The natural history of group I introns

Peik Haugen, Dawn M. Simon and Debashish Bhattacharya

Department of Biological Sciences and Roy J. Carver Center for Comparative Genomics, University of Iowa, 312 Biology Building, Iowa City, IA 52242-1324, USA

There are four major classes of introns: self-splicing group I and group II introns, tRNA and/or archaeal introns and spliceosomal introns in nuclear pre-mRNA. Group I introns are widely distributed in protists, bacteria and bacteriophages. Group II introns are found in fungal and land plant mitochondria, algal plastids, bacteria and Archaea. Group II and spliceosomal introns share a common splicing pathway and might be related to each other. The tRNA and/or archaeal introns are found in the nuclear tRNA of eukaryotes and in archaeal tRNA, rRNA and mRNA. The mechanisms underlying the self-splicing and mobility of a few model group I introns are well understood. By contrast, the role of these highly distinct processes in the evolution of the 1500 group I introns found thus far in nature (e.g. in algae and fungi) has only recently been clarified. The explosion of new sequence data has facilitated the use of comparative methods to understand group I intron evolution in a broader context and to generate hypotheses about intron insertion, splicing and spread that can be tested experimentally.

Introduction

Mobile genetic elements have had a major impact on eukaryotic genomes [1]. Group I introns are small RNAs (they range in size from 250–500nt) and comprise one of the four distinct types of introns (Box 1). These elements are found in a wide variety of organisms (e.g. in fungi, algae and in many other unicellular eukaryotes), genes (i.e. protein, rRNA and tRNA coding genes) and genomes throughout the tree of life. Group I introns spread efficiently at the DNA level into intronless cognate sites by a process termed homing (see Glossary). The success of group I introns in genomes is also a result of their ability to self-splice from RNA transcripts, which potentially renders them neutral to the host. The structure, folding and autocatalysis of a few model group I introns have been rigorously studied, and we now have a detailed understanding of how these elements act as catalytic RNAs (ribozymes) during splicing of precursor RNA [2–4] (Box 2). The mechanism of group I intron spread in DNA has also been studied in detail. Homing endonuclease genes (HEGs) that invade non-critical regions (i.e. terminal loops) of group I introns promote intron mobility by encoding highly site-specific homing endonucleases (HEs). HEGs are themselves selfish genetic elements and the vast majority of group I introns in nature do not contain a HEG (see the following section for intron life cycle). HEs and putative HEs (encoded by HEG-like sequences) are divided into four families based on the presence of conserved amino acid motifs: these are the His-Cys box, LAGLIDADG, GIY-YIG and HNH families. Differences in the endonuclease sequences suggest independent origins of the four families of HE and putative HE proteins [5]. Despite extensive knowledge about group I intron biochemistry, understanding the complex and often unpredictable evolutionary history of group I introns and their associated HEGs has proven to be a formidable challenge. The explosive growth of DNA sequence data in GenBank (http://www.ncbi.nlm.nih.gov/GenBank/GenBankstats.html) has, however, turned the tide enabling a more

Box 1. Types of introns

Four major classes of introns are recognized based on their splicing mechanism. These are autocatalytic group I and group II introns, spliceosomal introns, tRNA and/or archaeal introns. Group I introns are widespread in protot nuclear rDNA genes, fungal mitochondria, bacteria and bacteriophages. These pre-RNA insertions self-splice by a well-characterized and distinctive two-step pathway that relies on an external guanosine nucleotide as the cofactor. Group II introns are found in bacterial and organellar genomes and self-splice through a pathway that is different from group I introns but is similar to the mechanism of spliceosomal intron removal. In these introns, rather than using an external guanosine, the 2′-OH of an adenine residue within the intron is the nucleophile. The shared splicing mechanism suggests that the small nuclear (sn)RNA components of the spliceosome are derived from a group II intron. Spliceosomal introns are the most common insertions found in nuclear pre-mRNA genes. The tRNA introns are found in eukaryotic nuclei and in Archaea and are enzymatically removed by a cut-and-rejoin mechanism that requires ATP and an endonuclease. This pathway is completely different from that of spliceosomal introns.

Glossary

Homing: the process by which an intron (or intein) spreads into the homologous position in an allele that lacks the intron.

