The spread of LAGLIDADG homing endonuclease genes in rDNA

Peik Haugen and Debashish Bhattacharya*

Department of Biological Sciences and Center for Comparative Genomics, University of Iowa, 210 Old Biology Building, Iowa City, IA 52242-1324, USA

Received January 29, 2004; Revised and Accepted March 10, 2004

ABSTRACT

Group I introns that encode homing endonuclease genes (HEGs) are highly invasive genetic elements. Their movement into a homologous position in an intron-less allele is termed homing. Although the mechanism of homing is well understood, the evolutionary relationship between HEGs and their intron partners remains unclear. Here we have focused on the largest family of HEGs (encoding the protein motif, LAGLIDADG) to understand how HEGs and introns move in rDNA. Our analysis shows the phylogenetic clustering of HEGs that encode a single copy of the LAGLIDADG motif in neighboring, but often evolutionarily distantly related, group I introns. These endonucleases appear to have inserted into existing introns independent of ribozymes. In contrast, our data support a common evolutionary history for a large family of heterologous introns that encode HEGs with a duplicated LAGLIDADG motif. This finding suggests that intron(double-motif) HEG elements can move into heterologous sites as a unit. Our data also suggest that a subset of the double-motif HEGs in rDNA originated from the duplication and fusion of a single-motif HEG encoded by present-day ribozymes in LSU rDNA.

INTRODUCTION

Group I introns that encode homing endonuclease genes (HEGs) are mobile genetic elements that can rapidly spread in genetic crosses between intron-containing and intron-minus strains [reviewed by Jurica and Stoddard (1)]. Phylogenetic studies have been critical in establishing the evolutionary history of HEGs and their associated introns and in clarifying how they have spread in genomes (2–8). The HEG-mediated mobility of group I introns is known as ‘homing’ because the intron is faithfully inserted into homologous sites in intron-less alleles [reviewed by Belfort and Perlman (9)]. An alternative model to explain group I intron mobility is called reverse splicing, in which the ribozyme recognizes and integrates into a homologous or ectopic RNA site independent of HEG activity (10–12).

The homing endonuclease (HE) proteins that confer intron mobility are divided into four families based on the presence of conserved amino acid motifs; these are the LAGLIDADG, His-Cys box, GIY-YIG and HNH families. Differences in the sequence motifs suggest independent origins of the four families (13). Crystal structures are available for several representatives of the LAGLIDADG family (14,15), and for one member of the His-Cys box family (16). Of the HEGs encoded by group I introns, the LAGLIDADG family has by far the largest number of known representatives [i.e. >200 members (13)]. In each of its members, the conserved LAGLIDADG motif is present in one or two copies (Fig. 1A). The single-motif HEGs like I-Cre recognize the DNA target sequence as a homodimer, whereas the double-motif HEGs function as monomers with one LAGLIDADG motif in each domain. Consistent with this observation, the function of the homodimer requires greater overall symmetry in the DNA target sequence (i.e. palindromic and pseudo-palindromic target sites) than the equivalent target of the relatively more divergent double-motif HEGs, which contain polymorphisms that distinguish the two domains (1). A prediction of HEG movement based on these observations is that single-motif HEGs should be more restricted in their ability to spread to ectopic sites, whereas double-motif HEGs may be more successful at invading more divergent target sites.

In addition to mediating intron spread, HEGs are mobile elements that can move in a ribozyme-independent fashion. For example, in a group I intron in the T4 phage, a flanking sequence very similar to that recognized by the HE was found at the site of insertion of the intron-encoded HEG (17). This observation is consistent with a recent invasion of the HEG into the group I intron. Moreover, the phylogeny of all nuclear His-Cys box HEGs and their associated introns support HEG mobility either into homologous or heterologous introns (8). Typical hallmarks of intron-independent mobility of HEGs include
the ‘switching’ of HEGs between different intron peripheral loops and between sense and antisense strands of intron DNA (8).

Here we set out to establish the evolutionary history of the LAGLIDADG HEGs and their associated introns in rDNA to understand the origin of HEGs, the dynamics of HEG mobility, and how these correlate with the evolutionary history of the associated group I introns. In order to do this, we aligned a total of 82 group I introns with LAGLIDADG HEGs that are inserted into organellar and bacterial rDNA (see Fig. 1B). By reconstructing the HEG and intron phylogenies and comparing them to intron/HEG distribution in rDNA, we were able to deduce a model to explain the origin and spread of these mobile elements.

