MITOSCISSOR: A Useful Tool for Auto-Assembly of Mitogenomic Datasets in the Evolutionary Analysis of Fishes

Libertas Academica
FREEDOM TO RESEARCH

Zheng Sun[1,2], Yuanzhi Cheng[2] and Junbin Zhang[1]

[1]College of Fisheries and Life Science, Shanghai Ocean University, Shanghai, P.R. China. [2]Institute of Oceanology, Chinese Academy of Sciences, Qingdao, P.R. China.

ABSTRACT: As a result of the development of rapid and efficient sequencing technologies, complete sequences of numerous mitochondrial genomes are now available. Mitochondrial genomes have been widely used to evaluate relationships between species in several fields, including evolutionary and population genetics, as well as in forensic identification and in the study of mitochondrial diseases in humans. However, the creation of mitochondrial genomes is extremely time consuming. In this paper, we present a new tool, MITOSCISSOR, which is a rapid method for parsing and formatting dozens of complete mitochondrial genome sequences. With the aid of MITOSCISSOR, complete mitochondrial genome sequences of 103 species from *Tetraodontiformes* (a difficult-to-classify order of fish) were easily parsed and formatted. It typically takes several days to produce similar results when relying upon manual editing. This tool could open the .gb file of Genbank directly and help us to use existing mitogenomic data. In the present study, we established the first clear and robust molecular phylogeny of 103 tetraodontiform fishes, a goal that has long eluded ichthyologists. MITOSCISSOR greatly increases the efficiency with which DNA data files can be parsed and annotated, and thus has the potential to greatly facilitate evolutionary analysis using mitogenomic data. This software is freely available for noncommercial users at http://www.filedropper.com/mitoscissor.

KEYWORDS: mitochondrial genome, evolution, Perl script, phylogeny

CITATION: Sun et al. MITOSCISSOR: A Useful Tool for Auto-Assembly of Mitogenomic Datasets in the Evolutionary Analysis of Fishes. *Evolutionary Bioinformatics* 2015:11 115–120 doi: 10.4137/EBO.S22340.

RECEIVED: December 02, 2014. **RESUBMITTED:** February 10, 2015. **ACCEPTED FOR PUBLICATION:** February 16, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

TYPE: Methodology

FUNDING: This work was supported by the National Key Technology R&D Program 2011BAD13B01 from the Ministry of Science and Technology; Chinese National Funds 40776089 and 31272662; Shanghai Universities First-class Disciplines Project of Fisheries; and "Dongfang" Projects from Shanghai Municipal Education Commission and the Shanghai Education Development Foundation. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: jb-zhang@shou.edu.cn

Introduction

Mitochondria are intracellular organelles that have their own DNA, distinct from the nuclear genomes of the cells in which they reside. Throughout evolution, mitochondrial genomes (mitogenomes) have accumulated many small genetic changes that differ between species, and which allow biologists to infer the phylogenetic relationships between species, and even between subpopulations within a species.

The analysis of mitochondrial DNA (mtDNA) has contributed enormously to our understanding of evolution. The rapid rate of mtDNA sequence divergence compared to that of the nuclear genome (due to a relatively high nucleotide substitution rate and inefficient DNA repair in the mitochondria) allows the discrimination of recently diverged lineages.[1] Mitochondrial genes that are frequently used for these analyses include cytochrome b (*CYTB*), the *12S* and *16S* rRNA genes, and the NADH dehydrogenase (*ND*) genes.[2–4] However, in some cases, the use of a single gene or several genes can result in phylogenetic trees with low support or contradictory topologies. In addition, organisms can show different types of variations in these frequently used markers, making it difficult to draw reliable conclusions.[5] Because of these factors, the use of complete mitogenome sequences for phylogenetic analysis has increased in the past decade, and such analyses have been proven to be more informative than shorter sequences and to provide better phylogenetic resolution.[3,5]

