each subgrouping, along with the total consensus sequence for group II intron ORFs. Conserved motifs of group II intron ORFs (subdomains 0–7 of the RT domain, the X domain and the Zn domain) are labeled according to previous studies (7,9,32) taking into account the boundaries of conservation seen in our alignment. [Subdomains 0 and 2A are named in accordance with non-LTR RTs (7); subdomain 0 is conserved only between group II intron and non-LTR RTs; subdomain 2A is weakly conserved among non-LTR, group II intron and retron RTs.] We have also defined two spacer regions which are sites for amino acid insertions in many of the ORFs. These sites are between subdomains 4 and 5 of the RT domain (the 4/5 spacer) and between subdomain 7 and domain X (the 7/X spacer). The 4/5 spacer varies from 1 to 179 amino acids for different RTs, while the 7/X spacer ranges from 0 to 235 amino acids.

Variations between the consensus sequences in Figure 2 reflect evolutionary divergence of the lineages of ORFs. The most striking differences are in domain X. Domain X is highly conserved in the mitochondrial and euglenoid groups (21 and 38 positions conserved out of 105), less conserved in the algal group (15 positions) and poorly conserved in the bacterial group (five positions). In concluding that domain X is 'poorly conserved' in bacteria, we note that RT subdomains 0–7 are equally conserved within each of the mitochondrial and bacterial groups (90 and 94 residues respectively). Interestingly, the conserved positions in domain X are not shared among the groups, further suggesting that domain X is the most rapidly evolving region of the ORF.

Phylogenetic analysis

Phylogenetic analysis was based on RT subdomains 0–7 and domain X. Although plant *nad1*I4 ORFs do not contain subdomains 0 and 1, their inclusion improved resolution for the rest of the tree and did not affect placement of *nad1*I4 ORFs (not shown). All other positions that were not unambiguously alignable among all group II intron ORFs were omitted from analysis, including the domain 4/5 spacer and all idiosyncratic insertions (see Fig. 3 for listing of insertions). In total, the alignable sequence used to construct the tree was 260 amino acids, of which 235 sites were informative. The tree was rooted with RTs of four subclasses of non-LTR elements [*D.m.*RD2, *D.m.*jockey, *C.e.*RT1 and human L1 (7)], using only the alignable amino acids in RT subdomains 0–7 (202 positions). Since domain X sequence is not present in the outgroups, we confirmed that inclusion of domain X in the analysis did not significantly affect the branching pattern. We also confirmed that the internal topology of an unrooted tree was the same as the rooted tree (not shown).

Phylogenetic trees were derived by neighbor-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML) algorithms (Materials and Methods). The phylogenetic model derived from a neighbor-joining algorithm and rooted with non-LTR RTs is presented in Figure 3 along with bootstrap values for NJ and MP analyses. ML analysis was consistent with NJ and MP analysis but gave lower statistical support. The topology of the tree is divided into two major clusters which we term the mitochondrial and chloroplast-like lineages. Members of the mitochondrial lineage include all known fungal, liverwort and plant mitochondrial group II intron ORFs, as well as several ORFs of bacteria and brown algal mitochondria. The base of the mitochondrial lineage is defined by a node with 96% bootstrap support (NJ). Bootstrap support within the mitochondrial lineage does not support a specific branching order among the subgroupings; however, a cluster of four liverwort intron ORFs (*atpA*I1, *atpA*I2, *cob1*I3 and *cox1*I1 ORFs) appear to have given rise to the *nad1*I4 (*mat*R) family of ORFs in higher plants.

Table 1. Mitochondrial group II intron-encoded ORFs and related ORFs

Species name[a]	Organism class	Host gene	Intron	ORF domains[b]	YxDD[c]	Size[d]	Frame[e]	Accession number[f]
Saccharomyces cerevisiae (*S.c.*)	Yeast	*cox1*	I1	RT-X-Zn	+	778	Fusion	V00694
Saccharomyces cerevisiae (*S.c.*)	Yeast	*cox1*	I2	RT-X-Zn	+	785	Fusion	V00694
Kluyveromyces lactis (*K.l.*)	Yeast	*cox1*	I1	RT-X-Zn	+	786	Fusion	X57546
Schizosaccharomyces pombe (*S.p.*)	Yeast	*cob1*	I1	RT-X-Zn	+	807	Fusion	X54421
Schizosaccharomyces pombe (*S.p.*)	Yeast	*cox1*	I1	RT-X-Zn	+	787	Fusion	SP0251292
Schizosaccharomyces pombe (*S.p.*)	Yeast	*cox2*	I1	RT-X-Zn	+ (FADD)	778	Fusion	SP0251293
Venturia inaequalis (*V.i.*)	Fungus	*cob1*	I1	RT-X	+	707	Fusion	AF004559
Allomyces macrogynus (*A.m.*)	Fungus	*cox1*	I3	RT-X-Zn	+	785	Fusion	U41288
Podospora anserina (*P.a.*)	Fungus	*cox1*	I1	RT-X-Zn	+	787	Fusion	X55026
Podospora anserina (*P.a.*)	Fungus	*cox1*	I4	RT-X-Zn	+	789	Fusion	X55026
Podospora anserina (*P.a.*)	Fungus	*ND5*	I4	RT-X	– (YANV)	779	Fusion	X55026
Podospora comata (*P.co.*)	Fungus	*cox1*	I1	RT-X-Zn	+	783	Fusion	Z69899
Podospora curvicolla (*P.cu.*)	Fungus	*ND5*	I1	RT-X	– (YANV)	745	Fusion	Z69898
Neurospora crassa (*N.c.*)	Fungus	*cox1*	I1	RT-X-Zn	+	835	Fusion	X14669
Cryphonectria parasitica (*C.p.*)	Fungus	N.D.[g]	N.D.[g]	RT-X-Zn	+	778	N.D.[g]	AF218567
Marchantia polymorpha (*M.p.*)	Liverwort	*atpA*	I1	RT-X-Zn	+	1064	Fusion	M68929
Marchantia polymorpha (*M.p.*)	Liverwort	*atpA*	I2	RT-X	+	909	Fusion	M68929
Marchantia polymorpha (*M.p.*)	Liverwort	*atp9*	I1	RT-X	– (YADN)	771[h]	Fusion	M68929
Marchantia polymorpha (*M.p.*)	Liverwort	*cox1*	I1	RT-X-Zn	– (YAGN)	914[i]	Fusion	M68929
Marchantia polymorpha (*M.p.*)	Liverwort	*cox1*	I2	RT-X-Zn	+	836[j]	Fusion	M68929
Marchantia polymorpha (*M.p.*)	Liverwort	*cob1*	I3	RT-X-Zn	+	949	Fusion	M68929
Marchantia polymorpha (*M.p.*)	Liverwort	*cox2*	I2	RT-X	+	743	Fusion	M68929
Marchantia polymorpha (*M.p.*)	Liverwort	SSU rDNA	I1	RT-X-Zn	+	502[k]	Free	M68929
Arabidopsis thaliana (*A.t.*)	Green plant	*nad1*	I4	RT(2–7)-X	+	656	Free	X98300
Glycine max (*G.m.*)	Green plant	*nad1*	I4	RT(2–7)-X	+	668	Free	U09988
Oenothera berteriana (*O.b.*)	Green plant	*nad1*	I4	RT(2–7)-X	+	655	Free	M63034
Solanum tuberosum (*S.t.*)	Green plant	*nad1*	I4	RT(2–7)-X	– (YADN)	674	Free	AJ003130
Triticum aestivum (*T.a.*)	Green plant	*nad1*	I4	RT(2–7)-X	+	678	Free	X57965
Vicia faba (*V.f.*)	Green plant	*nad1*	I4	RT(2–7)-X	+	665	Free	M30176
Zea mays (*Z.m.*)	Green plant	*nad1*	I4	RT(2–7)-X	+	653	Free	U09987
Porphyra purpurea (*P.p.*)	Red alga	LSU rDNA	I1	RT-X-Zn	+	544	Free	AF114794
Porphyra purpurea (*P.p.*)	Red alga	LSU rDNA	I2	RT-X-Zn	+	546	Free	AF114794
Pylaiella littoralis (*P.li.*)	Brown alga	LSU rDNA	I1	RT-X-Zn	+	308	Free	Z48620
Pylaiella littoralis (*P.li.*)	Brown alga	LSU rDNA	I2	RT-X-Zn	+	318	Free	Z48620
Pylaiella littoralis (*P.li.*)	Brown alga	*cox1*	I1	RT-X-Zn	+	765	Free	Z72500
Pylaiella littoralis (*P.li.*)	Brown alga	*cox1*	I2	RT-X-Zn	+	810	Free	Z72500
Pylaiella littoralis (*P.li.*)	Brown alga	*cox1*	I3	RT-X-Zn	+	748	Free	Z72500
Free-standing ORFs								
Marchantia polymorpha (*M.p.*)	Liverwort	-	Orf732	RT-X	+ (FADD)	732	Free	M68929
Porphyra purpurea (*P.p.*)	Red alga	-	rtl	X-Zn	–	211	Free	AF114794