### MATERIALS AND METHODS

**Retrieval of group I intron and LAGLIDADG HEG sequences**

The Comparative RNA web site (http://www.rna.icmb.utexas.edu/) was used to identify large group I introns in the rDNA of organellar and bacterial genomes. Introns known to encode HEGs with the LAGLIDADG motif were retrieved from the
National Center for Biotechnology Information (NCBI) GenBank (18), along with the HE sequences. In some cases, introns from closely related species were excluded to minimize the size of the dataset. Representatives of HEs from the different rDNA insertion sites were then used as queries in tblastn searches at the NCBI web site (http://www.ncbi.nlm.nih.gov/BLAST/) to identify HEGs or HE pseudogenes that were not annotated on the Comparative RNA web site or in GenBank. The LAGLIDADG HEs and their associated group I introns that were identified in this study are listed in Table S1 (see Supplementary Material).

The alignment and phylogeny of HE protein sequences

Prior to phylogenetic inference, the LAGLIDADG HEs were divided into two separate datasets based on previous work (4). The first (dataset one) comprised single-motif HEs that could be unambiguously aligned with a subset of double-motif HEs (Clade 1 in Fig. 1C) (3,19). This group of proteins was readily aligned by using as guides the I-CreI and I-MsoI crystal structures (14) in combination with the alignment published in Dalgaard et al. (4). The proteins from the LSU rDNA position L2593 (by convention, the numbering reflects the Escherichia coli genic position) were significantly different from others at the C-terminus and were therefore used as the outgroup in the phylogenetic analysis of dataset one. The L1943 HE, although clearly related to the Clade 1 type of sequences (and the other single-motif HEs, see alignment in Supplementary Material), was excluded from the analyses because the inclusion of this divergent sequence significantly lowered the bootstrap support at several nodes in the tree and the position of the HE could not be reliably determined regardless of the outgroup or the phylogenetic method that was used. Dataset two that contained the remaining double-motif HEs (10 sequences could not be aligned to either dataset and were excluded; see Supplementary Material, Table S1) were aligned and represent sequences equivalent to Clade 2 (Fig. 1C). Twenty-nine characters (from a total of 96 aa) from the highly divergent middle region of both the N- and C-terminal halves of the HEs were aligned using CLUSTAL W (20), whereas the more conserved regions were aligned manually (alignment provided as Supplementary Material). Because double-motif HEs have a significantly higher degree of variation (e.g. 2.4× higher average pairwise sequence divergence for Clade 2 double-motif proteins versus all single-motif proteins under the WAG + Γ evolutionary model [calculated with TREE-PUZZLE 5.0 (21)]) than single-motif HEs (4), dataset two was far more challenging and produced a less robust alignment. The N- and C-terminal regions were treated as separate entries in those cases in which HEs encoded two LAGLIDADG motifs. A subset of sequences from dataset one was included in dataset two as outgroup HEs [i.e. Ascocalyx abietina and Sclerotinia sclerotiorum (569), Cordyceps jezoensis (1210), Agrocybe aegerita (1224), Monomastix sp. CL-1997 and Acanthamoeba castellanii (L1951), Chlamydomonas agloformis, Chlamydomonas reinhardtii, Monomastix sp. CL-1997, Acanthamoeba castellanii and Chlorella vulgaris (L2593)].

