BMC Evolutionary Biology

Short-term sequence evolution and vertical inheritance of the *Naegleria* twin-ribozyme group I intron

Odd-Gunnar Wikmark[1], Christer Einvik[1,2], Johan F De Jonckheere[3] and Steinar D Johansen*[1,4]

Address: [1]Department of Molecular Biotechnology, RNA Research Group, Institute of Medical Biology, University of Tromsø, N-9037 Tromsø, Norway, [2]Department of Pediatrics, University Hospital of North Norway, N-9038 Tromsø, Norway, [3]Protozoology Laboratory, Scientific Institute Public Health, B1050 Brussels, Belgium and [4]Department of Fisheries and Natural Sciences, Bodø University College, N-8049 Bodø, Norway

Email: Odd-Gunnar Wikmark - Odd-Gunnar.Wikmark@fagmed.uit.no; Christer Einvik - Christer.Einvik@fagmed.uit.no; Johan F De Jonckheere - jdjonckh@ben.vub.ac.be; Steinar D Johansen* - steinar.johansen@fagmed.uit.no

* Corresponding author

Abstract

Background: Ribosomal DNA of several species of the free-living *Naegleria* amoeba harbors an optional group I intron within the nuclear small subunit ribosomal RNA gene. The intron (Nae.S516) has a complex organization of two ribozyme domains (NaGIR1 and NaGIR2) and a homing endonuclease gene (NaHEG). NaGIR2 is responsible for intron excision, exon ligation, and full-length intron RNA circularization, reactions typical for nuclear group I intron ribozymes. NaGIR1, however, is essential for NaHEG expression by generating the 5' end of the homing endonuclease messenger RNA. Interestingly, this unusual class of ribozyme adds a lariat-cap at the mRNA.

Results: To elucidate the evolutionary history of the Nae.S516 twin-ribozyme introns we have analyzed 13 natural variants present in distinct *Naegleria* isolates. Structural variabilities were noted within both the ribozyme domains and provide strong comparative support to the intron secondary structure. One of the introns, present in *N. martinezi* NG872, contains hallmarks of a degenerated NaHEG. Phylogenetic analyses performed on separate data sets representing NaGIR1, NaGIR2, NaHEG, and ITS1-5.8S-ITS2 ribosomal DNA are consistent with an overall vertical inheritance pattern of the intron within the *Naegleria* genus.

Conclusion: The Nae.S516 twin-ribozyme intron was gained early in the *Naegleria* evolution with subsequent vertical inheritance. The intron was lost in the majority of isolates (70%), leaving a widespread but scattered distribution pattern. Why the apparent asexual *Naegleria* amoebae harbors active intron homing endonucleases, dependent on sexual reproduction for its function, remains a puzzle.

Background

Naegleria is a common genus of soil and freshwater free-living amoeba of the vahlkampfiid family [1]. *Naegleria* apparently lack a sexual reproduction cycle since meiosis never has been observed or proven experimentally. Subsequently, a number of genetically-defined variants have been isolated in nature and proposed as distinct species [1,2]. A typical *Naegleria* amoeba cell contains a distinct

nucleus with a large and predominant nucleolus containing as much as 3000–5000 copies of an approximately 14-kb ribosomal DNA (rDNA) plasmid [3,4]. Each rDNA molecule carries a single transcription unit for the ribosomal RNA (rRNA) genes. Some *Naegleria* isolates have been reported to contain group I intron insertions at conserved sequence sites, both within the small subunit (SSU) and large subunit (LSU) rRNA genes [5]. Introns have been noted at position 516 in SSU rDNA (i.e. a position that is homologous to corresponding position in the *E. coli* rRNA gene) and at positions 1921, 1926, 1949, and 2563 in LSU rDNA [5-11].

Group I introns are autocatalytic genetic elements carrying a ribozyme domain responsible for the intron self-splicing reaction, and occasionally a homing endonuclease gene (HEG) encoding an endonuclease protein directly involved in intron mobility at the DNA level [12,13]. A group I splicing ribozyme possess a well-defined three-dimensional structure organized into three functional domains (catalytic domain, folding domain, and substrate domain) by approximately ten paired RNA segments named P1–P10 [14,15]. The most common and best characterized of the *Naegleria* rDNA introns is Nae.S516. Group I introns at position 516 in SSU rDNA are relatively common among eukaryotic microorganisms with more than 250 cases reported so far [16,17], and with both lateral and vertical inheritance patterns compared to that of host rDNA. A widespread distribution and structural diversity among the S516 group I introns have been noted, including several complex introns carrying HEGs [16,18].

A typical *Naegleria* S516 intron has a twin-ribozyme organization and represents the most complex class of all group I introns known [19]. Nae.S516 consists of a small group I-like mRNA capping ribozyme (NaGIR1) and a HEG domain, both inserted into the P6b segment of a regular group IC1 splicing ribozyme (NaGIR2). Expression and functional aspects of the *Naegleria* S516 intron have been reported. The splicing ribozyme (NaGIR2) is responsible for the autocatalytic activity that generates intron excision and exon ligation, as well as full-length intron RNA circle formations [7]. The ability to form full-length intron RNA circles is a general property of nuclear group I introns and could be important in RNA mobility at the RNA level, or even as an intermediate in the expression of the homing endonuclease [11,20,21]. The *Naegleria* HEG (NaHEG) encodes a 245 amino acid protein that belongs to the His-Cys box family of homing endonucleases [22,23]. The intron endonuclease recognizes and binds to a 19-bp DNA sequence flanking the S516 rDNA site and cleaves the DNA generating a five-nucleotide 3' staggered end [24,25]. In general, group I intron endonucleases promote intron homing at the DNA-level by generating a

double-stranded break in the intron-less target DNA, followed by invasion of the donor intron-containing allele and DNA repair using the intron-containing allele as template [21]. However, sexual mating is the biological framework for nuclear group I intron homing [26,27] and it is unclear why the apparently asexual *Naegleria* contains and maintains homing introns.

The expression of the NaHEG is dependent on a functional NaGIR1 ribozyme, which defines the 5' end of the homing endonuclease mRNA by internal processing and modification of the excised Nae.S516 intron [7,28,29]. Primer extension analyses of both cellular RNA from *Naegleria* and *in vitro* transcribed intron RNA [7,28] are consistent with the formation of a tiny lariat cap structure between nucleotide 1 and 3 of the messenger, as recently reported in the related DiGIR1 ribozyme [30]. Thus, the NaGIR1 capping ribozyme represents a novel class of ribozymes possessing a new catalytic function, which is reflected in its unique RNA architecture [29,31].

The complex and unique structural organization of the *Naegleria* twin-ribozyme intron makes it interesting to investigate the evolutionary origin of the different intron domains, as well as the inheritance pattern within the *Naegleria* genus. Here we report several new intron variants and have performed sequence and phylogenetic analyses providing new insight into fundamental questions such as intron structure, intron-host biology, and the origin and evolution of intron HEG and ribozyme domains. The *Naegleria* twin-ribozyme intron serves as an attractive model system in the characterization of evolutionary processes behind a recently gained, but vertically inherited, selfish genetic element.

