Hindawi Publishing Corporation
Dataset Papers in Biology
Volume 2013, Article ID 854869, 12 pages
http://dx.doi.org/10.7167/2013/854869

Dataset Paper

Detection of Introns in Eukaryotic Small Subunit Ribosomal RNA Gene Sequences

Dipankar Bachar,[1,2] Laure Guillou,[3,4] and Richard Christen[1,2]

[1] *CNRS UMR 7138, Systématique, Adaptation, Évolution, Université de Nice-Sophia Antipolis, Parc Valrose, BP 71, 06108 Nice Cedex 02, France*
[2] *UMR 7138, Systématique, Adaptation, Évolution, Université de Nice-Sophia Antipolis, Parc Valrose, BP 71, 06108 Nice Cedex 02, France*
[3] *CNRS UMR 7144, Laboratoire Adaptation et Diversité en Milieu Marin, Place Georges Teissier, 29680 Roscoff, France*
[4] *Station Biologique de Roscoff, Université Pierre et Marie Curie-Paris 6, Place Georges Teissier, 29680 Roscoff, France*

Correspondence should be addressed to Richard Christen; christen@unice.fr

Received 27 April 2012; Accepted 20 May 2012

Academic Editors: A. G. de Brevern, D. R. Flower, and M. L. Raymer

The gene encoding SSU-rRNA sequences is the tool of choice for phylogenetic analyses and environmental biodiversity analyses of bacteria, Archaea but also unicellular Eukaryota. In Eukaryota, gene sequences may often be interrupted by long or several introns. Searching in GenBank release 188, we found descriptions of 3638 such sequences. Using a database of 180 000 SSU-rRNA sequences well annotated for taxonomy and a C++ program written for that purpose, we computed the presence of 18 691 introns (among which the 3638 described introns). Filtering on length and sequence quality, 3646 sequences were retained. These introns were clustered; clusters were analyzed for the presence of single or multiple clades at various levels of taxonomic depth, allowing future analyses of horizontal transfers. Various analyses of the results are provided as tabulated files as well as FASTA files of described or computed introns. Each sequence is annotated for cellular location (nuclear, chloroplast, and mitochondria), positions at which they were found in the SSU-rRNA sequences and taxonomy as provided by GenBank.

1. Introduction

The gene that microbiologists use to determine the taxonomic affiliations of microbes using molecular methods needs to meet a number of requirements. It has to be conservative in its function and present in every organism analyzed. Often the presence of conserved domains is required, allowing the design and use of universal PCR primers. Finally, sequences of most known organisms must be available in the public databases (i.e., International Nucleotide Sequence Database Collaboration (INSDC) between Japanese, European, and American nucleotide databases, resp., DDBJ, ENA, and GenBank, http://www.insdc.org/). A gene meets these requirements to a high degree: the gene for one of the RNA subunits that together form the ribosome, also known as the small subunit ribosomal RNA (SSU-rRNA) gene. For that reason, the gene encoding the SSU-rRNA serves as a prominent tool for phylogenetic and environmental biodiversity analyses of bacteria, Archaea but also unicellular Eukaryota [1–3].

SSU-rRNA gene sequences may contain numerous self-splicing introns of variable lengths [4–23]. The SSU-rRNA genes can thus be enlarged to up to 3.5 kb. Introns have rarely been identified in bacterial SSU-rRNA gene sequences (see one example in *Thiomargarita namibiensis* [5]), but they are often present in SSU-rRNA gene sequences of Eukaryota (see the aforementioned part). Such length heterogeneity of SSU-rRNA gene sequences has so far seldomly been considered when constructing SSU-rRNA-based clone libraries or more recently for deep sequencing of SSU-rRNA amplicons. The detection of elongated SSU-rRNA gene sequences may have implications for analyses in molecular ecology and may cause systematic biases. When dealing with clone libraries, the usually low number of sequences allows for a manual analysis, and introns in new sequences can be easily

detected and not taken into account. However, with the advent of Next Generation Sequencing (NGS) technologies, hundreds of thousands of sequences are generated during each experiment, and these data need to be analyzed via automated pipelines. If a sequence from a given organism differs from the sequence of the same organism present in the database via intron insertions/deletions, proper identification will frequently fail. Also, NGS pipelines apply filters on amplicon lengths, and a bias due to the PCR amplification step will discriminate against longer sequences. As a result, some groups of organisms may be simply discarded because the domains amplified become too long due to the insertion of introns.

In this paper, we assessed the frequency of such introns using the release 188 of GenBank, for every sequence that contained at least one domain identified as a partial or complete SSU-rRNA sequence of eukaryotic origin. We also identified the origin of these sequences as nuclear, mitochondria, or chloroplast encoded. In a first step, we selected and extracted subsequences that had been annotated to contain introns. We then used these sequences to identify and extract putative introns from the entire set of SSU-rRNA gene sequences of eukaryotic origin. Finally, for each category, we provide the longest described intronic sequence. This dataset can be used to feed automated pipelines dealing with NGS datasets.

