Group I introns and inteins: disparate origins, convergent parasitic strategies

Rahul RaghavanΨ and Michael F. Minnick*

Division of Biological Sciences
The University of Montana
Missoula, Montana 59812

Running title: Convergent evolution of group I introns and inteins

*Corresponding author. Mailing address: Division of Biological Sciences, The University of Montana, Missoula, Montana 59812. Phone: 406-243-5972. Fax: 406-243-4184. E-mail: mike.minnick@mso.umt.edu.

Ψ Present address: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721
Genomes of most organisms harbor DNA of foreign origin that serves no known function. Since these elements may not contribute to a host’s fitness but utilize host resources for their perpetuation, it is appropriate to consider them as genetic parasites (4). With the advent of sequencing technologies, a wide variety of parasitic elements have been discovered in bacteria from all environments, including obligate intracellular pathogens (100) that were thought to be shielded from horizontal gene transfer (HGT). Detection of a parasitic genetic element in a genome represents only a snapshot of the continuing and dynamic interplay between the host’s attempts to purge the element and the element’s ability to persist. These adaptable genetic parasites have evolved mechanisms to overcome defenses (75) of the cellular machinery to ultimately invade, colonize and replicate within the host. Their success is very evident in the human genome, which consists mostly of such apparently superfluous DNA (65). Even compact bacterial genomes packed with functional genes contain mobile genetic elements (100), underscoring their universality in nature.

A number of parasitic genetic elements are found in bacterial genomes including transposons, insertion sequences, prophages, introns, inteins and intervening sequences. While bacteria, especially pathogenic bacteria, are well studied, their parasitic genetic elements have not received as much attention. In the past few years while studying the obligate intracellular pathogen *Coxiella burnetii*, we came to appreciate the intimate relationship shared between bacterial hosts and parasitic elements (92, 93). In addition, interesting new studies have shed light on the evolutionary histories of group I introns, inteins and homing endonucleases (9, 109) and infused excitement into the field. This mini-review, which focuses on the biology and evolution of group I introns and inteins
found in bacteria, is an attempt to catalyze interest among bacteriologists for these fascinating genetic parasites.

### Structure and Biology

#### Group I introns

Introns are non-coding, intragenic regions that are removed from precursor RNA to form the mature RNA by splicing the exons (coding regions that flank introns) together. They are much more common in eukaryotes than in bacteria (50). Introns are classified into four groups: group I, group II/group III, spliceosomal and tRNA/archaeal introns based on splicing mechanisms (47). Spliceosomal introns are found in eukaryotes and utilize spliceosomes (large protein-RNA complexes) for splicing (70), whereas tRNA introns splice with the help of specialized enzymes (74). Group I and group II introns are able to self-splice - using different mechanisms - without the aid of any proteins and are thus referred to as ribozymes (104). Self-splicing by a group I intron from *Tetrahymena thermophila* was one of the first ribozymes to be described, in the early 1980s (61).

Ribozymes are considered legacies of a primordial RNA world, where RNA possessed both information-encoding and catalytic properties, before the advent of DNA and protein-based life forms (35).

Group I introns are small RNAs (~250 to 500 nt in size) that have invaded protein-, rRNA- and tRNA-encoding genes in a variety of organisms, including algae, fungi, lichens, some lower eukaryotes and a few bacteria (47). While the first bacterial group I intron was not discovered until 1990 (63, 116), the recent availability of inexpensive and accurate whole genome sequencing technologies have made it possible to identify these mobile elements in a number of bacterial species from diverse ecosystems. All bacterial
group I introns analyzed to date have been shown to self-splice. An exception is an intron of *Simkania negevensis*, which reportedly remains unspliced in the mature 23S rRNA (33). Also, akin to the scenario in eukaryotes, *C. burnetii* and some *Synechococcus* strains contain multiple introns interrupting the same gene (45, 93).

