MitoFish and MitoAnnotator: A Mitochondrial Genome Database of Fish with an Accurate and Automatic Annotation Pipeline

Wataru Iwasaki,*[1,2] Tsukasa Fukunaga,[1,2] Ryota Isagozawa,[2] Koichiro Yamada,[3] Yasunobu Maeda,[1] Takashi P. Satoh,[4] Tetsuya Sado,[5] Kohji Mabuchi,[1] Hirohiko Takeshima,[1] Masaki Miya,[5] and Mutsumi Nishida*[1]

[1]Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Chiba, Japan

[2]Department of Computational Biology, The University of Tokyo, Kashiwa, Chiba, Japan

[3]RNAi Incorporated, Bunkyo-ku, Tokyo, Japan

[4]Collection Center, National Museum of Nature and Science, Tsukuba, Ibaraki, Japan

[5]Department of Zoology, Natural History Museum and Institute, Chiba, Japan

***Corresponding author:** E-mail: iwasaki@aori.u-tokyo.ac.jp; mnishida@aori.u-tokyo.ac.jp.

Associate editor: Claudia Russo

Abstract

Mitofish is a database of fish mitochondrial genomes (mitogenomes) that includes powerful and precise de novo annotations for mitogenome sequences. Fish occupy an important position in the evolution of vertebrates and the ecology of the hydrosphere, and mitogenomic sequence data have served as a rich source of information for resolving fish phylogenies and identifying new fish species. The importance of a mitogenomic database continues to grow at a rapid pace as massive amounts of mitogenomic data are generated with the advent of new sequencing technologies. A severe bottleneck seems likely to occur with regard to mitogenome annotation because of the overwhelming pace of data accumulation and the intrinsic difficulties in annotating sequences with degenerating transfer RNA structures, divergent start/stop codons of the coding elements, and the overlapping of adjacent elements. To ease this data backlog, we developed an annotation pipeline named MitoAnnotator. MitoAnnotator automatically annotates a fish mitogenome with a high degree of accuracy in approximately 5 min; thus, it is readily applicable to data sets of dozens of sequences. MitoFish also contains re-annotations of previously sequenced fish mitogenomes, enabling researchers to refer to them when they find annotations that are likely to be erroneous or while conducting comparative mitogenomic analyses. For users who need more information on the taxonomy, habitats, phenotypes, or life cycles of fish, MitoFish provides links to related databases. MitoFish and MitoAnnotator are freely available at http://mitofish.aori.u-tokyo.ac.jp/ (last accessed August 28, 2013); all of the data can be batch downloaded, and the annotation pipeline can be used via a web interface.

Key words: **phylogenetics, database, genome annotator, fish, mitochondrion, high-throughput sequencing.**

Article

Introduction

Genetic information provides a foundation for the protection and management of biological diversity and enables researchers to decipher the evolutionary histories of diverse biological species. For metazoans, one of the most useful source of information is mitochondrial DNA (Gissi et al. 2008). As a notable example, the *barcode* sequence of the cytochrome c oxidase subunit I (*COX1*) gene in mitogenomes is a particularly useful tool for species identification as it has been exhaustively collected in the Barcode of Life project (Ratnasingham and Hebert 2007). The availability of large ranging primers, the species-level diversity, and the compactness of the *COX1* barcode sequences (~650 bp) represents considerable advantages for the effective identification of species. However, it is widely known that information obtained from a single gene is often insufficient for resolving branches of phylogenetic trees (Miya and Nishida 2000; Arnason et al. 2002; Pacheco et al. 2011).

The MitoFish database collects complete mitogenomic data of fish, i.e., vertebrates excluding tetrapods. Fish occupy an important position in the evolution of vertebrates and the ecology of the hydrosphere, which covers approximately 70% of the Earth's surface. Whole mitogenomic sequencing was first introduced into the phylogenetic study of fish at the end of the 20th century (Miya and Nishida 1999). Since then, the advantages of mitogenomic sequencing in evolutionary research has been demonstrated in many studies (e.g., Miya et al. 2003; Ramsden et al. 2003). As a publicly accessible, specialized sequence database of fish mitogenomes, MitoFish has received an average of more than 30,000 unique accesses annually since its launch in 2004. In recent years, it has become much easier to sequence whole mitogenomes at a reasonable cost and in unprecedented volumes due to the emergence of the high-throughput DNA sequencing technologies. Today, a benchtop-type sequencer is capable of sequencing dozens of mitogenomes

Table 1. Comparison Table of Mitogenomic Databases.

Database	Taxonomic Coverage	Sequence Data Type	Availability of Re-annotation Pipeline	Update Frequency/ Last Update[a]	Reference
GOBASE	Eukaryotes	Complete + partial	—	June 2010	O'Brien et al. (2009)
MamMiBase	Mammals	Protein coding genes only	—	June 2010	Vasconcelos et al. (2005)
METAMiGA	Metazoans	Complete	—	Daily	Feijao et al. (2006)
MitoZoa	Metazoans, excluding placozoans	Complete + nearly complete	Semiautomatic	December 2011	D'Onorio de Meo et al. (2012)
MitoFish	Fish (vertebrates, excluding tetrapods)	Complete + partial	Fully automatic	Monthly	

[a]The last update dates were checked on 25 June, 2013.

in a single run; thus, the volume of mitogenomic information is expected to grow rapidly in the near future.