2. Methodology

A database of sequences was built as follows. Reference sequences were retrieved from GenBank release 188 (152 177 733 sequences) using a series of 175 keywords. 674 146–688 113 subsequences were retrieved using ACNUC [24]. Filtering out subsequences shorter than 800 nt and longer than 8000 nt provided 237 967 sequences. Filtering for Eukaryotic origin provided 145 968 sequences. These sequences were contained in 145 163 GenBank files which were used for the analyses of SSU-RNA gene sequences. Introns were retrieved from the same query but without filtering on sequence length, that is, using 471 263 GenBank files (entries) (598 851 rRNA subsequences).

A given entry may contain a single SSU rRNA gene sequences, sequences of several genes, or a complete genome. When extracting SSU rRNA gene sequences and looking, we had for a long time considered two different cases.

(i) The rRNA is solely described as to be joined, with no description of introns as in accession number EF689886 (939 occurrences found):

```
rRNA    join(<1..213,559..660,712..146,1790..2004,
        2394..2407,2721..2904,3130..>3246)
        /product ="18S ribosomal RNA"
```

(ii) Both join and introns are described as in accession number U54637 (2187 occurrences found):

```
rRNA    join(650..2046,3427..3926)
        /product = "small subunit ribosomal RNA"
intron  2047..3426
        /note = "small subunit rRNA; group-IC2"
```

This led us to build a database of SSU-rRNA annotated in the following way:

(i) _U|: sequence simply annotated as SSU-rRNA;

(ii) _R|: sequence described by a join (see the following) and the corresponding sequence of that of rRNA (a consequence of the older default case for protein coding sequences);

(iii) _G|: the genomic sequences (rDNA gene) corresponding to the _R| sequences including putative introns.

During the time course of this study, we discovered several other possibilities:

(1) a simple description of the presence of intron(s) such as that of accession number AM231292, 512 occurrences, but the join to describe the rRNA sequence is missing:

```
gene    <1..>3533
        /gene = "18S rRNA"
rRNA    <1..>3533
        /gene = "18S rRNA"
        /product = "18S ribosomal RNA"
intron  569..1716
        /gene = "18S rRNA"
        /note = "S529 group I intron"
intron  2837..3365
        /gene = "18S rRNA"
        /note = "S1389 group I intron"
```

(2) some strange annotations such as those of accession number X80710, for which exons are described, but within introns of length 0:

```
rRNA    join(1..528,529..1134,1135..1743)
        /product = "18S ribosomal RNA"
exon    1..528
        /number = 1
exon    529..1134
        /number = 2
exon    1135..1743
        /number = 3
```

(3) still another strange case such as that of accession number AB016009, which creates two separate sequences for the SSU rRNA gene:

```
rRNA    <1..1618
        /product = "18S ribosomal RNA"
intron  1619..2115
        /note = "Group IC1 intron (site 1389);
        Insertion site is site 1389 corresponding
        to E. coli 16S
        rDNA"
rRNA    2116..>2254
        /product = "18S ribosomal RNA"
```

(4) note finally peculiar cases such as that of accession number AF274109, in which introns are described as spliced sequences:

source	1..2126
	/organism = "Ochrolechia parella"
	/mol_type = "genomic DNA"
	/db_xref = "taxon:129506"
rRNA	join(<1..384,443..1469, 1825..>2126)
	/product = "18S ribosomal RNA"
intron	join(385..442,1470..1824)

(5) the creation of chimeric sequence accession number AB259428, because the sequences extracted are a concatenation of $5'$ and $3'$ ends of the molecule, which could suggest that an intron has been found in between, when this is only a series of N inserted by the authors:

source	1..525
	/organism = "Didymium panniforme"
	/mol_type = "genomic DNA"
	/specimen_voucher = "TNS-M-Y-16880 (TNS)"
	/db_xref="taxon:385452"
rRNA	join(<1..233,334..>525)
	/product = "small subunit ribosomal RNA"
gap	234..333
	/estimated_length = unknown

(6) finally some sequences such as AJ506970 which are quite difficult to treat:

source	1..1906
	/organism = "Physcia dimidiata"
	/mol_type = "genomic DNA"
	/specimen_voucher = "Mayrhofer & Litterski 13.932 (GZU)"
	/db_xref = "taxon:116814"
gene	1..1906
	/gene = "nrSSU"
gene	order(<1..507,825..1262,1476..1906, AJ507615.1:1..12, AJ507615.1:206..565)
	/gene = "nrSSU"
rRNA	join(<1..507,825..1262,1476..1906, AJ507615.1:1..12,AJ507615.1:206..565)
	/gene = "nrSSU"
	/product = "18S ribosomal RNA"
exon	1..507
	/gene = "nrSSU"
	/number = 1
intron	508..824
	/gene = "nrSSU"
	/number = 1
exon	825..1262
	/gene = "nrSSU"
	/number = 2
intron	1263..1475
	/gene = "nrSSU"
	/number = 2
exon	1476..1906
	/gene="nrSSU"
	/number = 3

Annotation extends the sequence over two or more accession numbers. In our present annotation and extraction pipeline, the _U| sequence will cover the extension over the various accession numbers, while the _G| sequence will be extracted solely from the sequence contained in the primary accession number (the one contained in the entry).

Each of these different annotations was used to build a database on described intronic sequences.

Reference taxonomy was built as follows. rRNA sequences of 800 nt or more were used to build a taxonomic database, using manual analyses and with homogenization of the taxonomy to 8 fields from domain (nuclear or organelle), kingdom, phylum, and so forth to genus and finally species. This database was then used to estimate the origins of introns retrieved. Note however that in most files provided in the dataset, the original GenBank taxonomy was kept.