Homing endonuclease: a highly specific endonuclease protein that is typically encoded by self-splicing introns. They promote intron mobility through homing.

Introns: intervening sequences in genes that are removed from precursor RNA in a process termed splicing.

Monophyletic: a group of organisms (or molecular sequences) which includes the most recent common ancestor of all of its members and all of the descendants of that most recent common ancestor. Monophyletic groups are also called clades.

Phylogeny: the evolutionary relationships among organisms or genes.

Reverse splicing: the reversal of the forward self-splicing reaction in which a free intron integrates into another RNA molecule. This term is sometimes used to describe the process by which a group I intron becomes stably integrated into the genome following the reverse splicing reaction and subsequent reverse transcription and reintegration into the genome.


Corresponding author: Bhattacharya, D. (dbhhattar@blue.weeg.uiowa.edu).
Box 2. Group I intron secondary structure and splicing

The typical secondary structure of a group I intron consists of approximately ten paired elements (P1–P10, plus the optional P11–P17; Figure Ia) that are organized into three domains at the tertiary structure level [4]. The intron recognizes the 5' exon sequence by a 4–6nt base pairing [the internal guide sequence (IGS) interaction]. Approximately 100 nucleotides constitute the central catalytic core of the intron RNA (the shaded regions in Figure Ia). Group I introns are removed from precursor RNAs through a two-step splicing reaction (Figure Ib). Many introns can self-splice in vitro as naked RNAs (i.e. no proteins are required for intron splicing), whereas others need protein factors to facilitate correct folding of the ribozyme core. A prerequisite for splicing is the binding of an exogenous guanosine (exoG) cofactor to a pocket in the catalytic core of the intron, called the G-binding site. During the first step of splicing, the cofactor attacks the 5' splice site (SS) and attaches to the intron resulting in the release of the upstream exon. The exoG leaves the G-binding site and is replaced by the last splice site (a). Secondary structure and self-splicing of group I introns. Successful catalysis is dependent on the correct folding of the intron.

![Figure 1](http://www.rna.icmb.utexas.edu/) Box 2. Group I intron secondary structure and splicing

The typical secondary structure of a group I intron consists of approximately ten paired elements (P1–P10, plus the optional P11–P17; Figure Ia) that are organized into three domains at the tertiary structure level [4]. The intron recognizes the 5' exon sequence by a 4–6nt base pairing [the internal guide sequence (IGS) interaction]. Approximately 100 nucleotides constitute the central catalytic core of the intron RNA (the shaded regions in Figure Ia). Group I introns are removed from precursor RNAs through a two-step splicing reaction (Figure Ib). Many introns can self-splice in vitro as naked RNAs (i.e. no proteins are required for intron splicing), whereas others need protein factors to facilitate correct folding of the ribozyme core. A prerequisite for splicing is the binding of an exogenous guanosine (exoG) cofactor to a pocket in the catalytic core of the intron, called the G-binding site. During the first step of splicing, the cofactor attacks the 5' splice site (SS) and attaches to the intron resulting in the release of the upstream exon. The exoG leaves the G-binding site and is replaced by the last splice site (a).

**Group I intron distribution in nature**

The distribution of group I introns on the tree of life is shown in Figure 1 (derived from [http://www.rna.icmb.utexas.edu/][6]). We accounted for intron redundancy by representing with single entries, sets of sequences from taxa with identical names that represent population samples (e.g. 59 introns at the same site in different isolates of the green alga *Closterium ehrenbergii*) or different cultivated strains. As of June 8 2004, a total of ~1400 group I introns were reported in eukaryotic genomes. Of these, 800 are in the nucleus at 47 different sites in small subunit (SSU) ribosomal (r)RNA and 44 sites in large subunit (LSU) rRNA, 220 are in mitochondrial genes and 370 are in plastid DNA. Nuclear group I introns are limited to rDNA genes, whereas in the organelles, they are found in rRNA, tRNA and protein coding regions. Group I introns have a sporadic and highly biased distribution in nature and many microbial eukaryotes are rich in these sequences (Figure 1). However, group I introns appear to be absent in animals, except for three cases. Two are found in the mitochondrial genome of the sea anemone *Metridium senile* and one is in the mitochondrial gene cytochrome c oxidase subunit I of the coral *Acropora tenuis*. Group I introns have not yet been found in Archaea.