In addition to the insights that mtDNA can provide on the differences between closely related organisms, the sequencing of mitogenomes is much more cost effective than the analysis of entire nuclear genomes. Thus, the popularity of the complete mitogenome as a phylogenomic marker is increasing.[6–9]

Comparative analysis of mitochondrial genomes that have been sequenced and annotated thus far has resolved many controversial phylogenetic issues, including inconsistencies regarding both lower level and higher level relationships.[9] However, creation of these DNA datasets is time consuming when these mitogenomic sequences must be manually parsed. For example, the conventional method for manually constructing such mitogenome matrices involves the following steps: (1) Fasta sequences of all the selected mitochondrial genomes are downloaded from GenBank or other public databases. (2) The sequence of every gene in each mitochondrial genome is determined and then saved as an individual fasta file. A large number of files are generated during this process, and the compilation is cumbersome. For example, the mitogenomic analysis for 100 vertebrate species will generate 3,700 files, since there is on average 37 genes in one vertebrate mitochondrial genome. (3) All of the sequences are then arranged in the same orientation using sequence-editing software such as BioEdit or DNASTAR, and the stop codons are removed from protein-encoding genes. (4) Finally, the protein-coding genes, rRNA genes, and tRNA genes are concatenated head to tail to form a single supergene for the evolutionary analysis. Obviously, parsing mitogenomic sequences is very time consuming, and the workload to analyze so many genes is enormous without the use of processing tools.

Here we present a useful Perl script named MITOSCISSOR, which automatically executes the parsing work for mitogenomic data. As an example, we applied MITOSCISSOR for creating mitogenomic data sets for 103 Tetraodontiform fishes. Phylogenetic relationships within this order were also inferred, providing new insights into the relationships within this order. It should be noted here that MITOSCISSOR is limited to datasets containing mitogenomes with the same overall gene order. This problem will be solved with the growth of data. The development of MITOSCISSOR will increase the ease of investigating positive selection in marine fish (and other organisms), by its ability to quickly parse out genic regions of interest.

Materials and Methods

Taxonomic sampling. Mitogenome sequences of 103 Tetraodontiform fishes and 9 fish species from other orders were employed in the present study. The corresponding GenBank accession numbers (Supplementary File 1) were searched in the MitoZoa database (http://srv00.ibbe.cnr.it/mitozoa/), and the complete L-strand nucleotide sequences from these mitogenomes (.gb file) were downloaded using the NCBI batchentrez tool (http://www.ncbi.nlm.nih.gov/sites/batchentrez).

MITOSCISSOR implementation. MITOSCISSOR was written in the Perl language. Activeperl (http://www.perl.org/get.html) should be installed before using MITOSCISSOR in Windows OS.

The MITOSCISSOR software and all the GenBank files should be saved together in a single folder. Once these steps are performed, the command line interpreter cmd.exe would be started and automatically linked to the working directory. Users then doubleclick “MitoScissor.exe”, and the window interface would be presented, as shown in Figure 1. Processing could be easily done according to the instructions.

In the Start Menu (operation system such as Windows OS) search box, users could type in “command prompt” or “cmd”, then carry out the following operations >dir (“dir” command shows there are three files in the working directory). Subsequently, the software could be executed by typing a simple command line, “perl MitoScissor.pl sequence.gb outputfile”. Note that “printall” is an optional parameter used in cases where the Genbank file does not meet recognition models and the operator wants to output all sequences that can be identified.

In this procedure, the user should choose an appropriate module model. The organization of the mitochondrial genome is generally conserved among vertebrates, but some variations between higher taxonomic levels have been found.[10,11] Therefore, we designed different recognition models with regard to specific taxa. Here, we chose the module of Fish.

In the current version of MITOSCISSOR, a Windows OS interface (shown in Fig. 1) was designed to provide greater operational convenience for users.