Results and discussion
Widespread but sporadic distribution of Nae.S516 introns within the *Naegleria* genus
Sequence analysis of ITS-rDNA from 70 natural isolates of *Naegleria* (Table 1) was performed to gain insight into the genetic relationships among strains and species. A phylogenetic tree based on the NJ method is presented in Figure 1, and corroborates previous finding of six main clusters of *Naegleria* isolates [1,2]. Most clusters appear monophyletic with high bootstrap and Bayesian supports, and Cluster 5 is the most prominent with 30 annotated *Naegleria* isolates (Figure 1). A closer inspection of the SSU rRNA identified intron sequences inserted at position S516 in 21 of 70 strains analyzed. All introns, except one [8], belong to the highly complex twin-ribozyme group I intron family [7,19]. Group I introns at position 516 in SSU rDNA are relatively common among eukaryotic microorganisms [19]. Interestingly, the *Naegleria* 516 introns (Nae.S516; for intron nomenclature see [32]), show a widespread but scattered distribution among the
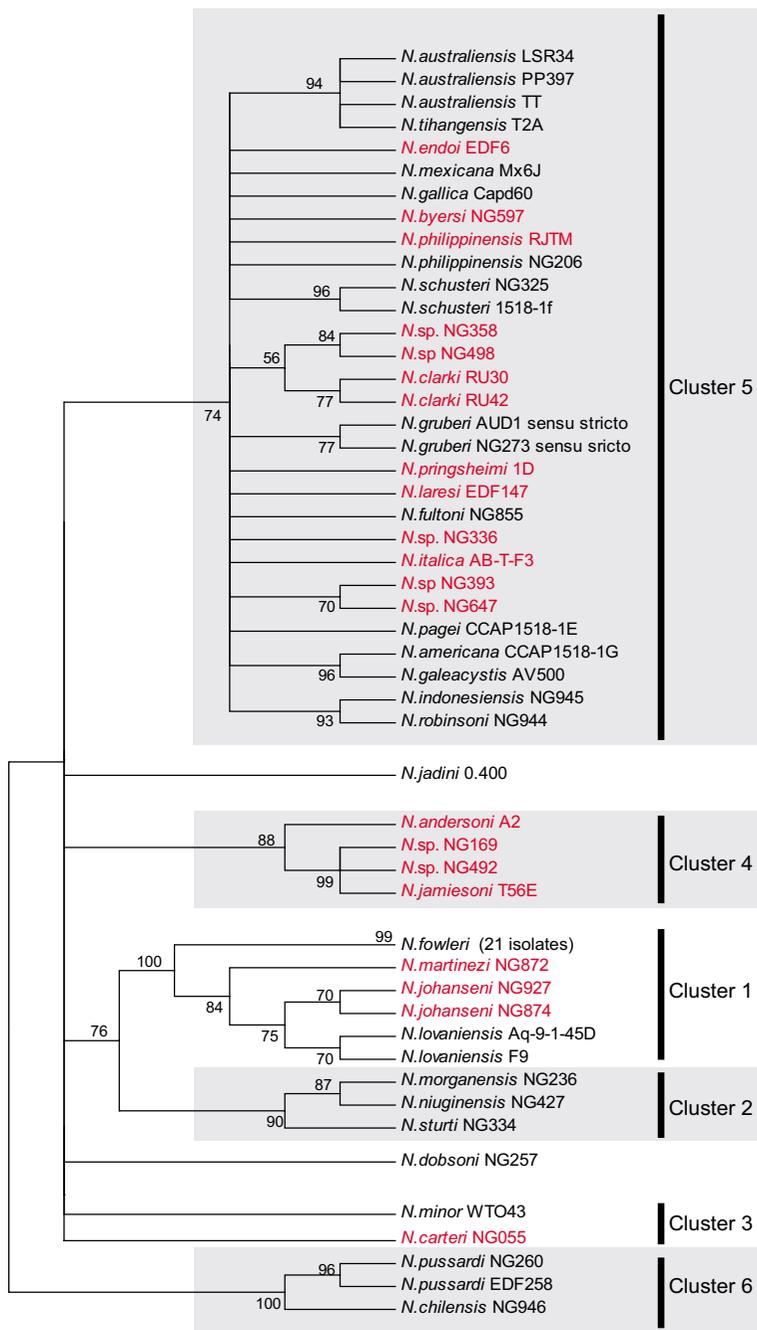
Figure 1
Distribution of the Nae.S516 introns among *Naegleria* isolates. NJ phylogenetic tree of the ITS-rDNA based on 287 nucleotide positions and 70 *Naegleria* isolates. NJ bootstrap values above 50% are shown at the branches. The six clusters of *Naegleria* isolates (Cluster 1–6) are indicated according to [1,2]. *Naegleria* isolates known to contain (21 isolates) or lack (49 isolates) the Nae.S516 intron are shown in red or black, respectively. All introns are approximately 1,3 kp in size (corresponding to a twin-ribozyme organization), except *N. byersi* NG597 which is 375 bp (only NaGIR2). The *N. fowleri* branch in Cluster 1 represents 21 distinct isolates.
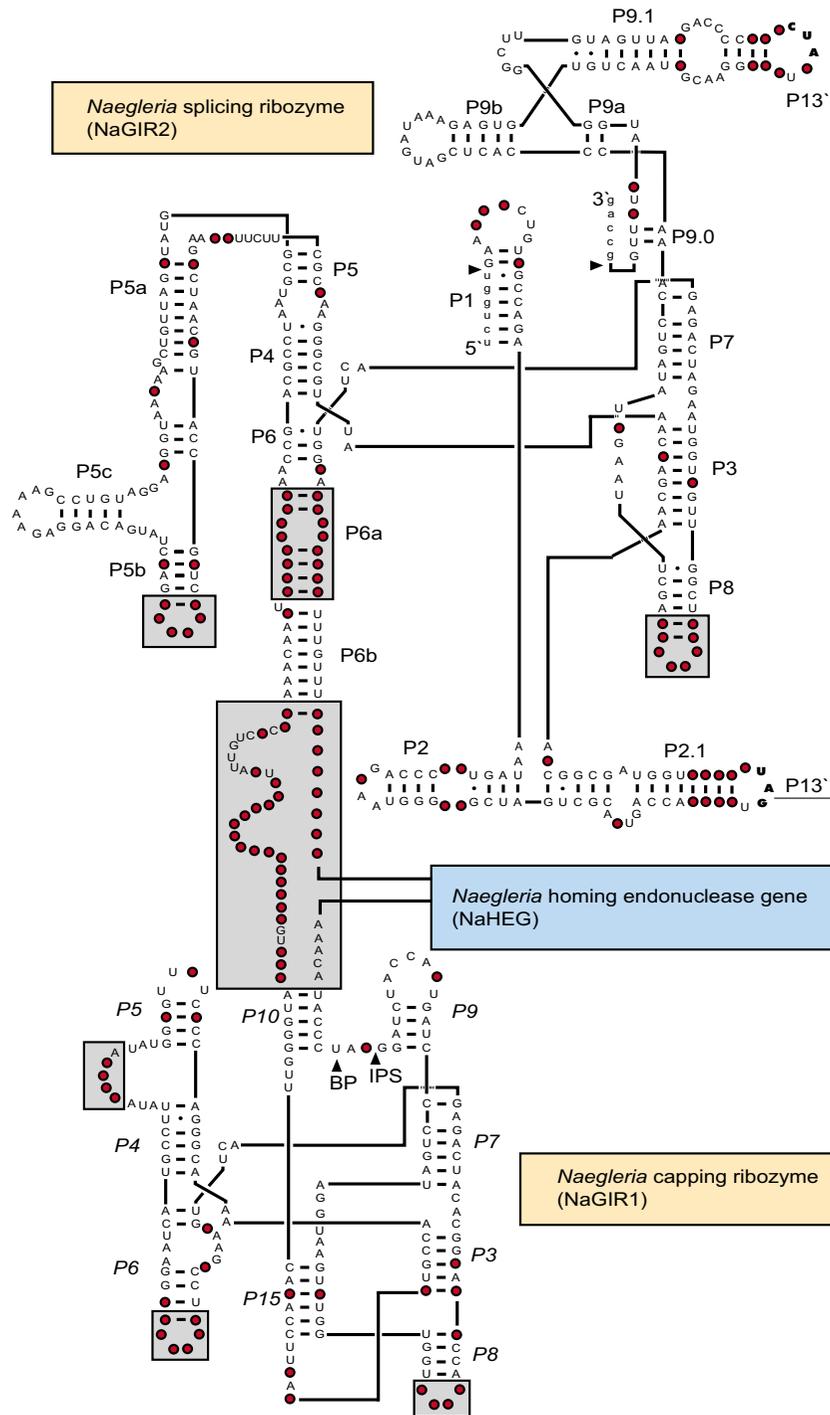
Figure 2
Consensus structure diagram of the Nae.S516 twin-ribozyme intron. The structure diagram is based on the 13 completely sequenced Nae.S516 introns from distinct natural isolates (see Table 1) and is folded according to previously reported models [7,19] with some modifications. Invariant nucleotide residues are presented as uppercase letters. Red filled circles represents a variable position in one or more intron sequences, and regions with size and structural variations are boxed. Nae.S516 contains the two distinct ribozymes NaGIR1 and NaGIR2, and the homing endonuclease gene NaHEG. IPS, internal processing site; BP, branch point nucleotide (U).

```
M V S I K Q K V T R L R N K L R S R S T R N G P K T I T V K I D R K S F K N G Y D P L V D T I D Y G Y S K M A   Nja
. . . . . . . . . . . . . . . . . . A . . . L S . H . . . . . . . . . . . . . . . . . . . . . . . N . . . . . . . . . .   Nan
. . . . T K M . R S . . K R . . V . . . T K R . . . . . . . . . . L E R . . . . . . . . . . K . . . . . . . . . .   Nph
. . T . . . M . R T . . . V . A S . . T K R . . . . . . . . . . L E . . . . . . . . . . K N . . . . . . . . . .   Npr
. . T . R K M . R . . . . K R . . I . . . T K R . . . . . . . . . . E . . . . . . . . . . K S . . . . . . . . . .   Nit
. . T . R K I . K . . . . . . . A . . . T K R . I . . . . . . . . E . . . . . . . . . . K . . . . . . . N . . .   RU30
. . T . R K I . K . . . . . . . A . . . T K R . I . . . . . . . . E . . . . . . . . . . K . . . . . . . N . . .   RU42
. . T . R K I . K . . . . . . . A . . . T K R . R . . . . . . . . E . . . . . . . . . . N S . . . . . . N . . .   NG358
. . T . . K M . R . . . . . . . A . . T K H . . I . . . . . . . . . . . . . . . . . . . N . . . . . . . . . . .   NG393
. . T . R K I . K . . . . . . . A . . . T K R . R . . . . . . . . E . . . . . . . . . . N S . . . . . . N . . .   NG498
. . T . . K M . R . . . . . . . A . . T K H . . I . . . . . . . . . . . . . . . . . . . N . . . . . . . . . . .   NG647
. . . . . . E . . K K . . K R . L T G . . S R Y . . . . . . . . . . E . . . D . . . . . N I . . . . . . . . . .   Nca
. . K V . - - I N . I S S . D K T S T . T . S S . . . . . . . . . . E E . L D . . . . . I K . . . . . . . . . .   Nma

K I T V N K N D Q L A K L K N C K Q A V N I F N E W L S N R - - - K G D K G R S G K Q K P Y C F D E L K K L D   Nja
. . . . . . . . . . . . . . . . K . . . . . . . . . . . - - - . . . . . . . . . . . . . . . . . . . . . . . .   Nan
. . . . A . C . R . S . . G . R . . K . . . . . . . . . L . . - - . . . . . . . . . . . . . . . . . . . . . .   Nph
. . . . P G . . R . S . . G . R . T . R . V . . D . . . T . - - - . . E . . . . . . . . F . . . . . . E . . .   Npr
. . . . A . C . R . S . . N . R . . K . . . . . . . . . H - - - . . . . . . . . . . . . . . . . . . . . . . .   Nit
