Current Bacterial Gene Encoding Capsule Biosynthesis Protein CapI Contains Nucleotides Derived from Exonization

Yong Wang[1], Xia-Fang Tao[2], Zhi-Xi Su[3], A-Ke Liu[3], Tian-Lei Liu[1], Ling Sun[1], Qin Yao[2], Ke-Ping Chen[2] and Xun Gu[3,4]

[1]School of Food and Biological Engineering, Jiangsu University, Zhenjiang, China. [2]Institute of Life Sciences, Jiangsu University, Zhenjiang, China. [3]School of Life Sciences, Fudan University, Shanghai, China. [4]Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA, USA.

ABSTRACT: Since the proposition of *introns-early* hypothesis, although many studies have shown that most eukaryotic ancestors possessed intron-rich genomes, evidence of intron existence in genomes of ancestral bacteria has still been absent. While not a single intron has been found in all protein-coding genes of current bacteria, analyses on bacterial genes horizontally transferred into eukaryotes at ancient time may provide evidence of intron existence in bacterial ancestors. In this study, a bacterial gene encoding capsule biosynthesis protein CapI was found in the genome of sea anemone, *Nematostella vectensis*. This horizontally transferred gene contains a phase 1 intron of 40 base pairs. The nucleotides of this intron have high sequence identity with those encoding amino acids in current bacterial *CapI* gene, indicating that the intron and the amino acid-coding nucleotides are originated from the same ancestor sequence. Moreover, 5′-splice site of this intron is located in a GT-poor region associated with a closely following AG-rich region, suggesting that deletion mutation at 5′-splice site has been employed to remove this intron and the intron-like amino acid-coding nucleotides in current bacterial *CapI* gene are derived from exonization. These data suggest that bacterial *CapI* gene contained intron(s) at ancient time. This is the first report providing the result of sequence analysis to suggest possible existence of spliceosomal introns in ancestral bacterial genes. The methodology employed in this study may be used to identify more such evidence that would aid in settlement of the dispute between *introns-early* and *introns-late* theories.

KEYWORDS: horizontal gene transfer, Blast search, phylogenetic analysis, intron, exonization

CITATION: Wang et al. Current Bacterial Gene Encoding Capsule Biosynthesis Protein CapI Contains Nucleotides Derived from Exonization. *Evolutionary Bioinformatics* 2016:12 303–312 doi: 10.4137/EBO.S40703.

TYPE: Short Report

RECEIVED: August 05, 2016. **RESUBMITTED:** September 18, 2016. **ACCEPTED FOR PUBLICATION:** September 22, 2016.

ACADEMIC EDITOR: Liuyang Wang, Associate Editor

PEER REVIEW: Eight peer reviewers contributed to the peer review report. Reviewers' reports totaled 2152 words, excluding any confidential comments to the academic editor.

FUNDING: This study was supported by the Scientific Research Promotion Fund for the Talents of Jiangsu University (no. 09 JDG029), the National Natural Science Foundation of China (no. 31572467), the National Basic Research Program (973) of China (no. 2012CB114604), and the Project Funded by Priority Academic Program Development of Jiangsu Higher Education Institutions. 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: ywang@ujs.edu.cn

Introduction

Introns are nucleotide sequences interrupting the coding regions (exons) in a gene, which are frequently seen in eukaryotic genes, and are occasionally identified in prokaryotic rRNA and tRNA genes but are not found in all current prokaryotic mRNA genes. Regarding the origination and evolution of introns, there have been two contrary hypotheses, ie, the *introns-early* and the *introns-late* theories. Introns-early theory proposes that introns already existed in ancestral prokaryotes and intron loss allowed the current organisms to have intronless or intron-poor genomes.[1–3] Introns-late theory holds that introns are an innovation of eukaryotes and intron gain has been a continuous process during the evolution of eukaryotes.[4,5] In the past decades, despite data accumulation and continuous debating, this dispute has not yet been settled.

At the time of proposition, introns-late theory had more evidence, because all current prokaryotic genes are free of spliceosomal introns, being supportive to *introns are an innovation of eukaryotes*, and intron number and length in eukaryotes increase with the complexity of organisms, being supportive to *intron gain has been a continuous process during evolution of eukaryotes*.[5–7] In contrast, introns-early theory had no strong evidence at first, because no introns were found in all prokaryotic protein-coding genes and no data were obtained to show the existence of intron-rich genomes in ancestral eukaryotes. Later, as more and more investigations were conducted, substantial evidence has been found to favor the introns-early theory (see Ref. 8 for detailed review). For instance, it was found that the ancestral eukaryotic forms contained intron-rich genomes[9–11] and evolution of eukaryotic genes primarily involves intron loss with only a few episodes of intron gain.[12–14] However, all these findings are only supportive to *intron loss allowed the current organisms to have intronless or intron-poor genomes* for introns-early theory. So far, no evidence has been obtained to support *introns already existed in ancestral prokaryotes*. In fact, although existing tRNA and rRNA

genes of prokaryotes have been found to possess introns,[15–18] all existing mRNA genes of prokaryotes are free of introns. Therefore, whether ancestral prokaryotic protein-coding genes had introns becomes the focusing point of argument between the *introns-early* and *introns-late* theories. However, investigating whether ancestral prokaryotic mRNA genes had introns is confronted with certain difficulty, because ancestral prokaryotes are not available today.