With the help of MITOSCISSOR, mitogenomic datasets were created in three simple steps. (1) GenBank accession numbers for complete mitochondrial genomes were searched and listed. (2) The corresponding GenBank files (hereafter called inputfile.gb) containing the complete mitochondrial genome sequences were obtained from public databases. (3) The downloaded GenBank files and MITOSCISSOR were placed in the same directory. Once these steps have been performed, the cmd.exe (a command in Windows OS) is opened, and MITOSCISSOR is run by typing the simple command, “perl MitoScissor.pl sequence.gb outputfile” (the process is shown in Fig. 2).

Tests of positive selection. The PAML package12 was used to estimate parameters in substitution models and to reveal whether positive selection was occurring. In PAML,

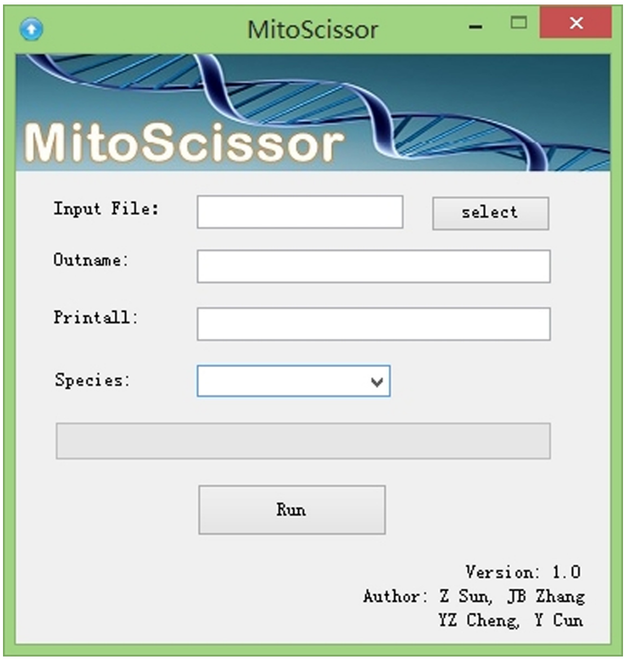

Figure 1. Window OS interface of MITOSCISSOR.

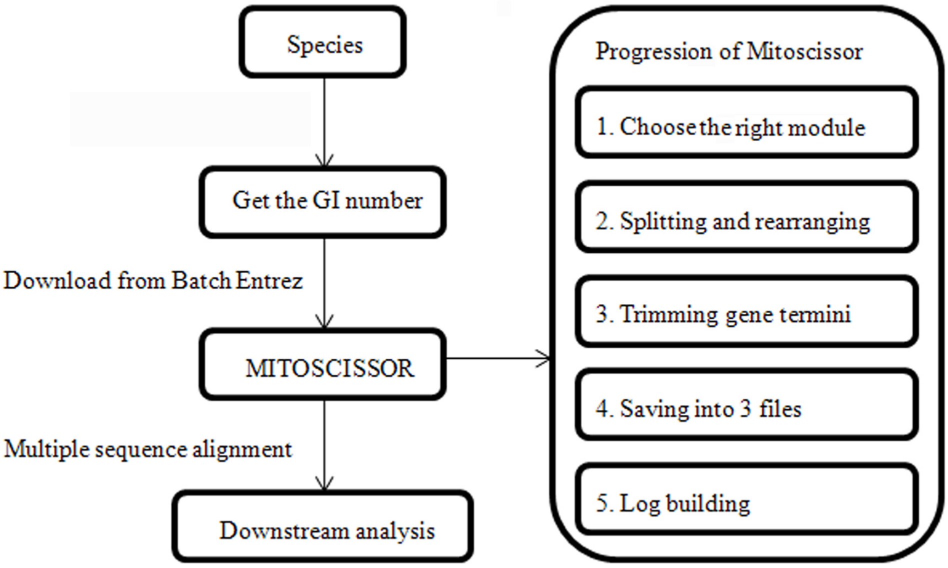
Species
Get the GI number
Download from Batch Entrez
MITOSCISSOR
Multiple sequence alignment
Downstream analysis
Progression of Mitoscissor
1. Choose the right module
2. Splitting and rearranging
3. Trimming gene termini
4. Saving into 3 files
5. Log building