R . . L Q . S . R . S . . D . R . . K . . . . . . . . C - - - . . . . . . . . . F V . . . . . . . . . . . . .   RU30
R . . L Q . S . R . S . . D . R . . K . . . . . . . . C - - - . . . . . . . . . F V . . . . . . . . . . . . .   RU42
R . . L Q . S . R . S . . D . R . . K . . . . . . . . . - - - . . . . . . . . . F V . . . . . . . . . . . . .   NG358
. . . . T Q S . R . S . . S . R . M . R . . . . . . . . - - - . . . . . . . . . F . . . . . . . . . . . . . .   NG393
R . . L Q . S . R . S . . D . R . . K . . . . . . . . . - - - . . . . . . . . . F V . . . . . . . . . . . . .   NG498
. . . . T Q S . R . S . . S . R . M . R . . . . . . . . - - - . . . . . . . . . F . . . . . . . . . . . . . .   NG647
. . . . A . K . R . S . . D S R . . K . . . . . . . T . - - - . E . . . . . . . . . . . . . . . . . . . . . .   Nca
. . . . Q K * D Q L * K I N * * * T A K N V . . S . . N S K G K A S . T . . Q . . * Q L P Y I H E E . . . F .   Nma

V C K H E F G E C L I G A A S K T K S G F K V R F M N N K G S D S Y V H H V S V F A K S T C E N C I H S R K M   Nja
. . . . . . . . . . . . . . . . . N . . . . . . . . . . . D . . . . . . . . . . . . . N . . . . . . . . . . .   Nan
. . E . . . . . . . . . . . . . . N . . . . . . . . . . . L . . . . . . . . . . A . . N . . . N . . . . T V E .   Nph
. . . . . . . . . . . . . . . . . T . . . . . . . . . . . . K . . . . . . . . . A . . N . . N . D . . . . V E .   Npr
. . E . . . . . . . . . . . . . . N . . . . . . . . . . . L . . . . . . . . . . A . . N . . N . . . . . V E .   Nit
I . N . . . . . . . . . . . N . . A . . . . . L . K Q . . . . . A . . N . . A . . . . . . . . . V E .   RU30
I . N . . . . . . . . . . . N . . A . . . . . L . K Q . . . . . A . . N . . T . . . . . . I E .   RU42
I . N . . . . . . . . . . . N . . A . . . . . L . K Q . . . . . A . . N . . T . . . . . . I E .   NG358
I . E . . . . . . . . . . . N . . . . . . . . L . K Q . . . . . A . . N . . T . . . . . . V E .   NG393
I . N . . . . . . . . . . . N . . A . . . . . L . K Q . . . . . A . . N . . T . . . . . . I E .   NG498
I . E . . . . . . . . . . . N . . . . . . . . L . K Q . . . . . A . . N . . T . . . . . . V E .   NG647
I . . . . . . . . . . . . . N . . . . . . . . . . . . . . . . . . . . A . . N . . N V . . . . L E .   Nca
I . . . . . . . . . . . . . N . . . . . . . . K . . . . . . . L . N . K A . E S . . . L E .   Nma

L E T V S S S K K D P D A R T I S H L C G N G G C A R P G H L R I E K K S V N D E R T H C H F L L R R S Q S V   Nja
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . . . . . . . . . . . . . . . . L   Nan
. . . . . . . T . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . R   Nph
. . . . . . . S E . . . . . . . . . . . . S . . . . . . . . . . . . . . . . . . . . . . . . . . . L   Npr
. . . . . . . . E . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . R   Nit
. . . . . . . N E . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . I   RU30
. . . . . . . N E . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . I   RU42
. . . . . . . S E . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . I   NG358
. . . . R . . E N . . . . . . . . . . . . . . . . . . . . . . . . . . N . . . . . . . . . . . . . R   NG393
. . . . . . . S E . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . I   NG498
. . . . R . . E N . . . . . . . . . . . . . . . . . . . . . . . . . . N . . . . . . . . . . . . . R   NG647
. . . . . . . R . D . . . . . . . . . . . . . . . . . . . . . . . . . . C . . . . . . . . . . . . L   Nca
I . . . . A R . T E D . . . . . . . . . C . . . . . . . . . . S . K T . . . . . . . . V . . N . K . Q   Nma

A Q S E M I R L A C P H T P R C F V N L Y K I N K P Y Y   N.jamiesoni T56E
. . . . . . . . . . . . . . . . . . I . . . . . . .   N.andersoni PPMFB-6
N . . . . . . . . . . . . . . S . K . . . M . . . . T . . .   N.philippinensis RJTM
H . . . . . . . . . . . . . . . K . . . . I . . . . . . .   N.pringsheimi 1D
N . . . . . . . . . . . . . . S . K . . . M . . . . T . . .   N.italica AB-TF-3
N . . . . . . . . . . . . . . . K . . . . I . . . . T . . .   N.clarki RU30
N . . . . . . . . . . . . . . . K . . . . I . . . . T . . .   N.clarki RU42
N . . . . . . . . . . . . . . . K . . . . I . . . . T . . .   N.sp. NG358
H . . . . . . . . . . . . . . . S . . . . I . . . . T . . .   N.sp. NG393
N . . . . . . . . . . . . . . . K . . . . I . . . . T . . .   N.sp. NG498
H . . . . . . . . . . . . . . . S . . . . I . . . . T . . .   N.sp. NG647
H . . . . M . . . . . . . . . . . . . . I . . . . . . .   N.carteri NG055
E . . . . . M . . . . . . E . K . . . . I . . . . T . . .   N.martinezi NG872
```