Introns were computed as follows. We developed a special C++ program that allows to search a file of sequences within another file of sequences, similarly to Blast, except that a Needleman-Wunsch global alignment is used. A search is done with the following parameters:

(i) n which is the maximal number of best hits to return,

(ii) p which is the minimum percentage of similarity to return a hit,

(iii) similarity which is computed from the alignment, not taking into account terminal gaps.

After different trials, parameters $n = 30$ and $p = 90$ were retained. Every SSU-rRNA sequence was searched for the presence of introns; for each sequence introns found were sorted by decreasing length and successively removed from the sequence when present. When overlapping introns are found, only the longest one is thus taken as a true intron. We are aware that this approach is still crude, but in absence of inspection of the present results by experts, it is difficult to improve the algorithm without sound proofs of false positive and false negative results. Checking the algorithm on protein-coding sequences could be an alternate solution. However checking that was done on sequences that had been described as containing introns showed the results to be good (see the results). These results were then analyzed using a series of dedicated Python scripts that produced the dataset provided.

598 851 entries were retrieved from GenBank release 188 (April 2012), which contained annotations identifying at least one subsequence as being an SSU-rRNA gene sequence. In this paper, the term *entry* will be used to refer to every information available for a given accession number. A given entry may contain a single SSU-rRNA gene sequence, sequences of

several genes, or a complete genome. When extracting SSU-rRNA gene sequences, we considered three different cases depending on how SSU-rRNA sequences are annotated for introns, and each sequence was labeled as follows:

(i) _U|: for sequences simply annotated as SSU-rRNA,

(ii) _R|: for sequences described by a join (implicit description), which implies that the gene contains intron(s), or with introns described in the features (explicit description),

(iii) _G|: for the genomic sequences (rDNA gene sequence) corresponding to the _R| sequence including putative introns.

We thus built a database of 184 827 sequences of 800 nt or more (_U|: 182 107 sequences; _G| and _R|: 1360 sequences, some of which may contain several introns). Extraction of introns—either explicitly or implicitly—described led us to build a database of 3638 intronic sequences for intronic sequences of 40 nt or more and not having at least 10 consecutive *N* (to which we then added some more sequences as described previously; total sequences: 3646); see Table 1.

We also extracted the annotation that described the type of each intronic sequence (see Table 2).

As shown in Table 2, the vast majority of introns are described as group I introns. We used the Vienna RNA Web-Servers (http://rna.tbi.univie.ac.at/) to visualize predictions of secondary structures of these sequences, but we were not able to find any obvious conserved secondary structure among a given category (group I or group II), but we did not use any extensive search as described, for example, in [25].

We clustered these sequences as described previously and looked at the clusters obtained. Figure 1 shows an example of the result.

The results shown in Figure 1 are a typical example of what could be seen: in a given cluster, some sequences differ for the boundaries of the intronic sequences. It is remarkable that for sequences submitted by the same author(s), boundaries are identical but differ from the boundaries found by different authors. We can also observe that the group to which an intron is assigned can be more or less precise.

Described introns are predominantly found within the 50–600 nt range. Only 5 introns had a length of more than 2000 nt (Table 3).

The following describes the computation for retrieval of undescribed introns. We next used these 3638 sequences to analyze our database of 184 827 SSU-rRNA sequences with a length of 800 nt or more for the presence of such intronic sequences. Since some of the _G| (annotated in GenBank for the presence of introns) could not be retrieved, we manually added back some sequences that had been filtered out because they either were too short or contained too many *N*; this provided a new database of 3638 intronic sequences that we again used to find introns in rRNA-SSU sequences. Numbers found are described in Table 4.

Introns found in _G| sequences are interesting to test the efficiency of our algorithm, since we would expect for each _G| sequence to retrieve introns as described in the GenBank annotations (usually and for finding introns, the authors compared the genomic sequence to the rRNA sequence). Results of this analysis were as follows: in 1555 cases, we exactly retrieved the boundaries described in the entries. However, in 357 cases, there was a discrepancy. Dataset Item 29 (Table) shows what occurred. In most cases, the difference lied in one or a few positions at either or both ends. This problem results from the extreme difficulty to exactly determine intron boundaries in non-protein-coding sequences. We looked for a confirmation of this by clustering annotated introns allowing from 5 to 20 differences between sequences (taking or not external gaps as differences). Looking at aligned sequences confirmed the results of Table 1 (see Figure 2).

The _RC sequences are puzzling, since we did not expect to retrieve many "new" introns from rRNA sequences which had been carefully studied for the presence of introns. We randomly and manually analyzed a small number of cases, but we were not able to decide whether or not these "new" introns were real cases or artifacts. The main problem lies in the methods section of the publications associated with the sequences. In no case, we were able to find a proper description of how introns had been identified, that is, either by comparisons with public sequences of closely related organisms or by comparison between the genomic sequence and the direct sequencing of the rRNA sequence. Since rRNA sequencing is tedious, we expect the first case to be true in most cases.

Table 1

(a) Annotations used to extract intronic sequences.

Annotation	Number
Sequences with both join and introns	2187
Sequences with only introns	512
Sequences with only join	939

(b) Genomic localization of described introns.