Figure 2. Flow-process diagram in the software MITOSCISSOR.

there is a basal ratio model that assumes a single average ratio of nonsynonymous versus synonymous substitution rates ($\omega = dN/dS$) for all nonsynonymous substitutions. Other site models treat the ω ratio for any codon of a gene as a random variable from a statistical distribution, and allow ω values to vary among codons.[13] Positive selection is defined as presence of some codons at which $\omega > 1$, with 1 meaning neutral selection and <1 purifying selection. Likelihood ratio tests (LRTs), in which twice the log likelihood difference between the maximum log-likelihood estimates of the constrained and unconstrained models is distributed as an asymptotic χ^2 distribution, were used to assess gene evolution under positive selection.[13]

Sequence data and phylogenetic analyses. Sequences were then subjected to multiple alignments using MAFFT.[14] Gblocks was adopted to trim both the ambiguous alignments and highly diverged regions of the alignment via the "more stringent selection" setting.[15] Pairwise nucleotide differences in evolutionary distance were plotted in order to examine the pattern of sequence substitution with DAMBE.[16] Modeltest 3.06[17] was used to determine the most appropriate nucleotide substitution models with Akaike information criterion (AIC).[18] Analyses of phylogenetic relationships were conducted in a Bayesian framework using the software package MrBayes3.1.2.[19] Bayesian analyses were run five times for 50 million generations to ensure the convergence of Markov chain Monte Carlo (MCMC) chains. MCMC chains were run until convergence was reached, as determined by the standard deviation of the split frequencies dropping below 0.01.

Results and Discussion

It took less than one minute for MITOSCISSOR to assemble the 103 Tetraodontiform mitochondrial genomes that we had selected and to split these genomes into four partitions, each with their own fasta file. The partitions were as follows: protein genes, tRNA genes, rRNA genes, and individual genes. The corresponding fastafiles were output_proteingene.fa, output_rRNA.fa, output_tRNA.fa, and individual gene.fa. The results of this analysis are shown in Supplementary Files 2–5. The size of the Tetraodontiform mitochondrial genomes we examined ranged from 16,418 bp to 16,649 bp (shown in Supplementary File 6). The software conserves codon position order in the case of protein-coding genes that have an incomplete start or stop codon for analyses requiring codon-partitioning by inserting gaps.

The accuracy of molecular phylogeny primarily hinges on the selection of unsaturated genes.[20,21] Iss and Iss.c values were calculated with the aid of the program EVOLVER in the PAML package.[12] The index score (Iss = 0.4361) was significantly lower than the critical score (Iss.c = 0.7529) ($P < 0.001$), meaning that the mitogenomic datasets did not have any significant substitution saturation. Given our dataset did not show significant saturation, this made it suitable to examine the phylogenetic relationships among the Tetraodontiforms.

All nine Tetraodontiform families were clearly resolved as monophyletic with high support and were clearly revealed in Bayesian analysis (GTR+I+G model), and sister-group relationships were established (the minimum posterior probability value = 0.91) (Fig. 3). The topological inference from phylogenetic relationships in this study showed that subfamily Tetraodontinae is paraphyletic, which is in agreement with phylogenetic analyses based on comparative skeletal anatomy.[22,23] *Tetraodontiformes* represent one of the most recently evolved orders of teleost fishes. Members of this order display a variety of specialized body shapes, and show traits of extreme reductive evolution, with loss of anal fin spines as well as simplified cranial and skeletal structures.[22] Their unusual and diverse morphological characteristics make these fishes difficult to classify, and the phylogenetic relationships