━━━ **His-Cys Box**

🟩 **Zinc binding His & Cys**

🟧 **Homologous to I-*Ppo* I Asn119 Mg²⁺-Coord. site**

Figure 3

Alignment of *Naegleria* homing endonuclease sequences. Identical residues compared to the *N. jamiesoni* T56E endonuclease sequence are shown by dots and deletions by dashes. Functional important residues involved in zinc binding and catalysis are indicated. Divergent regions within the *N. martinezi* NG872 sequence due to reading-frame shifts (*) are boxed. For structural comparisons to the I-*Ppo*I homing endonuclease, see [23].
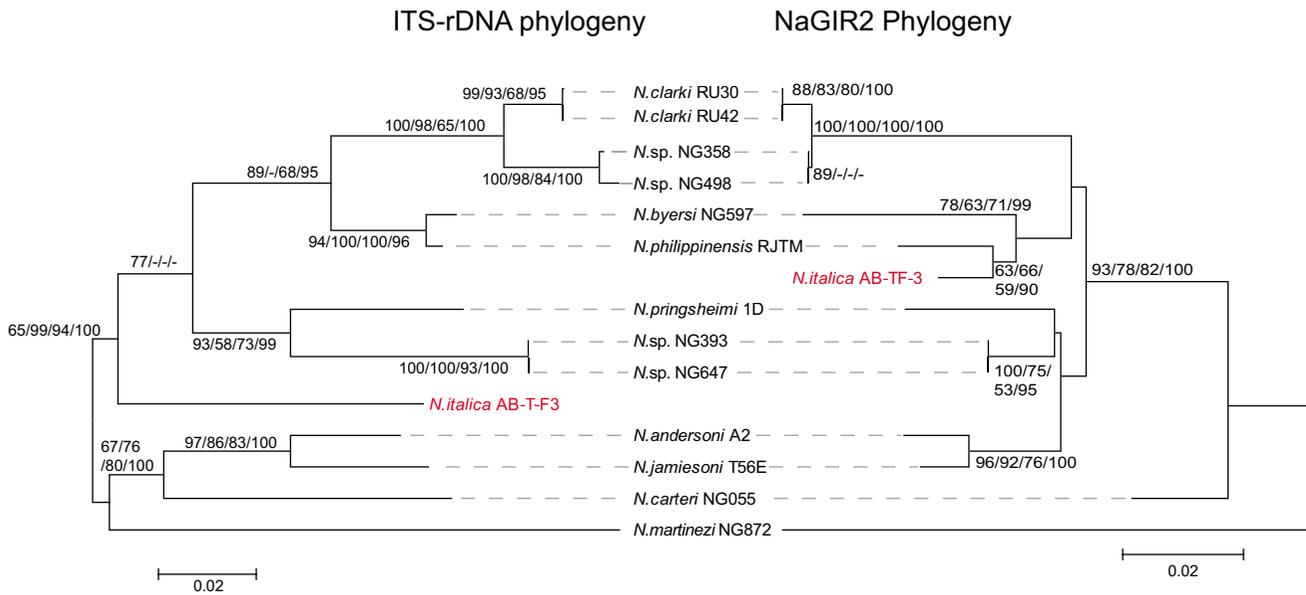
Figure 4
Phylogeny of NaGIR2 and ITS-rDNA. The NaGIR2 and ITS-rDNA NJ trees were generated from 356 and 415 nt datasets. Bootstrap values from 2000 replicates and Bayesian posterior probability values, all over 50%, are shown at branches. The values are from, in order, NJ, MP, ML, and BAY analyses. See Materials and Methods for more detailed analytical parameters. ITS-rDNA and NaGIR2 phylogenies are congruent, except for *N. italica* AB-T-F3 (marked in red). See Table 1 for information about sequence accession numbers.

Naegleria isolates (Figure 1). All clusters, except the early diverging Cluster 6 [2], harbor isolates that carry the group I intron. Furthermore, no linkages could be noted between the presence of the S516 intron and environmental factors such as optimal growth temperature, pathogenesis, habitats, and geographical origin of *Naegleria* isolates.

Structural features and sequence variability of intron domains

Fourteen of the 21 introns were selected for more detail structural characterizations (see Table 1). All introns, but one (*N. byersi* NG597), possess the twin-ribozyme group I intron organization previously reported [7,16], and an updated RNA consensus structure diagram is shown in Figure 2. The intron consists of three functional domains, identified as distinct intron structures. The autocatalytic *Naegleria* splicing ribozyme (NaGIR2) is responsible for intron excision and exon ligation during precursor rRNA processing in the *Naegleria* nucleolus, as well as for the generation of circular intron RNAs [7,28]. NaGIR2 represents a typical group IC1 intron with clear structural resemblance to the well-studied *Tetrahymena* ribozyme [14,15]. The consensus structure (Figure 2) is strongly supported by compensatory base substitutions among the *Naegleria* introns. The most prominent differences between the *Naegleria* and *Tetrahymena* ribozymes are the

lack of a P9.2 segment in *Naegleria*, the presence of an optional tetra-loop in L5b, and a large sequence insertion (approximately 950 nt) in P6b harboring the homing endonuclease gene (NaHEG) and the capping ribozyme NaGIR1 (Figure 2).

NaGIR1 is a group I-like ribozyme with an evolved biological role in intron NaHEG expression [29,30], likely to generate a lariat cap-structure at the 5' end of the *Naegleria* homing endonuclease messenger [30]. The three-dimensional architecture of NaGIR1 is related to that of bacterial tRNA group I intron ribozymes [18,29,33], but with a unique catalytic core organization that contains the novel pseudoknot segment P15 [7,29,31,34]. As seen from Figure 2, most of the core nucleotides are highly conserved among the various *Naegleria* capping ribozymes. Variable regions are almost exclusively located in the terminal loops of P6 and P8, the internal loop junction J5/4, and sequences flanking NaGIR1 and NaHEG.

The third intron domain consists of the NaHEG which codes for a 245 amino acid (aa) His-Cys endonuclease that recognizes and cleaves the intron lacking allele sequence in rDNA [24,25]. An alignment of the derived amino acid sequence from the studied *Naegleria* introns is presented in Figure 3. The amino acid identities between pairs vary from 81 % to 100 %, with the positive charged