[a]Species abbreviations used in this manuscript are shown in parentheses.
[b]The presence of conserved domains: RT, reverse transcriptase domain; X, domain X (putative splicing function); Zn, Zn domain (nuclease activity). In cases where the complete RT domain is not present, subdomains are indicated in parentheses (0–7).
[c]The presence of a catalytic YxDD motif in subdomain 5 of the RT. Deviations from YxDD are indicated, with FxDD considered to be functional since it is common in other RTs (32).
[d]Size (amino acids) of the ORF. For ORFs in-frame with the upstream exon, the length was calculated based on the first amino acid fully coded within the intron. However, these sizes are inaccurate since the encoded proteins are processed at their N-termini.
[e]Fusion: ORF is translated in-frame with the upstream exon. Free: start codon for ORF is contained within the intron.
[f]Accession numbers are for DNA database entries except where noted.
[g]No data. The complete flanking sequence was not reported.
[h]Published size is 681 amino acids; a single frameshift in domain X replaces 34 C-terminal amino acids with 124 amino acids similar to other domain X sequences (8).
[i]Published size is 887 amino acids; a single termination readthrough adds 27 amino acids with similarity to other Zn domains.
[j]Published size is 742 amino acids; a single termination readthrough adds 94 amino acids with similarity to the Zn domain (9).
[k]Published size is 501 amino acids; two frameshifts and a termination readthrough add 72 amino acids with similarity to the Zn domain (9).

Table 2. Chloroplast group II intron-encoded ORFs and related ORFs[a, b]

Species name	Organism class	Host gene	Intron	ORF domains	YxDD	ORF size (amino acids)	Frame	Accession number
Oocystacea sp. (*O.s.*)	Green alga	*petD*	I1	RT-X-Zn	+	608	Free	S05341[c]
Scenedesmus obliquus (*S.o.*)	Green alga	*petD*	I1	RT-X-Zn	+	608	Free	P19593[c]
Bryopsis maxima (*B.m.*)	Green alga	*rbcL*	I1	RT-X	+	478[d]	Free	X55877
Pyrenomonas salina (*P.s.*)	Cryptomonad	*cpn60*	I1	RT-X	– (YANN)	452	Free	X81356
Euglena gracilis (*E.g.*)	Euglenoid	*psbC*	I4	RT-X	– (YGYY)	458	Free	X70810
Euglena gracilis (*E.g.*)	Euglenoid	*psbD*	I8	RT-X	– (LSSD)	506	Free	X70810
Euglena deces (*E.d.*)	Euglenoid	*psbC*	I4	RT-X	– (FSNT)	453	Free	Z99833[e]
Euglena myxocylindracea (*E.m.*)	Euglenoid	*psbC*	I4	RT-X	– (YGYH)	436	Free	Z99835[e]
Euglena viridis (*E.v.*)	Euglenoid	*psbC*	I4	RT-X	– (FGYF)	468	Free	Z99836[e]
Lepocinclis buetschlii (*L.b.*)	Euglenoid	*psbC*	I4	RT-X	– (VGEN)	451	Free	Z99834.1
Free-standing ORF								
Astasia longa (*A.l.*)	Euglenoid	-	ORF456	RT-X	– (YGYF)	456	Free	X14385

[a]See notes for Table 1 for a general description of column entries.
[b]The sequence Z99832 (*Cryptoglena pigra*) is not shown in the table. Its 102 amino acids are homologous to the C-terminus of other euglenoid *psbC*I4 intron-encoded proteins but the DNA sequence encoding the N-terminus was not reported.
[c]Accession number is for the protein database entry. The DNA sequence was reported by Kück (49).
[d]The published size is 274 amino acids; three frameshifts add 204 amino acids with similarity to group II intron-encoded proteins (8).
[e]Translation of the DNA database entry was reported by Doetsch *et al.* (50).

The chloroplast-like lineage is defined by a node of 75% bootstrap support (NJ), and is divided into the algal group and the euglenoid group. The algal group is a highly heterogeneous collection of ORFs from algal chloroplasts, algal mitochondria and bacteria. The euglenoid group consists mostly of related *psbC*I4 ORFs which are found in group III introns, a degenerate form of group II introns (33). Members of the euglenoid group were omitted from the NJ phylogenetic calculation because of their extreme divergence (~25% identity to algal ORFs; Fig. 2), but we include them in Figure 3 with dotted lines because MP analysis consistently placed these ORFs in the chloroplast-like lineage with substantial bootstrap support (>70%), and also because BLAST searches suggested that their closest relatives are algal ORFs (6/6 of the best matches, data not shown).

Intron ORFs that do not belong to the mitochondrial or chloroplast-like lineages are bacterial, and comprise four well-defined clades, each with 100% bootstrap support. The four bacterial groups are positioned at the base of the two main lineages, but bootstrap support for their positions is quite weak due to low sequence conservation among group II intron ORFs, and also because of low conservation between group II intron ORFs and the outgroup RTs (average of 21% identity). Rooting the tree with retron RTs (NJ) or non-LTR RTs (MP) resulted in similar trees, but with <50% support for all basal nodes. These alternate trees predicted the earliest branching ORFs to be bacterial group C (NJ, retron RT outgroup) and bacterial group D (MP, non-LTR RT outgroup). Thus, it is not possible with this data set to accurately predict the position of the root or the basal branching order.

Variation in spacer elements supports the phylogenetic groupings

Spacer elements were omitted from the phylogenetic analyses because they could not be aligned for all sequences; however, they provide additional support for some groupings of the phylogenetic tree. Figure 3 tabulates the lengths of the spacers for all ORFs in the phylogenetic tree. The domain 4/5 spacer is 1–38 amino acids in bacterial and chloroplast-like groups, 42–101 amino acids in the fungal group, and 176–179 amino acids in the *mat*R family. This pattern supports the mitochondrial clade and also suggests, since the outgroups have short spacers (6–25 amino acids), that the most primitive group II intron ORFs also had a short spacer between domains 4 and 5, which was expanded when the ORF migrated to mitochondria and was expanded even further in the transfer to plant mitochondria. A similar scenario is seen for the 7/X spacer. The 7/X spacer is 0–7 amino acids in the bacterial groups A, B, C and D, 0–19 amino acids in the chloroplast-like group, and 19–35 amino acids in the fungal mitochondrial group. Notably, the 7/X spacer is 150–235 amino acids in both the liverwort subcluster and the *mat*R family, suggesting that an expansion of the 7/X spacer occurred in liverwort before the ORF was passed on to higher plants.

Because of the unusually large size of the insertions (235 amino acids = 26 kDa), we examined the position of the spacers in the crystal structure of HIV-RT (6). Insertions in the 4/5 spacer would be predicted to produce an extension of the finger domain and would probably not interfere with the active site in domain 5. Similarly, the insertions in 7/X spacer would be located in a tether region near the thumb domain, and could reasonably extend away from the active site of the RT.