All group I introns share a conserved secondary structure (Fig. 1A), which consists of paired elements (P) that assist in self-splicing by using a guanosine (or GMP or GTP) as a cofactor (110). P4-P5-P6 and P3-P7-P9 form two separately folding helices within the core. Helix P3-P7-P9 contains the binding site for the guanosine (G-binding site, GBS) and is the minimal catalytic domain required for splicing (51). P1 and P10 are complementary to 5’ and 3’ exons, respectively, and are collectively termed the internal guide sequence (IGS) (113). Based on secondary structure, group I introns are classified further into 13 subgroups (78, 105).

In the first step of splicing (Fig. 1B), the 3’-OH group of an exogenous guanosine bound to GBS carries out a nucleophilic attack on the 5’ splice site, which is marked by a conserved G•U wobble pair within P1. After the first step, this guanosine is covalently bound to the free 5’ end of the intron and leaves the GBS, allowing the conserved terminal guanine (ΩG) to occupy the GBS and mark the 3’ splice site. In the second splicing step, the 3’-OH group of the free 5’ exon attacks the 3’ splice site, in a reaction that is chemically equivalent to the reverse of step 1, resulting in ligation of 5’ and 3’ exons and release of the intron (103). Excised introns have been observed to circularize, but the significance of this property is not clearly understood (82). An exception to the splicing mechanism described above was observed in an intron (Cbu.L1917) located in the 23S rDNA of *C. burnetii* (93). This intron has a 3’ terminal adenine in place of the
otherwise conserved guanine. Consequently, Cbu.L1917 has a retarded rate of self-splicing in vitro (91).

Group I introns are mainly found inserted in tRNA and rRNA genes of bacteria. Increasingly, they are being found in a variety of protein-coding genes, including recombinase A and ribonucleotide reductase (77, 108). In bacteriophages, they are seen in both tRNA genes and in some protein-coding genes like DNA polymerase, ribonucleotide reductase and thymidylate synthase (45). A bias towards disrupting structural RNA genes in bacteria could be due to the coupling of transcription and translation, which might prevent the ribozyme from attaining its optimum tertiary structure required to splice efficiently (84). However, introns that require concordant translation for efficient splicing are also known (96, 99), showing that introns adapt to their environment. Sexual reproduction in eukaryotes brings intron-containing and intronless alleles of the host gene together, providing an opportunity for introns to spread by homing (see below). Even though rampant HGT provides ample opportunity for the movement of introns and inteins, a lack of sexual reproduction is commonly invoked as an explanation for the apparent scarcity of group I introns in bacteria when compared to mitochondria and chloroplasts of lower eukaryotes (29). Another possible reason for this phenomenon is the observed inhibition of bacterial growth caused by these elements. For example, group I introns of Tetrahymena and Coxiella, when expressed in E. coli were found to associate with ribosomes, inhibit translation and retard bacterial growth (83, 92). Due to this low-fitness trait, group I introns, like any other gene that decreases reproductive success of the host would presumably be lost from the population by negative selection.
Various group I introns have evolved to associate with other parasitic elements. Twintrons (30), where two distinct group I introns are found associated with each other, and IStrons (77, 107), where a group I intron and an insertion sequence (IS) are seen together, are two such cases. Rarely, spliceosomal introns are also found within group I introns (48). As we will see later, group I introns also associate with endonucleases.

**Inteins**

Similar to ribozymes, another iconoclastic discovery was that of inteins (internal proteins). These elements are transcribed and translated together with the host protein, but self-excite, leaving the flanking sequences (exteins) spliced together. The first intein was discovered in yeast vacuolar ATPase genes in 1990 (56). Since then, hundreds of inteins have been discovered in Bacteria, Archaea and Eukaryotes. An intein database, InBase has been established to provide information on all known inteins (87). In bacteria, inteins are found inserted in a variety of conserved proteins including DNA polymerase, helicase, gyrase, recombinase A and ribonucleotide reductase; whereas in bacteriophages inteins have been reported in DNA polymerase and ribonucleotide reductase (87). The ratio of intein size to host protein size varies widely with some inteins four times as large as the host protein and others only one-tenth the size of the host protein (87).