NaGIR1 phylogeny　　　　　　　　　　　　　　　　NaHEG phylogeny
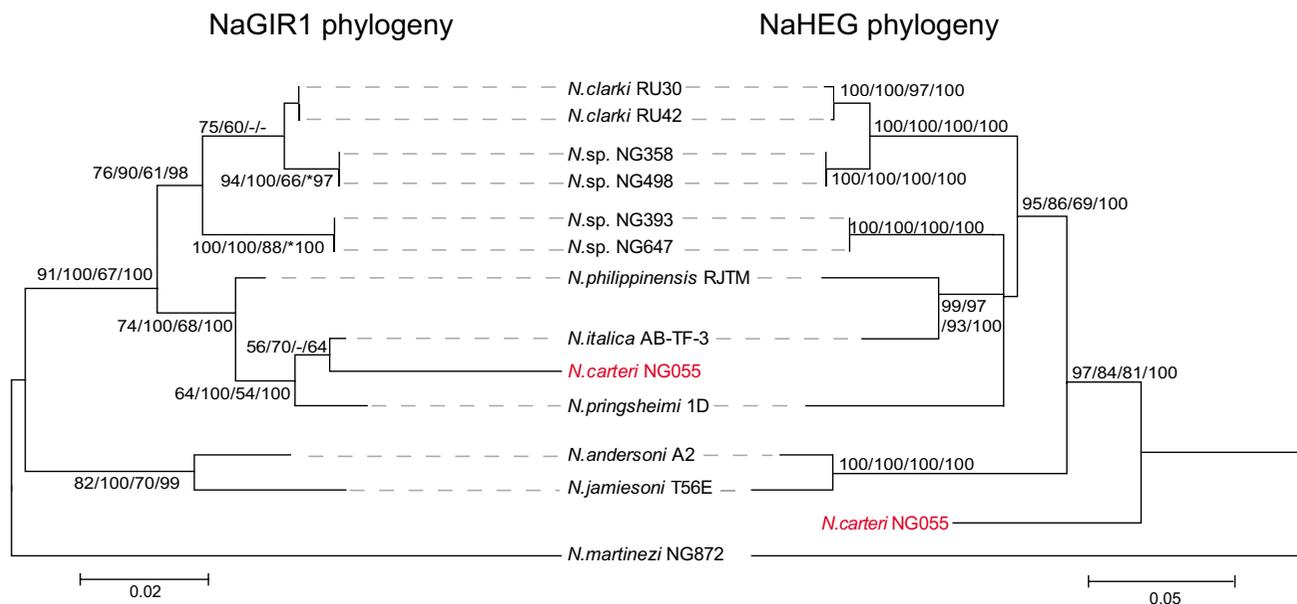
Figure 5
Phylogeny of NaGIR1 and NaHEG. The NaGIR1 and NaHEG NJ trees were generated from 230 and 747 nt datasets. Bootstrap values from 2000 replicates and Bayesian posterior probability values, all over 50%, are shown at branches. The values are from NJ, MP, ML, and BAY analyses, respectively. See Materials and Methods for more detailed analytical parameters. NaGIR1 and NaHEG phylogenies are congruent, except for *N. carteri* NG055 (marked in red). See Table 1 for information about sequence accession numbers.

N-terminal region (first approximately 80 aa) as the most variable part. The N-terminal region contain arginine and lysine rich sequences that resembles known RNA binding domains [35,36], indicating that the NaHEG encodes a complex protein with both endonuclease and RNA binding functions. All three intron domains at the RNA-level (NaGIR1, NaGIR2, endonuclease mRNA) are possible targets for intron protein RNA binding. Both NaGIR1 and NaGIR2 are known to fold correctly and to be catalytically active *in vitro* without assistance of proteins [7,31,34], suggesting no essential maturase function of the intron encoded protein. However, the intron protein could still be able to bind to it own messenger, which is predicted to be highly structured [7]. This possibility remains to be further experimentally explored. Residues previously noted to be essential for endonuclease active site definition, catalysis, and zinc coordination (Figure 3) [25] are highly conserved. The non-synonymous to synonymous substitution rates (dN/dS) were calculated for the NaHEG sequences and found in all cases to be below one (data not shown), indicating purifying selection. Interestingly, the intron present in *N. martinezi* NG872 isolate contains hallmarks of a degenerated NaHEG, seen as multiple frame shifts and small indels (Figure 3).

The splicing ribozyme of Nae.S516 intron is vertically inherited in **Naegleria**
NaGIR2 represents the splicing-ribozyme domain of Nae.S516 and thus corresponds to the sequences present in a prototype group I intron such as the *Tetrahymena* intron. We have previously performed phylogenetic analysis of various nuclear S516 group I introns, including five NaGIR2 variants, and recognized the *Naegleria* introns as a monophyletic clade among the group IC1 introns [16]. Here we extend the analysis to include 14 NaGIR2 variants representing the different isolates of *Naegleria*. Furthermore, we also include a host rDNA analysis based on their corresponding ITS-rDNA sequences. The intron phylogeny was based on 356 sequence positions within NaGIR2, strictly aligned according to the structure diagram in Figure 2. Different methods (NJ, MP, ML, and BAY) were used to build the phylogenetic trees, and all trees were essentially identically in topology. Similarly, the ITS-rDNA phylogeny was based on 415 sequence positions using the same set of tree building methods described above in intron analysis. Figure 4 shows representative NJ trees of both the ITS-rDNA and NaGIR2 phylogenies with overall congruent branching patterns and significant bootstrap and Bayesian supports. The only exception is *N. italica* AB-TF-3 (Figure 4), that could represent a recent horizontal intron transfer. However, the *N.*

italica ITS-rDNA branch topology is only poorly supported in NJ analysis (77 %) and without bootstrap and Bayesian supports in the MP, ML, and BAY analyses, respectively. Thus, we infer there are no experimental support of horizontal intron transfer and conclude that the NaGIR2 domain, representing the Nae.S516 intron, is vertically inherited within the *Naegleria* genus.

The capping ribozyme NaGIR1 and its downstream NaHEG are evolutionary linked

Divergent evolutionary histories of group I splicing ribozymes and their HEGs have been described in several nuclear group I introns [37-39]. The *Naegleria* twin-ribozyme intron is highly unusual among group I introns in that it contains two distinct structural domains (NaGIR1 and NaHEG) inserted into the same peripheral region (P6) of NaGIR2 (see Figure 2). To address the relationships among the variants of NaGIR1 and NaHEG, as well as between the different intron domains (NaGIR1, NaGIR2, and NaHEG) we performed phylogenetic analyses based on the sequence alignments. Figure 5 presents representative NJ trees of NaGIR1 and NaHEG. The phylogenetic trees, built by the MP, ML, and BAY methods, were essentially identical. The NaHEG tree was based on 747 aligned positions and possesses significant bootstrap and Bayesian supports. The total size of the NaGIR1 domain is less than 250 nucleotides in size and thus only 230 positions could be included in the analysis. However, we found the tree topology to be well supported in bootstrap and posterior probability analyses. Interestingly, the NaGIR1, NaHEG, and NaGIR2 (compare Figures 4 and 5) phylogenies were found to have congruent branching topologies consistent with a co-evolutionary pattern of the domains within introns. But there is one clear exception in *N. carteri* NG055 (see Figure 5), suggesting a homologous recombination-like event between natural sequence intron variants.

A *Naegleria* S516 group I intron with only NaHEG or only NaGIR1 insertions has never been observed, suggesting a strong linkage between the domains. Both structural and functional data give further support to a close linkage between the NaGIR1 and NaHEG domains. Jabri and Cech [34] showed that the RNA structure essential for NaGIR1 catalysis includes nucleotide residues within the NaHEG coding region. Functional experiments in yeast conclude that expression of NaHEG, and subsequent endonuclease activity in yeast extracts, is completely dependent on a functional NaGIR1 ribozyme [28]. Thus, the NaGIR1 and NaHEG domains have to be considered as one functional unit within the Nae.S516 intron.

Gain of a L5b GNRA tetraloop in NaGIR2 during Naegleria *evolution*

One of the best-studied tetraloop receptor interaction is the L5b-P6a tertiary structure in the *Tetrahymena* group I intron ribozyme [40]. Here, the GAAA loop in L5b specifically binds to the 11-nt receptor motif CCUAAG-UAUGG within the helical stem of P6a by docking into the shallow groove. The L5b-P6a interaction in *Tetrahymena* is essential for an efficient folding of the P4–P6 domain, and subsequently the folding and activity of the splicing ribozyme.

Two of the most variable parts in the *Naegleria* GIR2 splicing ribozyme are L5b and P6a (Figure 2). A closer inspection of the 13 twin-ribozyme intron sequences identify tetra-, penta-, and hexaloops in L5b of 8, 4, and 1 introns, respectively. All tetraloops belong to the GNRA family (N = A, G, C or U; R = A or G) known to specifically interact with receptor sequences. To address the distribution pattern of the L5b tetraloops among the various *Naegleria* intron isolates a phylogenetic analysis based on 1370 positions, representing the complete twin-ribozyme introns, was performed. Essential identical trees were obtained from the NJ and MP methods, and a representative NJ tree is shown in Figure 6A. Interestingly, introns harboring a L5b GNRA tetraloop cluster together with high bootstrap support, suggesting that a L5b tetraloop was gained late the evolution of the *Naegleria* genus. The only exception appears the L5b pentaloop of the *N. philippinensis* RJTM intron (Figure 6A), but this example may represent a secondary loss of a GNRA tetra-loop (e.g. GUAA to AUAAA).