Annotation	Number
Mitochondria encoded	50
Nuclear encoded	3560
Chloroplast encoded	36

Table 2: Category of well-annotated intronic sequences.

Intron Type	Number of Sequences
Group I intron	1461
Group IA2 intron	1
Group IA3 intron	1
Group IC intron	1
Group IC1 intron	42
Group IC2 intron	3
Group IE intron	5
Group II intron	10
Group IIA1 intron	8
Group IIB1 intron	10

Table 3: Longest introns described (>2000 nt). CG: complete genome; rRNA: only rRNA(s); fragment: long genomic fragment; 1: join only; 2: intron only; 3: both.

Accession Number and Genomic Locations	Cell Location	Clade	Species	Genome Fragment	Description of Intron
M68929.153546.157925	mito-SSU	Viridiplantae	*Marchantia polymorpha*	CG	3
HQ292070.249.3107	mito-SSU	Fungi	*Ophiostoma minus*	rRNA	3
HQ292073.249.3106	mito-SSU	Fungi	*Ophiostoma minus*	rRNA	3
HQ292072.249.3106	mito-SSU	Fungi	*Ophiostoma minus*	rRNA	3
HQ343317.195.3047	Nuclear	Fungi	*Ophiostoma torulosum*	rRNA	3
X91267.36.2342	Nuclear	Viridiplantae	*Scenedesmus pupukensis*	rRNA	3
AF029891.2413.4594	mito-SSU	Fungi	*Cryphonectria parasitica*	fragment	3
EF689875.1460.3551	Nuclear	Fungi	*Verrucaria marmorea*	rRNA	1

Table 4: Type of sequence and number of introns found. _UC: introns found in sequences that had no description of introns; _GC: introns found in genomic sequences, described for presence of introns; _RC: introns found in sequences after removal of introns as described in annotations.

Sequence Type Found	Number of Introns
_UC	15 822
_GC	2162
_RC	607

Table 5: Taxonomy and location of described introns.

Number of Introns	Cell Location	Clade
33	chloro-SSU	Viridiplantae
2	mito-LSU	Fungi
7	mito-LSU	Viridiplantae
1	mito-LSU	Stramenopiles
40	mito-SSU	Fungi
1	mito-SSU	Haptophyceae
1	mito-SSU	Ichthyosporea
7	mito-SSU	Viridiplantae
9	Nuclear	Alveolata
253	Nuclear	Amoebozoa
1	Nuclear	Euglenozoa
1928	Nuclear	Fungi
24	Nuclear	Heterolobosea
1	Nuclear	Metazoa
10	Nuclear	Rhizaria
338	Nuclear	Rhodophyta
968	Nuclear	Viridiplantae
13	Nuclear	Stramenopiles

Since new intronic sequences are differently annotated according to the kind of sequences they extracted, a conservative approach would be not to take into account the _RC sequences as real introns. On the other hand, keeping every intronic sequence would allow identification of yet unknown introns in the genomes of a given clade (Table 5).

We then clustered the introns found and analyzed clusters of 10 or more sequences which we extracted as FASTA

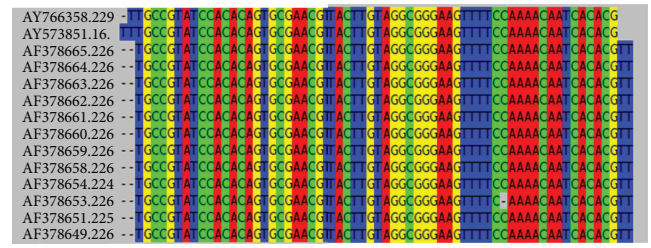


Figure 1: Comparisons of 5′ and 3′ boundaries of sequences submitted by different authors. AY766358 was submitted in 2007 by Milstein and Oliveira as "group I intron." AY573851 was submitted in 2004 by Teasdale, Klein, West, and Mathieson as "group ICI intron." Remaining sequences were submitted in 2002 by Broom et al. as "putative group IC1 intron."

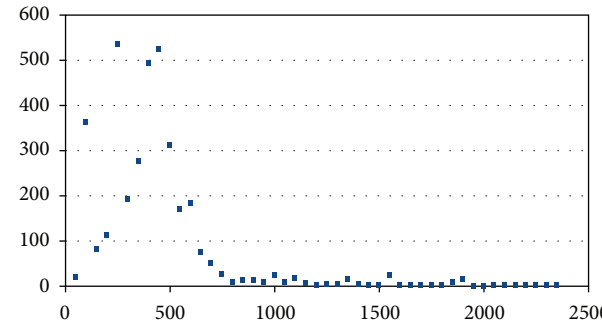
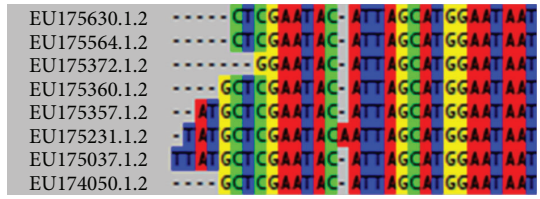


Figure 2: Distribution length of described SSU-rRNA introns.