A minimal evolution (ME) phylogenetic tree was inferred from dataset one. The program TREE-PUZZLE 5.0 was used to calculate the pairwise distances under the WAG evolutionary model (22) and the gamma (Γ) distribution of rate variation across sites (with eight categories). Next, the WAG + Γ distance matrix was used to infer the ME tree by using Fitch in the PHYLIP V3.6a3 (23) program package. Global rearrangements and 10 random additions of sequences were used in the tree search. TREEVIEW 1.6.6 (24) was used to produce the tree image. Support for nodes in the ME tree was calculated with bootstrap analysis and Bayesian inference. Neighbor joining (NJ)-bootstrap values were calculated by first generating 500 bootstrap replicates using Seqboot (PHYLIP). These replicates were used as input for two different distance matrix analyses. In the first analysis, NJ–WAG + Γ trees were made from each bootstrap replicate as described above and the consensus tree was made using Consense (in the PHYLIP package). In the second analysis, the 500 replicates were used as input for NJ–JTT+ Γ tree reconstruction. A maximum parsimony (MP) bootstrap analysis (200 replicates) was done using MEGA 2.1 (25) (gap = use all sites, CNI level = 3, random addition of trees = 10, weighting method = standard parsimony). In addition, an ME bootstrap analysis (2000 replicates, Poisson correction) was done using MEGA 2.1 (pairwise deletion, CNI level = 2, initial tree = NJ, max. number of trees to retain = 1). Finally, we used Bayesian analysis [MrBayes V3.0b4 (26)] with the WAG + Γ model to calculate posterior probabilities for monophyletic groups. Metropolis-coupled Markov chain Monte Carlo from a random starting tree was initiated in each Bayesian inference and run for 2 million generations. Trees were sampled each 100 cycles. Four chains were run simultaneously, of which three were heated and one was cold, with the initial 50 000 cycles (500 trees) being discarded as the ‘burn-in’. Stationarity of the log likelihoods was monitored in each analysis to verify convergence by 2 million cycles (results not shown).

For dataset two, we generated the 50% majority rule consensus tree calculated from the Bayesian analysis to summarize the evolutionary relationship of these divergent HEs. Because the evolutionary distances between the sequences could not be calculated under the WAG + Γ and JTT + Γ evolutionary models, bootstrap analysis of these data was done using the ME method with Poisson corrected distances (2000 replicates, as described for dataset one) and with maximum parsimony (200 replicates, as described for dataset one).

The alignment and phylogeny of group I intron sequences

The group I ribozyme RNA sequences were aligned strictly on the basis of homologous secondary structures (P1–P9). The distantly related introns with LAGLIDADG HEs from mitochondrial cytochrome c oxidase subunit I (cox1) were used as the outgroup in this analysis [Catalpa fargesii (AJ223411), Clerodendrum trichotomum (AJ223414), Cucumis sativus (AJ223416), Digitalis purpurea (AJ223415), Vinca rosea (AJ223423)]. The consensus tree from the Bayesian analysis of the DNA sequences (GTR + I + Γ model) was used to represent the intron phylogeny. Support for groups in this tree was tested with a NJ–Jukes–Cantor bootstrap analysis (2000 replicates) using PAUP* 4.0b8 (27) and with maximum parsimony analysis (100 replicates) using MEGA, as described above.
RESULTS
Phylogeny of LAGLIDADG HEGs in rDNA
Figure 2 shows a minimum evolution tree of dataset one that includes rDNA HEGs from Clade 1 (Fig. 1C) plus others that could be readily aligned with these sequences. This tree shows that single-motif HEGs that are inserted into introns at the same rDNA position form monophyletic groups even though they may originate from distantly related taxa or are found in different organellar genomes and in prokaryotes. Exceptions to this finding are the C.vulgaris and Thermotoga subterranea
introns that are positioned outside the L1951 and L1917 clades, respectively (albeit without bootstrap support). The clustering of endonucleases on the basis of a shared insertion site rather than based on taxonomic relationships is consistent with the origin of many of these HEGs through lateral transfers between species and genomes. For the double-motif HEGs (Clade 1 HEGs in Fig. 2), the N- and C-terminal halves form sister groups of each other, suggesting that these sequences originated by gene duplication (followed by fusion) in a common ancestor.

The second protein dataset (dataset two) contains the Clade 2 rDNA HEGs (see Fig. 1C), which encode the highly divergent double-motif proteins from different insertion sites. A subset of protein sequences that represents the breadth of dataset one (11 HEGs in total, see Materials and Methods) was used as the outgroup in this analysis (for simplicity, denoted as ‘Outgroup HEGs’ in Fig. 3). Figure 3 shows a Bayesian majority rule (50%) consensus tree of the protein sequences. As expected for a small and highly divergent dataset, many of the nodes are weakly supported with respect to both Bayesian posterior probabilities and bootstrap values. Some of the branches that define major lineages do, however, have significant support and show, as in dataset one (see Fig. 2), that lineages are primarily defined by intron insertion site and that the N- and C-terminal halves of the double-motif HEGs form sister clades (i.e. see Clades 1 and 2 in Figs 2 and 3).