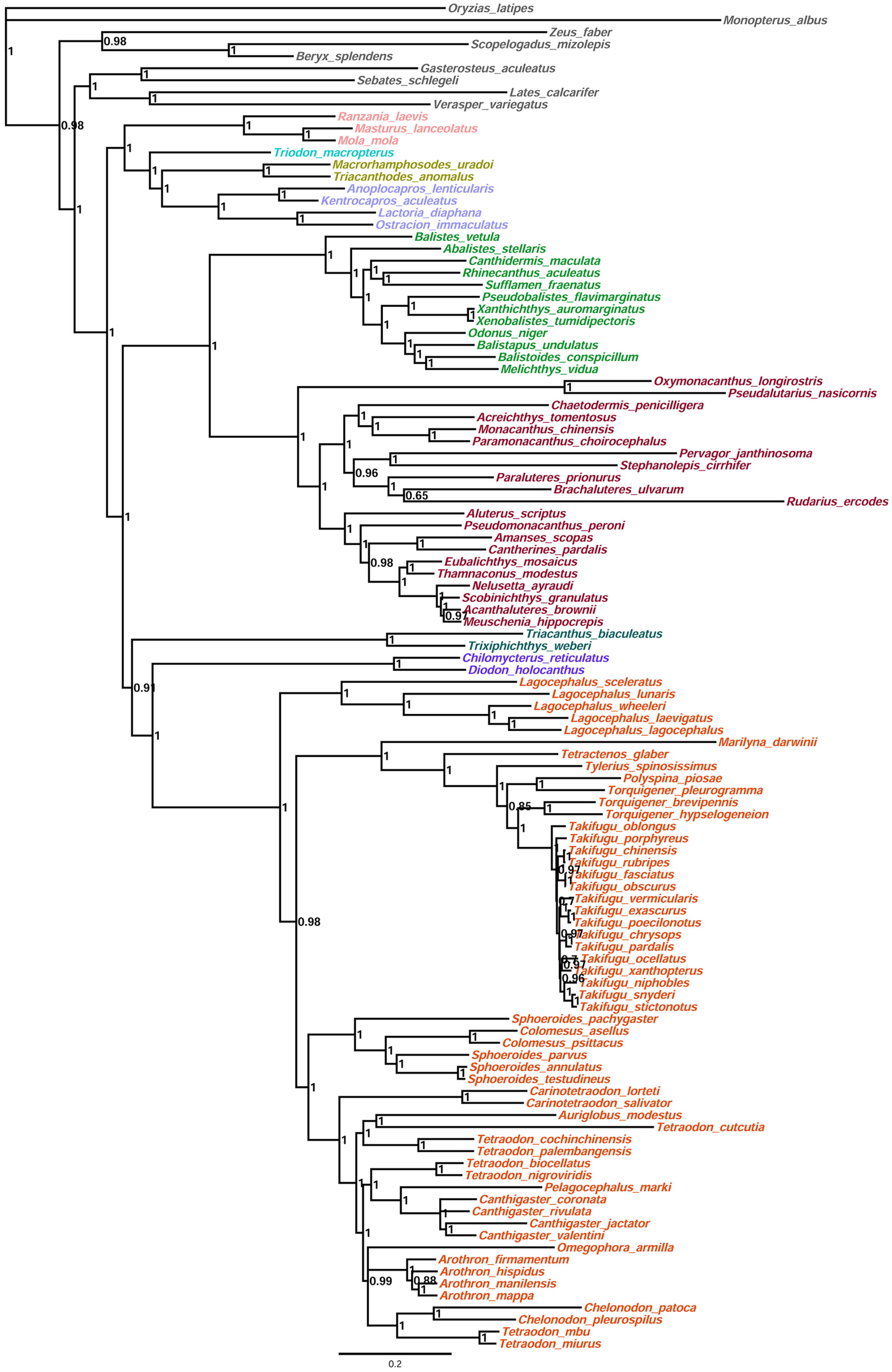

Figure 3. Bayesian phylogenetic tree of the order *Tetraodontiformes* based on mitogenomic datasets using Mrbayes3.1.2. The node values indicate the Bayesian probabilities for each phylogenetic clade. The scale bars represent codon substitutions per site.