Although metazoan mitogenomes have diverse structures (Boore 1999), vertebrate mitogenomes are typically circular, approximately 16 kb in length, and encode 13 protein-coding genes, 22 transfer RNA (tRNA), and 2 ribosomal RNA (rRNA) genes. These genes are variously oriented between the two strands of the mitogenomes; typically, one strand contains the NADH dehydrogenase subunit 6 (*ND6*) gene and 8 tRNA genes, and the other strand contains the remaining genes. The gene order is highly conserved, with changes in position typically only observed for tRNAs. Although there are exceptions, repeated genes and gene regions are rarely observed. The annotation of these mitogenomes is intrinsically difficult because, for example, sometimes the structures of mitochondrial tRNAs degenerate, protein-coding genes adopt divergent start/stop codons, and genetic elements overlap. The divergent start codons include ATG, GTG, TTG, ATA, ATT, CTG, TTA, ATC, and ACG, and the stop codons include TAA, TAG, AGA, AGG, TA-, AG-, and T–, where "-" denotes immature stop codons that require the post-transcriptional addition of A bases (Satoh 2006).

New Approaches

Here, we provide a novel report of the MitoFish database, including its recent major updates. The updates include a pipeline named MitoAnnotator, which automatically annotates fish mitogenomes rapidly and accurately. The annotation process likely represents a severe bottleneck for future mitogenomic efforts due to the production of an overwhelming amount of data. In addition, RefSeq (Pruitt et al. 2012), the most comprehensive database for mitogenomes, is known to contain many incorrect mitogenomic annotations (Bernt et al. 2013), which can lead to inaccurate research results. These errors result not only from human errors committed during manual annotation steps (e.g., the annotations of the strands of mitochondrial genes are sometimes reversed) but also from the aforementioned intrinsic difficulties of annotating mitogenomes. MitoAnnotator was developed to overcome these difficulties and can also be used for the re-annotation of previously sequenced fish mitogenomes. MitoFish also contains re-annotations of already sequenced fish mitogenomes that researchers can use as standardized references when they encounter annotations that are likely to be erroneous in public databases or when they conduct large-scale comparative mitogenomic studies. With the added functionality of MitoAnnotator, MitoFish serves as a regularly updated mitogenomic database equipped with a re-annotation function (table 1).

MitoFish and MitoAnnotator are freely available at http://mitofish.aori.u-tokyo.ac.jp/ (last accessed August 28, 2013); all of the data can be batch downloaded, and the MitoAnnotator pipeline can be used via a web interface.

Results and Discussion

Database Content—Overview

The principal content of MitoFish is fish mitogenomic sequence data. The database now contains more than 1,000 complete fish mitogenomic sequences and is regularly updated by incorporating RefSeq updates every month. In addition, MitoFish provides precise mitogenomic annotations, which can be readily adopted in a wide range of studies. In addition to mitogenomes, MitoFish contains partial mitogenomic sequence data, which are updated monthly by incorporating GenBank updates (Benson et al. 2013). MitoFish includes a total of 17,000 source fish species, which is more than half of the number of currently valid fish species in the world (Nelson 2006; Froese and Pauly 2013).

The sequence data are associated with taxonomic information (i.e., orders, families, genera, and species). In addition, information on the sample voucher and registration institution is provided wherever available. For users who need more information on taxonomy, fish habitats, phenotypes, or life cycles, MitoFish provides links to related databases such as FishBase, NCBI Taxonomy, Integrated Taxonomic Information System, and the Catalog of Fishes.

Typical Users and User Interface

MitoFish is accessed via a web browser. Figure 1 shows a screenshot of the home page. The vertical menu bar on the right side allows users to move to the four main functions of MitoFish: species/taxonomy search, sequence similarity search, batch data download, and fish mitogenome annotation. The first two searches can also be performed directly from the home page.

We assume four primary types of users that correspond to these four functions. The first type includes users who are

Fig. 1. MitoFish home page. A vertical menu bar on the right-hand side allows users to access the main functions of MitoFish. The fish species/taxonomy search and sequence similarity searches can also be performed directly from the home page.

interested in particular species or groups of fish and will search for that subset of mitogenomic data. Such users can easily access pages regarding the species/taxa of interest via the species/taxonomy search function. On the mitogenome page of each species (fig. 2), a picture of the fish and a visual representation of its annotated circular mitogenome aid visual recognition, and users are able to download the mitogenomic sequence of the species along with annotation data. In addition, taxonomic information and links to external databases are summarized on the same page to aid further analysis. Users can apply the downloaded mitogenomes to their analysis directly or can, for example, construct polymerase chain reaction (PCR) primers using the complete mitogenomic data to sequence their own samples for further molecular evolutionary analysis.

The second user type includes researchers who possess fish mitochondrial DNA sequence data and want to identify the species or infer the evolutionary background. For such users, MitoFish provides a sequence similarity search function. When a user inputs a nucleotide sequence, MitoFish runs a