Phylogeny of group I introns in rDNA that encode LAGLIDADG HEGs
To understand HE spread, we reconstructed the evolutionary history of the group I introns associated with these endonucleases. Figure 4A shows a Bayesian 50% majority rule consensus tree that was inferred from the intron data. The position of the HEG in the intron RNA structure, the rDNA intron insertion site, and schematic figures representing the different group I intron subclasses are shown in this figure. Consistent with previous studies (8,28–32), introns inserted into the same rDNA site are generally more closely related to each other. More interesting, however, is the strong support for the monophyly of introns from seven different insertion sites (boxed clade in blue field in Fig. 4A). These introns belong to the group IC2 subclass and all contain HEGs inserted in the P9 paired element.

DISCUSSION
Distribution of LAGLIDADG HEGs in rDNA
We assembled a dataset of 82 group I introns with LAGLIDADG HEGs that are inserted into organellar and bacterial rDNA and aligned the HE sequences and the group I ribozyme sequences to infer their phylogenies. The distribution of group I introns that encode HEGs with one or two copies of the LAGLIDADG motif is shown in Figure 1B. The single-motif HEGs are remarkably restricted in their distribution and are clustered in one region of LSU rDNA (i.e. five out of six positions occur between L1917 and L1951). In contrast, the double-motif HEGs form clusters (e.g. S1210, S1224, S1247 and L1931, L1939, L1949) but overall are more broadly distributed in both SSU and LSU rDNA. The non-random distribution of rDNA group I introns has previously been demonstrated through statistical analysis of linear rDNA sequences (33) and by mapping introns on the tertiary structures of the ribosome subunits (34).

Two separate lineages of proteins with two LAGLIDADG motifs in rDNA
A previous study of LAGLIDADG proteins demonstrated that HEGs encoded by group I introns in rDNA are found primarily in two clades that are relatively distantly related to each other (4; see Fig. 1C). We have focused on these two clades and added sequences currently found in GenBank to study HEG evolution and spread. Based on our analyses, we present a putative evolutionary hypothesis for the origin of the Clade 1 HEs (Fig. 4B). This model is based on the phylogenetic analysis shown in Figure 2 that indicates the single-motif HEG cluster (i.e. L1917–L1951) to represent the ancestral pool from which the Clade 1 double-motif HEs may have originated (through gene duplication and fusion). The clustered HEs are most likely accounted for by an initial HEG insertion into a group I intron in one of the sites between L1917–L1951, followed by HEG spread into introns in neighboring sites. Thereafter, one of the single-motif HEs underwent a gene duplication followed by fusion, thus creating an ancestral Clade 1-type double-motif HEG (open arrow in Fig. 4B). The basal position of the L1931 and L1939 double-motif HEs in both the N- and C-terminal halves of Clade 1 identifies them as potentially the earliest divergence after HEG duplication. Interestingly, HEs at intron site L1931 also exist as single-motif endonucleases that are closely related to Clade 1. The L2593 HEs are the most distantly related to the Clade 1 double-motif HEs and are therefore least likely to be candidates for the direct ancestor of these endonucleases.

Given that the origin of the Clade 1 HEs is reasonably clear, then what is the evolutionary source for the Clade 2 proteins? It has previously been suggested that double-motif HEs can be accounted for by one or more gene duplication events involving an I-CreI-like protein HEG (13,35–37). There are, however, numerous group I introns found at six rDNA insertion sites (i.e. L1917–L1951, L2593; Fig. 1B) that encode single-motif HEGs, and any of these could potentially have undergone gene duplication and fusion, thereby creating the double-motif Clade 2 progenitor. Our analyses do not allow us to conclusively identify the ancestor of Clade 2 HEs.

