


# An Intron of Invertebrate Microphthalmia Transcription Factor Gene Is Evolved from a Longer Ancestral Sequence

Jun-Ming Mao<sup>1</sup> , Yong Wang<sup>2</sup>, Liu Yang<sup>2</sup>, Qin Yao<sup>1</sup> and Ke-Ping Chen<sup>1</sup>

<sup>1</sup>School of Life Sciences, Jiangsu University, Zhenjiang, China. <sup>2</sup>School of Food and Biological Engineering, Jiangsu University, Zhenjiang, China.

Evolutionary Bioinformatics  
Volume 17: 1–7  
© The Author(s) 2021  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1176934320988558



**ABSTRACT:** Introns are highly variable in number and size. Sequence simulation is an effective method to elucidate intron evolution patterns. Previously, we have reported that introns are more likely to evolve through mutation-and-deletion (MD) rather than through mutation-and-insertion (MI). In the present study, we further studied evolution models by allowing insertion in the MD model and by allowing deletion in the MI model at various frequencies. It was found that all deletion-biased models with proper parameter settings could generate sequences with attributes matchable to 16 invertebrate introns from the microphthalmia transcription factor gene, whereas all insertion-biased models with any parameter settings failed to generate such sequences. We conclude that the examined invertebrate introns may have evolved from a longer ancestral sequence in a deletion-biased pattern. The constructed models are useful for studying the evolution of introns from other genes and/or from other taxonomic groups. (C++ scripts of all deletion- and insertion-biased models are available upon request.)

**KEYWORDS:** Mutation, insertion, deletion, sequence simulation, evolution model

**RECEIVED:** August 21, 2020. **ACCEPTED:** December 29, 2020.

**TYPE:** Short Report

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the National Natural Science Foundation of China (Nos. 31872425 and 31861143051).

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Yong Wang, School of Food and Biological Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang, 212013, China. Email: ywang@ujs.edu.cn

## Introduction

The discovery that eukaryotic genes are interrupted by intron sequences is an important milestone of biological research.<sup>1–3</sup> Soon after this discovery, the debate about early or late emergence of introns began and continues even today.<sup>4–6</sup> The “introns-early” theory states that introns emerged early in ancestral prokaryotes and continuous intron sequence loss is the main event involved in evolution of prokaryotes and eukaryotes.<sup>7–10</sup> This theory is supported by existence of intronless or intron-poor genomes in extant organisms.<sup>11,12</sup> The “introns-late” theory posits that introns emerged late in ancestral eukaryotes and continuous intron gain is the main event involved in the evolution of eukaryotes.<sup>13,14</sup> This theory is supported by the existence of a higher number and size of introns in more complex organisms.<sup>15,16</sup> In the past decades, considerable evidence has been found to support the introns-early theory. For instance, ancestral eukaryotic genomes have a much higher intron density than extant eukaryotic genomes,<sup>17–21</sup> and intron loss has occurred predominantly during the evolution of eukaryotic lineages.<sup>22–26</sup> However, these findings are mainly based on the presence or absence of introns among the surveyed organisms. They constitute sufficient evidence merely for reduction in intron number during the evolution of eukaryotic genomes. For an intron that exists in all surveyed organisms, the reason of an increase in its size according to the complexity of the organism has not been clearly explained. This remains a challenging question against the introns-early theory.

Owing to the high variation of intron sequences, studies on intron size evolution have been confined within relatively small taxonomic groups. This is mainly because homology between

intron sequences only exists among organisms that belong to the same order/family. Therefore, studies on intron size variation have been conducted only in a few lineages including fungi, nematodes, fruit flies, pigeons, peas, and a carnivorous plant genus,<sup>27–32</sup> in which intron sizes have been found to change in a strong deletion-biased pattern. As no homologous intronic segments are available for examination of the presence or absence of a specific intronic segment, the study of size variation of an intron in organisms belonging to different phyla/classes requires novel approaches, such as sequence simulation. Previously, we have constructed evolutionary models to simulate the evolution of an intron in organisms from 7 classes of chordates.<sup>33</sup> We found that introns in various chordate species could evolve from a longer ancestral sequence through base deletion, and the existence of longer introns in higher organisms could be attributed to a lower efficiency in base deletion. In the present study, the same approach was used to simulate the evolution of an intron from 16 invertebrate species using re-constructed deletion- and insertion-biased evolution models. Testing results from the execution of all re-constructed evolution models suggested that the surveyed invertebrate introns were evolved in a deletion-biased pattern as well.