Although ancestral prokaryotes are not available today, it is possible to find gene samples of ancestral prokaryotes for study because some prokaryotic genes might have been deposited in eukaryotes through horizontal gene transfer at ancient time. Horizontal gene transfer is an evolutionary phenomenon that involves the transfer of genes between different species,[19,20] and various criteria have been established to identify and validate horizontally transferred genes (HTGs) between certain organisms.[21] In recent years, as public databases have accumulated a large number of nucleotide and protein sequences, basic local alignment search tool (Blast) searches and phylogenetic analysis have been frequently used to identify HTGs between prokaryotic and eukaryotic genomes[22–25] and three bacterial protein-coding regions horizontally transferred into bdelloid rotifers have been found to possess spliceosomal introns.[22] However, no further analysis has been conducted to analyze origination of these introns.

In present study, we employed Blast searches and phylogenetic analyses to identify intron-containing bacterial HTGs harbored in eukaryotes and further analyzed origination and evolution of the introns in HTGs. As a result, a bacterial HTG containing one intron was found in sea anemone. Nucleotides of this intron have high sequence identity with those that encode amino acids in current bacterial gene. Further analyses revealed that this intron is originated from the donor bacterium but not created by the recipient sea anemone, suggesting that bacterial *CapI* gene contained intron(s) at ancient time.

Materials and Methods

Primary protein Blast search. In our study, protein Blast (Blastp) search, phylogenetic analysis, gene structure, and sequence comparison were employed to identify bacterial HTGs harbored in eukaryotes (Fig. 1). First, a primary Blastp search was conducted to obtain eukaryotic proteins having high identity with bacterial proteins. For this search, all 7,518 protein sequences of a cyanobacterial strain named *Anabaena variabilis* ATCC 29413 were used to conduct Blastp searches against eukaryotic protein databases at http://blast.ncbi.nlm.nih.gov/Blast.cgi by setting *E* (expect) value at 1e−50 and limiting target organisms as *eukaryotes*. Each search result was examined to keep bacterial proteins with less than 10 hit eukaryotic proteins. Here, 10 is a value arbitrarily selected to reduce the number of eukaryotic hit proteins for later analysis by considering that a potential HTG should be confined within a limited number of eukaryotic species. The primary Blastp searches generated a list of eukaryotic proteins with high similarity to cyanobacterial proteins (ie, $E < 1e{-}50$).

Secondary Blastp search and calculation of alien index. In the second Blastp search, each of the eukaryotic proteins obtained from primary Blastp search was used to conduct Blastp searches at http://blast.ncbi.nlm.nih.gov/Blast.cgi by setting *E* value at 0.1 and limiting target organisms as *bacteria* and *eukaryotes*, respectively. *E* values of the best-hit bacterial and best-hit eukaryotic proteins were used to calculate the alien index (AI) of the concerned protein by using the formula developed by Gladyshev et al.[22], ie, AI = log[(*E* value of best-hit eukaryotic protein) + e−200] − log[(*E* value of best-hit bacterial protein) + e−200]. Here, if no hits were found, the *E* value was set to 1. Accordingly, a protein currently existing in eukaryotes with an AI value of over 45 is considered to be a potential HTG-encoded protein.[22]

Examination of intron locations. Whether the coding sequence of a specific eukaryotic protein has intron was found out by first using the protein accession number as query item and then selecting *Gene* as target database in GenBank (http://www.ncbi.nlm.nih.gov/). If two or more exons are shown in the resultant webpage, select *Sequence Text View* in *Tools* option to view location and phase information of specific introns.

Calculation of GT and AG contents. GT and AG contents are the presences of GT and AG among all checked two-letter locations in a given nucleotide sequence. For example, if GT is present six times in a nucleotide sequence of 100 bases, its GT content is calculated as 6.1% (ie, 6 divided by 99), because the 100 base sequence has 99 two-letter locations to check. Besides, if the four bases are completely randomly distributed in a nucleotide sequence, the normalized content of each of the 16 dinucleotide combinations will be equal or close to 6.25% (ie, 1 divided by 16).

Sequence alignment and phylogenetic analysis. Multiple sequence alignment of nucleotides and proteins was conducted using Muscle program[26] embedded in MEGA 5.10.[27] The aligned protein sequences were used to construct phylogenetic trees with minimum evolution algorithm in MEGA 5.10. The aligned nucleotide sequences were shaded using GeneDoc Multiple Sequence Alignment Editor and Shading Utility (Version 2.6.02)[28] and copied to rich text file for further annotation.