Inteins have a modular organization consisting of three functional domains: N- and C-terminal splicing domains and an optional endonuclease domain (Fig. 2 A) (90). The N-terminal domain is comprised of four motifs, A, B, N2 and N4. The C-terminal domain contains two motifs, F and G. The central endonuclease domain consists of four motifs, C, D, E and H. The N- and C-terminal motifs are involved in protein splicing and are conserved in most inteins, whereas motifs C, D, E and H are absent in a number of inteins.
(referred to as mini-inteins). The first and last amino acids of the intein and the first amino acid of the C-terminal extein are involved in the splicing reaction (39). The first amino acid (motif A) in all inteins is Cys or Ser; the terminal amino acid is a conserved Asn, and the first amino acid that follows the intein is Cys, Ser or Thr. Intein splicing involves four successive nucleophilic displacements (Fig. 2B) (87). The first step is an N/O/S acyl shift where the OH or SH side chain of the amino-terminal Ser or Cys attacks the carbonyl carbon of the preceding amino acid to generate an ester/thioester intermediate linking the N-terminal extein to the side chain of the first intein amino acid, thereby breaking the peptide bond between N-extein and the intein. The second step is a transesterification, where the OH or SH side chain of the first C-extein amino acid attacks the N-terminal ester/thioester bond formed in step 1. This results in the transfer of the N-extein to the side chain of the first C-extein amino acid, forming a branched intermediate.

In the third step, the peptide bond between the intein and C-extein is broken by cyclization of the conserved C-terminal Asn to form a succinimide, resulting in intein excision. The N-extein is now attached via an ester bond to the side chain of the first C-extein amino acid. In the final step, the ester bond rapidly undergoes an acyl rearrangement to the thermodynamically more stable, normal peptide bond. In most inteins, the amino acid preceding the terminal Asn is an His, which is thought to assist in Asn cyclization. Part or all of the remaining residues are also important for proper intein folding to generate the active site (21). Some non-canonical inteins have been identified with variations in structure and splicing chemistries (3, 42, 88).

In some cases, an intein and its host protein (e.g., DnaE) are split into two separate fragments (32, 115). The N- and C-termini of DnaE containing the N- and C-termini of
the intein, respectively, are encoded on two separate genes, *dnaE*-n and *dnaE*-c, located on different parts of the genome. Functional DnaE protein is recreated from the two fragments by the *trans*-splicing activity of the split-intein (76). Split-inteins have also been found in other enzymes like ribonucleotide reductase, DNA ligase, gp41 and IMP dehydrogenase (23).