The primary function of GNRA tetraloops is to participate in long-range RNA-RNA interactions by specific binding to a receptor motif. A variety of receptor motifs, ranging from 4 to 12 nt, have been recognized experimentally [41-43]. Figure 6B presents secondary structure diagrams of the various NaGIR2 P6a regions and their corresponding L5b loops. A prominent difference in the P6a structure is noted among introns possessing L5b GNRA tetraloops compared to those with penta- or hexaloops. Introns with GNRA loops contain a less tightly base-paired P6a stem with several proposed exposed residues (see Figure 6B) compared to the P5b penta- or hexaloop containing introns (compare *N. clarki* RU30 and *N. carteri* NG055). We speculate that the exposed residues in P6a could be involved in RNA-RNA interactions as GNRA receptors. However, these sequences do not fit any known consensus motifs, suggesting that new motifs are yet to be experimentally identified.

Conclusion

Evolutionary aspects on the structural organization of *Naegleria* twin-ribozyme group I introns have been
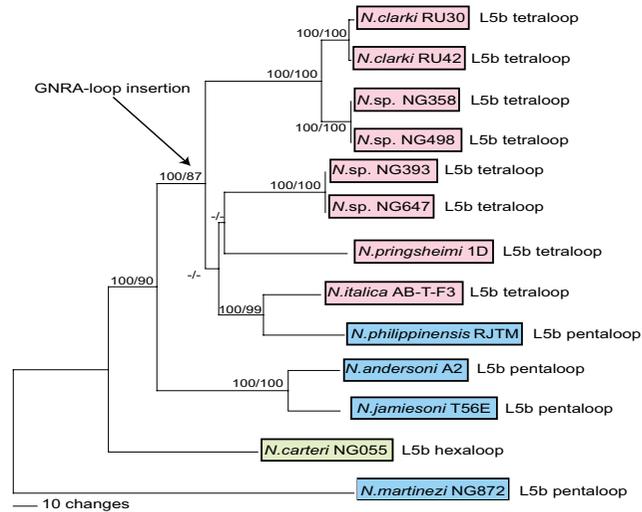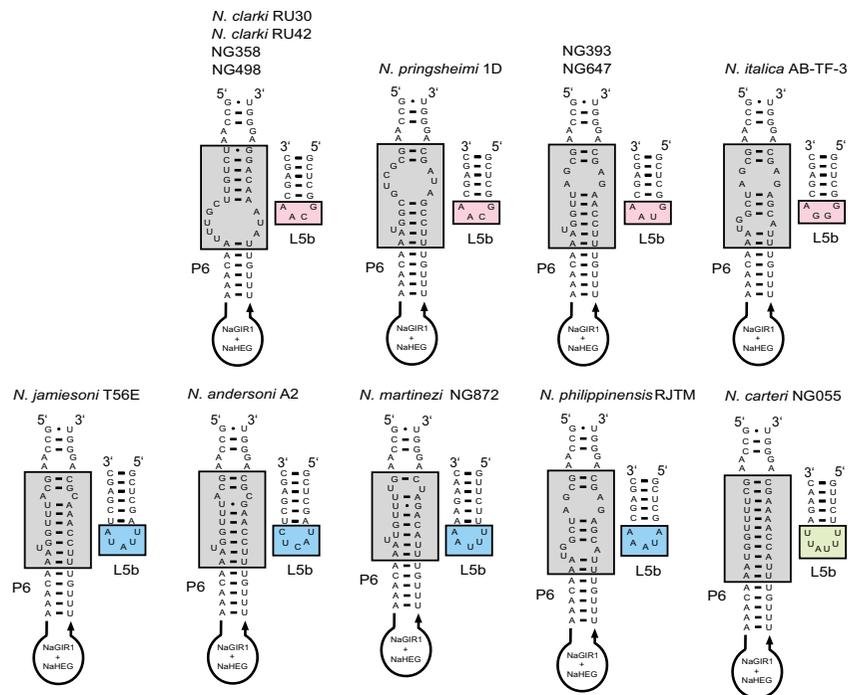
Figure 6
Structural variations in L5b and P6a of NaGIR2. A) NJ phylogenetic tree of the 13 complete Nae.S516 introns based on 1370 positions. NJ and MP bootstrap values are indicated at the branches. The presence of L5b tetra- (red), penta- (blue), or hexaloops (green) in L5b in NaGIR2 are indicated. All tetra-loops belong to the GNRA-loop family. B) Structure diagrams of P6a (grey boxes) and L9b representing the 13 twin-ribozyme Nae.S516 introns.

reported previously [19]. Whereas the NaGIR2 splicing ribozyme appears related to other eukaryote rDNA group IC1 intron [16], the NaGIR1 capping ribozyme has recently evolved from a bacterial tRNA group I introns [29]. Here we present phylogenetic evidence of a vertical inheritance pattern of the Nae.S516 intron in *Naegleria* that includes each of the domains NaGIR1, NaGIR2, and NaHEG, and corroborates a previous study based on 5 intron sequences [6]. Based on the reported distribution pattern and phylogeny, we propose the following vertical inheritance scenario for the Nae.S516 evolution. 1) A pre-organized twin-ribozyme group I intron was gained in the rDNA early in evolution of the *Naegleria* genus, but after the Cluster 6 branching (see Figure 1). 2) Once established, the *Naegleria* intron co-evolved along with its host rDNA by maintaining intron activities including intron splicing, endonuclease mRNA capping, and homing endonuclease DNA cleavage. 3) The intron was subsequently lost (see Figure 1) by sporadic deletions in most isolates (70 %). 4) Degradation of the NaHEG is initiated due to loss of biological function, and subsequent selection pressure (e.g. *N. martinezi* NG872), resulting in complete deletion of the NaHEG as well as its regulatory NaGIR1 capping ribozyme (e.g. *N. byersi* NG597) [8]. 5) The remaining introns have to improve and adjust their functions by continuous sequence evolution in order to be maintained within rDNA. Here, a recent gain of a GNRA tetraloop receptor in the P4–P6 domain would facilitate folding of the splicing ribozyme (Figure 6).

What is the biological role of a functional NaHEG in *Naegleria* S516 introns? There are only two reported examples addressing the biological role of nuclear group I intron HEGs in experimental settings. In sexual matings between intron-containing and intron-lacking strains of either the myxomycetes *Physarum polycephalum* or *Didymium iridis*, group I introns were shown to be mobile due to the double-strand-break-repair pathway induced by intron-encoded homing endonucleases [26,27]. In both cases the homing endonucleases were found to cleave the group I intron lacing alleles in a highly sequence specific manner, resulting in unidirectional transfers of introns into the intron-lacking strains. This process is dependent on a sexual reproduction of the host organism, which is apparently absent in *Naegleria*. However, the *Naegleria* intron endonucleases possess hallmarks linked to a function in intron homing. First, sequence comparisons show that the *Naegleria* enzymes belong to the same His-Cys homing endonuclease family as the known homing endonucleases I-*Ppo*I and I-*Dir*I encoded by the mobile *Physarum* and *Didymium* group I introns [10,11,25]. Second, the *Naegleria* endonucleases cleave only intron lacking alleles flanking the intron insertion site at the SSU rRNA gene [24,25]. Finally, artificial expression of the *Naegleria* endonuclease and its intron in yeast generate intron homing intermedi-

ates consistent with a homing endonuclease function [28]. Interestingly, *Naegleria* may occasionally perform sexual reproduction in nature since Pernin and co-workers [44,45] reported evidence for genetic exchange in *N. lovaniensis*, including chromosomal recombination. Both haploid and diploid strains of the *N. gruberi* NEG isolate have been described based on both amoeba DNA content and UV-sensitivity [46,47]. Perhaps the observed recombination-like feature of NaGIR1 in *N. carteri* NG055 (see Figure 5) is a result of rare sexual mating. This possibility remains to be experimentally explored.