Results

HTGs in *Nematostella vectensis* and *Batrachochytrium dendrobatidis*. From an extensive screening using 7,518 *A. variabilis* proteins as query to identify intron-containing genes horizontally transferred from bacteria into eukaryotes (Fig. 1), we found that one protein sequence, ie, XP_001618246.1, of the sea anemone *N. vectensis* (designated as NvHTG1) is coded by an intron-containing HTG of bacteria. Besides, XP_006683402.1 of the chytrid *B. dendrobatidis* (designated as BdHTG1) is also coded by HTG of bacteria because they meet the multiple criteria established

to validate horizontal gene transfer events.[21] First, these two proteins have very high sequence identity with bacterial capsule biosynthesis protein CapI. Blastp search using NvHTG1 and BdHTG1 as query sequences revealed that both *E* values of their best-hit bacterial proteins are 0.0 among which NvHTG1 has 85% sequence identity and 99% query coverage, and BdHTG1 has 79% sequence identity and 97% query coverage with bacterial CapI, while their best-hit eukaryotic proteins have *E* values of 7e−115 and 1e−145, respectively. After using the formula developed by Gladyshev et al.[22] to calculate AI, it was found that NvHTG1 and BdHTG1 have the AI value of 198 and 127, respectively, far beyond the set threshold value (45) for being potential HTGs.

Second, NvHTG1 and BdHTG1 have very close phylogenetic relationship with bacterial CapI protein. This is evident as shown in the evolutionary tree constructed using these two proteins with their best-hit 100 bacterial and best-hit 100 eukaryotic proteins (Fig. 2 and Supplementary File 1). In the evolutionary tree, NvHTG1 and BdHTG1 are located in the phyletic clade formed by CapI proteins of CFB group bacteria and firmicutes, respectively (Fig. 2, clades I and II), demonstrating that *NvHTG*1 and *BdHTG*1 have the latest common ancestor with bacterial *CapI* gene. Besides, protein sequence XP_001618247.1 of *N. vectensis* (designated as NvHTG2), which is linked with NvHTG1, and protein sequences XP_006683399.1 and XP_006683401.1 of *B. dendrobatidis* (designated as BdHTG2 and BdHTG3), which are linked with BdHTG1 (Fig. 3), are also found to have the latest common ancestor with correspondent bacterial genes. This was observed by using NvHTG2, BdHTG2, and BdHTG3 as query sequences to conduct Blastp searches and constructing evolutionary trees with their best-hit bacterial

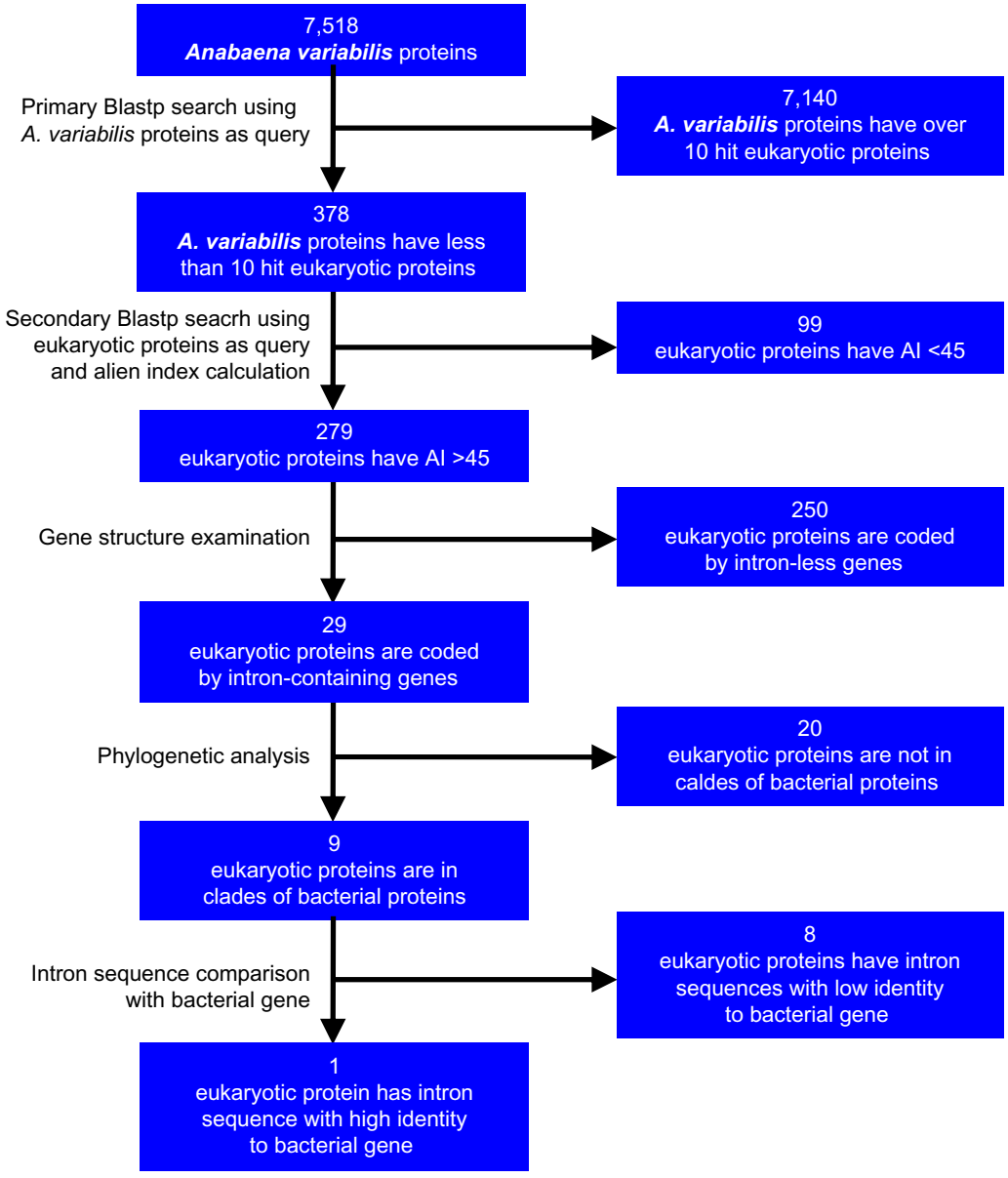

Figure 1. Flowchart of the steps and result in searching bacterial HTGs harbored in eukaryotes.