between tetraodontiform families are still debated among ichthyologists.[23] Earlier studies have investigated the osteology, ontogeny, and mycology of these fishes in an attempt to clarify their interrelationships, but without clear results.[23,24] Additionally, none of the molecular phylogenetic analyses thus far published has been in full agreement with the conclusions drawn from these studies. The nuclear locus *RAG1* and the mitochondrial genes *12S* and *16S* rRNA were used by Holcroft[25] and Alfaro et al.[26] to determine representative tetraodontiform lineages and estimate their relationships. Neither study provided a clear resolution about the basal relationships with confidence, except for two sister-group relationships. Yamanoue et al.[27] tried to reconstruct the phylogeny of *Tetraodontiformes* using complete mitochondrial genomes, but their comparison of 25 mitogenomes failed to uncover clear relationships between the six families examined (Triacanthidae, Balistidae, Monacanthidae, Tetraodontidae, Diodontidae, and Molidae). Some ichthyologists have proposed that large taxon sample sizes would improve the effectiveness of such analyses, increasing the average accuracy of reconstructed phylogenies and reducing the inference of erroneous taxon bipartitions.[28–31] In the present study, using the complete mitogenome of more than Tetraodontiform fishes, we inferred a very robust evolutionary relationship of Tetraodontiform fishes compared to previous researches employing single or several genes.

Detecting positive selection is generally difficult because the signal may be swamped by the ubiquitous negative selection, but the proposition that mitochondrial genes have been subjected to non-neutral evolution has been gaining support in the last decades.[4,20] Many studies have uncovered some incidences of positive selection in several genes in the fish. Examples include the cytochrome c oxidase subunit (*COX2*) gene in tunas and billfishes[3,4,32]; the NADH dehydrogenase 2 (*ND2*), NADH dehydrogenase 4 (*ND4*), and NADH dehydrogenase 5 (*ND5*) genes in Pacific salmon[21]; and the *ND4* and *ND5* genes in Cobitoidea fishes.[22] However, the comprehensive display of all mitochondrial genes has been rarely reported due to lack of large-scale sequence analysis.[4] In the present study, LRT analysis (P-value < 0.05) suggested that positive selection occurred in nine genes [ATPase 6 (*ATP6*), ATPase 8 (*ATP8*), *CYTB*, *COX2*, *ND2*, *ND3*, *ND4*, *ND4L*, and *ND5*] for tetraodontiform fishes, with *ATP6*, *ATP8*, *ND4*, and *ND5* genes showing the highest proportions of codons under significant selective pressure (Fig. 4). Positive selection of *CYTB*, *ND4*, and *ND5* was found to occur in at least two independent evolutionary branches (Table 1), decreasing the likelihood of artifactual constructs resulting from short branching and indicating that these lineages have gone through adaptive molecular evolution.

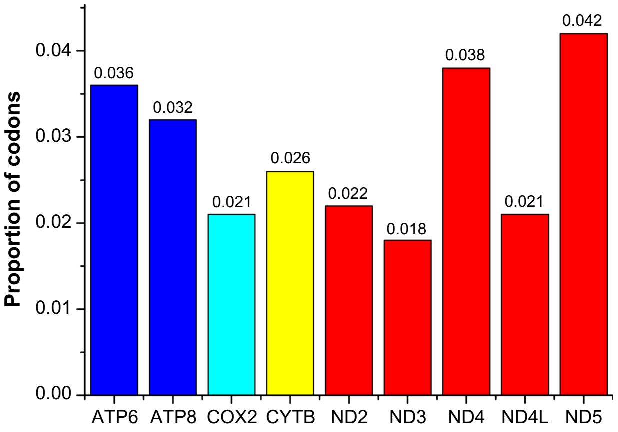

Figure 4. Bar chart showing the proportion of codons under significant positive selection in each gene for Tetraodontiform fishes.

Table 1. Lineages of Tetraodontiform fishes under positive selection detected to by branch analysis, gene, ratio of *dS*/*dN* and *P*-value.