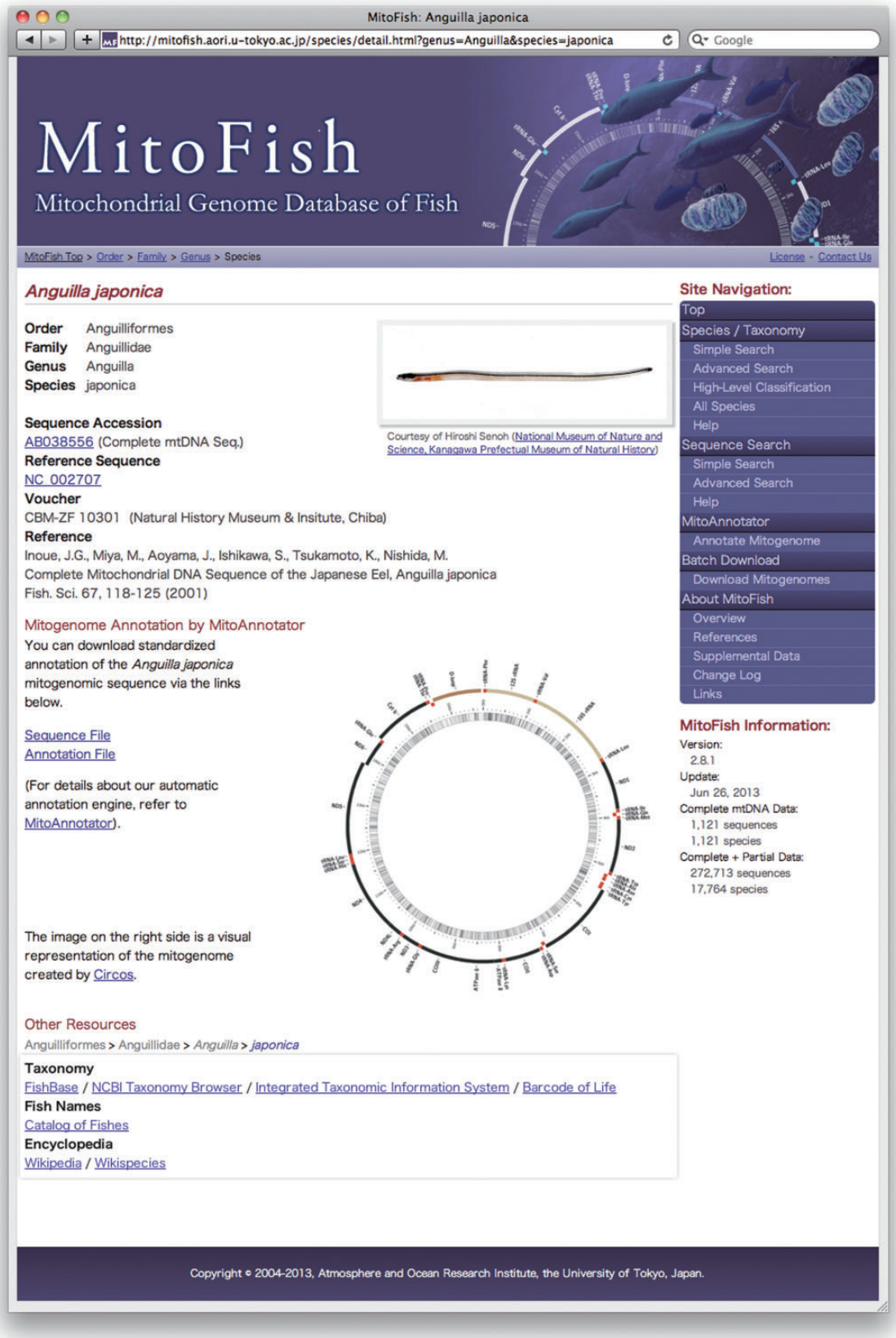

Fig. 2. Mitogenome page of individual species. The mitogenome page of each species includes a picture of the fish and a visual representation of the annotated circular mitogenome to aid visual recognition. Users can download mitogenomic sequences and the associated annotation data from the links. Information on sample vouchers and registration institutions is also provided. To facilitate further analysis, taxonomic information and links to external databases are comprehensively summarized.

BlastN search (Camacho et al. 2009) against the mitogenomic or mitogenomic + partial mitochondrial sequence database and outputs sequences similar to the sequence provided by the user. Links to each mitogenomic data page are included in the search result page to allow users to easily download sequences similar to the input sequence for further analysis.

The third user type includes those who are interested in comparative mitogenomics. MitoFish provides precise and standardized annotations that are batch downloadable. For example, users can convert the annotations into concatenated gene sequences or synteny structure data to conduct large-scale analyses of mitogenomic evolution.

Finally, the fourth user type refers to users who have sequenced fish mitogenomes and wish to annotate their sequences as easily as possible. MitoAnnotator includes functions intended for this user type. These functions are described in the following sections.

MitoAnnotator: Mitochondrial Genome Annotation Pipeline

MitoAnnotator is a pipeline for automatically annotating fish mitogenomic sequences with a high degree of accuracy. The high-quality, in-house manual annotation of 250 fish mitogenomes (Satoh 2006) was incorporated into the development of the pipeline and enhanced the performance of MitoAnnotator. When sequences are provided by the user in the conventional FASTA format, MitoAnnotator provides a full mitogenomic annotation without any user input in approximately 5 min. This rapid response is highly desirable in an annotation pipeline in the current era of high-throughput sequencing.

As described above, a vertebrate mitogenome typically contains 13 protein-coding genes; 22 tRNA genes (two tRNAs each for serine and leucine and one tRNA for each of the other 18 amino acids); 2 rRNA genes; and 1 control region or "d-loop", which is a non-coding region for replication and transcription control (Boore 1999). MitoAnnotator automatically finds these 38 elements and outputs their coordinates and strands as described below (see fig. 3 for an overview).

A vertebrate mitogenome is a circular molecule that can be represented arbitrarily in linear representations (e.g., in FASTA format). In addition to the two complementary sequences of a given circular DNA molecule, the start position can be chosen arbitrarily. We followed a convention that places a $tRNA^{Phe}$ gene at the first position, as $tRNA^{Phe}$ genes are typically located immediately after the control region in vertebrate mitogenomic sequences (Boore 1999). Accordingly, MitoAnnotator first locates the $tRNA^{Phe}$ gene within the input sequence and adjusts the coordinates to place the $tRNA^{Phe}$ gene in the first position. The $tRNA^{Phe}$ gene is detected using MiTFi (Juhling et al. 2012) with an e-value threshold of 1e−5. If the $tRNA^{Phe}$ gene is not found (e.g., when a partial mitogenomic sequence is provided), the coordinate adjustment is not conducted, and the original sequence is directly fed into the subsequent steps.