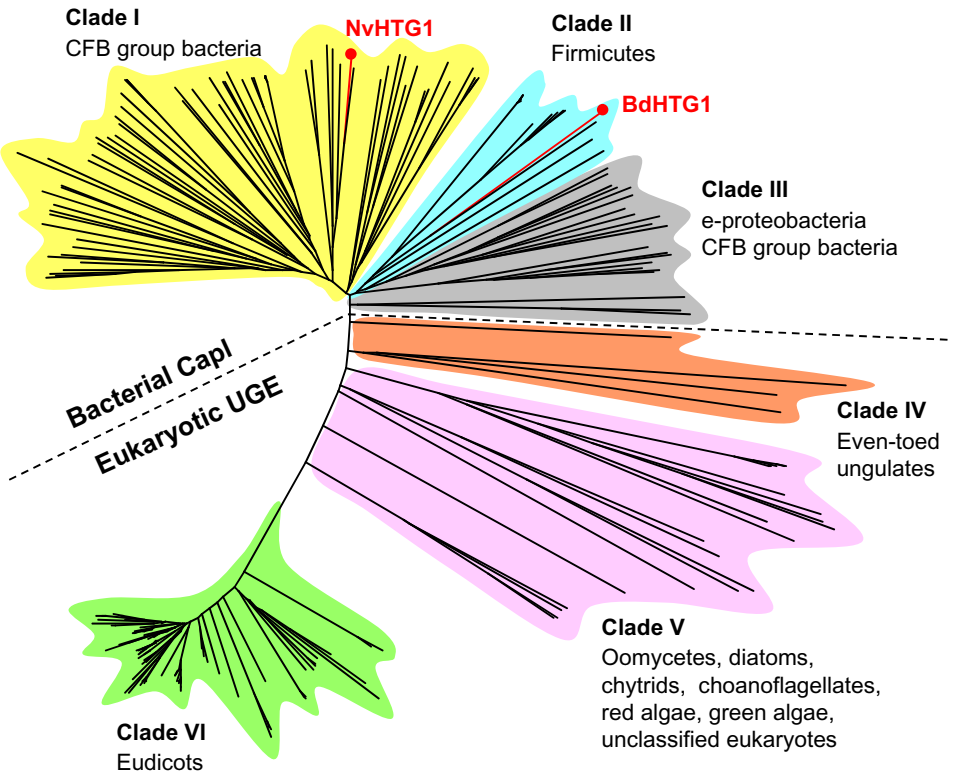

Figure 2. Phylogenetic relationship between bacterial capsule biosynthesis protein CapI and eukaryotic UGE. Shown here is a minimum evolution tree constructed using two HTGs (NvHTG1 and BdHTG1) harbored in *Nematostella vectensis* (*Nv*) and *Batrachochytrium dendrobatidis* (*Bd*), 100 bacterial CapI protein sequences, and 100 eukaryotic UGE protein sequences. Phylogenetic clades I–III are formed by bacterial CapI protein together with NvHTG1 and BdHTG1, and phylogenetic clades IV–VI are formed by eukaryotic UGE protein. Clades I and II are circumscribed to indicate locations of NvHTG1 and BdHTG1 among bacterial CapI sequences of CFB group bacteria and firmicutes, respectively. Clade III includes CapI proteins from e-proteobacteria and CFB group bacteria. Clades IV–VI include UGE proteins from even-toed ungulates, a variety of eukaryotes, and eudicots. From clades II to V, more and more sequences are originated from stem of the tree, showing increased divergence between sequences in the same clade. Multiple sequence alignment of all protein sequences is shown in Supplementary File 1. Total length of the alignment is 795. Amino acids from 219 to 585 in the alignment are used for above tree construction.

and best-hit eukaryotic protein sequences. Each of the three constructed trees had the similar topology with Figure 2 (data not shown).

Third, gene locations of these two proteins and their neighboring proteins are conserved in bacteria. Figure 3 shows gene locations of two NvHTGs, three BdHTGs, and their neighboring proteins. In *N. vectensis*, the coding region of NvHTG2 is on upstream of NvHTG1. Sequence comparison revealed that NvHTG2 is the C-terminal portion of bacterial UDP-glucose 6-dehydrogenase (UGD). It was found that homologs of these two linked NvHTGs are also linked in 28 bacterial strains, though in various patterns (Fig. 3A). In *B. dendrobatidis*, XP_006683402.1 is flanked by BdHTG2 and BdHTG3 of which best-hit bacterial proteins are the two members of glycosyltransferase family 2 (designated as GT2a and GT2b). Our examinations to the locations of homologous bacterial genes of these three linked BdHTGs displayed that four strains of firmicutes have similar conserved gene locations (Fig. 3B). Taken together, the conservatism of gene locations between NvHTGs, BdHTGs, and many other bacterial strains allows us to determine the direction of horizontal gene transfer, ie, from bacteria into the ancestor of sea anemones and chytrids.[21]