Evolution of domain structures and activities of the ORFs

Figure 3 tabulates additional intron properties, including the presence of ORF structural domains, a YADD motif at the polymerase active site, and a group II intron RNA structure. It appears that the Zn domain was lost many times during evolution, since ORFs without the Zn domain are scattered throughout the tree. Lack of the Zn domain is expected to

Table 3. Bacterial group II intron-encoded ORFs and related ORFs[a, b]

Species name	ORF name[c]	Host gene	Locus[d]	Intron[e]	ORF domains	YxDD	Size (amino acids)	Frame	Accession number
Anabaena sp. (*A.s.*)	ORF439	None[f]	Tas transposable element	+	RT(0,4-7)-X-Zn	+	439	Free	U13767
Bacillus anthracis (*B.a.*-07)	PX01–07	PX01–08/PX01–06[g]	Virulence plasmid PX01	+	RT-X-Zn	+	602	Free	AF065404
Bacillus anthracis (*B.a.*-23)	PX01–23	PX01–24/ORFX[g]	Virulence plasmid PX01	+	RT-X	+	461	Free	AF065404
Bacillus halodurans (*B.h.*)	ORF1	None[f]	Chromosome	+	RT-X	+	418	Free	AB031210
Bacillus megaterium (*B.m.*)	iepA	None[f]	Class II transposon	+	RT-X-Zn	+	588	Free	AB022308
Calothrix sp. (*C.s.*)	ORF2	ORF1	Unknown	+	RT-X-Zn	+	584	Free	X71404
Clostridium difficile (*C.d.*)	Unnamed	ORF14	Conjugative transposon Tn5397	+	RT-X-Zn	+	609	Free	X98606
Escherichia coli (*E.c.*B)	IntB	ORFH	H-repeat (Rhs)	+	RT-X	+	416	Free	X77508
Escherichia coli (*E.c.*D)	IntD	IS629[h]	Chromosome	+[h]	RT-X	+	448	Free	D37918
Escherichia coli (*E.c.*-0157)	L7072	None[i]	Plasmid p0157	+	RT-X-Zn	+	574	Free	AF074613
Lactococcus lactis (*L.l.*)	LtrA	Relaxase	Conjugative transfer plasmid pRS01	+	RT-X-Zn	+	599	Free	U50902
Pseudomonas alcaligenes (*P.a.*)	ORFX6	N.D.[j]	Plasmid RP4	+	RT-X	+	490	Free	U77945
Pseudomonas putida (*P.p.*)	MatP1	N.D.[j]	Plasmid PRA500	+	RT-X	+	473	Free	AF101076
Pseudomonas sp. (*P.s.*)	ORF494	None[f]	κγ element in Tn5040	+	RT-X	+	494	Free	PSY18999
Serratia marcescens (*S.ma.*)	RetA	None[f]	Plasmid R471a	–[k]	RT-X	+	495	Free	AF027768
Shigella flexneri (*S.f.*)	SfiA	IS629-like ORF	She pathogenicity island	+	RT-X	+	431	Free	U97489
Sinorhizobium meliloti (*S.me.*)	ORF RmInt1	ORF B	ISRm2011–2	+	RT-X	+	415	Free	Y11597
Sphingomonas aromaticivorans (*S.a.1*)	MatRa	Replication primase	PNL1 plasmid	+	RT-X-Zn	+	633	Free	AF079317
Sphingomonas aromaticivorans (*S.a.2*)	ORF404	ORF392/ORF416	PNL1 plasmid	+	RT-X-Zn	+	571	Free	AF079317
Streptococcus pneumoniae (*S.p.*)	Unnamed	None[f]	Capsular polysaccharide biosynthetic locus	+	RT-X	+	425	Free	AF030367

[a]For a general description of column entries see notes to Table 1.
[b]Sequences omitted from the table were as follows. Highly divergent sequences which may be degenerate remnants were: D90902 [*Synechocystis sp.*, 508 amino acids, RT(0–7)]; and D64002 [*Synechocystis sp.*, 521 amino acids, RT(0–7)]. Fragments corresponding to incompletely sequenced DNA were: S35081 (*Azotobacter vinelandii*; 86 amino acids); Z47187.1 (*Calothrix sp.*, 120 amino acids); S35080 (*Calothrix sp.*, 161 amino acids); and AAD29837 (*Pseudomonas putida*, 216 amino acids). Fragments corresponding to (presumably) non-functional remnants of group II intron ORFs were: Z98756 (*Mycobacterium leprae*, 161 amino acids); AL021428 (*Mycobaterium tuberculosis*, 235 amino acids); AE000069 (*Rhizobium sp.* NGR234, 133 amino acids); AL049661 (*Streptomyces coelicolor*, 145 amino acids); AF0746111 (*Yersinia pestis*, 156 amino acids); BAA17969 (*Synechocystis sp.*, 150 amino acids); and S43481 (*E. coli*, 121 amino acids). AF006691 (*Pseudomonas putida*, ~210 amino acids) is a fragment found untranslated in the GenBank DNA sequence database.
[c]The ORF name listed in the publication or database entry.
[d]The locus of the ORF, if known.
[e]The presence of an intron structure surrounding the ORF was evaluated by folding the sequence into a consensus group II intron structure (N.Toor and S.Zimmerly, unpublished).
[f]The intron does not appear to be inserted into an ORF although a very small ORF cannot be ruled out.
[g]No host gene is annotated in the GenBank entry, but the intron probably interrupts neighboring ORFs.
[h]Sequence has not been reported for the 5′ end of the intron, including the upstream exon (IS629-like ORF) and 680 bp of the intron. Otherwise, this sequence is virtually identical to the *S.f.* intron.
[i]The intron is located between ORFs 7070 and 7073; a 5′ extension of the 7073 ORF could include the intron.
[j]No data. Complete flanking sequence was not reported.
[k]Intron domains 5 and 6 are clearly not present; the intron structure may be degenerated.