Lineages in the tree of life that are particularly intron-rich are fungi, plants and the red and green algae. These taxa contain ~90% of all group I introns that have been identified to date. However, DNA sequencing efforts (in particular for organelle genomes) have been biased towards particular groups of genes and organisms (e.g. in animals and plants (of the 623 sequenced, or partially sequenced, mitochondrial genomes, 555 are from animals; [http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html][7]), and it is possible that larger pools of introns remain to be discovered (e.g. in excavates, a group of potentially anciently diverged flagellates that bear a unique cluster of flagellae; [http://microscope.mbl.edu/][8]) as the taxonomic breadth of sequencing programs increase. In contrast to the situation in eukaryotes, group I introns are rare in bacteria. Only fifty introns representing nine different insertion sites occur in this prokaryotic lineage (i.e. four in tRNA genes, four in rRNA genes and one in the recA gene [6–8]). Group I introns also occur infrequently in viruses [9] and phages [10,11] and are not shown in Figure 1.

**Group I intron mobility**

*The evolutionary fates of group I introns*

The sporadic distribution of group I introns suggests that many of these mobile elements have undergone horizontal transfer (through homing or reverse splicing) into different species and genes. The potential evolutionary fates of laterally transferred group I introns are loss from the lineage, vertical inheritance and/or movement within the
lineage with partial loss or fixation (i.e. stable inheritance with no losses and no apparent mobility within the lineage). An additional level of complexity is that many group I introns appear to be mosaic genetic elements. Unlike their counterparts in group II introns, HEGs in group I introns are less likely to share the same evolutionary history as the catalytic RNA [12]. In fact, a group I intron-encoded HEG from one intron (e.g. intron one) can invade another group I intron at the homologous (same) or heterologous (other) position, independent of the group I ribozyme encoded by intron one (see the following section).

The intron life-cycle and homing
In 1999 Goddard and Burt published a model of intron evolution that involved cyclical gain and loss (Figure 2 [13]). Analysis of the yeast omega group I ribozyme showed that a group I intron with a full-length HEG invades an intron-minus population and spreads into all homologous (same) sites through homing (see Ref. [5] for mechanistic details of homing). Once the intron becomes fixed in the population, the HEG no longer has a biological function (i.e. the HEG function is to confer intron mobility), accumulates mutations and is eventually inactivated or lost. Unable to spread, the intron is also destined to be lost over time and will not reappear until the same intron with the full-length HEG is reintroduced through gene flow.

To evaluate if the ‘omega’ life-cycle applies to other HEG-associated group I introns, it is necessary to examine many different introns from natural populations. If multiple populations of an organism that carries a HEG-containing intron are sampled, then one would expect to find introns with full-length HEGs, degenerate HEGs, introns lacking HEGs in addition to examples of intron absence. This pattern fits well with what has been reported previously [6,11,14–20], although every step in the intron cycle is not always found in each case (probably as a result of insufficient sampling). An extensive study of SSU rDNA (S)516 and S1506 (the numbering reflects the homologous position in Escherichia coli rDNA) group I introns from geographically distinct populations of red algae identified individuals lacking introns, those containing introns without a HEG and those containing introns with various deletions and/or mutations in the HEG [17]. In summary, the Goddard and Burt intron life-cycle model, which was established with the yeast omega intron, is supported by the intron and HEG distribution in red algae and is consistent with the intron distribution in other naturally occurring populations. This suggests that recurrent gain and loss is likely to be a general
feature of HEG-associated group I introns. Introns do however manage to ‘escape’ from the cycle by inserting into a new genic position (discussed in the following section).

Introns in the same insertion site are evolutionarily related
Reconstruction of group I intron phylogeny is made possible by the existence of the conserved RNA secondary structures (P1–P9) that can be used to guide the alignment of these sequences. The alignments are used as input to computer programs (e.g. PAUP*, PHYLIP, MrBayes and MEGA) to infer evolutionary trees. The core region used in these analyses is short in length (100–250 nt) but, as shown in Figure 3, it still encodes sufficient phylogenetic signal to resolve many nodes in the intron trees.