Exonized nucleotides in bacterial *CapI* gene. Our above analyses verify that proteins NvHTG1 and BdHTG1 are coded by bacterial horizontally transferred *CapI* gene. Nevertheless, coding region of NvHTG1 is predicted to have a phase 1 intron of 40 base pairs (bp), while that of BdHTG1 has no intron (Fig. 3). Where does the intron in *NvHTG1* gene come from? Was it brought by the donor bacterium during horizontal gene transfer or inserted/created by the recipient sea anemone posterior to the gene transfer? A comparison on genomic DNAs between *NvHTG1* and bacterial *CapI* genes led us to conclude that this intron sequence should have been brought by the donor bacterium. In the aligned genomic DNA sequences of *NvHTG1* and *CapI* genes, the intron in *NvHTG*1 is located at sites 218–257 (Fig. 4A). The predicted intron length (40 bp) is not a multiple of 3. Therefore, if the *NvHTG1* gene is to yield a mature mRNA sequence with correct reading frame, an intron should be removed from its pre-mRNA. This validates the existence of an intron at or near that location. As we can see, apart from the presence of GT at

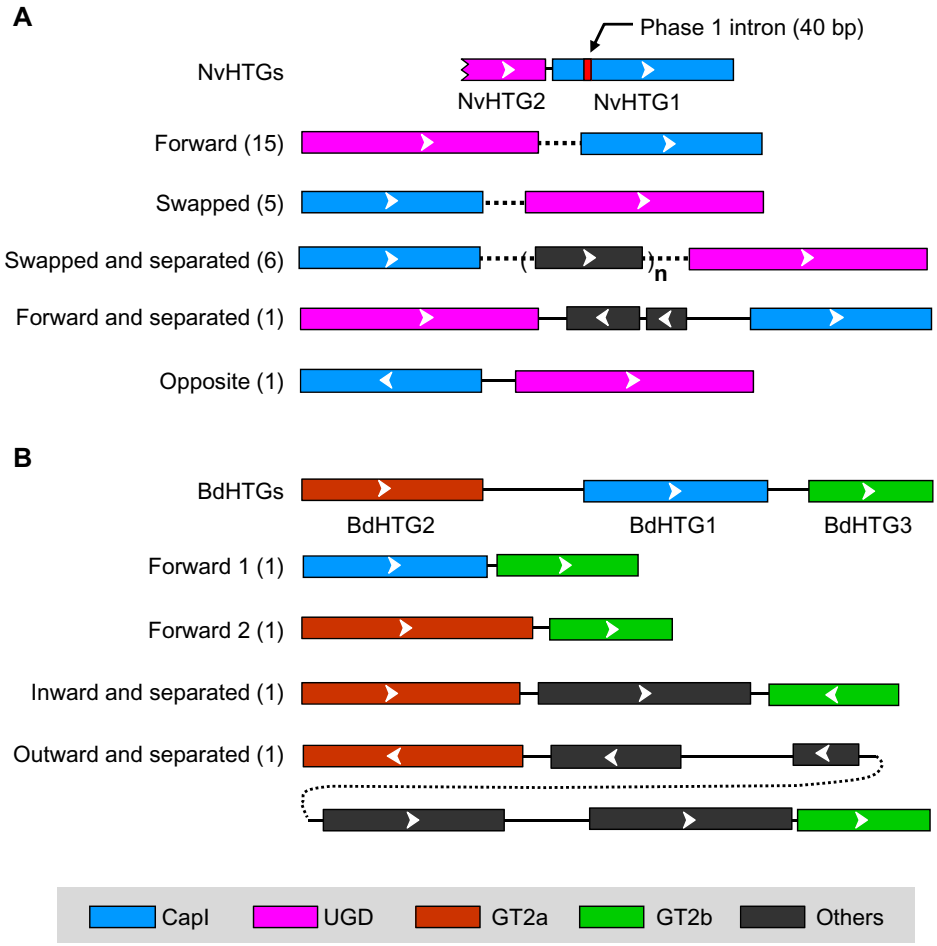

Figure 3. Comparison of protein coding regions between HTGs and bacteria. (**A**) Linked locations of coding regions for two NvHTGs, bacterial CapI (blue) and bacterial UGD (magenta). (**B**) Linked locations of coding regions for three BdHTGs, bacterial CapI (blue), bacterial GT2a (brown), and GT2b (green). Value in brackets indicates number of bacterial strains having the correspondent genomic structural pattern. Bars and solid lines are drawn to scale according to the lengths of protein and nucleotide sequences, respectively. Dotted lines represent different lengths in different bacterial strains. Bars of the same color (except black) indicate homology between protein sequences ($n = 1–4$).

site 218, it is also present at sites 200 and 203 (Fig. 4A), which displays that these two GTs may also be used as 5′-splice site because the resultant intron will be increased by 18 and 15 bp, both of which do not interrupt the correct reading frame of mature mRNA. Moreover, the predicted intron has high sequence identity with the correspondent region (sites 221–257) of bacterial *CapI* gene. It would be very unlikely that this intron sequence was inserted by the recipient sea anemone, otherwise it should not be able to have such high sequence identity with nucleotides of *CapI* gene at sites 221–257 (Fig. 4A). Thus, these 40 bp nucleotides should have been carried and sent to the recipient organism by the donor bacterium during horizontal gene transfer.