## Materials and Methods

### *Invertebrate introns and their attributes*

In invertebrates, the coding sequence for bHLH (basic helix-loop-helix) motif of the microphthalmia transcription factor (MITF) has a conserved phase 1 intron in the basic region.



**Table 1.** Sixteen species selected to represent different phylum/class of invertebrates.

PHYLUM	CLASS	SPECIES	INTRON (BP)
Porifera (sponges)	Demospongiae	<i>Amphimedon queenslandica</i>	1201
Cnidaria (cnidarians)	Anthozoa (anthozoans)	<i>Nematostella vectensis</i> (starlet sea anemone)	567
Platyhelminthes (flatworms)	Trematoda	<i>Schistosoma haematobium</i>	828
Nemertea (ribbon worms)	Pilidiophora	<i>Notospermus geniculatus</i>	1342
Priapulida (priapulids)	Priapulimorpha	<i>Priapulid caudatus</i>	971
Annelida (annelid worms)	Polychaeta (polychaetes)	<i>Hydroides elegans</i> (calcareous tube worm)	380
Mollusca (mollusks)	Bivalvia (bivalves)	<i>Crassostrea gigas</i> (Pacific oyster)	1188
Brachiopoda (lampshells)	Not available	<i>Phoronis australis</i>	1878
Echinodermata (echinoderms)	Echinoidea (sea urchins)	<i>Strongylocentrotus purpuratus</i> (purple sea urchin)	1061
Hemichordata (hemichordates)	Enteropneusta (acorn worms)	<i>Saccoglossus kowalevskii</i>	490
Tunicata (tunicates)	Ascidiacea (sea squirts)	<i>Ciona intestinalis</i> (vase tunicate)	672
Arthropoda (arthropods)	Arachnida (arachnids)	<i>Latrodectus Hesperus</i> (western black widow)	3104
	Merostomata (horseshoe crabs)	<i>Limulus polyphemus</i> (Atlantic horseshoe crab)	1168
	Branchiopoda	<i>Daphnia pulex</i> (common water flea)	1335
	Malacostraca	<i>Hyalella azteca</i>	2008
	Insecta (true insects)	<i>Danaus plexippus</i> (monarch butterfly)	259

This intron has 259 to 3104 base pairs (bp) in the 16 species selected to represent different phyla/classes of invertebrates (Table 1). These invertebrate introns are considered to evolve from a common ancestral sequence, because each invertebrate species has only 1 MITF gene and the nucleotides flanking this intron are highly conserved (Figure 1). Therefore, they are eligible targets for this study which focused on simulation of intron evolution from one common ancestral sequence.

The multiple sequence alignment obtained using Muscle program<sup>34</sup> has very few conserved sites among these introns (Supplemental Figure S1), based on which no sequence insertion or deletion can be identified. Accordingly, the phylogenetic tree constructed using MEGA 5.2 software<sup>35</sup> has very low bootstrap values at branching nodes (Figure 2), based on which no clear evolutionary inference can be made. Therefore, sequence simulation was conducted toward these introns by following the method described in our previous report.<sup>33</sup> These 16 invertebrate introns were found to have an  $L_{MSA}$  (size of multiple sequence alignment) value of 3434bp, an  $R_{T92+G+I}$  (ratio of transition to transversion under the Tamura 3 parameter model<sup>36</sup> with gamma distribution and invariant sites) value of 1.95, a  $\bar{D}$  (overall mean distance) value of 1.425, an  $SE_{\bar{D}}$  (standard error of the overall mean distance) value of 0.119, and a  $TS_{ML}$  (topology score of the constructed ML tree) of 32.