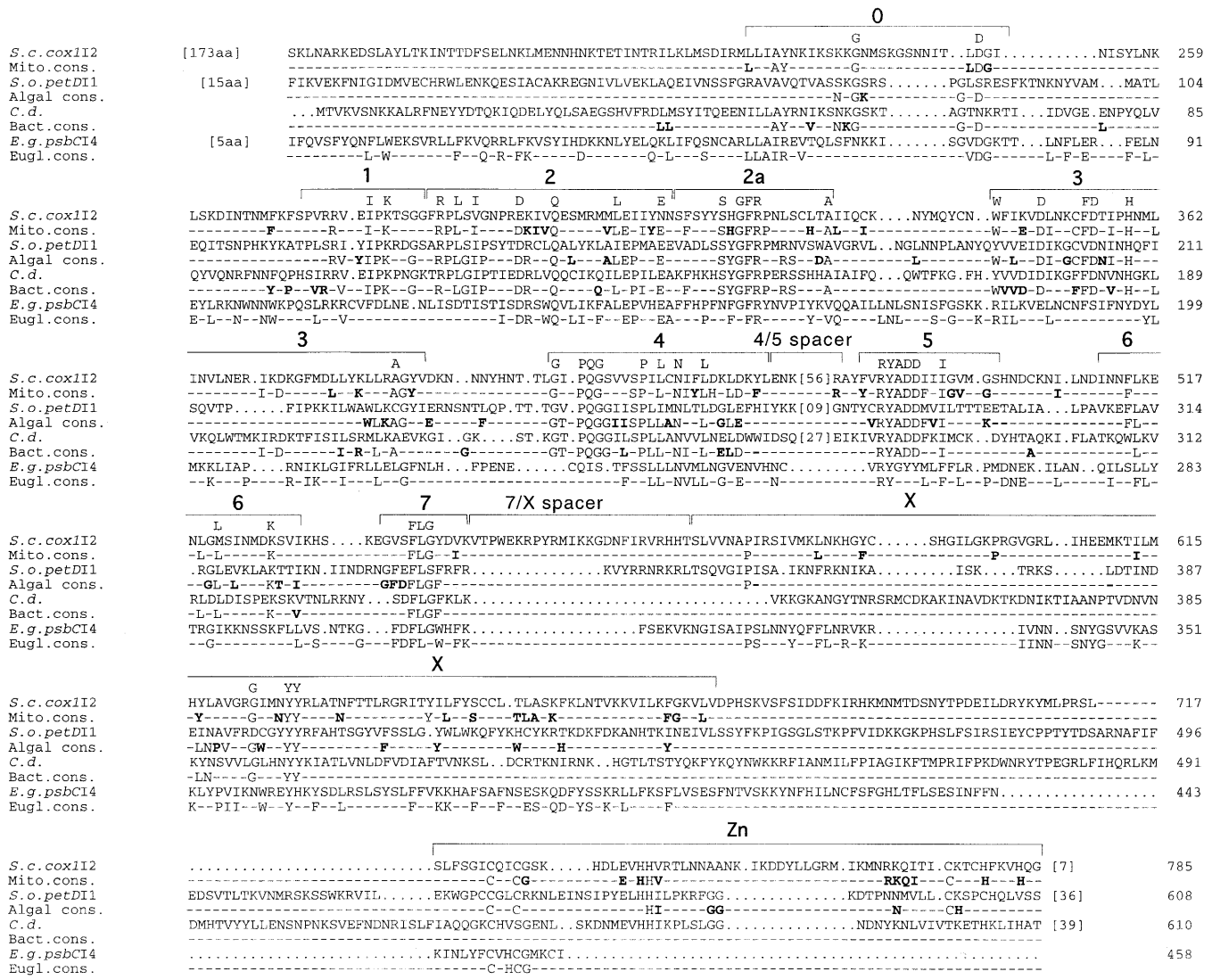


Figure 2. Alignment of RT amino acid sequences. One example sequence is shown for each of the major phylogenetic groupings (mitochondrial, algal, bacterial and euglenoid; see Fig. 3). Below each example sequence is the consensus sequence for that group, defined as the positions conserved in at least 75% of the members. The overall consensus is presented above the alignment, and is defined as residues common to the mitochondrial, algal and bacterial consensus sequences. Domains 0–7, X and Zn are marked as described in the text. The locations of spacers between RT domains 4 and 5 and between domains 7 and X are indicated (see text and Fig. 3). Residues marked in bold are consensus sequence positions specific to one group. In cases of closely related sequences (e.g. plant *nad1*I4 introns), only one example was included in the consensus sequence calculation. Sequences used in the calculation were: mitochondrial group (*Z.m.nad1*I4, *M.p.cox1*I1, *M.p.cob1*I3, *M.p.atp*AI1, *M.p.atp*AI2, *S.c.cox1*I2, *S.c.cox1*I1, *M.p.atp*9I1, *S.p.cox*2I1, *S.p.cox1*I1, *A.m.cox1*I3, *P.a.cox1*I4, *M.p.SSU*I1, *P.a.cox1*I1, *S.a.1*, *L.l.*, *S.p.cob1*I1, *V.i.cob1*I1, *P.li.cox1*I3, *P.a.ND5*I4, *M.p.*ORF732, *M.p.cox1*I2, *P.li.cox1*I2, *P.li.cox1*I1, *N.c.cox1*I1 and *M.p.cox2*I2); algal group (*P.p.LSU*I1, *C.s.*, *B.m.rbcL*I1, *B.a.*-07, *S.o.petD*I1, *P.li.LSU*I1, *P.li.LSU*I2, *P.s.* and *E.c.*-0157); bacterial group (*S.f.*, *S.p.*, *P.a.*, *B.m.*, *C.d.*, *B.a.*-23, *S.me.*, *B.h.* and *E.c.B*); and euglenoid group (*E.v.psbC*I4, *A.l.*ORF456, *E.g.psbC*I4, *E.d.psbC*I4, *L.b.psbC*I4, *E.m.psbC*I4 and *P.s.cpn60*I1).

reduce the efficiency of mobility, but not necessarily block mobility in all organisms. For the yeast intron *S.c.cox1*I2, deletion of the Zn domain reduces mobility to essentially undetectable levels (34), although the bacterial *P.a.* and *S.me.* introns are mobile *in vivo* despite lacking a Zn domain (18,35). We note that the Zn domain is missing in most of the putatively early branching bacterial ORFs, suggesting that bacterial introns in general do not require a Zn domain. Some bacterial ORFs without Zn domains may reflect ancestral ORFs that predate acquisition of the Zn domain.

The YADD motif was also lost several times in evolution. The YADD motif is absent in the euglenoid lineage and in four mitochondrial ORFs. Loss of the YADD motif in yeast introns *S.c.cox1*I1 and *S.c.cox1*I2 eliminates RT activity, but only reduces *in vivo* mobility to 40% of wild-type levels, which has been explained by an alternative mobility mechanism based on double-strand break repair (36). In contrast to organellar intron ORFs, all bacterial ORFs contain the YADD motif, which suggests that intron survival in bacteria requires RT activity.

Evolution of the Zn domain

To address the spotty distribution of the Zn domain in Figure 3, we phylogenetically-analyzed the Zn domain alone. The data set included 77 amino acid positions with 59 informative sites,

and ambiguously aligned residues were not excluded. Outgroups used to root the tree were the *E.coli* colicin E7 and *Pseudomonas aeruginosa* pyocin S1, members of the larger nuclease family to which the Zn domain belongs (9). The phylogenetic tree derived from NJ analysis is shown in Figure 4A. Bootstrap support for the branching order is poor due to low sequence conservation and the short sequence analyzed; however, the Zn domain of the mitochondrial lineage is separated from other Zn domains by a node with 79% bootstrap support. Figure 4B shows a sequence alignment of Zn domains for selected ORFs along with 50% consensus sequences for mitochondrial and chloroplast-like lineages. The Zn domains of bacterial group B and chloroplast-like groups are seen to be similar to the nuclease motifs of colicin and pyocins in their lengths, and slightly in their sequences (pink shading). In contrast, the Zn domain of the mitochondrial lineage is significantly expanded, and has additional conserved positions near its C-terminus (NRKQIPLC). This data is consistent with the possibility of acquisition of the Zn domain in bacteria from a bacterial family of nucleases, and subsequent expansion of the domain in mitochondria. Taking into account the low resolution in the phylogenetic analysis, there is little indication for 'swapping' of Zn domains among ORFs, and the spotty distribution of the Zn domain may be due to domain loss alone.

Horizontal versus vertical inheritance

The phylogenetic model in Figure 3 predicts both horizontal and vertical inheritance of group II intron ORFs. Vertical inheritance is suggested for the euglenoid *psbC*I1 and plant *nad1*I4 families of ORFs, since the introns of each family are confined to the same DNA location. Furthermore, the *psbC*I1 introns are unlikely to be mobile because the ORFs lack mobility-related motifs (Fig. 2). Mobility competence of *nad1*I4 ORFs is less clear since the ORFs contain a YADD motif. Still, vertical inheritance seems likely because of large insertions within the ORF, and because of the apparent agreement between ORF phylogeny and species phylogeny, with monocots and legumes each forming subgroupings.

Other than these two intron families, there is little evidence for strict vertical inheritance. The only other ORFs located in identical genomic sites in different species are: *K.l.cox1*I1 and *S.c.cox1*I2; *A.m.cox1*I3 and *P.a.cox1*I4; and *S.f.* and *E.c.D*. As described in the introduction, *K.l.cox1*I1 and *S.c.cox1*I2 probably represent a horizontal transfer event. In the case of *A.m.cox1*I3 and *P.a.cox1*I4 ORFs, the ORF amino acid sequences are 54% identical while *cox1* amino acid sequences are 68% identical. Although this would be consistent with vertical inheritance, vertical inheritance is not certain since the introns have been reported only in these two distantly related fungi (Fig. 5), and not in other sequenced fungal genomes such as *Schizosaccharomyces pombe*, *S.cerevisiae* or *Pichia canadensis*. The introns *S.f.* and *E.c.D* probably represent a horizontal transfer, since the introns are 99.6% identical in total DNA sequence (intron and ORF) while their IS629-like exon DNA sequences are 91% identical and, furthermore, *E.c.D* is present in only a minority of *E.coli* strains (15).

The predicted ORF phylogeny in Figure 3 suggests multiple horizontal transfers between fungi and liverwort, between fungi and brown algae, among fungi, and among bacteria. Within the fungal and bacterial groups of ORFs, there is little correspondence between ORF phylogeny and species phylogeny (compare Figs 3 and 5). For example, of the three *S.pombe* ORFs, two are found only in subsets of *S.pombe* strains (37,38); the *S.p.cox2*I1 ORF is most closely related to a liverwort ORF, while *S.p.cox1*I1 and *S.p.cob1*I1 ORFs are more related to other fungal ORFs than to the other *S.pombe* ORFs. Of the nine *M.p*. intron ORFs, four are related and possibly diverged within liverwort, while at least two (*M.p.cox1*I2 and *M.p.SSU*I1) are more related to fungal and brown algal ORFs than the other *M.p.* ORFs, suggesting that they were the result of horizontal transfers. The bacterial intron ORFs *B.a.*-07 and *B.a.*-23 are phylogenetically distant but are located 10 kb apart on the same plasmid. The three ORFs in *E.coli* are also phylogenetically distant. All of these examples are most easily explained by a high frequency of horizontal transfers, although vertical inheritance cannot be ruled out in all cases. Horizontal transfers may be the rule for the time-frame represented in the phylogenetic tree, while long term vertical inheritance might occur only for ORFs that have lost mobility functions but retained splicing function.

Horizontal transfers appear to be relatively infrequent between fungi and bacteria. The only clear example is the set of three bacterial ORFs found in the mitochondrial lineage. Because of low resolution for branching order within the mitochondrial lineage, it is not possible to predict whether these ORFs were the earliest branching in the lineage, or reflect a horizontal transfer from mitochondria to bacteria. In either case, a horizontal transfer is implicated. The exception for long distance horizontal transfers is in the chloroplast-like lineage, which contains representatives from Gram-positive bacteria, Gram-negative bacteria, cyanobacteria, chloroplasts of green algae and euglenoids, and mitochondria of red and brown algae. These ORFs appear to have transferred horizontally at rates exceeding other phylogenetic classes. The extent of horizontal transfers may reflect unique properties of the lineage such as independence from factors of the host organism.