An important insight into group I intron evolution that has come from phylogenetic analyses is that sequences from homologous genic sites are closely related (monophyletic) even though they can reside in distantly related organisms (e.g. eukaryotes and prokaryotes) [21]. It is unlikely that this result is explained by a single origin of each intron before the eukaryotic radiation (or the split of bacteria and eukaryotes) followed by widespread loss in different lineages. Introns are highly divergent sequences and it is therefore unlikely that they remain conserved across the millions or billions of years that separate taxa such as prokaryotes and eukaryotes. A more parsimonious scenario is that group I introns are laterally transferred across the tree of life resulting in their sporadic and highly biased distribution. If the second idea is correct, then the observation that intron movement nearly always occurs to the homologous site in different organisms suggests that strong restrictions exist to guide introns to their target (i.e. high sequence specificity in homing).

Intron movement through reverse splicing
Group I introns do however occasionally escape to heterologous genic sites, but the mechanism(s) responsible for these invasions have proven elusive. Because of the limited sequence requirement (4–6nt) for group I ribozymes to locate their target region, reverse splicing of free intron RNAs into target RNAs provides a more plausible model for intron spread into ectopic sites than homing, which relies on a 15–45nt recognition sequence for the intron-encoded homing endonuclease. Whether this pathway has an important role in group I intron spread, however, is unknown because of the lack of any direct evidence of reverse-splicing-mediated intron movement from genetic crosses. Furthermore, in contrast to homing that has been shown to be highly efficient, it is likely that reverse splicing, with its reliance on chance integration followed by two additional steps (i.e. reverse-transcription and then recombination), would be less effective in promoting intron movement. Rare reverse splicing events are therefore most likely to be recognized in the context of a broadly sampled intron phylogeny. A final consideration is that rDNA exists as a multi-copy gene family necessitating that alleles containing transferred introns must rise to high frequency (presumably through concerted evolution or less parsimoniously, repeated reverse splicing events) in individuals and in populations to ensure survival and fixation.

In spite of these constraints, conclusive experimental evidence exists that group I introns can fulfill one crucial step in the reverse splicing mobility pathway, full integration into foreign RNA in vivo [22,23]. In one analysis, 19 variants of the Tetrahymena intron, all with different internal guide sequences were expressed in E. coli [23]. Partial reverse splicing at 69 sites and complete integration at one novel site in the 23S rRNA were found with this approach. Phylogenetic analyses provide support for another prediction of the reverse splicing pathway, statistical support for the close relationship of introns at ectopic rDNA sites (e.g. S1046–S1052, S289–S1199 [18]). In these cases, sequence similarity is limited to the proximal 5′ flanking exon in each pair of related introns. This
finding is consistent with the 4–6nt sequence requirement in the 5’ exon that is predicted for the RNA-based reverse splicing and integration pathway. In addition, HEGs have not been found within the introns at these insertion sites making it less likely that endonucleases mediated the invasions of new sites at the DNA level. Although we still lack conclusive proof of reverse splicing-mediated intron mobility, accumulating experimental and comparative data are in favor of this mechanism having contributed to the spread of group I introns.

**Intron-independent HEG mobility**

It has been known for some time that sequences encoding HEGs insert into the peripheral regions of ribozymes and mobilize them through homing. We now have a more detailed picture of the intron-HEG relationship and the evolutionary history of the two partners turns out to be more complicated than previously thought. The most direct approach to identify intron-independent HEG movement is through the comparison of intron and HEG trees. Unfortunately, this approach is stymied by the rare occurrence of intact HEGs in nature (i.e. they decay rapidly following intron fixation [20]). HEG mobility can, however, leave clear evolutionary footprints (Figure 4). For example, HEGs that are closely related to the His-Cys box family are inserted into different peripheral loops or encoded on different strands of homologous group I introns (e.g. introns inserted after positions 516, 943 and 1506 in the gene that encodes the nuclear SSU rRNA [20]). Movement to different peripheral loop-regions within the intron provides evidence for HEG mobility within a single ribozyme lineage. Furthermore, closely related HEGs are found in phylogenetically distantly related group I introns that are located at adjacent rDNA sites (sites L1923, L1925 and L1926), suggesting that local HEG mobility into heterologous introns occurs [20]. A study of rDNA introns that encode HEGs of the LAGLIDADG family also identified related HEGs in adjacent introns, in some cases, resulting from HEG mobility alone or, in other cases, as a result of the mobility of the intron–HEG element as a unit [24].