Existence of eukaryotic introns in highly varied regions. As shown in Figure 4, the genomic DNA sequences of NvHTG1, BdHTG1, and 10 bacterial *CapI* genes have two highly varied regions. Although NvHTG1 has an intron of which 5′-splice site is located in the first highly varied region (Fig. 4A), no intron is found in the second highly varied region (Fig. 4B). Because the first highly varied region is where an intron possibly existed, could the second highly varied region be possible location of another intron? In order to find out the answer of this question, we examined the intron locations in coding regions of all 100 eukaryotic UDP-glucuronate 4-epimerase (UGE) proteins, which are homologous to bacterial CapI as shown in Figure 2. It is found that 9 of the 100 eukaryotic UGEs have coding regions containing introns in their open reading frames. Among them, coding regions of SaUGE and CsUGE have a phase 1 intron in this region (Fig. 5A), and coding regions of SaUGE and CrUGE have a phase 0 intron in the second highly varied region (Fig. 5C).

Discussion

Ever since the proposition of *introns-early* hypothesis, the absence of evidence to show that ancestral bacterial protein-coding genes had introns has long been the major obstacle to

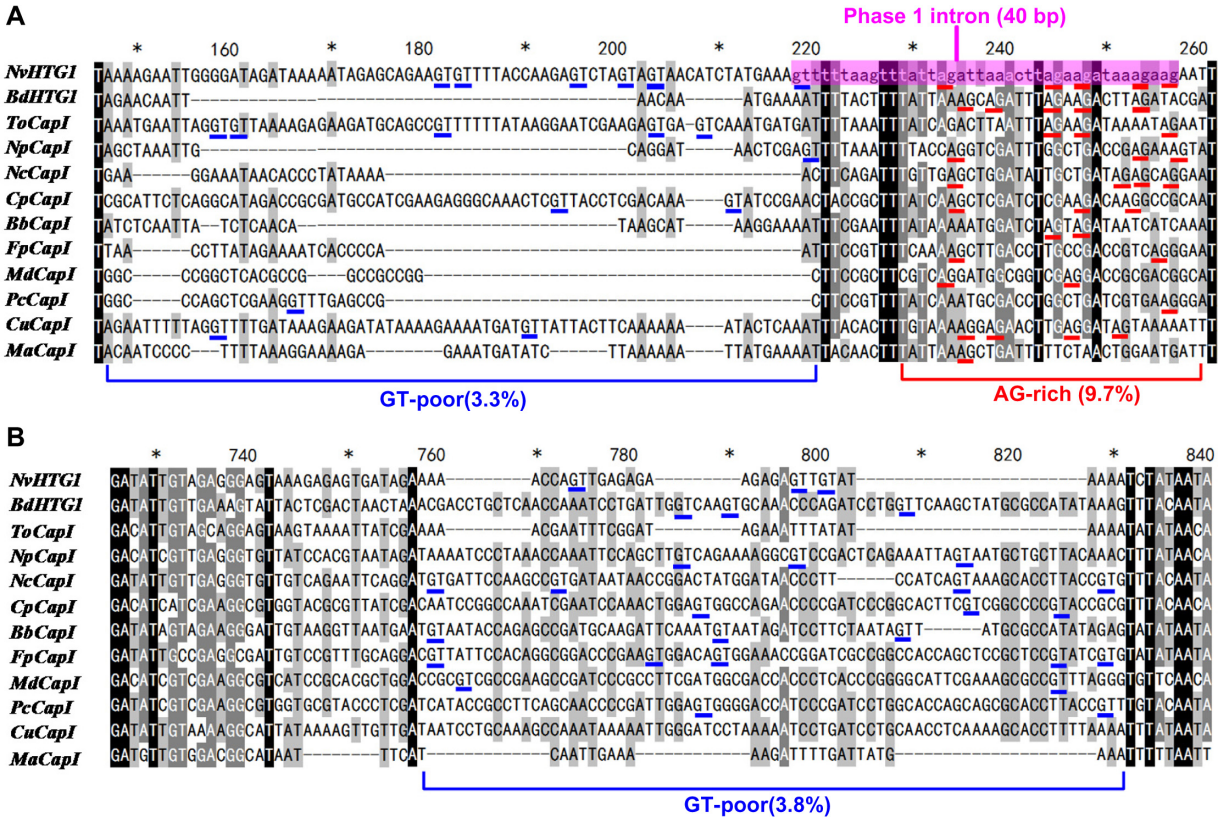

Figure 4. Multiple sequence alignment of genomic DNAs from NvHTG1, BdHTG1, and ten bacterial *CapI* genes. (**A**) Aligned nucleotide sequences from sites 146 to 261. (**B**) Aligned nucleotide sequences from sites 726 to 840. Please refer to Supplementary File 2 for aligned sequences containing all sites.
Notes: Sequence names are abbreviated using two letters to indicate species name followed by gene name. *Nv* stands for *Nematostella vectensis*, a species of sea anemones. *Bd* stands for *Batrachochytrium dendrobatidis*, a species of chytrids. *To* stands for *Tenacibaculum ovolyticum*, a strain of CFB group bacteria. *Np* stands for *Nostoc punctiforme*, a strain of cyanobacteria. *Nc* stands for *Neptuniibacter caesariensis*, a strain of g-proteobacteria. *Cp* stands for *Chlorobaculum parvum*, a strain of green sulfur bacteria. *Bb* stands for *Bacillus bombysepticus*, a strain of firmicutes. *Fp* stands for *Franconibacter pulveris*, a strain of enterobacteria. *Md* stands for *Methyloversatilis discipulorum*, a strain of b-proteobacteria. *Pc* stands for *Pelobacter carbinolicus*, a strain of d-proteobacteria. *Cu* stands for *Campylobacter ureolyticus*, a strain of e-proteobacteria. *Ma* stands for *Methanococcus aeolicus*, a strain of euryarchaeotes (archaebacteria).