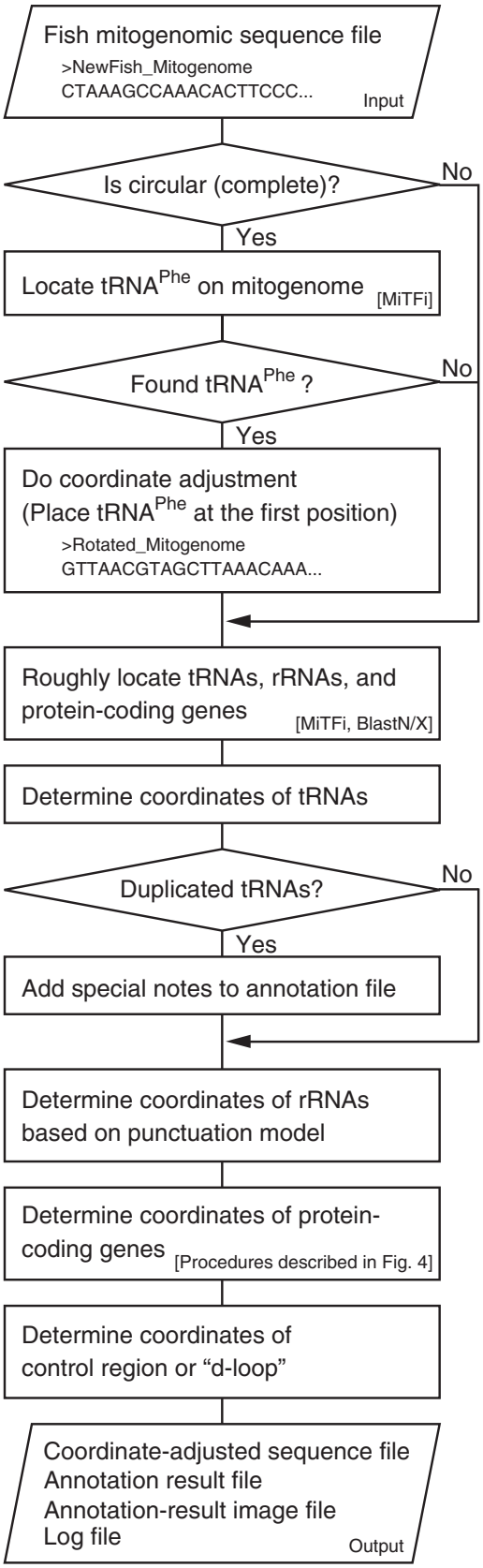

FIG. 3. Overview of the MitoAnnotator pipeline. Please refer to the main text and figure 4 for the details of each procedure.

MiTFi is a tool for accurately locating tRNA genes within mitogenomic sequences. Other than MiTFi, tRNAscan-SE (Lowe and Eddy 1997) is the most commonly used tool for locating tRNA genes in prokaryotic and eukaryotic genomic DNA sequences. However, tRNAs encoded in mitogenomes sometimes have exceptional structures and cannot be discovered using general methods. For example, mitochondrial tRNAs can have incomplete cloverleaf structures lacking otherwise highly conserved loops (Anderson et al. 1981) or arms (Arcari and Brownlee 1980; de Bruijn et al. 1980). Consequently, tRNAscan-SE finds mitochondrial tRNA genes with high specificity but with low sensitivity (Juhling et al. 2012). To identify tRNA genes in mitogenomic sequences with greater sensitivity, a second tool, named ARWEN (Laslett and Canback 2008), employs a heuristic algorithm that first searches for hairpin structures to avoid overlooking degenerate structures. However, in compensation, the heuristics of ARWEN result in a substantial false discovery rate. MiTFi addresses this need through covariance models (a special case of stochastic context-free grammars designed for modeling RNA consensus sequence and structure) developed from known vertebrate mitochondrial tRNAs and uses Infernal (Nawrocki et al. 2009) as its search engine.

MitoAnnotator then determines the precise coordinates of the 22 tRNA genes, 2 rRNA genes, 13 protein-coding genes, and the control region. First, MitoAnnotator searches for their approximate positions. Here, MiTFi is applied again, but this time to locate the remaining 21 tRNAs. Because exceptional mitogenomes containing multiple tRNAs have been reported in vertebrates, we allowed MiTFi to accept multiple tRNAs with e-values below 1e−5. Next, BlastX and BlastN (Camacho et al. 2009) are applied to identify protein-coding genes and rRNA genes, respectively, against a fish mitochondrial gene database created from 250 fish mitogenomes (Satoh 2006). To our knowledge, the duplication of protein-coding genes and rRNA genes has not been reported in fish mitogenomes, whereas some avian families are reported to have tandem duplications of their mitogenomic regions that include protein-coding genes (Sammler et al. 2011). Thus, we chose to permit only hits with the lowest e-values for these two gene categories.

Next, to obtain their precise coordinates, the following steps are conducted. For tRNAs, the 3′ termini of some tRNA annotations in RefSeq lack a base after the accepter stems (and before the CCA tails). MitoAnnotator consistently includes the + 1 base to standardize the annotations. In case multiple tRNA genes are identified, each of them is included in the output annotations, and special notes are added for the tRNA genes with the lowest e-values. In general, duplicated tRNA genes become redundant and quickly degenerate during evolution because of their weak functional constraints. In mitochondrial genomes, however, duplicated tRNA genes may be maintained as punctuation markers that keep their flanking elements intact through splicing (Mabuchi et al. 2004). Thus, the annotation of degenerated tRNA genes is very important because they may retain biological functions and play an important role in deciphering the evolution of mitogenomic structures.