**Did a gene duplication event trigger the spread of LAGLIDADG HEGs?**

Homing endonucleases that contain the conserved LAGLIDADG motif represent the largest family of HEGs and are widespread in nature, occurring in several different types of mobile elements (i.e. group I and II introns, inteins and archaeal introns) and as freestanding open
reading frames (ORFs) [25]. Comparative studies suggest that the successful spread of LAGLIDADG HEGs was driven by a gene duplication and fusion event followed by the rapid evolution of the DNA-binding domain [24,26,27]. Approximately 40 group I intron-associated LAGLIDADG HEGs contain one copy of the motif (i.e. LAGLIDADG motif), whereas the remainder contain two copies (>200 cases in total [5,24]). Interestingly, the single-motif HEGs are restricted to introns located at six insertion sites in LSU rDNA; five out of six of these are clustered between L1917 and L1951. By contrast, the double-motif HEGs are more widely distributed in rDNA introns and in introns in other types of genes.

The difference in the success of these HEGs in spreading into novel introns probably reflects how they recognize target sequences. The single-motif HEs function as homo-dimers and, therefore, need a high degree of symmetry in the DNA target sequence, whereas the double-motif HEs function as monomers. The N- and C-terminal halves of the monomeric HEs (i.e. each half corresponds to one protein in the dimers of single-motif HEs) have accumulated many amino acid substitutions and might have a less stringent requirement for DNA target symmetry [26]. Most double-motif LAGLIDADG HEGs in rDNA can be traced to one of two lineages [24]. One of these (designated ‘Clade 1’ in Ref. [24]) appears to be derived from a duplication and fusion event involving a single-motif HEG in an intron located between the rDNA positions L1917–L1951, whereas the origin of the other group of double-motif HEGs (‘Clade 2’) is unclear. In summary, evolution has found an effective solution to target site restriction in LAGLIDADG HEGs. Duplication
and divergence of the ancestral single-motif HEGs to form a two-motif unit appears to have freed them to explore a larger set of host introns and genic sites.

The role of proteins in group I intron splicing

**HEGs can avoid loss by gaining maturase activity**

A fascinating aspect of the group I intron-HEG relationship is that some endonucleases also function as matu-

rases (i.e. they assist in intron folding [28]). The existence of these bifunctional proteins suggests that the gain of maturase activity has provided the HEs and/or their associated introns with an evolutionary advantage. So what advantage could maturase activity offer? One hypothesis is that the gain of a HEG-encoded DNA (typically ~1kb in size) in the group I intron might lower the efficiency of self-

splicing. The maturase activity conferred by the HEG could therefore be selected for because it rescues or sufficiently improves splicing ability to compensate for the presence of the ORF [28]. Another intriguing possibility is that maturase activity evolves (or is retained) because it gives the HEG an opportunity to escape inactivation or loss [29] (i.e. endonuclease function will be lost if the intron is fixed in the population, whereas maturase activity might be retained to ensure correct intron excision). This scenario appears to have played out recently in *Saccharomyces cerevisiae* in which two non-adjacent amino acid changes activate endonuclease activity in an intron-encoded LAGLIDADG maturase [30,31]. Other important insights have come from structural and biochemical studies of the I-AnI HE protein [29,32,33]. In this case, the RNA and DNA-binding domains of the protein are non-overlapping (i.e. they are located at different places in the protein structure). It is therefore likely that the maturase activity evolved as a secondary adaptation. In summary, there appears to be a dynamic relationship between ribozyme, HE and maturase activities in HEG-associated group I introns. The relative evolutionary importance of these processes reflects primarily the potential for intron spread in populations or species and the extent of dependence on the maturase for intron splicing.