DISCUSSION

In this paper we present a compilation and phylogenetic analysis of group II intron ORFs, and suggest a model for the evolutionary history for mobile group II introns. Our data are consistent with previously published phylogenetic trees, the most detailed of which were reported by Ferat *et al.* (15) (NJ, 26 ORFs), (24) (NJ, 19 ORFs) and (39) (NJ, 14 ORFs). The major differences are that our analysis is expanded to include many more sequences, particularly bacterial sequences, our tree is rooted, and we include a detailed analysis of differences in ORF structure among the phylogenetic groupings.

Group II intron ORFs in bacteria

We have uncovered numerous bacterial ORFs and ORF fragments which had not been specifically reported as group II introns in the literature. The expanded data set confirms the earlier observation that bacterial intron ORFs are mainly found in mobile DNAs (15). In 18/20 ORFs and 5/6 ORF fragments where the locus of the intron is known, the ORF is found in a plasmid or mobile DNA. The location of bacterial introns in mobile DNAs is distinct from introns in mitochondria and chloroplasts, where the introns typically lie in housekeeping genes. Bacterial introns are also distinct because they are

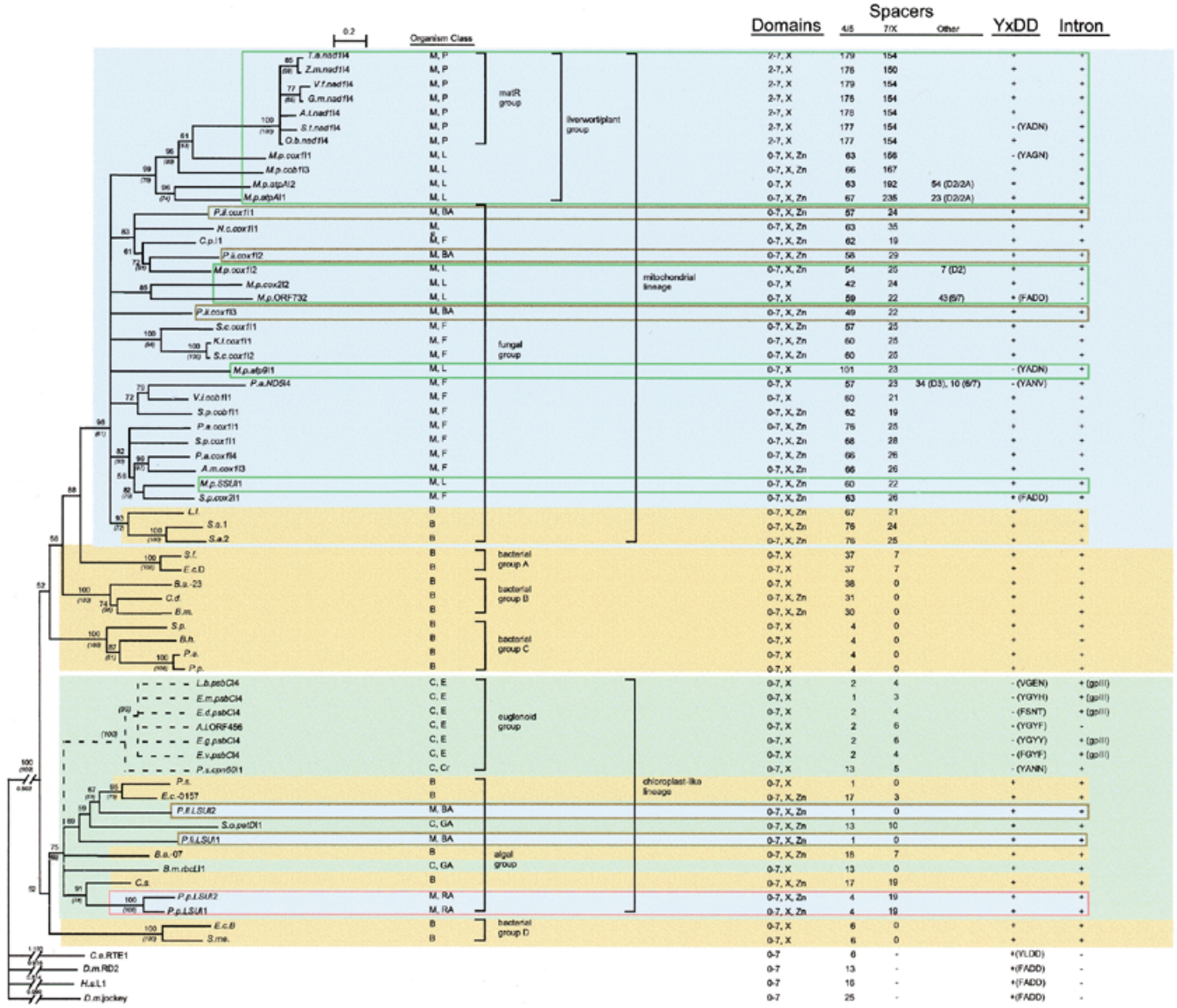


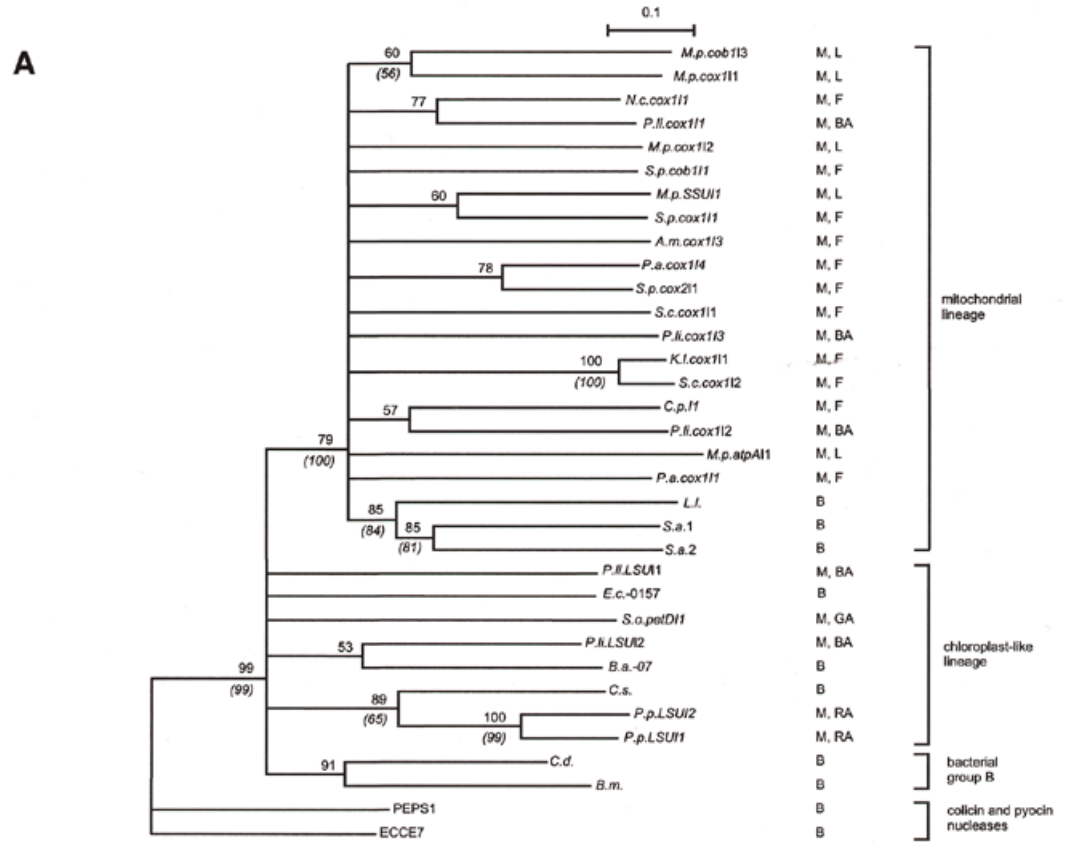
Figure 3. Phylogenetic model of group II intron ORF relationships. The phylogenetic estimate was based on RT subdomains 0–7 and domain X, and was calculated by a neighbor-joining algorithm (PHYLIP; see Materials and Methods). The tree was rooted with four RTs of non-LTR retroelements: *Caenorhabditis elegans* RTE1 (accession number AF025462), *Drosophila melanogaster* RD2 (X51967), *Homo sapiens* L1 (U93574) and *D.melanogaster* jockey (M22874) (8). Bootstrap values are expressed as percentages, and were derived from 1000 (NJ) or 100 (MP) samplings, with MP values shown in italics and parentheses. Nodes with <50% support are collapsed. The predicted approximate location of euglenoid ORFs is shown with dotted lines (see text). Juxtaposed with the inferred phylogenetic relationships are properties of the introns, including protein domains present (subdomains 0–7 of the RT domain, domain X, Zn domain), the size of spacer segments between conserved motifs (see Fig. 2 for spacer definitions), idiosyncratic insertions, the presence of the YADD motif or a functional substitute (see Table 3 footnote), and the presence of a group II intron structure (see Table 3). Euglenoid ORFs are found in group III introns; *P.s.cpn60*I1 is reported to be a twintron (46), but the published RNA structure is probably incorrect. Abbreviations and color codings are: M (mitochondria; blue), C (chloroplast; green), B (bacteria; yellow), P (higher plant; green outline), L (liverwort; green outline), F (fungus; no outline), BA (brown alga; brown outline), GA (green alga; no outline), RA (red alga; pink outline), E (euglenoid; no outline), Cr (cryptopmonad; no outline).