### Design of evolution models

In our previous report,<sup>33</sup> mutation-and-deletion (MD) and mutation-and-insertion (MI) models were designed to simulate consecutive deletion and consecutive insertion events,

respectively. In the present work, we introduced an insertion event in the MD model and a deletion event in the MI model at various frequencies to construct deletion- and insertion-biased models, respectively. For example, the MD90/10 model allows for 90% chances of base deletion and 10% chances of base insertion, while the MI90/10 model allows for 90% chances of base insertion and 10% chances of base deletion. Overall, 6 deletion- and 6 insertion-biased models were constructed. They are designated as MD100, MD90/10, MD80/20, MD70/30, MD60/40, MD55/45, MI100, MI90/10, MI80/20, MI70/30, MI60/40, and MI55/45. All models were constructed using the C++ computational language.

### Simulation of intron evolution

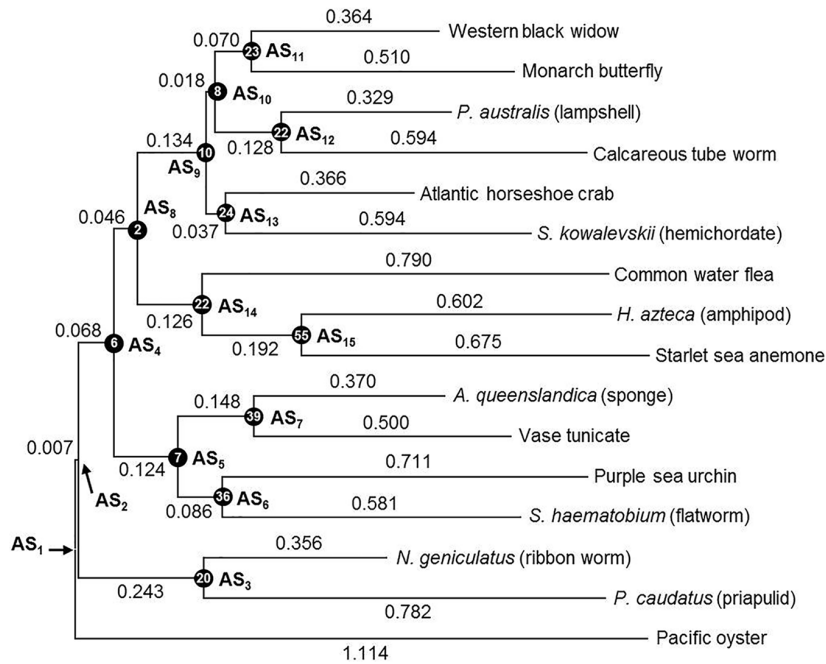
Each of the above-mentioned constructed evolution models was first tested using factors and levels designed in accordance with the  $L_{16}(4^*5)$  orthogonal table (Table 2). For model testing, the phylogenetic tree of the 16 invertebrate introns (Figure 2) was referenced to determine the evolution steps for all 16 sequences intended for generation. Based on statistical analysis of the results of the orthogonal test, further tests were conducted against each model to confirm whether the model-generated sequences had attributes that matched the 16 invertebrate introns. Please refer to our previous report<sup>33</sup> for detailed operational procedures.

### R value of the constructed models

The transition to transversion ratio ( $R$ ) was set to 2.0 in all deletion- and insertion-biased models because the  $R$  value of