BRANCH/LINAGE	GENE	*dS/dN*	*P*-VALUE
Molidae	*ND4*	0.0362	0.013
Balistidae	*ATP8* *CYTB* *ND4*	0.0183 0.0142 0.0451	0.029 0.038 0.012
Monacanthidae	*ND5* *ATP8* *ND5*	0.0169 0.0225 0.0112	0.032 0.043 0.028
Tetraodontidae	*COX2* *CYTB* *DN4* *ND3* *ND5*	0.0167 0.0236 0.0511 0.0235 0.0308	0.041 0.035 0.001 0.018 0.027

Conclusions

We have used MITOSCISSOR to analyze 103 Tetraodontiform mitogenome sequences and established the first robust molecular phylogeny with clear resolution for interfamilial relationships in the order *Tetraodontiformes*. MITOSCISSOR offers a simple and efficient approach to simultaneously parsing numerous mitochondrial genome files. We hope that MITOSCISSOR will dramatically increase the number of mitochondrial genomes that can be rapidly and accurately analyzed and compared, both in evolutionary biology and other fields of study. The highlights of our study are as follows:

1. We compiled a new software, MITOSCISSOR, for parsing and formatting complete mitochondrial genome sequences. MITOSCISSOR was much more reliable and efficient compared to manual editing.
2. We established the first robust molecular phylogeny with clear resolution for Tetraodontiform fishes.
3. Positive selection was detected in nine mitochondrial genes or Tetraodontiform fishes.

Acknowledgments

Thanks for the academic editor and reviewers for their comments, which were very helpful in improving the manuscript.

Author Contributions

Designed the software MITOSCISSOR: ZS, YZC, JBZ. Phylogentic analysis: ZS, YZC. Prepared the manuscript: JBZ, ZS. All authors reviewed and approved of the final version of the manuscript.

Supplementary Material

Supplementary File 1. Complete mitochondrial genome sequences used in this study.

Supplementary File 2. Sequences of protein genes for 103 Tetraodontiform fishes.

Supplementary File 3. Sequences of tRNA genes for 103 Tetraodontiform fishes.

Supplementary File 4. Sequences of rRNA genes for 103 Tetraodontiform fishes.

Supplementary File 5. Split individual genes for 103 mitochondrial genomes of Tetraodontiform fishes.

Supplementary File 6. General information about mitochondrial genomes employed in this paper.