**Intron and intein mobility through homing endonucleases**

One of the most successful parasitic genetic elements is the homing endonuclease (HE). HEs are simple and elegant parasitic elements that consist of a single gene that encodes a single protein and are inherited in a dominant, non-Mendelian manner (12). HEs are small (<40 kDa) proteins that recognize and cleave long DNA target sequences (usually 14-40 bp), that typically occur only once per host genome, thus minimizing any potential negative impact on the host (80). Based on conserved sequence motifs, HEs can be classified into several families (36, 62) that have evolved in parallel to achieve optimal balance of size, target sequence specificity and attenuated fidelity to allow for maximum success (97, 120). The LAGLIDADG family has one or two conserved motifs called the dodecapeptide motifs with a consensus LAGLIDADG sequence (22), and most HEs found to date belong to this family. The ββα-Me family contains two sub-families. The His-Cys box sub-family contains a ~30 amino acid region with two His and three Cys (54), whereas the H-N-H sub-family contains a ~30 amino acid region with conserved His and Asn residues (102). The GIY-YIG family contains conserved ‘GIY’ and ‘YIG’ tripeptides flanking a ~10 amino acid segment (60). Some HEs utilize catalytic domains acquired from other proteins (or vice versa). An HE that utilizes the PD-(D/E)-XK motif commonly employed by restriction endonucleases (120), another which uses a domain
similar to that of DNA resolvases (119), and a novel HE related to very-short patch repair
(Vsr) endonucleases (23) were recently described. Interestingly, an H-N-H endonuclease
makes up the cytotoxic domain of colicin E9, a group A colicin that kills bacteria by
nonspecific degradation of chromosomal DNA (81). In theory, any endonuclease that can
cause a double-strand break and initiate recombination through flanking-sequence
similarity can function as an HE. In fact, the restriction enzyme EcoR1 was
experimentally made to stimulate intron homing (27).
Insertion of an HE into a gene can potentially impair its function. In order to limit the
negative impact on a host and loss by negative selection, HEs tend to associate with other
self-splicing elements like group I introns and inteins that are nearly neutral to selection
(6). Together, HEs and introns/inteins have formed a successful, mutually beneficial
association wherein the HE provides mobility to the intron/intein and the intron/intein
provides a “safe haven” for HE. The process by which the composite element moves
from one site to another is called homing (Fig. 3A) (55). The homing mechanism requires
that the protein be translated. When an intron/intein-containing allele comes together
with an intron/intein-lacking allele, the HE protein binds to a homing site composed of
the flanking exon/extein sequences and cleaves it. The host repairs this double-strand
DNA break using homologous recombination between the alleles, which results in
insertion of the parasitic element into the target sequence. The site is now “immune” to
further HE cleavage because the inserted element disrupts the target sequence. Some HEs
belonging to the H-N-H sub-family create single-strand nicks instead of double-strand
breaks and also recognize intron-plus DNAs as substrates (41). The process by which
recombination and homing occur after a single-strand nick is not clearly understood. HEs
tolerate a degree of variation within their long recognition sequence, which enables them
to coevolve with the host target sequence (97) and move to ectopic sites (19). Another
reason for HEs’ success is their adaptability to a new host, which conceivably helps to
explain the divergent DNA binding regions observed between similar HEs from different
hosts (18, 73).

HEs go through a dynamic cycle that includes invasion, fixation, inactivation,
elimination and eventual reinvasion (Fig. 3B) (37-39). Once an HE-containing parasitic
element invades a new host, it spreads to all the individuals in that host’s population and
becomes fixed in that population. Once fixed, the HE becomes non-functional (since
there are no available target sequences), starts to degenerate and eventually becomes lost
(13). In fact, a large number of group I introns and inteins have lost their respective HE
genes. The intron/intein itself will maintain its sequence because any change will affect
its splicing ability, thus negatively impacting the host. Eventually, the whole element is
lost from the population by a precise deletion event. The parasitic genetic element
reappears in the population only through a new HGT event; a critical process for the
long-term maintenance of a parasitic genetic element in a population. Some exceptions to
this ‘homing cycle’ model have been previously described (38). To prevent being purged
from a genome, some HEs utilize an intriguing strategy: they have evolved a maturase
function (98). Maturase activity promotes intron splicing by stabilizing RNA folding. To
function as a maturase, HEs have evolved an RNA-binding site in addition to their DNA-
binding site, showing their adaptability (72). Some HEs confer beneficial functions upon
their hosts. VDE (also known as PI-SceI), an HE found inserted in a self-splicing intein in
the VMA-1 gene of Saccharomyces cerevisiae, is one of the main regulators of the host’s
high-affinity glutathione transporter (79). HO (F-SceII), a freestanding HE in *S. cerevisiae*, mediates mating-type switching (14). An intron-encoded HE was shown to provide a selective advantage to intron-containing *Sulfolobus acidocaldarius* cells over cells without the intron (1), and I-HmuI, an intron-encoded HE of *Bacillus subtilis* phage SP82, is required for exclusion of DNA from the related phage SPOI, in the progeny of mixed infections (41).