Regarding rRNAs, we adopted the tRNA punctuation model as a principal criterion (Ojala et al. 1980). According to this model, the processing of flanking tRNAs directly produces both the 5′ and 3′ ends of rRNAs (and mRNAs) or results in the location of rRNA genes and tRNA genes on mitogenomes without gaps between them. However, our analysis of the large fish mitogenome dataset suggested that this model would not be applicable to some rRNAs. Therefore, in the event that the tRNA punctuation model resulted in exceptionally long 12S rRNA genes (more than 1,000 bp) or 16S rRNA genes (more than 1,850 bp), we employed the BlastN search results directly to annotate the rRNA genes. In such cases, it is intrinsically difficult to determine rRNA gene structures from genomic sequences alone; transcribed rRNAs must be directly sequenced.

Protein-coding genes require more complex rules (fig. 4). First, MitoAnnotator continually extends the end positions of the BlastX-hit regions by three bases until a vertebrate mitochondrial stop codon is found (i.e., TAA, TAG, AGA, or AGG [Osawa et al. 1989]). When the search does not find a stop codon before the succeeding element, we allowed the ATP synthase Fo subunit 8 (*ATP8*), *COX1*, NADH dehydrogenase subunit 4L (*ND4L*), subunit 5 (*ND5*), and *ND6* genes to overlap with the succeeding element. For the remaining eight protein-coding genes, MitoAnnotator allows the genes to overlap with the succeeding element only if their coding strands are different. If the coding strands are the same or if the above procedures also cannot identify a stop codon in the succeeding element, then MitoAnnotator sets the −1 position of the start position of the succeeding element as the end position. Next, the start positions are determined (fig. 4*B*). Because mitochondrial genomes are typically very efficient and have little space between the coding elements (Ojala et al. 1980), MitoAnnotator chooses the mitochondrial start codon (i.e., ATG or GTG [Desjardins and Morais 1991]) that is farthest from the stop codon and that does not overlap with the preceding element. In this case, the following exceptions were introduced by again referring to our in-house dataset. First, for the NADH dehydrogenase subunit 1 (*ND1*), subunit 2 (*ND2*), subunit 3 (*ND3*), and *ND5* genes, one-base overlaps with the preceding element are allowed. Second, for the ATP synthase Fo subunit 6 (*ATP6*) and *ND4* genes, overlaps of up to 20 bases are allowed. Third, if the above criteria fail to identify a start codon, a search for special start codons is permitted for the following four genes: CTG for *ATP6*, TTG for *ND1*, ATA and ATT for *ND3*, and TTG for cytochrome c oxidase subunit II (*COX2*). Fourth, if the start codons found are contiguous (e.g., ATGGTGATG), the last ATG codon is accepted as the start site. In the case that the contiguous start codons do not contain ATG, the last codon is chosen. Fifth, if no start codons are identified using the above criteria, MitoAnnotator searches for the farthest TTG, ATA, ATT, CTG, TTA, ATC, or ACG within + 30 bp from the end position of the preceding element. Sixth, if this process also fails to identify a start codon, MitoAnnotator sets the first in-frame position after the preceding element as the start position (i.e., + 1, + 2, or + 3 from the end position of the preceding