Evidence for the spread of introns with HEGs into heterologous sites
To relate the phylogeny of HEs to the evolutionary history of group I introns, we generated a Bayesian 50% majority-rule...
Figure 2. Phylogeny of Clade 1 HEs inferred with a ME-WAG + \( \Gamma \) analysis. The thick branches denote >95% Bayesian posterior probability (WAG + \( \Gamma \) model) for groups to the right. The bootstrap values above the branches on the left of the slash-marks are inferred from an NJ-JTT + \( \Gamma \) analysis, whereas those on the right are from an NJ-WAG + \( \Gamma \) analysis. The bootstrap values shown under the branches on the left of the slash-marks are inferred from an ME±Poisson analysis, whereas those on the right are from an MP analysis. The branch lengths are proportional to the number of substitutions per site (see scale bar). The insertion sites of the introns containing these HEs are shown, as is the sister-group relationship of the double-motif HEs in Clade 1 (blue field). The open arrow indicates the timing of the putative gene duplication event. HEGs that are inserted into P8 of group IA3 introns (orange), P8 of group IB4 introns (blue) or into P6 of group IB4 introns (reddish purple) are shown in different colors. The asterisks associated with the \( \text{T.papilionaceus} \) L1939 HE indicates that it is inserted into a group I intron that is distantly related to the remaining introns associated with HEs in Clade 1 (see Fig. 4A). Introns are inserted into bacterial (B), mitochondrial (M) and chloroplast (C) rDNA genes.
Figure 3. Phylogeny of Clade 2 HEs represented by a 50% majority rule consensus tree generated using Bayesian inference (WAG + Γ model). The thick branches denote >95% posterior probability for groups to the right. The bootstrap values above the branches are from an MP analysis, whereas those below the branches are from an ME-Poisson analysis. The branch lengths are proportional to the number of substitutions per site (see scale bar). The insertion sites of the introns containing these HEs are shown, as is the sister-group relationship of the double-motif HEs in Clade 2. Introns are inserted into bacterial (B), mitochondrial (M) and chloroplast (C) rDNA genes.

Figure 4. Phylogeny of rDNA group I introns that encode LAGLIDADG HEs, and a putative model for the origin and spread of a subset of double-motif LAGLIDADG HEGs (Clade 1) in rDNA genes. (A) The 50% majority rule consensus tree generated using Bayesian inference (GTR + I + Γ model). The thick branches denote >95% posterior probability for groups to the right. The bootstrap values above the branches are from an MP analysis, whereas those below the branches are from an NJ-Jukes-Cantor distance analysis. The branch lengths are proportional to the number of substitutions per site (see scale bar). The intron insertion sites and the helical position of the group I introns are shown as are schematic representations of the different intron subclasses. The monophyletic group I intron lineage that contains the Clade 1 double-motif HEs is boxed in the blue field. The question mark associated with the Chlamydomonas pallidostigmatica S793 intron indicates that the position of the HEG insertion in the intron RNA structure is unclear. Our secondary structure prediction of the intron indicates that the HEG might be located in the P7.1-P7.2 elements (data not shown). The color codes indicate the three different intron/single-motif HEG constellations inserted between positions L1917 and L1951 (see Fig. 2 legend). Introns are inserted into bacterial (B), mitochondrial (M) and chloroplast (C) rDNA genes. (B) Based on (A) and Fig. 2, the prediction is that there have been intron-independent transfers (at least two) of single-motif HEGs into pre-existing introns located between L1917 and L1951. Group I intron/double-motif HEG combinations thought to have been transferred as a unit into heterologous rDNA sites are boxed in the blue field, whereas the exception is marked with an asterisk (due likely to an intron-independent HEG transfer).
consensus tree of the ribozyme sequences (Fig. 4A) and mapped the HEG insertions on the intron secondary structures. The single-motif HEGs at positions L1917–L1951 are inserted in the paired RNA-elements P6 of group IB4 introns (L1917, L1923, L1943), P8 of group IB4 introns (L1931) or P8 of group IA3 introns (L1951). The three different intron/HEG constellations are shown in matching colors in Figures 2 and 4. And because the intron tree does not support monophyly of these introns (they belong to the distantly related intron subclasses IB and IA), we infer that the HEG distribution at
neighboring sites (L1917–L1951) can most easily be explained by the independent movement of an ancestral HEG into pre-existing rDNA introns. In other words, the group I introns at sites L1917–L1951 are physically clustered on the rDNA gene but are phylogenetically distantly related to each other, whereas the HEGs these introns encode are not only in close proximity in the gene but are also evolutionarily closely related. This strongly suggests the movement of at least a fraction of these HEGs into introns at neighboring rDNA sites.