Another mechanism of intron mobility involves reverse splicing. In this process, excised intron-RNA base-pairs with host RNA sequences that are complementary to its IGS, followed by integration into the transcript by a reversal of the splicing process. The intron then becomes inserted into the corresponding gene through reverse transcription followed by recombination. Reverse splicing has been demonstrated in the lab (95) and inferred from intron transposition patterns (8, 49).

**Disparate origins, convergent parasitic strategies**

Both group I introns and inteins arose from preexisting molecules with autocatalytic abilities (Fig. 4). The progenitor of group I introns is thought to be a prebiotic catalytic RNA (66). These primordial ribozymes were part of the ‘RNA world’, which predates the current DNA- and protein-based biology. Some of the relics of the era where RNA functioned as a catalyst, are found strewn across the contemporary biosphere: RNase P, ribosomes, spliceosomes, telomerase and self-splicing introns (16). Although RNA is a good informational molecule and a powerful catalyst, it is not clear how these molecules replicated themselves in the pre-protein world. Recently, Vicens and Cech showed that group I introns have the potential to polymerize RNA chains by forming 3’, 5’ phosphodiester bonds, a key milestone in the search for a prebiotic replicase (109). In
addition, earlier in vitro evolution studies demonstrated that group I introns can catalyze RNA ligation (52, 118). Taken together, these observations suggest that modern group I introns arose when the self-splicing activity of a primordial replicase-like ribozyme was exploited for mobilization of the molecule as a parasitic genetic element.

Intein, the protein analog of a group I intron, is thought to have evolved from an ancient protein domain possessing the ability to self-cleave. A number of proteins involved in important biological processes are known to self-cleave (86). Among them, Hedgehog developmental proteins found in eumetazoans are evolutionarily related to inteins (43, 58). The self-cleaving C-terminal portion (Hog domain) of these proteins has a domain called Hint (Hedgehog and intein) that shares structural, sequence and biochemical similarities with inteins (58). The Hog domain autocatalyzes its cleavage and attaches a cholesterol moiety to the N-terminal Hedge domain. The cleaved Hedge domain with the attached cholesterol is secreted from the cell and serves as a signaling molecule (11). In addition to inteins and Hedgehog proteins, domains related to Hint have been identified and termed bacterial intein-like domains (BILs). Three types of BILs have been identified so far (BIL-A, -B, -C) from diverse bacteria, including human and plant pathogens, and predatory bacteria (2, 26). Each type of BIL is as different from one another as they are from Hedgehog-Hint and intein (11). BILs are hypothesized to generate host protein diversity and aid in host microevolution (2, 26). Since the initial N-S/O acyl shift common to all Hint domains is employed by numerous other self-cleaving proteins involved in diverse biological processes (117), it is presumed that the progenitor Hint domain arose by positive selection for some advantageous biological function.

Duplication of the primordial Hint domain followed by a loop exchange would have
resulted in a protein that autocleaves but leaves the host protein intact by splicing together the N- and C- terminal flanks (69); setting the stage for the evolution of a parasitic protein.

The primary requirement of a successful genetic parasite is to cause no harm to the host, which both group I introns and inteins accomplish by splicing out at either the RNA or protein stages, respectively. This attribute aids in their maintenance only to a point, because the lack of benefit to the host puts them under constant “threat” of being purged from the compact bacterial genome, which is biased towards deletion. To counter loss by deletion, group I introns and inteins have evolved convergent strategies to improve mobility, maximize insertion-site availability and minimize the probability of being lost from a genome.