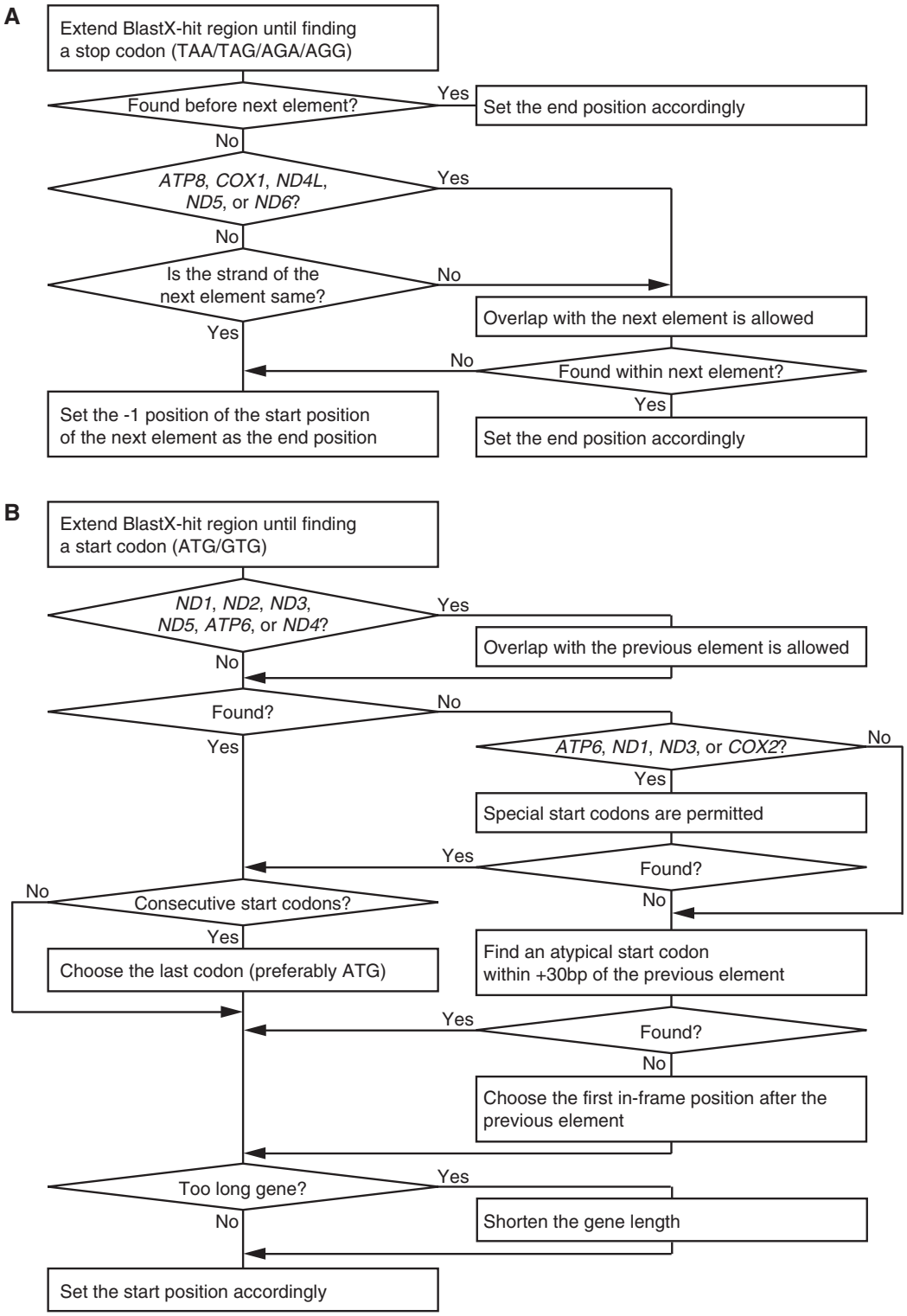

Fig. 4. Workflows to determine the coordinates of protein-coding genes. Workflows to determine the end position (*A*) and the start position (*B*) of protein-coding genes are presented.

element, depending on the reading frame). Seventh, if the identified genes are unusually long (over 220 bp for *ATP8*, 750 bp for *COX2*, 800 bp for *COX3*, 1000 bp for *ND1*, 355 bp for *ND3*, and 535 bp for *ND6*), the furthest start codon within the coding frame (within 170–220 bp for *ATP8*, 650–750 bp for *COX2*, 700–800 bp for *COX3*, 900–1,000 bp for *ND1*, 305–355 bp for *ND3*, and 435–535 bp for *ND6*) is chosen as the start site. Eighth, if the seventh rule cannot find a start codon, the last start codon that makes the length of the gene closest to the threshold lengths is chosen.

Finally, the control region is annotated if the above procedures provide any interval region longer than 600 bp.

Performance of MitoAnnotator

The performance of the MitoAnnotator pipeline was thoroughly examined using additional mitogenomic data from 42 newly sequenced fish (Miya et al. 2013; Sado T and Miya M, unpublished data; the list of species names and accession numbers is provided in the Material and Methods section) that were not included in the 250-fish mitogenomic dataset. The annotations were conducted very efficiently, taking approximately 5 min per mitogenome, and two expert curators (T.P.S. and T.S.) examined the results manually. T.P.S. had sequenced more than 200 fish mitogenomes in 14 years and T.S. had sequenced more than 400 in 8 years. MitoAnnotator correctly annotated all of the 42 mitogenomes, not only identifying the existence of the 38 mitogenomic elements but also precisely locating their start and stop positions at the single-nucleotide level.

Some fish mitogenomes have unusual structures (Inoue et al. 2003b). To evaluate whether MitoAnnotator can also correctly annotate exceptional cases, we downloaded and evaluated 10 mitogenomes known to have exceptional structures. Their species names, RefSeq accession numbers, and the characteristic parts of their mitogenomic structures are as follows: *Aspasma minima* (NC_008130, T-CR-I-Q-F-P-12S); *Aulostomus chinensis* (NC_010269, T-P-CR-ND1-M-ND2-W-Q-F-12S); *Ceratias uranoscopus* (NC_013882, 16S-ND1-Q-CR-L-I-M-M-ND2); *Chauliodus sloani* (NC_003159, W-A-N-Y-C-C-C-C-C-COI); *Conger myriaster* (NC_002761, ND5-CYB-T-CR-ND6-E-P); *Sigmops gracilis* (NC_002574, ND6-CYB-E-P-T-CR); *Coelorinchus kishinouyei* (NC_003169, ND6-CYB-T-P-E-CR); *Cryptopsaras couesii* (NC_013880, 16S-NC-I-ND1-Q-L-M-ND2); *Diaphus splendidus* (NC_003164, ND1-I-M-Q-ND2); and *Eurypharynx pelecanoides* (NC_005299, highly rearranged structure [Inoue et al. 2003b]). The experts confirmed that the annotations were correct. For example, a tandem repeat of five $tRNA^{Cys}$ genes on the *C. sloani* mitogenome was correctly annotated.

Some tools already exist to facilitate the annotation of metazoan mitogenomes. DOGMA (Wyman et al. 2004) is a pioneering tool in this field that helps researchers annotate mitochondrial and chloroplast genomes in a semiautomatic manner. Because DOGMA uses rather simple approaches to identify coding and noncoding genes, it requires users to manually check the results. MITOS (Bernt et al. 2013) is an automated pipeline for the de novo annotation of metazoan mitogenomes and comes closest to what MitoAnnotator achieves. The biggest difference between MITOS and MitoAnnotator lies in their running times. MITOS requires more than 1 h to annotate one mitogenome, whereas MitoAnnotator requires 5 min. Second, annotations by MITOS have been found to have many inconsistencies with annotations performed by experts, most frequently for the stop codons of protein-coding genes. We suppose that this is most likely because MITOS requires lengths of protein-coding genes too strictly to be multiples of three. For example, in the annotation of the exceptional *C. sloani* mitogenome, MITOS employed CTT as a stop codon for the cytochrome b (*CytB*) gene, which is unlikely. Other types of inconsistencies observed using MITOS are summarized in table 2. Third, only MitoAnnotator offers a coordinate adjustment function, which is an important feature for the end user. In MITOS, if a mitogenomic genetic element overlaps the boundary between the head and tail of a linear representation of a mitogenomic sequence, it is not annotated. In contrast, users can create a standardized annotation using the coordinate adjustment function of MitoAnnotator immediately after obtaining an assembled mitogenomic sequence. Last but not least, we envision that the framework of MitoAnnotator can also be applied to other groups of vertebrates once sufficient amounts of high-quality manual annotation data are obtained and the pipeline is appropriately modified.