REFERENCES

1. Harrison RG. Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. *Trends Ecol Evol*. 1989;4(1):6–11.
2. Baumsteiger J, Kinziger AP, Aguilar A. Life history and biogeographic diversification of an endemic western North American freshwater fish clade using a comparative species tree approach. *Mol Phylogenet Evol*. 2012;65(3):940–52.
3. Dalziel AC, Moyes CD, Fredriksson E, Lougheed SC. Molecular evolution of cytochrome c oxidase in high-performance fish (*teleostei*: *Scombroidei*). *J Mol Evol*. 2006;62(3):319–31.
4. Menezes AN, Viana MC, Furtado C, Schrago CG, Seuanez HN. Positive selection along the evolution of primate mitogenomes. *Mitochondrion*. 2013;13(6):846–51.
5. Havird JC, Santos SR. Performance of single and concatenated sets of mitochondrial genes at inferring metazoan relationships relative to full mitogenome data. *PLoS One*. 2014;9(1):e84080.
6. Inoue JG, Kumazawa Y, Miya M, Nishida M. The historical biogeography of the freshwater knifefishes using mitogenomic approaches: a mesozoic origin of the Asian notopterids (*Actinopterygii*: Osteoglossomorpha). *Mol Phylogenet Evol*. 2009;51(3):486–99.
7. Bar-Yaacov D, Blumberg A, Mishmar D. Mitochondrial-nuclear co-evolution and its effects on OXPHOS activity and regulation. *Biochim Biophys Acta*. 2012;1819(9–10):1107–11.
8. Lavoue S, Miya M, Inoue JG, Saitoh K, Ishiguro NB, Nishida M. Molecular systematics of the gonorynchiform fishes (*Teleostei*) based on whole mitogenome sequences: implications for higher-level relationships within the Otocephala. *Mol Phylogenet Evol*. 2005;37(1):165–77.
9. Duchene S, Archer FI, Vilstrup J, Caballero S, Morin PA. Mitogenome phylogenetics: the impact of using single regions and partitioning schemes on topology, substitution rate and divergence time estimation. *PLoS One*. 2011;6(11):e27138.
10. Boore JL. Animal mitochondrial genomes. *Nucleic Acids Res*. 1999;27(8):1767–80.
11. Pereira SL. Mitochondrial genome organization and vertebrate phylogenetics. *Genet Mol Biol*. 2000;23:745–52.
12. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586–91.
13. Yang Z. *Computational Molecular Evolution*. Vol 21. Oxford: Oxford University Press; 2006.
14. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*. 2008;9(4):286–98.
15. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000;17(4):540–52.
16. Xia X, Xie Z. DAMBE: software package for data analysis in molecular biology and evolution. *J Hered*. 2001;92(4):371–3.
17. Posada D. jModelTest: phylogenetic model averaging. *Mol Biol Evol*. 2008;25(7):1253–6.
18. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, eds. *Selected Papers of Hirotugu Akaike*. New York: Springer; 1998:199–213.
19. Huelsenbeck JP, Ronquist F. MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics*. 2001;17(8):754–5.
20. Liu SQ, Mayden RL, Zhang JB, et al. Phylogenetic relationships of the Cobitoidea (*Teleostei*: *Cypriniformes*) inferred from mitochondrial and nuclear genes with analyses of gene evolution. *Gene*. 2012;508(1):60–72.
21. Garvin MR, Bielawski JP, Gharrett AJ. Positive Darwinian selection in the piston that powers proton pumps in complex I of the mitochondria of Pacific salmon. *PLoS One*. 2011;6(9):e24127.
22. Skov P, Bennett M. Branchial vascular pathways in two species of *Tetraodontiformes* and the concept of secondary vessels and nutrient arteries. *Zoomorphology*. 2005;124(2):79–88.
23. Santini F, Tyler JC. A phylogeny of the families of fossil and extant tetraodontiform fishes (Acanthomorpha, *Tetraodontiformes*), upper cretaceous to recent. *Zool J Linn Soc*. 2003;139(4):565–617.
24. Winterbottom R. The Familial Phylogeny of the Tetraodontiformes (Acanthopterygii: Pisces): As Evidenced by Their Comparative Myology. Washington: Smithsonian Institution Press; 1974.
25. Holcroft NI. A molecular analysis of the interrelationships of tetraodontiform fishes (Acanthomorpha: *Tetraodontiformes*). *Mol Phylogenet Evol*. 2005;34(3):525–44.
26. Alfaro ME, Santini F, Brock CD. Do reefs drive diversification in marine teleosts? Evidence from the pufferfishes and their allies (Order *Tetraodontiformes*). *Evolution*. 2007;61(9):2104–26.
27. Yamanoue Y, Miya M, Matsuura K, Katoh M, Sakai H, Nishida M. A new perspective on phylogeny and evolution of tetraodontiform fishes (Pisces: *Acanthopterygii*) based on whole mitochondrial genome sequences: basal ecological diversification? *BMC Evol Biol*. 2008;8:212.
28. Zwickl DJ, Hillis DM. Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol*. 2002;51(4):588–98.
29. Pickett KM, Randle CP. Strange bayes indeed: uniform topological priors imply non-uniform clade priors. *Mol Phylogenet Evol*. 2005;34(1):203–11.
30. Ricklefs RE, Losos JB, Townsend TM. Evolutionary diversification of clades of squamate reptiles. *J Evol Biol*. 2007;20(5):1751–62.
31. Philippe H, Brinkmann H, Lavrov DV, et al. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 2011;9(3):e1000602.
32. Teacher AG, Andre C, Merila J, Wheat CW. Whole mitochondrial genome scan for population structure and selection in the Atlantic herring. *BMC Evol Biol*. 2012;12:248.