Invertebrate species	Upstream exon (partial)	Phase 1 intron	Downstream exon (partial)	Sequence number
<i>A. queenslandica</i> (sponge)	AAGAAAACAATCACAACATGA	1201	TTGAAAGAACACGGAAGA	XP_011406372.2
Starlet sea anemone	AAAAAGACAATCATAATATGA	567	TTGAGAGAACAGACGA	XP_001636474.1
<i>S. haematobium</i> (flatworm)	AAAAAGATAGTCATAATCGAA	828	TTGAACGTAAGCGACGA	XP_012797191.1
<i>N. geniculatus</i> (ribbon worm)	AAGAAGGACAATCATAATATCA	1342	TTGAAAGACGCGGCGA	NMRB01004045.1
<i>P. caudatus</i> (priapulid)	AAGAAGGACAATCACAACAAGA	971	TCGAGCGAAAAGCGCGG	XP_014672776.1
Calcareous tube worm	AAGAAGGATAACCATTAATCAGA	380	TTGAGAGACAGACCGG	LQRL01147342.1
Pacific oyster	AAGAAGACAACCATTAACATGA	1188	TTGAGAGAACAGGAAGA	XP_011447559.2
<i>P. australis</i> (lampshell)	AAGAAGACAGTCATAACATGA	1878	TTGAAAGACGCGAAGA	NMRA01000491.1
Purple sea urchin	AAGAAGGATAATCACAATATGA	1061	TTGAGCGAACAGGAAGA	XP_783071.4
<i>S. kowalevskii</i> (hemichordate)	AAAAAGATAACCATTAATATGA	490	TTGAGAGAACAGGCGGT	NP_001161587.1
Vase tunicate	AAGAAGGACAATCACAATATAA	672	TTGAGCGGACACGAAGG	NP_001071764.1
Western black widow	AAAAAGATAATCATAATAAAA	3104	TTGAAAGAACACGCGCA	JJRX02036501.1
Atlantic horseshoe crab	AAGAAGGATAATCACAATATGA	1168	TTGAAAGAACAGGAAGA	XP_013777631.1
Common water flea	AAGAAGGATAATCACAACATGA	1335	TTGAGCGGACGCGCAGG	FLTH01000050.1
<i>H. azteca</i> (amphipod)	AAGAAGACAACCATTAACATGA	2008	TTGAGCGACCCGTCCG	XP_018018547.1
Monarch butterfly	AAGAAGACAACCATTAATATGA	259	TCGAAACCCCGCTCGT	XP_032524945.1
<b>Consensus codons</b> → AAGAAAGACAATCATAATATGA			TTGAGAGAAGACGACGA	
<b>Amino acids</b> → K K D N H N M			I E R R R R	

**Figure 1.** Partial structure of microphthalmia transcription factor (MITF) gene in invertebrate. Shown here is a phase 1 intron flanked by conserved exon nucleotides encoding the basic region of bHLH motif from invertebrate MITF gene. The intron is located after A of the codon ATT which codes for isoleucine (I). Number between lines indicates intron length (base pairs). Invertebrate species is given in common name or abbreviated scientific name with taxon name in brackets. Intron locations were obtained by viewing gene structures linked to sequence numbers beginning with “XP” or “NP” at GenBank (www.ncbi.nlm.nih.gov). Intron locations of other sequence numbers were determined by manually comparing genomic sequences with those of known gene structures. Please refer to Table 1 for full scientific names of invertebrate species.



**Figure 2.** Phylogenetic tree of 16 invertebrate introns. The original maximum likelihood (ML) tree constructed using 16 invertebrate introns is shown. Invertebrate species is shown in common name or abbreviated scientific name with taxon name in brackets. Branch sizes are indicated by values above or below each branch. AS<sub>1</sub> to AS<sub>15</sub> indicate the locations of ancestral sequence No. 1 to 15. Numbers at nodes are bootstrap values obtained using 1000 replicates. Please refer to Table 1 for full scientific name and intron length of each invertebrate species.

the 16 invertebrate introns was 1.95, as determined by model testing using the MEGA 5.2 software.

*Statistical analysis*

The SPSS software (version 17.0) was used to perform all statistical analyses as described in our previous report.<sup>33</sup>

**Results**

*Testing of deletion-biased models*

The orthogonal tests (test nos. 1 to 16) for each deletion-biased model were repeated 10 times to obtain average attribute values of the model-generated sequences. Thereafter, the model parameters were optimized to perform more tests (test nos. 17

**Table 2.** Factor and level design for testing evolution models using  $L_{16}$  ( $4 \times 5$ ) orthogonal table.

EVOLUTION MODEL	LEVEL	FACTORS				
		$L_{AS1}$	$L_{AS15}$	$M_1$	$L_{ID}$	$M_{ID}$
Mutation-and-deletion	1	5000	3000	200	31-50	11-20
	2	6000	3250	400	71-90	21-30
	3	7000	3500	600	111-130	31-40
	4	8000	3750	800	151-170	41-50
Mutation-and-insertion	1	20	150	200	31-50	11-20
	2	40	200	400	71-90	21-30
	3	60	250	600	111-130	31-40
	4	80	300	800	151-170	41-50

Abbreviations:  $L_{AS1}$ , length of ancestral sequence 1;  $L_{AS15}$ , length of ancestral sequence 15;  $L_{ID}$ , length of bases inserted or deleted each time;  $M_{ID}$ , number of bases mutated each time;  $M_1$ , mutated bases per 1 branch length.