Table 2. Numbers of Genomes Whose Automatic Annotations Were Inconsistent with Annotations Performed by Experts for 42 Mitogenomes.

Category of inconsistent annotations[a]	Mito Annotator	MITOS
Annotation of additional genes	0	3[b]
Different start positions of protein-coding genes	0	42
Different stop positions of protein-coding genes	0	42
Different start positions of tRNA genes	0	0
Different stop positions of tRNA genes	0	3

[a]We excluded start/stop positions of rRNA genes from this comparison table because the annotation of rRNA genes is intrinsically difficult as described in the text.

[b]Each of the three additional genes predicted by MITOS was a second protein-coding gene copy located in the d-loop of each mitogenome. These genes were very short (the 105-bp *ATP8* gene of *Nesiarchus nasutus*, the 324-bp *ND6* gene of *Kali indica*, and the 438-bp *ND2* gene of *Diplospinus multistriatus*) and are likely to be misannotations.

In conclusion, MitoAnnotator is a fully automatic pipeline that efficiently annotates fish mitogenomes with high accuracy. Annotation results obtained from MitoAnnotator can be directly fed into public sequence repository services, thereby greatly reducing the efforts of researchers in annotating newly sequenced mitogenomes. In combination with MitoFish, we believe that MitoAnnotator will accelerate studies on fish evolution as data collection continues to become easier and less expensive.

Material and Methods

MitoFish Server

The server runs on a Linux operating system, and an Apache HTTP Server provides the web services. A MySQL database system stores information on each fish species. Perl and Ruby scripts process all of these data and the requests from users. All of these resources have been extensively used and are well supported. We have taken care to make MitoFish easily accessible via search engines; thus, search queries such as *fish mitogenome*, *fish mitochondrial genome*, and *fish mitochondria database* on google.com return MitoFish as the top hit as of June 2013.

Database Update

RefSeq and GenBank entries are downloaded every month to update MitoFish. For RefSeq, mitogenomic entries are batch

downloaded from the FTP URL ftp://ftp.ncbi.nih.gov/refseq/release/mitochondrion/ (last accessed August 28, 2013). All GenBank entries are downloaded, and those having the feature *organelle = mitochondrion* are selected. For both databases, sequence entries whose NCBI taxonomy classification entries are under Myxiniformes, Petromyzontiformes, Chondrichthyes, Actinopterygii, Coelacanthiformes, or Dipnoi are selected and incorporated into the mitogenomic and BLAST databases.

Mitochondrial Genomes

Newly sequenced mitogenomes from 42 diverse fish species were used in evaluating the performance of the MitoAnnotator pipeline. The 42 species are as follows (the International Nucleotide Sequence Database Collaboration accession numbers are provided): *Acanthocybium solandri (AP012945)*, *Anchoviella* sp. (AP012524), *Aphanopus carbo* (AP012944), *Ariomma indica* (AP012513), *Ariomma lurida* (AP012512), *Assurger anzac* (AP012508), *Benthodesmus tenuis* (AP012522), *Dionda episcopa* (AP012077), *Diplospinus multistriatus* (AP012523), *Epinnula magistralis* (AP012943), *Eumegistus illustris* (AP012497), *Euthynnus affinis* (AP012946), *Evoxymetopon poeyi* (AP012509), *Gempylus serpens* (AP012502), *Gymnosarda unicolor* (AP012510), *Hemitremia flammea (AP012078)*, *Icichthys lockingtoni* (AP012511), *Kali indica* (AP012500), *Luciocyprinus striolatus (AP012525)*, *Luxilus chrysocephalus (AP012079)*, *Macrhybopsis gelida (AP012080)*, *Margariscus margarita* (AP012081), *Microphysogobio yaluensis* (AP012073), *Nesiarchus nasutus* (AP012503), *Nocomis biguttatus* (AP012082), *Notropis atherinoides* (AP012083), *Notropis baileyi* (AP012084), *Opsopoeodus emiliae* (AP012085), *Pampus punctatissimus* (AP012516), *Peprilus burti* (AP012947), *Promethichthys prometheus* (AP012504), *Pteraclis aesticola* (AP012499), *Rastrelliger kanagurta* (AP012948), *Ruvettus pretiosus* (AP012506), *Sarda orientalis* (AP012949), *Scombrolabrax heterolepis* (AP012517), *Sphyraena japonica* (AP012501), *Tanakia tanago* (AP012526), *Taractes asper* (AP012498), *Tetragonurus atlanticus* (AP012515), *Tetragonurus cuvieri* (AP012514), and *Thyrsitoides marleyi* (AP012505). The extracted mitogenomes were amplified via the long PCR technique (Miya and Nishida 1999; Inoue et al. 2003a) and sequenced with the Sanger sequencing technique.

Acknowledgments

The authors thank Keiichi Matsuura for helpful discussion; Jun G. Inoue, Satoko Koide, and Tomoyuki Yamada for technical support; and the editor and four anonymous reviewers for their valuable comments. This work was supported by the Japan Society for the Promotion of Science (Grant Numbers 13556028, 19207007, 23370041, 23710231, 178087, 248046, and 258048) and the Japan Science and Technology Agency (CREST).

References

Anderson S, Bankier AT, Barrell BG, et al. (14 co-authors). 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290:457–465.