persuade the opponent and also to convince the proponent of this hypothesis. As a matter of fact, finding such evidence has confronted great difficulty because all current bacterial protein-coding genes are free of introns. And, of course, all ancestral bacteria are not available today. Fortunately, a certain number of horizontal gene transfer events occurred between bacterial and eukaryotic species at ancient time, most of which involved gene transfer from bacteria into eukaryotes. Such gene transfers allowed ancestral bacteria to *deposit* certain genes in ancestral eukaryotic cells, which have not undergone intensive genome streamlining like bacteria[2,29] and have thus kept the deposited bacterial genes more or less as they were received.

Our present study identified two and three linked bacterial genes that have been horizontally transferred into sea anemone and chytrid, respectively, which is evident because they meet multiple criteria established to identify and validate a horizontal gene transfer event.[21] Moreover, one of the genes (ie, *NvHTG1*) has a phase 1 intron of 40 bp, which has high sequence identity with bases encoding amino acids in current bacterial *CapI* genes. This high sequence identity strongly suggests that the 40 bp intron sequence was brought by the donor bacterium. Yet, it does not necessarily mean that ancestral bacterial *CapI* gene had intron in its form as in *NvTHG1*, because this 40 bp intron could have been created by the recipient sea anemone through base mutation or other genome amelioration activities. To address this concern, we conducted detailed examination to the genomic DNA sequences of NvHTG1, BdHTG1, and 10 bacterial *CapI* genes. Multiple sequence alignment displayed that their genomic DNA sequences are greatly conserved, with only two highly varied regions located at sites 147–220 and sites 758–831, respectively (Supplementary File 2 and Fig. 4). The 5′-splice site (GT) of NvHTG1 intron is located in the first highly varied region (Fig. 4A). This led us to suppose that this highly varied region results from frequent base mutation and/or base deletion in ancestral bacteria in order to remove the 5′-splice site. This is probably what happened indeed, because this region has a very low GT content. As we have calculated, the average GT content in all the 12 aligned genomic DNA sequences