to 24) according to the statistical analysis results of the orthogonal test. The results of orthogonal tests for all deletion-biased models are listed in Supplemental Tables S1 to S6. The effects of the model parameters on the attributes of the model-generated sequences are shown in Supplemental Figures S2 to S7, and the results of parameter optimization are listed in Supplemental Tables S7 to S12. By progressively adjusting parameter values, each deletion-biased model generated sequences with attributes that were not significantly different ( $P > 0.1$ ) with the 16 invertebrate introns (Table 3, upper half). However, the optimal value of a specific parameter varied considerably with the evolution model. For instance, the optimal  $M_1$  (mutated bases per 1 branch size) for models with less than 20% insertion frequency was 1200 bp, while that for models with 30%, 40%, and 45% insertion frequencies were 600, 200, and 800 bp, respectively. Additionally, the optimal  $L_{ID}$  (size of bases inserted or deleted each time) for models with less than 30% insertion frequency was below 50 bp, whereas that for models with 40% and 45% insertion frequency was above 111 bp. Although the optimal values for specific parameters were markedly different, a fine-adjusted combination of parameter values could always allow the deletion-biased models to generate sequences with attributes matched to the 16 invertebrate introns (Table 3, upper half). Therefore, we conclude that the surveyed invertebrate introns may have evolved from longer ancestral sequences (e.g., 5000 to 8000 base pairs) in a deletion-biased pattern.

### Testing of insertion-biased models

The orthogonal tests (test nos. 1 to 16) for each insertion-biased model were repeated 10 times to obtain average attribute values of the model-generated sequences. Then, the model parameters were optimized to perform more tests (test nos. 17-24) according to the statistical analysis results of the orthogonal test. The results of orthogonal tests for all insertion-biased models are

listed in Supplemental Tables S13 to S18. The effects of the model parameters on the attributes of the model-generated sequences are shown in Supplemental Figures S8 to S13, and the results of parameter optimization are listed in Supplemental Tables S19 to S24. By progressively adjusting parameter values, all insertion-biased models failed to generate sequences with attributes matched to the 16 invertebrate introns. Specifically,  $\bar{D}$  (overall mean distance) of the model-generated sequences was always significantly higher ( $P < .01$ ) than that of invertebrate introns (Table 3, lower half). In case that our orthogonal tests did not investigate the correlations between factors, the efficiency of parameter optimization might be lowered to some extent. Therefore, apart from the tests listed in Supplemental Tables S19 to S24, additional tests were conducted against each insertion-biased model using different parameter settings. However, all these tests provided negative results. Therefore, we conclude that the surveyed invertebrate introns may not have evolved from a shorter ancestral sequence (e.g., less than 80 base pairs) in an insertion-biased pattern.

### Discussion

Owing to a high variability in the number and size of introns, their evolution remains poorly understood. Intron variability results from multiple evolutionary events including intron gain, intron loss, intron slippage, DNA recombination, DNA transposition, and horizontal gene transfer.<sup>15,37-40</sup> While reduction in intron number has been observed during genome evolution in many eukaryotic lineages,<sup>17-26</sup> reduction of intron size has only been observed during genome evolution of organisms from different families/genera<sup>27-32</sup> and from different classes.<sup>33</sup> Our present work extends the study on intron size variation in organisms from different metazoan phyla. Theoretically, the evolution models established in this study can be used to test whether introns from other taxonomic groups evolve in a deletion- or insertion-biased pattern, because the phylogenetic tree formed by introns of interest is

Table 3. Attributes of sequences generated from MD and MI models using optimized parameters.