sometimes inserted outside of genes (20; Table 3), and the ORFs are frequently truncated, suggesting a higher degree of intron insertions at new locations followed by intron loss.

Are bacterial group II intron ORFs the oldest?

Our study is consistent with the theory that mobile group II introns originated in bacteria, but does not contribute substantial phylogenetic evidence toward it. The earliest branching ORFs in our analyses are bacterial with both NJ and MP algorithms, and with either non-LTR or retron RTs as outgroups, but the bootstrap values in all cases leave the phylogenetic support weak at best. We anticipate that, as more bacterial group II introns are reported, the bacterial cluster at the base of the tree will enlarge and perhaps the branching order will become more defined. Apart from the phylogenetic data, the observation that mitochondrial ORFs are expanded in three locations compared to the outgroup RTs suggests that ORFs of the mitochondrial lineage are not the earliest branching.

Were group II introns introduced to eukaryotes through the original organellar endosymbiont? Our evidence is consistent



B

Zn domain sequences

Mito.	*M.p.cob1I3*	RHPADRCSVIGCLE-----SKIEMHHIRALKRSGAGGPKMSIVNSSS-DRISGLEALHAALNRKQISLCREHHKAMH
	S.c.cox1I2	SLFSGICQICGSK------HDLEVHHVRTLNNAANK-IKDDYLLGRM----------IKMNRKQITICKTCHFKVH
	L.l.	RLKAKCCELCGTSD---ENTSYEIHHVNKVKNLK----GKEKWEMAM----------IAKQRKTLVVCFHCRHVI
	Mito. cons.	xxxxxxCxxCGSx------xxxEMHHVRxxxxxxx--xKxxxxxxx----------xAxNRKQIPLCxxCHxKxH
Chloro.	*S.o.petDI1*	EKWGPCCGLCRKNLE-INSIPYELHHILPKRFGG----------KDTPNNMVLLCKSPCHQLV
	P.li.LSUI1	RRDKYKCKVCNEYIC--GEDKVEIHHIKPKSLGG----------DDAISNNVVLHAECHKQLT
	C.s.	KKXLGKCSHCGLYFR--EDDLIEIDHIIPKSQGG----------KDVYDNLQALHRHCHDVKT
	Chloro. cons.	xxxxxxCxxCxxxxx--xxxxxExHHIxPKxxGG----------xDxxxNLxLLHxxCHxxxx
Bact.	*C.d.*	IAQQGKCHVSGENL---SKDNMEVHHIKPLSLGG----------NDNYKNLVIVTKETHKLIH
	B.m.	IAQYGKCAVTGVEL---GLDEWHCHHKTPYHLTK----------DDSYGNLMILHKSVHLLIH
Outgroups	ECCE7	RMKVGKAPKTRTQDVSGKRTSFELHHEKPISQNG----------GVYDMDNISVVTPKRHIDIH
	PEPS1	VMRDGGAPYVRESEQAGGRIKIEIHHKVRVADGG----------GVYNMGNLVAVTPKRHIEIH

Figure 4. Phylogenetic analysis of the Zn domain. (**A**) A phylogenetic tree of the Zn domain was derived by neighbor-joining analysis with 1000 bootstrap samplings, or by maximum parsimony with 100 bootstrap samplings (italics and parentheses). The tree was rooted with nuclease domains of *E.coli* colicin E7 (ECCE7; accession number 144375) and *P.aeruginosa* pyocin S1 (PEPS1; accession number Q06583). (**B**) Alignment of the Zn domains of selected group II intron ORFs and the nuclease domains of ECCE7 PEPS1. A 50% consensus sequence is shown for chloroplast-like and mitochondrial lineages. x represents residues with <50% conservation. Positions marked in bold show group-specific consensus sequences (mitochondrial, chloroplast-like and colicin/pyocin), or show agreement with the consensus sequence of another group (*C.d.* and *B.m.*). Pink shading indicates similarities between colicin/pyocin nuclease domain and bacterial or chloroplast-like Zn domains. The consensus sequence for the colicin/pyocin family of nucleases is a 100% consensus sequence according to Gorbalenya (9).

with this theory, but the inferred degree of horizontal transfer between bacteria and organelles is great enough that it would not have been necessary for the endosymbiont to have introduced a group II intron.

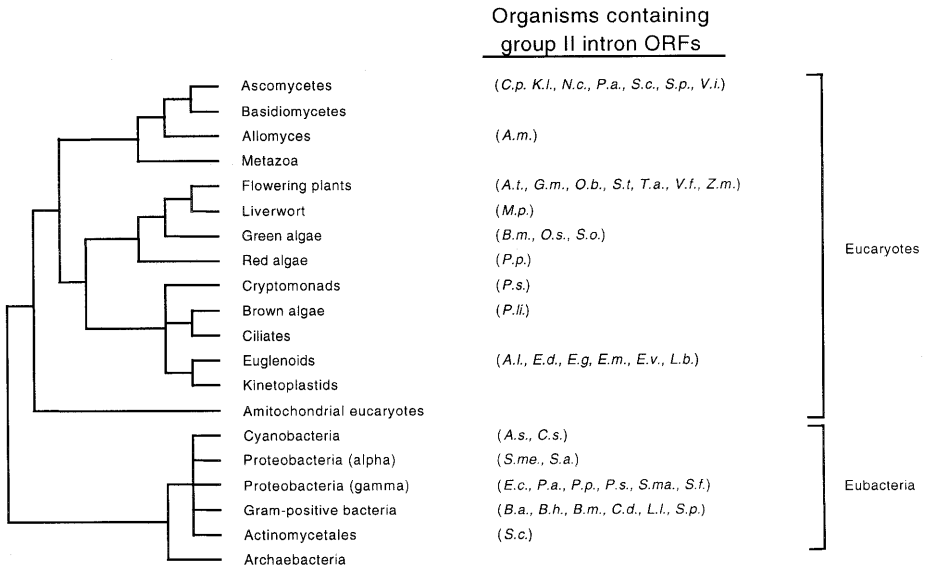


Figure 5. Predicted phylogenetic relationships for organisms containing group II intron ORFs, adapted from Pace (47) and Baldauf *et al.* (48) (not drawn to scale).

A model for the evolutionary history of group II intron ORFs

Our model for the evolutionary history of group II intron ORFs is shown in Figure 6. The oldest ORFs were probably bacterial and were compact in structure with only an RT and X domain. The Zn domain was acquired subsequently, which may have enhanced mobility. Introduction of mobile group II introns into chloroplasts and mitochondria may or may not have been mediated by the organellar endosymbionts. The ORF of the chloroplast-like lineage was essentially the same structurally as its bacterial ancestor except that the X domain became somewhat more conserved. In euglenoid chloroplasts, the *psbC*I4 ORFs diverged drastically and lost mobility but probably not splicing function. Intron ORFs in algal chloroplasts transmitted with high frequency across species and organellar boundaries, possibly because of inherent mobility properties such as independence from host cofactors. It is also possible that the chloroplast-like lineage originated outside chloroplasts and spread horizontally; the data are consistent with horizontal transfers from any source. In contrast to the chloroplast-like lineage, the mitochondrial lineage of ORFs changed significantly in structure from bacterial ORFs. These changes include the creation or expansion of the 4/5 and 7/X spacers, an increase in domain X conservation, and an increase in size and complexity of the Zn domain. Horizontal transfers among fungi, liverwort and brown algae were rampant. Possibly some intron ORFs transferred back to bacteria, although it is also possible that the *L.l*, *S.a.1* and *S.a.2* ORFs represent the earliest branches of the mitochondrial lineage. For a subgroup of liverwort ORFs the 7/X spacer expanded further, and one of these ORFs inserted into *nad1* of plants with concomitant expansion of the 4/5 spacer and loss of RT subdomains 0, 1 and the Zn domain, which together resulted in vertical inheritance.