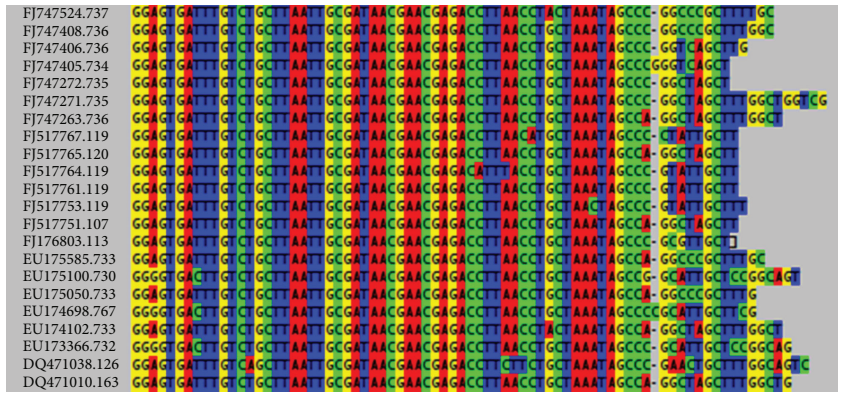
sequences. Differences observed at boundaries confirmed our previous observations and that finding boundaries by computation alone is even more difficult both because known introns may differ slightly at boundaries and because computation is difficult, even using a global alignment algorithm.

The problem appears more acute on the 3′ end, most likely due to a not accurate enough detection of the intron endings (Figure 3).

Our final analysis was to examine each cluster for taxonomic discrepancies. Because horizontal transfers of rRNA introns seem to be rare events [19, 26, 27], we expected that the taxonomy would be homogeneous within each cluster, at



(a)



(b)

Figure 3: Comparisons of $5'$ (a) and $3'$ boundaries (b) of intron sequences extracted by our pipeline and grouped in the same cluster.

least down to the family level. On the other hand, the presence of a given intron in distinct clades would suggest an old origin of this intron, followed by vertical transfer. Because the taxonomy associated with unicellular organism and provided by GenBank is often very poor, for sequences retrieved in various clusters we developed our own taxonomy, based on manual analyses (see previous explanation). These sequences were all annotated using 8 successive fields ranging from the domain (Eukaryota or Organelle) to the species level.

We provide a series of FASTA files in which only sequences extracted from abnormal clusters are shown; each sequence is annotated with our taxonomy. A series of files is thus provided, each with sequences clustered at a given level of dissimilarities and for discrepancies found at successive levels of the taxonomy. Note however that some clusters can be found abnormal when they are probably not. This is true in particular when a discrepancy is found as, for example, between "Eukaryota Opisthokonta" and "Eukaryota Eukaryota_X", the latter annotation resulting from our failing to properly assign an SSU-rRNA sequence. When such case is encountered, it can probably be manually solved.

3. Dataset Description

The dataset associated with this Dataset Paper consists of 69 items which are described as follows.

Dataset Item 1 (Nucleotide Sequences). The longest sequence in each group of introns (groups I, IA2, IA3, IC, IC1, IC2, IE, II, IIA1, and IIB1).

Dataset Item 2 (Nucleotide Sequences). Introns described in GenBank files (longer than 40 nt). Most introns containing long stretches of *N* have been removed (some were kept because they had unique sequences). Each sequence is described as follows: >accession.p1.p2|type|taxonomy, where accession is the accession number, p1 is the $5'$ position of the introns in the entries of sequences, p2 is the $5'$ position of the introns in the entries of sequences, type describes the intron origin, and taxonomy is the taxonomy provided by GenBank (| separated fields).

Dataset Item 3 (Nucleotide Sequences). Introns described in GenBank files (longer than 40 nt and sorted by increasing length).

Dataset Item 4 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (5 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 5 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (10 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 6 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (15 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 7 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (20 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 8 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (25 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 9 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (30 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 10 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (5 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences.

Dataset Item 11 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (10 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences.

Dataset Item 12 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (15 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences.

Dataset Item 13 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (20 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences.

Dataset Item 14 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (25 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences.

Dataset Item 15 (Nucleotide Sequences). Clusters of known introns at several dissimilarities (30 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences.

Dataset Item 16 (Nucleotide Sequences). Sequences for clusters of known introns containing 10 or more sequences of clusters of 5 differences *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence.

Dataset Item 17 (Nucleotide Sequences). Sequences for clusters of known introns containing 10 or more sequences of clusters of 10 differences *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence.

Dataset Item 18 (Nucleotide Sequences). Sequences for clusters of known introns containing 10 or more sequences of clusters of 15 differences *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence.

Dataset Item 19 (Nucleotide Sequences). Sequences for clusters of known introns containing 10 or more sequences of clusters of 20 differences *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence.

Dataset Item 20 (Nucleotide Sequences). Sequences for clusters of known introns containing 10 or more sequences of clusters of 25 differences *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence.

Dataset Item 21 (Nucleotide Sequences). Sequences for clusters of known introns containing 10 or more sequences of clusters of 30 differences *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence.

Dataset Item 22 (Nucleotide Sequences). Sequences for clusters of known introns containing at least 3 sequences at several dissimilarities (5 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 3 sequences are shown.

Dataset Item 23 (Nucleotide Sequences). Sequences for clusters of known introns containing at least 3 sequences at several dissimilarities (10 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 3 sequences are shown.