Methods
Naegleria *strains, DNA amplification, plasmid cloning, and DNA sequencing*
The following *Naegleria* isolates were DNA sequenced at ITS-rDNA and Nae.S516 intron regions in this study: *N. clarki* RU30 (ITS-rDNA and Nae.S516); *N. clarki* RU42 (ITS-rDNA and Nae.S516); *N. pringsheimi* 1D (ITS-rDNA and Nae.S516); *N. philippinensis* RJTM (ITS-rDNA and Nae.S516); *N. carteri* NG055 (ITS-rDNA and Nae.S516); *Naegleria* sp. NG647 (ITS-rDNA and Nae.S516); *Naegleria* sp. NG358 (ITS-rDNA and Nae.S516); *Naegleria* sp. NG393 (ITS-rDNA and Nae.S516); *Naegleria* sp. NG498 (ITS-rDNA and Nae.S516); *Naegleria* sp. NG169 (ITS-rDNA);*Naegleria* sp. NG336 (ITS-rDNA); *Naegleria* sp. NG491 (ITS-rDNA); *Naegleria* sp. NG492 (ITS-rDNA). The strains without designation are under revision and will be proposed proper species names based on phylogenetic analyses of ITS1-5.8S-ITS2 sequences (De Jonckheere et al. in preparation). A complete list of all 70 *Naegleria* isolates included in this study is presented in Table 1. DNA samples of *Naegleria* strains were prepared as described previously [2], dissolved in water, and applied as template in 50 μl standard PCR reactions. Amplified product of interest were plasmid cloned into the pGEM®-T Easy Vector System I (Promega). Individual clones where DNA purified and sequenced with the ABI PRISM BigDyeTerminator Cycle Sequencing Ready Reaction Kit (Perkin-Elmer), running on an ABI Prism 377 system (Perkin-Elmer), or using the sequencing service from MWG Biotech [48]. Two or more individual clones were sequenced from all introns and ITS-rDNA regions analysed. The following oligoprimers were used in Nae.S516 PCR amplifications and DNA sequencing analyses: OP 25 (5'-CTC GAA TTC GCT CTT GGA GCT GGA ATT A-3'), OP 26 (5'-ACG AAG CTT ATT TCT AAG CCT-3'), OP 28 (5'-CAG AGG AGT TTC TTA CCT ATC-3'), OP 131 (5'-AAA CGA ATT CTA TTG ATT AGT AGT-3'), OP 946 (5'-GAA TTG AAA AAG CTT GAT-3'), OP 1200 (5'-AAA CAA ATG CTA TTG ATC A-3'), OP 1201 (5'-GAA CGT CTA GAG ACT ACA CGG-3'), OP 1042 (5'-CGA TTT TCC ATG ATT TGG G-3'), OP 1043 (5'-ATA CCT CAA CAG AGG TCC-3'), OP 1044 (5'-GGA CGT CTA GAG ACT ACA CGG-3'), OP 1045 (5'-TGA TGC ACG TAC GAA TCG AGC-3'), OP

276 (5'-GGT AAA CAA ATC CCT GTT-3'), OP 823 (5'-TAA CCA TTT TGT ATG GGA-3'). Heteroplasmic rDNA alleles (intron-containing/intron lacing) were not observed. The following oligoprimers were used in ITS-rDNA PCR amplifications and DNA sequencing analyses: OP 918 (5'-AAC CTG CGT AGG GAT CAT TT-3'), OP 919 (5'-TTT CCT CCC CTT ATT AAT AT-3').

Sequence alignment and phylogenetic analysis
Multiple alignment of sequences were performed by using Megalign (Version 5.06) included in the Lasergene package from DNASTAR, Inc [49], Bioedit (Version 7.0.4.1; [50]), manuel refinements. Phylogenetic analyses and non-synonymous to synonymous substitution rates [51] were conducted using MEGA version 2.1 [52], PAUP* (Version 4.0 Beta) [53], and MrBayes (version 3.1) [54,55]. Trees were built with the methods of neighbor-joining (NJ) using different distance matrixes, maximum parsimony (MP) with the branch and bound search method, as well as Bayesian analyses (BAY) and maximum likelihood (ML) using different evolutionary models. The reliabilities of the tree topologies were evaluated by bootstrapping (NJ, MP, and ML), and posterior probability (BAY).

ITS and intron sequence analyses
Two different data sets of the internal transcribed rDNA spacer region (ITS-rDNA: ITS1-5.8S-ITS2) were used. In analysis with all the 70 *Naegleria* isolates, only 287 nucleotide positions could unambiguously be aligned due to high sequence variation in ITS2. However, when the analysis was restricted to the 14 intron-containing *Naegleria* isolates we extended the ITS-rDNA region to 415 nucleotide positions. Based on the multiple sequence alignment a NJ tree was constructed with the Kimura-2 evolutionary model of substitution (K2), with pair wise deletion of gaps and bootstrapped with 2000 replications with a cut-off value of 50%. Similarly, intron trees are constructed with NJ-K2 parameters. The robustness of the tree topologies were tested by the NJ-K2 parameter (first value), MP branch and bound search criteria (second value), and ML with the HKY+G model of substitution selected by Modeltest 3.7 [56], all from 1000 replicates. The last values where constructed by running 1000000 generations of Metropolis-coupled Markov chain Monte Carlo, and trees were sampled every 100 generations (average standard deviation of split frequencies below 0.01). A consensus tree was generated from the 75% last trees to find posterior probabilities (Burn-in value = 2500).

Authors' contributions
OGW did the sequencing of ITS-rDNA, and group I introns in collaboration with CE. OGW performed the phylogenetic analyses. JFD contributed with DNA preparations of *Naegleria* isolates. SDJ directed the research and wrote the paper in collaboration with OGW.

Acknowledgements

References
1. De Jonckheere JF: **A century of research on the amoebaflagellate genus *Naegleria*.** *Acta Protozool* 2002, **41**:309-342.
2. De Jonckheere JF: **Molecular definition and the ubiquity of species in the genus *Naegleria*.** *Protist* 2004, **155**:89-103.
3. Clark CG, Cross GA: **rRNA genes of *Naegleria gruberi* are carried exclusively on a 14-kilobase-pair plasmid.** *Mol Cell Biol* 1987, **7**:3027-3031.
4. Clark CG, Cross GA: **Small-subunit ribosomal RNA sequence from *Naegleria gruberi* supports the polyphyletic origin of amoebas.** *Mol Biol Evol* 1988, **5**:512-518.
5. Johansen S, Coucheron DH, Haugen P: **Group I introns in nuclear ribosomal DNA of *Naegleria*.** In *Ninth International meeting on the biology and pathogenicity of free-living amoebae proceedings* Edited by: Billot-Bonef S, Cabanes PA, Marciano-Cabral F, Pernin P, Pringuez E. Paris: John Libby Eurotext; 2001:254-260.
6. De Jonckheere JF: **Evidence for the ancestral origin of group I introns in the SSUrDNA of *Naegleria* spp.** *J Eukaryot Microbiol* 1994, **41**:457-463.
7. Einvik C, Decatur WA, Embley TM, Vogt VM, Johansen S: ***Naegleria* nucleolar introns contain two group I ribozymes with different functions in RNA splicing and processing.** *RNA* 1997, **3**:710-720.
8. De Jonckheere JF, Brown S: **Loss of the ORF in the SSUrDNA group I intron of one *Naegleria* lineage.** *Nucleic Acids Res* 1994, **22**:3925-3927.
9. De Jonckheere JF, Brown S: **Three different group I introns in the nuclear large subunit ribosomal DNA of the amoeboflagellate *Naegleria*.** *Nucleic Acids Res* 1998, **26**:456-461.
10. De Jonckheere JF, Brown S: **A novel ORF-containing group I intron with His-Cys box in the LSU rDNA of *Naegleria*.** *Acta Protozool* 2001, **40**:27-31.
11. Haugen P, De Jonckheere DF, Johansen S: **Characterization of the self-splicing products of two complex *Naegleria* LSU rDNA group I introns containing homing endonuclease genes.** *Eur J Biochem* 2002, **269**:641-1649.
12. Cech TR: **Self-splicing of group I introns.** *Annu Rev Biochem* 1990, **59**:543-568.
13. Haugen P, Simon DM, Bhattacharya D: **The natural history of group I introns.** *Trends Genet* 2005, **21**:111-119.
14. Lehnert V, Jaeger L, Michel F, Westhof E: **New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the *Tetrahymena thermophila* ribozyme.** *Chem Biol* 1996, **3**:993-1009.
15. Guo F, Gooding AR, Cech TR: **Structure of the *Tetrahymena* ribozyme: base triple sandwich and metal ion at the active site.** *Mol Cell* 2004, **16**:351-362.
16. Haugen P, Coucheron DH, Rønning SB, Haugli K, Johansen S: **The molecular evolution and structural organization of self-splicing group I introns at position 516 in nuclear SSU rDNA of myxomycetes.** *J Eukaryot Microbiol* 2003, **50**:283-292.
17. **Comparative RNA Web site** [http://www.rna.icmb.utexas.edu]
18. Bhattacharya D, Cannone JJ, Gutell RR: **Group I intron lateral transfer between red and brown algal ribosomal RNA.** *Curr Genet* 2001, **40**:82-90.
19. Einvik C, Elde M, Johansen S: **Group I twintrons: genetic elements in myxomycete and schizopyrenid amoeboflagellate ribosomal DNAs.** *J Biotechnol* 1998, **64**:63-74.
20. Nielsen H, Fiskaa T, Birgisdottir AB, Haugen P, Einvik C, Johansen S: **The ability to form full-length intron RNA circles is a general property of nuclear group I introns.** *RNA* 2003, **9**:1464-1475.
21. Birgisdottir ÅB, Johansen S: **Site-specific reverse splicing of a HEG-containing group I intron in ribosomal DNA.** *Nucleic Acids Res* 2005, **33**:2042-2051.