Group I introns can independently move from their insertion site to a new site in the same organism (transposition) or to the same site in a different organism (HGT) through reverse splicing (95, 114). A 4 – 6 bp sequence complementarity between the intron’s IGS and target RNA is all that is required to initiate reverse splicing, effectively providing any group I intron with a large pool of potential targets (114). However, successful integration of the intron into DNA will only occur if the target RNA along with the intron is reverse transcribed and subsequently undergoes recombination; surely a rare sequence of events that limits the spread of introns (29). Even when an intron reverse splices successfully and inserts into a new target DNA sequence, the forward splicing efficiency is much lower than at its natural site (89, 94), further restricting the spread of introns via reverse splicing. In the case of mini-inteins, it is not clear how they move from one insertion site to another. They do not employ recognizable transposition
mechanisms, and their coding DNA or RNA is not known to separate from the host sequence. They might be able to transpose to a new location along with their flanking sequences by a rare non-homologous recombination event. However, for this insertion to be successful, the intein must be in-frame, be active at its new site and the host protein should be functional even with the inserted intein-flanks (89). Many Group I introns and mini-inteins have solved these mobility limitations and broadened their potential target repertoire by linking up with endonucleases (25, 47).

The evolutionary steps that brought group I introns and inteins together with HEs to form composite mobile genetic elements have intrigued scientists since their discovery. Several studies have shown that HEs and their corresponding introns or inteins have separate phylogenetic histories (36, 46). Also, similar inteins and introns were found to contain different types of HEs; and closely-related HEs are known to be encoded within distantly-related introns (48). Moreover, ‘free-standing’ endonucleases that are mobile, even without being associated with an intron or an intein, have been found abundantly in some bacteriophage genomes (7, 101). Taken together, these observations suggest that mobile inteins and group I introns evolved by repeated, independent invasions of mini-inteins and endonuclease-free introns by HEs (25). Introns/inteins and HEs can come together by recombination. The newly formed bipartite intron/intein will be maintained only if intron/intein splicing is maintained and if the HE can promote the spread of its host element by homing. To this end, HEs are always found inserted in peripheral loops that do not play a role in intron splicing or in intein-domains that are not essential for splicing. Evolutionary forces that drive this process recently came to light when David Shub and colleagues showed that the propensity of introns and HEs to target highly
conserved sequences within conserved genes might have brought them together (9, 119) (Fig. 4).

The best strategy for parasitic elements to increase the prospect of finding an insertion site is to target DNA sequences that are most commonly encountered in a gene pool. Genes that play essential biological roles tend to be conserved across the biological spectrum and consequently serve as frequent insertion sites for parasitic genetic elements. Within these genes, group I introns and inteins insert into sequences that remain ‘static’ due to their functional importance (59, 97), further maximizing target availability. Only a rare, precise deletion event that removes the element and exactly recreates the host gene sequence will result in functional gene products, whereas an imprecise deletion can potentially mutate critical regions within the gene resulting in deleterious consequences to the host. Hence, targeting highly conserved sequences within conserved genes offers the additional benefit of minimizing the likelihood of the molecular parasite being purged from the host genome (28).

It has been known for some time that free-standing HEs can move between bacteriophage genomes through a process called intronless homing (7). An example is the free-standing HE (SegF) in T4 phage that targets and cleaves a conserved sequence within the adjacent gene 56 of related T2 phage during mixed infections. Repair of the double-strand break results in the replacement of T2’s gene 56 with that of T4’s gene 56 along with the insertion of SegF. In this manner, even though the HE is inserted in a less-conserved intergenic region, its maintenance and mobility are maximized by targeting a neighboring conserved sequence (28). As discussed above, introns and inteins also target conserved genes for insertion. Since bacteriophages have a limited repertoire of
conserved genes, it is inevitable that an intron/intein and an HE targeting the same
conserved sequence come together in the same phage genome. When this happens, the
intron/intein and the free-standing HE can move together from one host to another by a
process termed ‘collaborative homing’ (9, 119). A rare recombination event can insert the
HE into the intron/intein without affecting its splicing, thereby giving rise to a composite
parasitic element. This stable chimera can now efficiently spread through the population
by homing to cognate sites and rarely even to ectopic sites. Alternatively, in some cases
an HE might invade an intron due to the presence of a cleavage site within a peripheral
loop of the intron (48, 71). In this event, the HE must adapt quickly to recognize and
cleave intronless alleles of the target gene so as to facilitate homing, which in turn will
ensure its maintenance within the intron. The intron/intein provides a “safe haven” for the
HE where it can evolve quickly to “acclimate” to the new target sequences. HEs are fast-
evolving genes (18, 44) that are known to use a wide array of protein scaffolds (119, 120)
of diverse origins. Some HEs are even chimeras, with N- and C- termini from different
sources (7). Moreover, double- LAGLIDAG motif HEs that have evolved from single
motif ancestors are more successful at invading divergent target sites, thereby promoting
the spread of the composite genetic parasite to ectopic sites (44).