Dataset Item 24 (Nucleotide Sequences). Sequences for clusters of known introns containing at least 3 sequences at several dissimilarities (15 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 3 sequences are shown.

Dataset Item 25 (Nucleotide Sequences). Sequences for clusters of known introns containing at least 3 sequences at several dissimilarities (20 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 3 sequences are shown.

Dataset Item 26 (Nucleotide Sequences). Sequences for clusters of known introns containing at least 3 sequences at several dissimilarities (25 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 3 sequences are shown.

Dataset Item 27 (Nucleotide Sequences). Sequences for clusters of known introns containing at least 3 sequences at several dissimilarities (30 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 3 sequences are shown.

Dataset Item 28 (Table). Variations in sequence length of described introns.

Column 1: Minimum Length of Sequence
Column 2: Maximum Length of Sequence
Column 3: Number of Sequences

Dataset Item 29 (Table). Differences between boundaries described for introns (in GenBank entries) and boundaries found by computation.

Column 1: Described Introns
Column 2: Found $5'$
Column 3: Found $3'$

Dataset Item 30 (Nucleotide Sequences). Tabulated file for computed introns with description and sequence in which an intron was found with the following fields: ID of sequence in which an intron was found by computation, ID of sequence in which this intron had been described (accession.p1.p2: p1, p2 positions of introns in accession sequence), and sequences of introns found.

Dataset Item 31 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (5 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 32 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (10 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 33 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (15 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 34 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (20 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 35 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (25 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 36 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (30 differences between the longest sequences and other sequences), taking into account terminal gaps as differences.

Dataset Item 37 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (5 differences between the longest sequences and other sequences), taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 38 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (10 differences between the longest sequences and other sequences), taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 39 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (15 differences between the longest sequences and other sequences), taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 40 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (20 differences between the longest sequences and other sequences), taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 41 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (25 differences between the longest sequences and other sequences), taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 42 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (30 differences between the longest sequences and other sequences), taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 43 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (5 differences between

the longest sequences and other sequences), *not* taking into account terminal gaps as differences, but clustering not counting end gaps as differences.

Dataset Item 44 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (10 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences, but clustering not counting end gaps as differences.

Dataset Item 45 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (15 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences, but clustering not counting end gaps as differences.

Dataset Item 46 (Nucleotide Sequences). Clusters of computed introns at several dissimilarities (20 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences, but clustering not counting end gaps as differences.

Dataset Item 47 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (5 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 48 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (10 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 49 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (15 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 50 (Nucleotide Sequences). Sequences for clusters of computed introns containing at least 5 sequences at several dissimilarities (20 differences between the longest sequences and other sequences), *not* taking into account terminal gaps as differences. Each cluster is separated by an empty sequence. Only clusters with at least 5 sequences are shown.

Dataset Item 51 (Nucleotide Sequences). Introns found in genomes of chloroplasts.

Dataset Item 52 (Nucleotide Sequences). Introns found in genomes of mitochondria.

Dataset Item 53 (Nucleotide Sequences). Introns found in nuclear genomes.

Dataset Item 54 (Nucleotide Sequences). Introns found in nuclear genomes. Sequences are sorted by decreasing length.

Dataset Item 55 (Nucleotide Sequences). Sequences extracted from clusters that showed 5 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 2. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 56 (Nucleotide Sequences). Sequences extracted from clusters that showed 5 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 4. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 57 (Nucleotide Sequences). Sequences extracted from clusters that showed 5 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 5. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 58 (Nucleotide Sequences). Sequences extracted from clusters that showed 5 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 6. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum; …; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 59 (Nucleotide Sequences). Sequences extracted from clusters that showed 10 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 2. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 60 (Nucleotide Sequences). Sequences extracted from clusters that showed 10 differences or less between the longest sequences in cluster and other sequences in taxonomy

at level 3. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 61 (Nucleotide Sequences). Sequences extracted from clusters that showed 10 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 4. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 62 (Nucleotide Sequences). Sequences extracted from clusters that showed 10 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 5. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 63 (Nucleotide Sequences). Sequences extracted from clusters that showed 10 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 6. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 64 (Nucleotide Sequences). Sequences extracted from clusters that showed 15 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 2. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 65 (Nucleotide Sequences). Sequences extracted from clusters that showed 15 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 3. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 66 (Nucleotide Sequences). Sequences extracted from clusters that showed 15 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 4. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 67 (Nucleotide Sequences). Sequences extracted from clusters that showed 15 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 5. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 68 (Nucleotide Sequences). Sequences extracted from clusters that showed 15 differences or less between the longest sequences in cluster and other sequences in taxonomy at level 6. The taxonomic levels used to asses a discrepancy are as follows: level 1, nuclear or organelle; level 2, kingdom; level 3, phylum;…; level 7, genus; level 8, species. This is based on our own manually derived and homogeneous taxonomy for Eukaryota. There is no space in definition line.

Dataset Item 69 (Nucleotide Sequences). Sequence alignments that allowed to detect (compute) introns. Gene sequence is followed by intron sequence.

4. Concluding Remarks

Most of the introns contained in SSU-rRNA gene sequences have been described as group I introns, only 28 sequences being described as group II. Group I introns are a distinct class of RNA enzymes (ribozymes) characterized by a conserved RNA primary and secondary structures (http://www.rna.icmb.utexas.edu/) essential for splicing and are often capable of self-splicing. The sporadic and wide distribution of group I introns in the nuclear-encoded or organelle-encoded rRNA gene sequences of green and red algae, fungi, ciliates, and different amoebae suggests that these sequences have been highly successful in invading and maintaining in eukaryotic genomes [19, 28].