Arcari P, Brownlee GG. 1980. The nucleotide sequence of a small (3S) seryl-tRNA (anticodon GCU) from beef heart mitochondria. *Nucleic Acids Res.* 8:5207–5212.

Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu X, Janke A. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc Natl Acad Sci U S A.* 99:8151–8156.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Res.* 41: D36–D42.

Bernt M, Donath A, Juhling F, Externbrink F, Florentz C, Fritzsch G, Putz J, Middendorf M, Stadler PF. Forthcoming 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol.* 69:313–319.

Boore JL. 1999. Animal mitochondrial genomes. *Nucleic Acids Res.* 27: 1767–1780.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

D'Onorio de Meo P, D'Antonio M, Griggio F, Lupi R, Borsani M, Pavesi G, Castrignano T, Pesole G, Gissi C. 2012. MitoZoa 2.0: a database resource and search tools for comparative and evolutionary analyses of mitochondrial genomes in Metazoa. *Nucleic Acids Res.* 40: D1168–D1172.

de Bruijn MH, Schreier PH, Eperon IC, Barrell BG, Chen EY, Armstrong PW, Wong JF, Roe BA. 1980. A mammalian mitochondrial serine transfer RNA lacking the "dihydrouridine" loop and stem. *Nucleic Acids Res.* 8:5213–5222.

Desjardins P, Morais R. 1991. Nucleotide sequence and evolution of coding and noncoding regions of a quail mitochondrial genome. *J Mol Evol.* 32:153–161.

Feijao PC, Neiva LS, de Azeredo-Espin AM, Lessinger AC. 2006. AMiGA: the arthropodan mitochondrial genomes accessible database. *Bioinformatics* 22:902–903.

Froese R, Pauly D, editors. 2013. FishBase. World Wide Web electronic publication. [cited 2013 August 28] www.fishbase.org, version (04/2013).

Gissi C, Iannelli F, Pesole G. 2008. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity* 101:301–320.

Inoue JG, Miya M, Tsukamoto K, Nishida M. 2003a. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the "ancient fish". *Mol Phylogenet Evol.* 26:110–120.

Inoue JG, Miya M, Tsukamoto K, Nishida M. 2003b. Evolution of the deep-sea gulper eel mitochondrial genomes: large-scale gene rearrangements originated within the eels. *Mol Biol Evol.* 20:1917–1924.

Juhling F, Putz J, Bernt M, Donath A, Middendorf M, Florentz C, Stadler PF. 2012. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids Res.* 40:2833–2845.

Laslett D, Canback B. 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics* 24: 172–175.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955–964.

Mabuchi K, Miya M, Satoh TP, Westneat MW, Nishida M. 2004. Gene rearrangements and evolution of tRNA pseudogenes in the mitochondrial genome of the parrotfish (Teleostei: Perciformes: Scaridae). *J Mol Evol.* 59:287–297.

Miya M, Nishida M. 1999. Organization of the mitochondrial genome of a deep-sea fish, *Gonostoma gracile* (Teleostei: Stomiiformes): first example of transfer RNA gene rearrangements in bony fishes. *Marine Biotechnol.* 1:416–426.

Miya M, Nishida M. 2000. Use of mitogenomic information in teleostean molecular phylogenetics: a tree-based exploration under the maximum-parsimony optimality criterion. *Mol Phylogenet Evol.* 17: 437–455.

Miya M, Takeshima H, Endo H, et al. (12 co-authros). 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. *Mol Phylogenet Evol.* 26:121–138.

Miya M, Friedman M, Satoh TP, et al. (14 co-authros). Forthcoming 2013. Volutionary origin of the scombridae (Tunas and Mackerels): members of a paleogene adaptive radiation with 14 other pelagic fish familiess. *PLOS ONE*.

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25:1335–1337.

Nelson JS. 2006. Fishes of the world. Hoboken (NJ): John Wiley.

O'Brien EA, Zhang Y, Wang E, Marie V, Badejoko W, Lang BF, Burger G. 2009. GOBASE: an organelle genome database. *Nucleic Acids Res.* 37: D946–D950.

Ojala D, Merkel C, Gelfand R, Attardi G. 1980. The tRNA genes punctuate the reading of genetic information in human mitochondrial DNA. *Cell* 22:393–403.

Osawa S, Ohama T, Jukes TH, Watanabe K. 1989. Evolution of the mitochondrial genetic code. I. Origin of AGR serine and stop codons in metazoan mitochondria. *J Mol Evol.* 29:202–207.

Pacheco MA, Battistuzzi FU, Lentino M, Aguilar RF, Kumar S, Escalante AA. 2011. Evolution of modern birds revealed by mitogenomics: timing the radiation and origin of major orders. *Mol Biol Evol.* 28: 1927–1942.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 40:D130–D135.

Ramsden SD, Brinkmann H, Hawryshyn CW, Taylor JS. 2003. Mitogenomics and the sister of Salmonidae. *Trends Ecol Evol.* 18: 607–610.

Ratnasingham S, Hebert PD. 2007. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes.* 7:355–364.

Sammler S, Bleidorn C, Tiedemann R. 2011. Full mitochondrial genome sequences of two endemic Philippine hornbill species (Aves: Bucerotidae) provide evidence for pervasive mitochondrial DNA recombination. *BMC Genomics* 12:35.

Satoh TP. 2006. Comparative study on the structural features of fish mitochondrial genomes [doctoral thesis]. Tokyo (Japan): The University of Tokyo.

Vasconcelos AT, Guimaraes AC, Castelletti CH, et al. (23 co-authors). 2005. MamMiBase: a mitochondrial genome database for mammalian phylogenetic studies. *Bioinformatics* 21:2566–2567.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20:3252–3255.