Bacteriophages are an ancient and genetically non-homogenous group that is thought to
be the most abundant biological entity on earth (10, 15, 34). High rates of homologous
and non-homologous recombination, rapid evolution and profuse genetic exchange
provide phages with powerful means of innovation. Hence, they are considered to be the
‘start-up entities’ where several new bacterial genes originate (24). In the same vein,
phages most likely are the ‘melting-pot’ where inteins and group I introns associate with
free-standing HEs to form composite parasitic elements. For homing to take place, both intron-positive and intron-minus alleles of the host gene must come together. It is not clear how this could happen frequently enough in bacteria to facilitate the spread of group I introns and inteins (13). By contrast, bacteriophages have a global gene pool with intense horizontal exchange occurring among genetically related local groups of phages and low-grade exchange of sequences occurring over wide phylogenetic distances (34), making them ideal vehicles for the transfer of HEs, introns and inteins between phages, amongst bacteria, and between phages and bacteria. Further, in lysogeny the prophage provides a silent locus from which HEs can invade homologous host genes in bacteria (108). Introns and inteins also tend to target conserved genes that have both phage and bacterial copies; making insertions less toxic and improving the odds of homologous recombination (59, 77). In addition to transduction, a specialized conjugation system (111) and natural competence (17) might explain the observed abundance of inteins and introns, respectively, in *Mycobacterium* (87) and *Bacillus* (108) species.

**Harnessing genetic parasites**

Similar to inteins, group I introns have also been shown to mediate *trans*-splicing of exons contained on different RNAs (57). This property has potential use as a gene therapy tool. In fact, group I introns have been used to convert sickle β-globin transcripts into RNAs encoding γ-globin in an effort to treat sickle cell disease (64). *Trans*-splicing ribozymes also have the potential to be used for targeted gene delivery and as therapeutic cytotoxins (5, 57).

The protein splicing ability of inteins has been exploited as a biotechnology tool. Inteins can be used as tags to purify fusion proteins in place of traditional histidine tags. After
purification, the intein tag can be removed by utilizing the self-cleaving property of the intein (112). Other potential uses include the application of inteins for the semisynthesis of cytotoxic proteins (31) and for introducing nuclear magnetic resonance (NMR) labels into part of a large protein (85). The trans-splicing ability of split-inteins has been utilized for the synthesis of cyclic peptide libraries and in gene therapy (67, 106).

The ability to introduce specific double-strand DNA breaks makes HEs a very useful genetic tool. Some HEs are commercially available, e.g., I-Ceu I, I-Sce I and PI-Psp I (New England Bio Labs). HEs, along with other rare-cutting restriction enzymes, have been used to map bacterial genomes, especially to analyze chromosomal organization (68). HEs have been used to study double-strand break repair mechanisms in phages, yeasts, plants and mammalian cells and to study chromosomal repair systems in Drosophila (40, 53). Recently, artificial HEs were engineered with the aim of using them in human gene therapy (20).