Evolution of mobility activities?

Does the described progression in ORF structure correspond to development of mobility activities? We note that all mobile group II introns studied in any detail belong to the mitochondrial lineage, which differs in ORF structure from bacterial and chloroplast-like groups. Therefore, the 'classical' mobility properties associated with group II introns may not apply in all respects to the chloroplast-like and putatively early branching bacterial ORFs. There are several properties of mitochondrial group II intron ORFs which might not be as extensively developed for other families of group II intron ORFs. First, maturase activity may be less developed since domain X is seemingly poorly conserved in bacterial groups A, B, C and D, and somewhat less conserved in the chloroplast-like group. The *ltrA* protein of *L.l.* (mitochondrial lineage) binds to its intron very tightly and specifically with a K_d of ~0.25 pM (12,40). It is plausible that more primitive ORFs may lack such specialized binding properties. An alternative explanation is that the bacterial X domains might have evolved to interact with different intron RNA structures specific to their clade. In fact, the X domains are somewhat more conserved within some of the bacterial clades. Within bacterial group C, there are 21 absolutely conserved residues in domain X versus 106 in the RT domain, which can be compared to 21 versus 90 for mitochondrial ORFs (based on the 75% consensus). On the other hand, ORFs of bacterial group B have a ratio of only 8 versus 131 conserved residues, suggesting that the X domain is poorly conserved within at least one of the bacterial clades. A second potential difference in activities among group II introns is site-specificity, which is very high for mitochondrial introns due to a long recognition sequence [31 bp for *S.c.cox1*I2, (41); 35 bp for *L.l.* (42,43)]. In bacteria, less controlled mobility is suggested by the numerous fragmented ORFs and the occasional insertion of introns outside of genes. Finally, the efficiency of mobility is very high in the mitochondrial lineage [~90% for *S.c.cox1*I1, *S.c.cox1*I2 (36); 10–100% for *L.l.* (43,44)], but this may not be true of all bacterial introns. In the most extreme scenario, primitive group II introns may have less developed mobility functions across the board, including less efficient maturase activity, less site-specificity in insertion and lower mobility

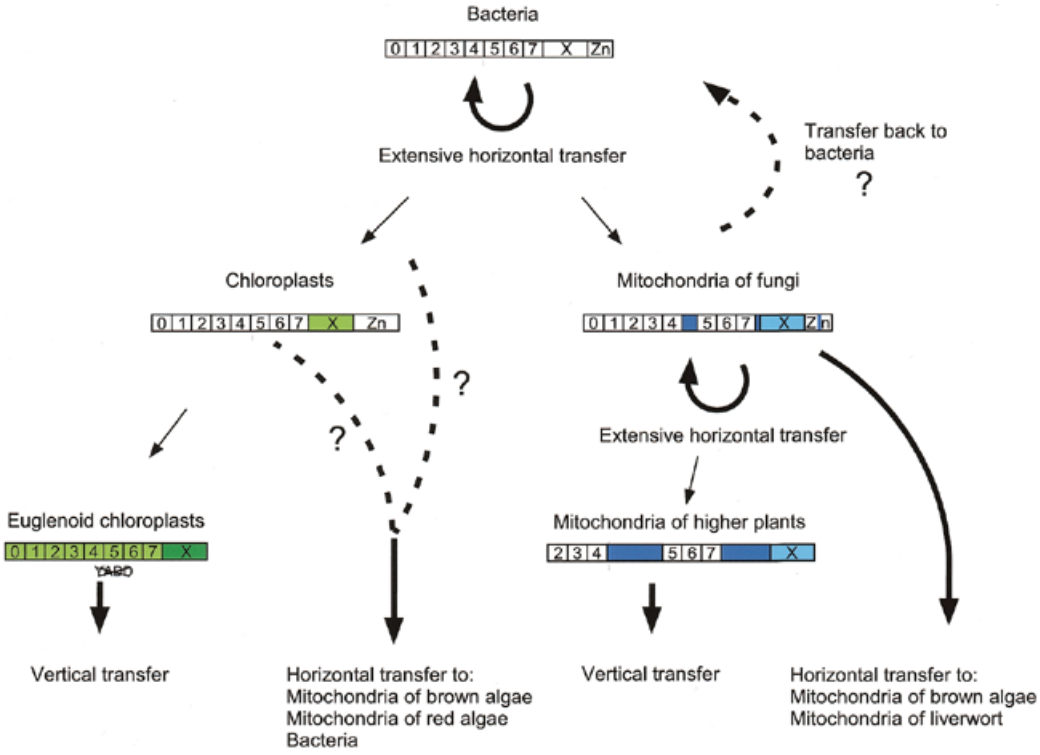


Figure 6. Model for the history of group II intron ORFs. Group II intron ORFs probably originated in bacteria where the introns transferred horizontally with high frequency. The ORFs were introduced to chloroplasts and mitochondria perhaps via the ancestral endosymbionts. In euglenoid chloroplasts, the introns lost mobility functions and were transferred vertically, while in the algal lineage the ORFs spread horizontally at a high rate. Introduction of group II intron ORFs to mitochondria resulted in an expansion of the subdomain 4/5 spacer, 7/X spacer and Zn domain, while domain X became highly conserved. The intron ORFs spread horizontally among mitochondria of fungi, liverwort and brown alga, and possibly transferred back to bacteria. In a subset of liverwort introns, the 4/5 and 7/X spacers were expanded further with concomitant loss of subdomains 0, 1 and the Zn domain, giving rise to the *mat*R family of introns in higher plants. Dotted arrows indicate uncertainty for the source or direction of horizontal transfer events. Colored shading of domains indicates development of group-specific motifs or sequence conservation.

frequency. The only experimental evidence addressing this issue comes from the *S.me*. intron, whose mobility properties are so far mostly consistent with introns of the mitochondrial lineage (35,45). Nevertheless, given these speculations, it is clear that group II introns in all lineages need to be investigated. At this point it is impossible to know to what extent characterized introns of the mitochondrial lineage represent all mobile group II introns.

ACKNOWLEDGEMENTS

We wish to thank Ted Chappell for help with proofreading, and Franz Lang, Ken Sanderson and Joyce Sherman for helpful comments on the manuscript. This work was supported by National Science and Engineering Research Council grant 203717-98. Salary support for S.Z. was from the Alberta Heritage Foundation for Medical Research.

NOTE ADDED AT REVISION

While this manuscript was under review, the review article 'Group II introns in the bacterial world' by F.Martinez-Abarca and N.Toro was published [*Mol. Microbiol*. (2000), **38**, 917–926]. This article also presents a compilation of bacterial group II introns and phylogenetically characterizes their ORFs.

Introns reported to GenBank since completion of this manuscript are: *Ralstonia eutropha* (accession number AF261712; related to bacterial group D ORFs); *Xylella fastidiosa* (AE003999; chloroplast-like); *Agrobacterium rhizogenes* (AP002086; chloroplast-like); *Pseudomonas putida* (Y18999; chloroplast-like); *Pavlova lutherii* (AF045691; mitochondrial).

REFERENCES

1. Nilsen,T.W. (1998) RNA–RNA interactions in nuclear pre-mRNA splicing. In Simons,R.W. and Grunberg-Manago,M. (eds), *RNA Structure and Function*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 279–307.
2. Lambowitz,A.M., Caprara,M., Zimmerly,S. and Perlman,P.S. (1999) Group I and group II ribozymes as RNPs: clues to the past and guides to the future. In Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World*. 2nd Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 451–485.
3. Eickbush,T.H. (1999) Mobile introns: retrohoming by complete reverse splicing. *Curr. Biol.*, **9**, R11–R14.
4. Zimmerly,S., Guo,H., Perlman,P.S. and Lambowitz,A.M. (1995) Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell*, **82**, 545–554.
5. Michel,F. and Ferat,J.-L. (1995) Structure and activities of group II introns. *Annu. Rev. Biochem.*, **64**, 435–461.
6. Kohlstaedt,L.A., Wang,J., Friedman,J.M., Rice,P.A. and Steitz,T.A. (1992) Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science*, **256**, 1783–1790.