MODEL	MODEL PARAMETERS				ATTRIBUTES OF GENERATED SEQUENCES						
	$L_{AS1}$	$L_{AS15}$	$M_1$	$L_{ID}$	$M_{ID}$	$L_{MSA}$	$R_{T92+G+H}$	$D$	$SE_5$	$TS_{ML}$	
Slls	/	/	/	/	/	3378 ± 73	1.92 ± 0.49	1.42 ± 0.11	0.138 ± 0.013	11.9 ± 2.4	
MD100	5000	3250	1200	11-30	41-50	3394 ± 67	1.81 ± 0.29	1.36 ± 0.17	0.129 ± 0.013	11.6 ± 1.7	
MD90/10	8000	3250	1200	11-30	21-30	3395 ± 85	2.01 ± 0.54	1.50 ± 0.11	0.137 ± 0.017	11.4 ± 2.0	
MD80/20	8000	3750	1200	31-50	21-30	3427 ± 69	1.77 ± 0.24	1.47 ± 0.14	0.144 ± 0.020	12.0 ± 2.5	
MD70/30	8000	3250	600	31-50	41-50	3407 ± 155	1.79 ± 0.39	1.52 ± 0.20	0.141 ± 0.017	10.7 ± 2.0	
MD60/40	8000	3750	200	111-130	21-30	3439 ± 86	1.69 ± 0.28	1.33 ± 0.19	0.140 ± 0.012	11.1 ± 1.3	
MD55/45	5000	3000	800	151-170	11-20	3420 ± 98	1.91 ± 0.47	1.50 ± 0.29	0.138 ± 0.015	11.2 ± 2.5	
MI100	80	250	400	111-130	11-20	3442 ± 98	2.06 ± 0.52	1.69 ± 0.16***	0.128 ± 0.006*	12.5 ± 2.5	
MI90/10	60	200	200	71-90	41-50	3457 ± 119*	1.74 ± 0.35	1.74 ± 0.23***	0.133 ± 0.013	9.7 ± 1.6**	
MI80/20	80	300	200	71-90	21-30	3431 ± 104	2.01 ± 0.39	1.70 ± 0.18***	0.132 ± 0.015	10.8 ± 2.3	
MI70/30	20	150	200	71-90	51-60	3421 ± 112	2.21 ± 0.58	1.77 ± 0.21***	0.135 ± 0.018	11.5 ± 2.0	
MI60/40	60	300	1200	31-50	21-30	3405 ± 126	1.82 ± 0.23	1.77 ± 0.25***	0.121 ± 0.015**	11.7 ± 2.3	
MI55/45	40	300	600	31-50	41-50	3415 ± 125	1.92 ± 0.37	1.78 ± 0.16***	0.130 ± 0.015	11.6 ± 1.8	

Abbreviations:  $D$ : overall mean distance;  $L_{AS1}$ : length of ancestral sequence 1;  $L_{AS15}$ : length of ancestral sequence 15;  $L_{ID}$ : length of bases inserted or deleted each time;  $L_{MSA}$ : length of multiple sequence alignment;  $M_{ID}$ : number of bases mutated each time;  $M_1$ : mutated bases per 1 branch length;  $R_{T92+G+H}$ : ratio of transition to transversion under Tamura 3 parameter model with gamma distribution and invariant sites;  $SE_5$ : standard error of the overall mean distance; Slls, sixteen invertebrate introns;  $TS_{ML}$ : topology score of the constructed ML tree.

This table lists the test result of No. 24 for each model. Please refer to Supplemental Tables S7 to S12 and S19 to S24 for test results of No. 17 to 23 of all evolution models. Attributes of Slls are obtained from allowing each of the sequence to mutate by only one base.

Data are presented as mean ± standard deviation (n = 10).

\*, \*\*, and \*\*\* indicate significant difference from independent t-test compared to Slls at  $P < .1$ ,  $P < .05$  and  $P < .01$  level, respectively.

only referenced for setting model parameters. While the overall trend of intron evolution is toward the loss of bases (i.e., shortening of introns), a question arises on intron length; why are introns longer in higher organisms compared to those in lower organisms? In our opinion, this is because lower organisms are more efficient in shortening introns. This is possible because, in general, lower organisms are reproduced more frequently than higher organisms; thus, they have more opportunities for genome reorganization.<sup>41,42</sup> However, further investigations are needed to compare the intron-shortening efficiency between lower and higher organisms.