Another interesting case of possible HEG mobility is the Clade 1 double-motif HEGs (Fig. 2). The introns in which these related HEGs are inserted are remarkably similar (except the L1939 intron in *Trimorphomyces papilionaceus*) and all have the HEG inserted in P9. The intron tree reflects the close relationship that is highly supported by both bootstrap and Bayesian analyses. We interpret the finding of closely related introns from seven different rDNA sites with related HEGs that are all inserted into P9 as potentially supporting a common evolutionary history of the intron and HEG sequences that have spread as a unit into neighboring and distant rDNA sites (S569, S1210, S1224, S1247 and L1931, L1939, L1949; see Figs 2 and 4). The lack of resolution in the intron tree, however, does not allow us to directly assess the congruence of these phylogenies (predicted under a model of strict co-inheritance). The L1939 HEG in a group IB intron is, however, most likely explained as a case of intron-independent HEG mobility.

**Concluding remarks**

Our data suggest that the most likely mechanism to explain the observed clustering of single-motif HEGs (L1917–L1951) is, at least in part, intron-independent mobility of HEGs into pre-existing ribozymes. We also found a potential case for the common evolutionary history of double-motif HEGs (Clade 1 in Fig. 2) and the introns in which they reside (boxed introns in the blue field in Fig. 4A), even though they are inserted at both neighboring and distant rDNA sites. This suggests that introns and HEGs can move into heterologous sites as a unit. Furthermore, our data suggest that the most likely ancestral source for a subset of double-motif HEs in rDNA (Clade 1, Figs 1C and 2) is a duplication and fusion event of a single-motif HEG encoded by an intron presently found between the LSU rDNA positions L1917–L1951. The progenitor of another subset of double-motif HEGs (Clade 2, Figs 1C and 3) remains unclear. A putative evolutionary hypothesis for the origin of most single-motif HEGs and a subset of the double-motif HEGs (Clade 1) is presented in Figure 4B. This model posits one gene duplication and fusion event and the spread of HEGs into both neighboring and distant rDNA sites.

Our study provides important insights into the mobility of single- and double-motif LAGLIDADG homing endonucleases in rDNA. We show that, as predicted, single-motif HEs spread less successfully into distant rDNA sites than their double-motif HE counterparts (Figs 1B and 4B). One possible explanation for the finding that single-motif HEGs spread primarily into neighboring sites is that it provides the HEG with an opportunity to escape to new introns that interrupt the same conserved rDNA target region [i.e. these HEs function as homodimers and need a partial symmetrical recognition sequence (19)] found in different populations or species. Importantly, movement into introns at neighboring sites affords the host protection from HE activity because the intron is likely to interrupt the HE recognition sequence (15–45 nt). It is conceivable that the HE adapts, over time, to the new neighboring target site where the intron is inserted to facilitate more efficient homing. In contrast, the monomeric double-motif HEs have a lower requirement for strictly palindromic sequences and are in fact more broadly distributed in rDNA (see Fig. 1B). In support of this idea, Figure 4B shows that the Clade 1 double-motif HEGs have invaded neighboring rDNA sites (S1210, S1224, S1247 and L1931, L1939, L1949), distant rDNA sites (S569) as well as protein genes (nad1, nad3, nad4L, cox1, atp6, see Fig. 1C). In conclusion, our findings illustrate the usefulness of comparative methods to clarify LAGLIDADG HEG evolution. We are aware, however, that these endonuclease sequences pose a significant challenge to evolutionary methods because of their high divergence rates, mobility and sporadic distribution. These issues are best addressed with a comprehensive data set, which is rapidly accumulating and which will allow us to test the hypotheses posited here.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at NAR Online.

**ACKNOWLEDGEMENTS**

This work was supported by grants DEB 01-0774 and MCB 01-10252 awarded to D.B. from the National Science Foundation and a grant from The Norwegian Research Council to P.H.

**REFERENCES**