7. Malik,H.S., Burke,W.D. and Eickbush,T.H. (1999) The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.*, **16**, 793–805.
8. Mohr,G., Perlman,P.S. and Lambowitz,A.M. (1993) Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res.*, **21**, 4991–4997.
9. Gorbalenya,A.E. (1994) Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. *Protein Sci.*, **3**, 1117–1120.
10. Shub,D.A., Goodrich-Blair,H. and Eddy,S.R. (1994) Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *Trends Biochem. Sci.*, **19**, 402–404.
11. Lambowitz,A.M. and Belfort,M. (1993) Introns as mobile genetic elements. *Annu. Rev. Biochem.*, **62**, 587–622.
12. Wank,H., SanFilippo,J., Singh,R.N., Matsuura,M. and Lambowitz,A.M. (1999) A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol. Cell*, **4**, 239–250.
13. Curcio,M.J. and Belfort,M. (1996) Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. *Cell*, **84**, 9–12.
14. Ferat,J.-L. and Michel,F. (1993) Group II self-splicing introns in bacteria. *Nature*, **364**, 358–361.
15. Ferat,J.-L., Le Gouar,M. and Michel,F. (1994) Multiple group II self-splicing introns in mobile DNA from *Escherichia coli*. *C. R. Acad. Sci. Paris*, **317**, 141–148.
16. Mills,D.A., McKay,L.L. and Dunny,G.M. (1996) Splicing of a group II intron involved in the conjugative transfer of pRS01 in lactococci. *J. Bacteriol.*, **178**, 3531–3538.
17. Mullany,P., Pallen,M., Wilks,M., Stephen,J.R. and Tabaqchali,S. (1996) A group II intron in a conjugative transposon from the gram-positive bacterium *Clostridium difficile*. *Gene*, **174**, 145–150.
18. Yeo,C.C., Tham,J.M., Yap,M.W.-C. and Poh,C.L. (1997) Group II intron from *Pseudomonas alcaligenes* NCIB 9867 (P25X): entrapment in plasmid RP4 and sequence analysis. *Microbiology*, **143**, 2833–2840.
19. Martinez-Abarca,F., Zekri,S. and Toro,N. (1998) Characterization and splicing *in vivo* of a *Sinorhizobium meliloti* group II intron associated with particular insertion sequences of the IS630-Tc1/IS3 retroposon superfamily. *Mol. Microbiol.*, **28**, 1295–1306.
20. Huang,C.-C., Narita,M., Yamagata,T., Itoh,Y. and Endo,G. (1999) Structure analysis of a class II transposon encoding the mercury resistance of the gram-positive bacterium *Bacillus megaterium* MB1, a strain isolated from Minamata Bay, Japan. *Gene*, **234**, 361–369.
21. Romine,F.M., Stillwell,L.C., Wong,K.-K., Thurston,S.J., Sisk,E.C., Sensen,C., Gaasterland,T., Fredrickson,J.K. and Saffer,J.D. (1999) Complete sequence of a 184-kilobase catabolic plasmid from *Sphingomonas aromaticivorans* F199. *J. Bacteriol.*, **181**, 1585–1602.
22. Hardy,C.M. and Clark-Walker,G.D. (1991) Nucleotide sequence of the *COX1* gene in *Kluyveromyces lactis* mitochondrial DNA: evidence for recent horizontal transfer of a group II intron. *Curr. Genet.*, **20**, 99–114.
23. Vogel,J., Börner,T. and Hess,W.R. (1999) Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic Acids Res.*, **27**, 3866–3874.
24. Fontaine,J.M., Goux,D., Kloareg,B. and Loiseaux-de Goër,S. (1997) The reverse-transcriptase-like proteins encoded by group II introns in the mitochondrial genome of the brown alga *Pylaiella littoralis* belong to two different lineages which apparently coevolved with the group II ribosyme lineages. *J. Mol. Evol.*, **44**, 33–42.
25. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic logical alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
26. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
27. Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
28. Strimmer,K. and von Haeseler,A. (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, **13**, 964–969.
29. Qiu,Y.-L., Lee,J., Bernasconi-Quadroni,F., Soltis,D.E., Soltis,P.S., Zanis,M., Zimmer,E.A., Chen,Z., Savolainen,V. and Chase,M.W. (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature*, **402**, 404–407.
30. Hilu,K.W. and Liang,H. (1997) The *mat*K gene: sequence variation and application in plant systematics. *Am. J. Bot.*, **84**, 830–839.
31. Zhang,L., Jenkins,K.P., Stutz,E. and Hallick,R.B. (1995) The *Euglena gracilis* intron-encoded *mat2* locus is interrupted by three additional group II introns. *RNA*, **1**, 1079–1088.
32. Xiong,Y. and Eickbush,T.H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.*, **9**, 3353–3362.
33. Copertino,D.W. and Hallick,R.B. (1993) Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. *Trends Biochem. Sci.*, **18**, 467–471.
34. Zimmerly,S., Guo,H., Eskes,R., Yang,J., Perlman,P.S. and Lambowitz,A.M. (1995) A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell*, **83**, 529–538.
35. Martinez-Abarca,F. and Toro,N. (2000) Homing of a bacterial group II intron with an intron-encoded protein lacking a recognizable endonuclease domain. *Mol. Microbiol.*, **35**, 1405–1412.
36. Moran,J.V., Zimmerly,S., Eskes,R., Kennell,J.C., Lambowitz,A.M., Butow,R. and Perlman,P.S. (1995) Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. *Mol. Cell. Biol.*, **15**, 2828–2838.
37. Schäfer,B. and Wolf,K. (1999) A novel group-II intron in the *cox1* gene of the fission yeast *Schizosaccharomyces pombe* is inserted in the same codon as the mobile group-II intron aI2 in the *Saccharomyces cerevisiae cox1* homologue. *Curr. Genet.*, **35**, 602–608.
38. Schäfer,B., Kaulich,K. and Wolf,K. (1998) Mosaic structure of the *cox2* gene in the petite negative yeast *Schizosaccharomyces pombe*: a group II intron is inserted at the same location as the otherwise unrelated group II introns in the mitochondria of higher plants. *Gene*, **214**, 101–112.
39. Burger,G., Saint-Louis,D., Gray,M.W. and Lang,B.F. (1999) Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*: cyanobacterial introns and shared ancestry of red and green algae. *Plant Cell*, **11**, 1675–1694.
40. Saldanha,R., Chen,B., Wank,H., Matsuura,M., Edwards,J. and Lambowitz,A.M. (1999) RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. *Biochemistry*, **38**, 9069–9083.
41. Guo,H., Zimmerly,S., Perlman,P.S. and Lambowitz,A.M. (1997) Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J.*, **16**, 6835–6848.
42. Mohr,G., Smith,D., Belfort,M. and Lambowitz,A.M. (2000) Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes Dev.*, **14**, 559–573.
43. Guo,H., Karberg,M., Long,M., Jones,J.P.,III, Sullenger,B. and Lambowitz,A.M. (2000) Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science*, **289**, 452–457.
44. Mills,D.A., Manias,D.A., McKay,L.L. and Dunny,G.M. (1997) Homing of a group II intron from *Lactococcus lactis subsp. lactis* ML3. *J. Bacteriol.*, **179**, 6107–6111.
45. Martinez-Abarca,F. and Toro,N. (2000) RecA-independent ectopic transposition *in vivo* of a bacterial group II intron. *Nucleic Acids Res.*, **28**, 4397–4402.
46. Maier,U.G., Rensing,S.A., Igloi,G.L. and Maerz,M. (1995) Twintrons are not unique to the *Euglena chloroplast* genome: structure and evolution of a plastome cpn60 gene from a cryptomonad. *Mol. Gen. Genet.*, **246**, 128–131.
47. Pace,N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science*, **276**, 734–740.
48. Baldauf,S.L., Roger,A.J., Wenk-Siefert,I. and Doolittle,W.F. (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, **290**, 972–977.
49. Kück,U. (1989) The intron of a plastid gene from a green alga contains an open reading frame for a reverse transcriptase-like enzyme. *Mol. Gen. Genet.*, **218**, 257–265.
50. Doetsch,N.A., Thompson,M.D. and Hallick,R.B. (1998) A maturase-encoding group III twintron is conserved in deeply rooted euglenoid species: are group III introns the chicken or the egg? *Mol. Biol. Evol.*, **15**, 76–86.