While deletion-biased patterns are followed by intron evolution in certain eukaryotic lineages, the ratio of deletion to insertion may vary considerably among different organisms. It is 3- and 6-fold higher among nematode and avian species, respectively.<sup>28,30</sup> It ranges from 1.2 to 9.0 in all deletion-biased models of the present study. Since all these models are proficient in simulating the evolution of the 16 invertebrate introns, we suggest that the intron deletion efficiency may be remarkably different among these invertebrate species. However, such differences can also result from certain inadequacies in designing evolution models. After examining our model designs, we conclude that they can be improved in 2 aspects. First, we may consider the effect of insertion/deletion on the phyletic clade formation of each intron. Second, we may use different sizes for stepwise insertion/deletion in simulation of the evolution of each intron. It is anticipated that the newly designed models will narrow the range of deletion to insertion bias to simulate the evolution of these invertebrate introns.

The deletion-biased evolution leads to the shortening of an intron but does not lead to its removal. It is advantageous in retaining functional introns and improving gene expression efficiency. Retention of a long intron lowers gene expression efficiency because it consumes substantial energy in both transcription and post-transcriptional processes. However, intron removal may lead to loss of important functional elements because an intron may be able to stimulate gene expression, regulate protein isoform formation, maintain RNA stability, or improve translation efficiency.<sup>43-46</sup> Therefore, in cases where an intron has an important regulatory function,<sup>47</sup> its shortening would be preferable compared to its complete removal. This probably explains the reason of maintenance of many introns in certain genes.

### Author Contributions

J-MM and LY performed the analysis. J-MM and YW wrote the manuscript with input from all authors. YW designed and constructed the models. YW, QY, and K-PC proposed and conceived the study. All authors approved the final version of the manuscript.

### ORCID iD

Jun-Ming Mao  <https://orcid.org/0000-0003-1866-5513>

### Supplemental Material

Supplemental material for this article is available online.

### REFERENCES

- Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA*. 1977;74:3171-3175.
- Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*. 1977;12:1-8.
- Gilbert W. Why genes in pieces? *Nature*. 1978;271:501.
- Koonin EV. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct*. 2006;1:22.
- Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. *Biol Direct*. 2012;7:11.
- Rogers SO. Integrated evolution of ribosomal RNAs, introns, and intron nurseries. *Genetica*. 2019;147:103-119.
- Doolittle WF. Genes in pieces: were they ever together? *Nature*. 1978;272:581-582.
- Gilbert W. The exon theory of genes. *Cold Spring Harb Symp Quant Biol*. 1987;52:901-905.
- Darnell JE, Doolittle WF. Speculations on the early course of evolution. *Proc Natl Acad Sci USA*. 1986;83:1271-1275.
- Stoltzfus A. Origin of introns—early or late. *Nature*. 1994;369:526-527; author reply 527-528.
- Nixon JE, Wang A, Morrison HG, et al. A spliceosomal intron in *Giardia lamblia*. *Proc Natl Acad Sci USA*. 2002;99:3701-3705.
- Simpson AG, MacQuarrie EK, Roger AJ. Eukaryotic evolution: early origin of canonical introns. *Nature*. 2002;419:270.
- Doolittle WF, Stoltzfus A. Molecular evolution. Genes-in-pieces revisited. *Nature*. 1993;361:403.
- Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr, Doolittle WF. Testing the exon theory of genes: the evidence from protein structure. *Science*. 1994;265:202-207.
- Logsdon JM Jr. The recent origins of spliceosomal introns revisited. *Curr Opin Genet Dev*. 1998;8:637-648.
- Penny D, Hoepfner MP, Poole AM, Jeffares DC. An overview of the introns-first theory. *J Mol Evol*. 2009;69:527-540.
- Carmel L, Wolf YI, Rogozin IB, Koonin EV. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res*. 2007;17:1034-1044.
- Csuros M, Rogozin IB, Koonin EV. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol*. 2011;7:e1002150.
- Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*. 2003;13:1512-1517.
- Roy SW, Irimia M. Splicing in the eukaryotic ancestor: form, function and dysfunction. *Trends Ecol Evol*. 2009;24:447-455.
- Wu B, Macielog AI, Hao W. Origin and spread of spliceosomal introns: insights from the fungal clade Zygomycota. *Genome Biol Evol*. 2017;9:2658-2667.
- Yang YF, Zhu T, Niu DK. Association of intron loss with high mutation rate in Arabidopsis: implications for genome evolution. *Genome Biol Evol*. 2013;5:723-733.
- Hooks KB, Delneri D, Griffiths-Jones S. Intron evolution in *Saccharomycetaceae*. *Genome Biol Evol*. 2014;6:2543-2556.
- Wang H, Devos KM, Bennetzen JL. Recurrent loss of specific introns during angiosperm evolution. *PLoS Genet*. 2014;10:e1004843.
- Sun Y, Whittle CA, Corcoran P, Johannesson H. Intron evolution in *Neurospora*: the role of mutational bias and selection. *Genome Res*. 2015;25:100-110.
- Roy SW. How common is parallel intron gain? Rapid evolution versus independent creation in recently created introns in *Daphnia*. *Mol Biol Evol*. 2016;33:1902-1906.
- Megarioti AH, Kouvelis VN. The coevolution of fungal mitochondrial introns and their Homing Endonucleases (GIY-YIG and LAGLIDADG). *Genome Biol Evol*. 2020;12:1337-1354.
- Konrad A, Brady MJ, Bergthorsson U, Katju V. Mutational landscape of spontaneous base substitutions and small indels in experimental *Caenorhabditis elegans* populations of differing. *Genetics*. 2019;212:837-854.
- Leushkin EV, Bazykin GA, Kondrashov AS. Strong mutational bias toward deletions in the *Drosophila melanogaster* genome is compensated by selection. *Genome Biol Evol*. 2013;5:514-524.
- Johnson KP. Deletion bias in avian introns over evolutionary timescales. *Mol Biol Evol*. 2004;21:599-602.
- Choi IS, Schwarz EN, Ruhlman TA, et al. Fluctuations in *Fabaceae* mitochondrial genome and content are both ancient and recent. *BMC Plant Biol*. 2019;19:448.
- Vu GTH, Schmutzer T, Bull F, et al. Comparative genome analysis reveals divergent genome evolution in a carnivorous plant genus. *Plant Genome*. 2015;8:doi:10.3835/plantgenome2015.04.0021.