22. Johansen S, Embley TM, Willassen NP: **A family of nuclear homing endonucleases.** *Nucleic Acids Res* 1993, **21**:4405.
23. Chevalier BS, Stoddard BL: **Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility.** *Nucleic Acids Res* 2001, **29**:3757-3774.
24. Elde M, Haugen P, Willassen NP, Johansen S: **I- *NjaI*, a nuclear intron-encoded homing endonuclease from *Naegleria*, generates a pentanucleotide 3' cleavage-overhang within a 19 base-pair partially symmetric DNA recognition site.** *Eur J Biochem* 1999, **259**:281-288.
25. Elde M, Willassen NP, Johansen S: **Functional characterization of isoschizomeric His-Cys box homing endonucleases from *Naegleria*.** *Eur J Biochem* 2000, **267**:7257-7266.
26. Muscarella DE, Vogt VM: **A mobile group I intron in the nuclear rDNA of *Physarum polycephalum*.** *Cell* 1989, **56**:443-454.
27. Johansen S, Elde M, Vader A, Haugen P, Haugli K, Haugli F: ***In vivo* mobility of a group I twintron in nuclear ribosomal DNA of the myxomycete *Didymium iridis*.** *Mol Microbiol* 1997, **24**:737-745.
28. Decatur WA, Johansen S, Vogt VM: **Expression of the *Naegleria* intron endonuclease is dependent on a functional group I self-cleaving ribozyme.** *RNA* 2000, **6**:616-627.
29. Johansen S, Einvik C, Nielsen H: **DiGIR1 and NaGIR1. Naturally occurring group I-like ribozymes with unique core organization and evolved biological role.** *Biochimie* 2002, **84**:905-912.
30. Nielsen H, Westhof E, Johansen S: **An mRNA is capped by a 2', 5' lariat catalyzed by a group I-like ribozyme.** *Science* 2005, **309**:1584-1586.
31. Einvik C, Nielsen H, Westhof E, Michel F, Johansen S: **Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites.** *RNA* 1998, **4**:530-541.
32. Johansen S, Haugen P: **A new nomenclature of group I introns in ribosomal DNA.** *RNA* 2001, **7**:935-936.
33. Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA: **Crystal structure of a self-splicing group I intron with both exons.** *Nature* 2004, **430**:45-50.
34. Jabri E, Cech TR: ***In vitro* selection of the *Naegleria* GIR1 ribozyme identifies three base changes that dramatically improve activity.** *RNA* 1998, **4**:1481-1492.
35. Weiss MA, Narayana N: **RNA recognition by arginine-rich peptide motifs.** *Biopolymers* 1998, **48**:167-180.
36. Chen Y, Varani G: **Protein families and RNA recognition.** *FEBS J* 2005, **272**:2088-2097.
37. Haugen P, Huss VAR, Nielsen H, Johansen S: **Complex group I introns in nuclear SSU rDNA of red and green algae: evidence of homing-endonuclease pseudogenes in the Bangiophyceae.** *Curr Genet* 1999, **36**:345-353.
38. Haugen P, Reeb V, Lutzoni F, Bhattacharya D: **The evolution of homing endonuclease genes and group I introns in nuclear rDNA.** *Mol Biol Evol* 2004, **21**:129-140.
39. Haugen P, Wikmark OG, Vader A, Coucheron DH, Sjøttem E, Johansen SD: **The recent transfer of a homing endonuclease gene.** *Nucleic Acids Res* 2005, **33**:2734-2741.
40. Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA: **Crystal structure of a group I ribozyme domain: principles of RNA packing.** *Science* 1996, **273**:1678-1685.
41. Jaeger L, Michel F, Westhof E: **Involvement of a GNRA tetraloop in long-range RNA tertiary interactions.** *J Mol Biol* 1994, **236**:1271-1276.
42. Costa M, Michel F: **Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with *in vivo* evolution.** *EMBO J* 1997, **16**:3289-3302.
43. Ikawa Y, Naito D, Aono N, Shiraishi H, Inoue T: **A conserved motif in group IC3 introns is a new class of GNRA receptor.** *Nucleic Acids Res* 1999, **27**:1859-1865.
44. Cariou ML, Pernin P: **First evidence for diploidy and genetic recombination in free-living amoebae of the genus *Naegleria* on the basis of electrophoretic variation.** *Genetics* 1987, **115**:265-270.
45. Pernin P, Ataya A, Cariou ML: **Genetic structure of natural populations of the free-living amoeba, *Naegleria lovaniensis*. Evidence for sexual reproduction.** *Heredity* 1992, **68**:173-181.
46. Fulton C: **Amebo-flagellates as research partners: the laboratory biology of *Naegleria* and *Tetramitus*.** In *Methods in Cell Physiology Volume 4*. Edited by: Prescott DM. New York: Academic Press; 1970:341-473.
47. Fulton C: ***Naegleria* : a research partner for cell and development biology.** *J Euk Microbiol* 1993, **40**:520-532.
48. **MWG Home**   [http://www.mwg-biotech.com]
49. **DNASTAR – Software for molecular biology**   [http://www.dnastar.com]
50. Hall TA: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucl Acids Symp Ser* 1999, **41**:95-98.
51. Nei M, Kumar S: *Molecular Evolution and Phylogenetics* Oxform University Press, New York; 2000.
52. Kumar S, Tamura K, Jakobsen IB, Nei M: **MEGA2: Molecular Evolutionary Genetics Analysis software.** Arizona State University, Tempe, Arizona, USA; 2001.
53. Swofford DL: **PAUP*: Phylogenetic analysis using parsimony (* and other methods).** In *Version 4.0b8* Sinauer, Sunderland, MA; 2002.
54. Huelsenbeck JP, Ronquist F: **MrBayes: Bayesian inference of phylogeny.** *Bioinformatics* 2001, **17**:754-755.
55. Ronquist F, Huelsenbeck JP: **MRBAYES 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19**:1572-1574.
56. Posada D, Crandall KA: **Modeltest: testing the model of DNA substitution.** *Bioinformatics* 1998, **14**:817-818.