Phylogenetic distribution of group I introns is widespread but often sporadic, which strongly suggests that these sequences are mobile genetic elements capable of some although rare horizontal transfers between evolutionarily distinct lineages or alternatively that they are old elements that can be often deleted.

Despite extensive literature and documentation, it was not our purpose to enter a detailed analysis of the *new* introns we found in clades in which they had not been yet described. This is why we separated our dataset in three categories.

(i) *_GC.* These sequences can be used in the future to improve our algorithm, provided that introns described in these entries have been properly described. Without any mention in the exact method used to identify introns that have been described, we were not able to pursue this issue.

(ii) *_UC.* These sequences are the most likely sequences containing genuine introns. These sequences could

be added to any database of SSU-rRNA sequences when biodiversity analyses are done from amplicons derived from rRNA sequences rather than rRNA gene sequences.

(iii) *_RC*. These sequences were unexpected (at least in such amount) but could not be presently explored for the same reasons as the _GC sequences.

We present a large series of files that could be helpful for future manual analyses of introns in SSU-rRNA sequences. We hope our analyses will be helpful for several studies.

(i) They will help in the analyses of intronic sequences themselves, in terms of structure, insertion sites, and so forth.

(ii) They will help in the analyses of intronic sequences in terms of history, that is, deciphering in terms of phylogeny, the time at which a given type of intron invaded the SSU-rRNA gene from another location and was then either kept of lost (using the sets of clusters having a heterogeneous taxonomy).

(iii) Comparative biodiversity analyses: such analyses can be based either on amplicons derived from gene sequences (DNA) or on amplicons derived from rRNA sequences following a reverse-transcription step. Comparisons of both datasets are, for example, used to compare total diversity (DNA based) and diversity of most active organisms (rRNA based). Searching introns in DNA-based sequences should help such comparisons, because we expect sequences with long introns to be easily retrieved as rRNA but no so as rDNA, which would lead to fallacious estimates of which species are active or not.

(iv) They will be helpful also in the building of well-annotated reference sequence databases of rDNA and rRNA sequences to be used for biodiversity analyses of NGS datasets for which a manual approach is clearly not feasible.

The next step envisioned in our analysis will be to find out PCR primers which are most often used to obtain amplicons of SSU-rRNA gene sequences, build a database of such amplicons, and analyze domains often used for diversity analyses and for which long introns are present in some clades. Indeed with the new NGS approaches of biodiversity, a filter on lengths is de facto applied and absence of a given clade in a sample could only result from long introns or multiple introns in the gene-amplified domain.

Dataset Availability

The dataset associated with this Dataset Paper is dedicated to the public domain using the CC0 waiver and is available at http://dx.doi.org/10.7167/2013/854869.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

The authors acknowledge the support of the Aquaparadox project financed by the Agence National de la Recherche programme Biodiversité and the Pôle Mer PACA.

References

[1] R. Christen, "Global sequencing: a review of current molecular data and new methods available to assess microbial diversity," *Microbes and Environments*, vol. 23, no. 4, pp. 253–268, 2008.

[2] D. A. Caron, P. D. Countway, A. C. Jones, D. Y. Kim, and A. Schnetzer, "Marine protistan diversity," *Annual Review of Marine Science*, vol. 4, pp. 467–493, 2012.

[3] X. Y. Zhi, W. Zhao, W. J. Li, and G. P. Zhao, "Prokaryotic systematics in the genomics era," *Antonie Van Leeuwenhoek*, vol. 101, no. 1, pp. 21–34, 2012.

[4] L. B. Harris and S. O. Rogers, "Evolution of small putative group I introns in the SSU rRNA gene locus of *Phialophora* species," *BMC Research Notes*, vol. 4, article 258, 2011.

[5] V. Salman, R. Amann, D. A. Shub, and H. N. Schulz-Vogt, "Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria," *Proceedings of the National Academy of Sciences*, vol. 109, no. 11, pp. 4203–4208, 2012.

[6] K. Takizawa, T. Hashizume, and K. Kamei, "Occurrence and characteristics of group 1 introns found at three different positions within the 28S ribosomal RNA gene of the dematiaceous *Phialophora verrucosa*: phylogenetic and secondary structural implications," *BMC Microbiology*, vol. 11, article 94, 2011.

[7] J. Kjems and R. A. Garrett, "Novel splicing mechanism for the ribosomal RNA intron in the archaebacterium *Desulfurococcus mobilis*," *Cell*, vol. 54, no. 5, pp. 693–703, 1988.

[8] U. Kück, I. Godehardt, and U. Schmidt, "A self-splicing group II intron in the mitochondrial large subunit rRNA (LSUrRNA) gene of the eukaryotic alga *Scenedesmus obliquus*," *Nucleic Acids Research*, vol. 18, no. 9, pp. 2691–2697, 1990.

[9] T. Aimi, T. Yamada, and Y. Murooka, "Group I self-splicing introns in both large and small subunit rRNA genes of *Chlorella*," *Nucleic acids symposium series*, no. 29, pp. 159–160, 1993.