In conclusion, self-splicing group I introns and inteins in bacteria originated from disparate self-cleaving sources, but evolved convergently to target conserved gene sequences and in turn to associate with homing endonucleases to maximize their persistence and spread. Bacteriophages might be the vessels where composite inteins and introns originate and the vehicles for their spread. With the current expansion of genomic data, more mobile genetic elements are being reported, and a concerted effort is needed to analyze and understand all of them. Decoding the unique biological and chemical properties of these intriguing elements will provide us with novel tools for industrial and medical applications.
ACKNOWLEDGMENTS

We thank the members of our lab for helpful discussions and technical assistance. Our work was supported by the NIH Rocky Mountain Regional Center of Excellence for Biodefense and Emerging Infectious Disease grant, U54 AI065357-040023 and NIH R21 AI078125. We are grateful to the reviewers for their comments, which improved the manuscript immensely.

FIGURE LEGENDS

**FIG. 1.** A. Predicted secondary structure of a group I intron (Cbu.L1951). Paired, conserved helices common to group I introns are designated P1 to P10. The 5'- and 3'-terminal intron bases are encircled. Intron sequence is in uppercase; 5'- and 3' exons are in lowercase and colored red and blue, respectively. P1 and P10 together form the IGS. The site of HE insertion in P8 is indicated as “HE” in green. Modified from (93). B. Mechanism of group I intron splicing. 5’ and 3’ exons are in red and blue, respectively. ΩG, terminal intron guanine. G*, exogenous guanosine. (1). Nucleophilic attack on 5’ splice site by the 3’-OH of G* in GBS. (2). Nucleophilic attack on 3’ splice site by the free 3’-OH of 5’ exon. (3). Free intron and spliced exons. Modified from (110).

**FIG. 2.** A. Intein modular structure. An intein with flanking exteins is shown. The N- and C-terminal splicing domains (in red) and the optional homing endonuclease (HE) domain (blue) with their conserved motifs are shown. Conserved amino acids Cys or Ser at the 5’ end of the intein, Asn at the 3’ end of the intein and Cys, Ser or Thr at the first position
on the 3' extein are also indicated. Modified from (87, 90). B. Intein splicing chemistry.

N- and C-exteins in red and blue, respectively. Modified from (39, 87).


FIG. 4. Convergence of evolutionary paths of group I introns, inteins and HEs.

REFERENCES


44. **Haugen, P., and D. Bhattacharya.** 2004. The spread of LAGLIDADG homing endonuclease genes in rDNA. Nucleic Acids Res. 32:2049-2057.


Fig. 1B. Raghavan & Minnick

Ω
(1) -GBS
OH-G
(2)
(3)
OH ΩG
-GBS
+
G*
ΩGG*
B
Fig. 2A. Raghavan & Minnick

Amino-extein

Carboxy-extein

Conserved motifs: A    N2      B    N4        C             D  E    H      F     G ... domain
Conserved 
amino acids:
Cys
or
Ser
His Asn Cys
“    “    “    “    “
Ser 
Thr

Fig. 2A. Raghavan & Minnick
Fig. 2B. Raghavan & Minnick

(1) N-O or S-O shift-

(2) Transesterification-

(3) Asn cyclization-

(4) O-N or S-N shift-
A

Intron / intein* allele

HE gene

-site cleavage

Intron / intein allele

Recombinational repair using intron/intein and HE as template

HE gene

Insertion of intron/intein and HE into new site
B. Homing Cycle-

Empty Target Site

- Intron / intein with functional HE
  - Homing (HGT)
- Intron / intein without HE or with dysfunctional HE
  - Precise Loss
- Degeneration
- Fixation

Intron / intein with HE from a donor gene or population

Retained as functional HE or maturase

Acquisition of new function
Prebiotic Various types of replicase endonucleases
Primordial Hint

Group I introns
Homing endonucleases
Intein

Convergent parasitic strategies

Target highly-conserved sequences within conserved genes

- Recombination -

Group I intron + HE
C S Intein + HE
HN

Spread by homing

Repeated gain and loss in populations (homing cycle)