The long-term evolution of self-splicing ability

Insights into the evolution of RNA structure and auto-

catalysis can be gained by comparing group I introns that have been vertically inherited for millions of years. These sequences provide potential examples of intron-host co-evolution and are candidates for introns that might have evolved a function that is beneficial to the host. Two outstanding examples of long-term vertical inheritance are the ancient group I intron inserted in the tRNA-Leu gene of cyanobacteria and plastids [34–36] and the S788 intron in the ascomycete fungi [37]. The tRNA-Leu intron has probably been vertically inherited for more than a billion years in plastids [38] and two-to-three billion years in the cyanobacterial ancestors of these organelles [34], whereas the S788 intron was probably acquired in ascomycetes ~400–600 million years ago [37]. Although there are some obvious differences in the evolutionary history of these two group I introns, there are also striking similarities. Most importantly, when self-splicing activity is mapped on the intron phylogeny (which generally recapitulates the host tree as a result of intron vertical inheritance), in both cases the early diverging ribozymes self-splice efficiently as naked RNAs. By contrast, the later
diverging introns can either only complete the first step of splicing [i.e. attack of the 5′ splice site by the exogenous guanosine which becomes covalently attached to the intron (the 5′ exon is freed)] or lack autocatalytic activity entirely (Figure 5). It is tempting to speculate that the partial or complete loss of activity in both riboazines is due to their dependence over evolutionary time on particular host proteins to facilitate splicing. Furthermore, the fixation of the tRNA-Leu intron in all plants combined with their loss of self-splicing ability suggests that a reciprocal dependence between the intron and the host might have evolved in the common ancestor of this highly successful lineage. Although unproven, this idea is supported by a study that shows chloroplast ribonucleoproteins (cpRNPs) associate with unspliced tRNA-Leu transcripts in Nicotiana tabacum [39]. The cpRNPs–tRNA complexes confer stability and ribonuclease resistance to the RNAs and potentially act as a scaffold for host-mediated splicing of the intron-containing tRNAs [39].

The S788 group I intron is also of great interest because derived members of this ribozone lineage have undergone major structural changes with many nucleotide substitutions in the central catalytic core, some of which are crucial for intron folding, and the loss of the larger peripheral P5abcd domain (Figure 5). Two distinct clades are identified in phylogenetic analyses of the S788 intron, an ancestral pool that contains the P5 extensions (+ P5abcd, Figure 5) and a later diverging clade that lacks this domain (−P5abcd) [37]. The loss of P5abcd is correlated with the inability to self-splice in vitro. This example therefore provides a direct link between group I intron structure evolution and autocatalysis in a vertically inherited intron.

Concluding remarks
The detailed mechanisms of intron autocatalysis and mobility have been established through rigorous biochemical analyses. Intron evolution is being clarified through the analysis of intron phylogeny, distribution, secondary structure and splicing in a diverse group of microbial eukaryotes. Bringing together these two disparate pools of information is essential to understand the natural history of group I introns. Recent work has successfully integrated structural and functional insights regarding group I introns into a broader evolutionary context. Examples of important contributions to our understanding of intron biology and evolution are the compilation of introns into a comprehensive database [6], development of the cyclical model for the gain and loss of HEG-associated introns [13], novel insights into the dynamic intron-HEG relationship [24,26,27,29] and the recognition that group I introns can be vertically inherited over long periods of evolutionary time and their interaction with the host cell can provide important insights into intron biology [34–37,39]. It will be important to investigate the role of host factors in intron splicing across the tree of life (e.g. mitochondrial tyrosyl-tRNA synthetase (Cyt-18, [31]), leucyl-tRNA synthetase (LeuRS, [40]) and mitochondrial RNA splicing 1 (Msr1, [41]) and the contribution of these factors to intron fixation and mobility. It is likely that host proteins generally have a role in the splicing of many group I introns; however, their broader impact on intron evolution is currently unknown. Another unanswered question is how group I introns are transferred between species. Viruses have often been suggested as possible vectors but evidence is still lacking. Finally, the true biological role(s) for group I intron RNAs remains to be clarified (for more details, see Ref. [42]).

Acknowledgements
This work was generously supported by grants awarded to D.B. from the National Science Foundation (MCB 0110252, DEB 0107754), a grant from The Norwegian Research Council to P.H., and Avis E. Cone and Stanley fellowships from the University of Iowa to D.S. We thank H. Joseph Runge (Iowa) for helpful discussions.

References
12 Toor, N. et al. (2001) Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. RNA 7, 1142–1152
40 Rho, S.B. et al. (2002) An inserted region of leucyl-tRNA synthetase plays a critical role in group I intron splicing. EMBO J. 21, 6874–6881
42 Nielsen, H. et al. (2003) The ability to form full-length intron RNA circles is a general property of nuclear group I introns. RNA 9, 1464–1475