[10] Y. Liu and M. J. Leibowitz, "Variation and in vitro splicing of group I introns in rRNA genes of *Pneumocystis carinii*," *Nucleic Acids Research*, vol. 21, no. 10, pp. 2415–2421, 1993.

[11] S. Johansen and V. M. Vogt, "An intron in the nuclear ribosomal DNA of *Didymium iridis* codes for a group I ribozyme and a novel ribozyme that cooperate in self-splicing," *Cell*, vol. 76, no. 4, pp. 725–734, 1994.

[12] D. Bhattacharya, S. Damberger, B. Surek, and M. Melkonian, "Primary and secondary structure analyses of the rDNA group-I introns of the *Zygnematales* (Charophyta)," *Current Genetics*, vol. 29, no. 3, pp. 282–286, 1996.

[13] M. L. Shinohara, K. F. LoBuglio, and S. O. Rogers, "Group-I intron family in the nuclear ribosomal RNA small subunit genes of *Cenococcum geophilum* isolates," *Current Genetics*, vol. 29, no. 4, pp. 377–387, 1996.

[14] M. K. Tan, "Origin and inheritance of group I introns in 26S rRNA genes of *Gaeumannomyces graminis*," *Journal of Molecular Evolution*, vol. 44, no. 6, pp. 637–645, 1997.

[15] C. Einvik, M. Elde, and S. Johansen, "Group I twintrons: genetic elements in myxomycete and schizopyrenid amoeboflagellate ribosomal DNAs," *Journal of Biotechnology*, vol. 64, no. 1, pp. 63–74, 1998.

[16] M. Hagen and T. R. Cech, "Self-splicing of the *Tetrahymena* intron from mRNA in mammalian cells," *The EMBO Journal*, vol. 18, no. 22, pp. 6491–6500, 1999.

[17] D. Bhattacharya, F. Lutzoni, V. Reeb, D. Simon, J. Nason, and F. Fernandez, "Widespread occurrence of spliceosomal introns in the rDNA genes of ascomycetes," *Molecular Biology and Evolution*, vol. 17, no. 12, pp. 1971–1984, 2000.

[18] A. Mavridou, J. Cannone, and M. A. Typas, "Identification of group-I introns at three different positions within the 28S rDNA gene of the entomopathogenic fungus *Metarhizium anisopliaese* var. *anisopliae*," *Fungal Genetics and Biology*, vol. 31, no. 2, pp. 79–90, 2000.

[19] N. Nikoh and T. Fukatsu, "Evolutionary dynamics of multiple group I introns in nuclear ribosomal RNA genes of endoparasitic fungi of the genus *Cordyceps*," *Molecular Biology and Evolution*, vol. 18, no. 9, pp. 1631–1642, 2001.

[20] P. Haugen, D. H. Coucheron, S. B. Rønning, K. Haugli, and S. Johansen, "The molecular evolution and structural organization of self-splicing group I introns at position 516 in nuclear SSU rDNA of myxomycetes," *The Journal of Eukaryotic Microbiology*, vol. 50, no. 4, pp. 283–292, 2003.

[21] E. W. Lundblad, C. Einvik, S. Rønning, K. Haugli, and S. Johansen, "Twelve group I introns in the same pre-rRNA transcript of the myxomycete *Fuligo septica*: RNA processing and evolution," *Molecular Biology and Evolution*, vol. 21, no. 7, pp. 1283–1293, 2004.

[22] O. G. Wikmark, P. Haugen, E. W. Lundblad, K. Haugli, and S. D. Johansen, "The molecular evolution and structural organization of group I introns at position 1389 in nuclear small subunit rDNA of myxomycetes," *The Journal of Eukaryotic Microbiology*, vol. 54, no. 1, pp. 49–56, 2007.

[23] E. M. del Campo, L. M. Casano, F. Gasulla, and E. Barreno, "Presence of multiple group I introns closely related to bacteria and fungi in plastid 23S rRNAs of lichen-forming *Trebouxia*," *International Microbiology*, vol. 12, no. 1, pp. 59–67, 2009.

[24] M. Gouy and S. Delmotte, "Remote access to ACNUC nucleotide and protein sequence databases at PBIL," *Biochimie*, vol. 90, no. 4, pp. 555–562, 2008.

[25] E. Freyhult, P. P. Gardner, and V. Moulton, "A comparison of RNA folding measures," *BMC Bioinformatics*, vol. 6, article 241, 2005.

[26] C. Wang, L. I. Zengzhi, M. A. Typas, and T. M. Butt, "Nuclear large subunit rDNA group I intron distribution in a population of *Beauveria bassiana* strains: phylogenetic implications," *Mycological Research*, vol. 107, no. 10, pp. 1189–1200, 2003.

[27] C. J. Jackson, R. C. Barton, C. G. Clark, and S. L. Kelly, "Molecular characterization of a subgroup IE intron with wide distribution in the large subunit rRNA genes of dermatophyte fungi," *Medical Mycology*, vol. 47, no. 6, pp. 609–617, 2009.

[28] D. Bhattacharya, T. Friedl, and G. Helms, "Vertical evolution and intragenic spread of lichen-fungal group I introns," *Journal of Molecular Evolution*, vol. 55, no. 1, pp. 74–84, 2002.