33. Wang GD, Wang Y, Zeng Z, et al. Simulation of chordate intron evolution using randomly generated and mutated base sequences. *Evol Bioinform Online*. 2020;16:1176934320903108.
34. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792-1797.
35. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 2011;28:2731-2739.
36. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*. 1992;9:678-687.
37. Stoltzfus A, Logsdon JM Jr, Palmer JD, Doolittle WF. Intron "sliding" and the diversity of intron positions. *Proc Natl Acad Sci USA*. 1997;94:10739-10744.
38. Logsdon JM Jr, Tyshenko MG, Dixon C, D-Jafari J, Walker VK, Palmer JD. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc Natl Acad Sci USA*. 1995;92:8507-8511.
39. de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc Natl Acad Sci USA*. 1998;95:5094-5099.
40. Wang Y, Tao XF, Su ZX, et al. Current bacterial gene encoding capsule biosynthesis protein CapI contains nucleotides derived from exonization. *Evol Bioinform Online*. 2016;12:303-312.
41. Koonin EV. Evolution of genome architecture. *Int J Biochem Cell Biol*. 2009;41:298-306.
42. Swan BK, Tupper B, Sczyrba A, et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA*. 2013;110:11463-11468.
43. Rose AB, Carter A, Korf I, Kojima N. Intron sequences that stimulate gene expression in *Arabidopsis*. *Plant Mol Biol*. 2016;92:337-346.
44. Marquez Y, Hopfer M, Ayatollahi Z, Barta A, Kalyna M. Unmasking alternative splicing inside protein-coding exons defines exitrans and their role in proteome plasticity. *Genome Res*. 2015;25:995-1007.
45. Thiele A, Nagamine Y, Hauschildt S, Clevers H. AU-rich elements and alternative splicing in the beta-catenin 3' UTR can influence the human beta-catenin mRNA stability. *Exp Cell Res*. 2006;312:2367-2378.
46. Tahmasebi S, Jafarnejad SM, Tam IS, et al. Control of embryonic stem cell self-renewal and differentiation via coordinated alternative splicing and translation of YY2. *Proc Natl Acad Sci USA*. 2016;113:12360-12367.
47. Parenteau J, Abou Elela S. Introns: good day junk is bad day treasure. *Trends Genet*. 2019;35:923-934.