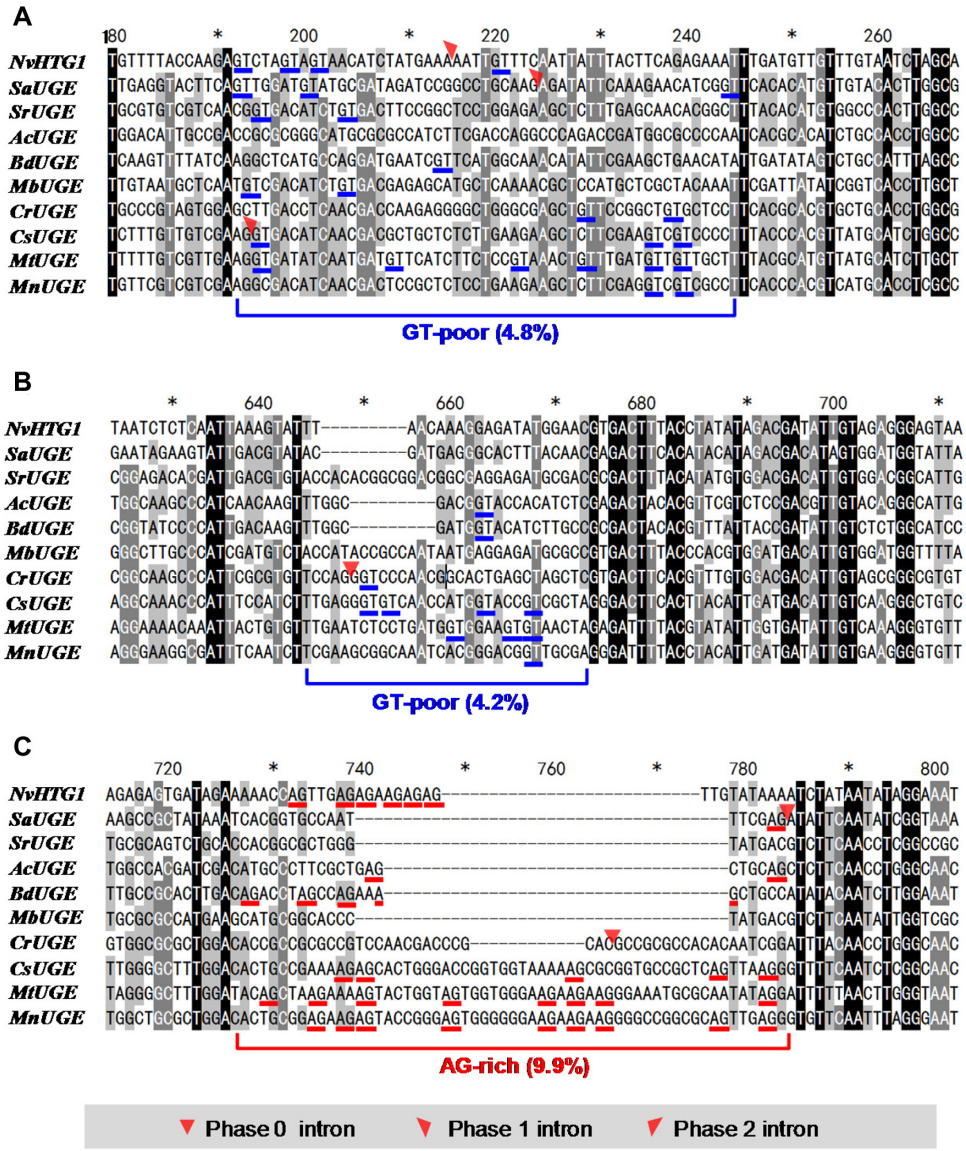

Figure 5. Multiple sequence alignment of exonic DNAs from NvHTG1 and nine eukaryotic UGE genes. (**A**) Aligned nucleotide sequences from sites 179 to 267. (**B**) Aligned nucleotide sequences from sites 624 to 712. (c) Aligned nucleotide sequences from sites 713 to 801. Please refer to Supplementary File 3 for aligned sequences containing all sites.

Notes: Sequence names are abbreviated using two letters to indicate species name followed by gene name. *Nv* stands for *Nematostella vectensis*, a species of sea anemones. *Sa* stands for *Sphaeroforma arctica*, a species of unclassified eukaryotes. *Sr* stands for *Salpingoeca rosetta*, a species of choanoflagellates. *Ac* stands for *Acanthamoeba castellanii*, a species of unclassified eukaryotes. *Bd* stands for *Batrachochytrium dendrobatidis*, a species of chytrids. *Mb* stands for *Monosiga brevicollis*, a species of choanoflagellates. *Cr* stands for *Chlamydomonas reinhardtii*, a species of green algae. *Cs* stands for *Camelina sativa* (false flax), a species of eudicots. *Mt* stands for *Medicago truncatula* (barrel medic), a species of eudicots. *Mn* stands for *Morus notabilis*, a species of eudicots. NvHTG1 sequence is added here for comparing the aligned nucleotides with Figure 4.

shown in Supplementary File 2 is 5.1%, a value slightly lower than the normalized dinucleotide content (6.25%). But in this region, the GT content is 3.3%. And if NvHTG1 is excluded since the intron is still present, this value (of the rest 11 intron-free sequences) is as low as 2.2%. What is more, nucleotides of *BdHTG1*, *NcCapI*, *BbCapI*, *FpCapI*, *MdCapI*, and *MaCapI* genes in this region have no GT at all. These six genes are typified by having relatively less nucleotides in this region, suggesting that their GTs had been removed through base deletion.

In association with the GT-poor region, there is an AG-rich region at sites 229–260 (Fig. 4A). The AG content in this region is 9.7%, being remarkably higher than the average AG content (5.9%) in all the 12 aligned sequences. The presence of GT-poor region followed by AG-rich region in bacterial *CapI* gene suggests that ancestral bacterial *CapI* gene had an intron at this location that had been removed through mutating the 5′-splice site. Therefore, although direct evidence is still not available to prove the existence of introns in ancestral bacterial genes, our data provide an example of possible intron existence in ancestral bacterial *CapI* gene. Removal of this hypothesized intron had left a GT-poor region, an AG-rich region, and exonized nucleotides in current bacterial *CapI* gene. This

can be considered as a new line of evidence to support the introns-early theory. In the past decades, while various lines of evidence were reported to support either the introns-early theory or the introns-late theory, many of the reported data were either questioned or integrated to reach a reconciled solution for the the introns-early versus introns-late dispute. For example, Hurst and McVean[30] regarded that the location of introns in the same position in both nuclear and organellar genes could be supportive evidence for both introns-early and introns-late hypotheses depending on whether intron insertion occurs at random or not. Koonin examined a great many of the data claimed to support either the introns-early hypothesis or the introns-late hypothesis[31–38] and proposed a compromise solution for the introns-early or introns-late disputes.[39] Koonin's solution could well explain the origin and evolution of introns during eukaryogenesis. Yet, the question of whether ancestral bacterial protein-coding genes have spliceosomal introns remains unclear. Our above analysis provides a new angle to address this question.

Our data also showed that coding regions of SaUGE, CrUGE, and CsUGE have introns in both the first and second highly varied regions (Fig. 5). Since bacterial CapI and eukaryotic UGE are homologous, intron existence in the second highly varied region in UGE suggests that ancestral bacterial *CapI* gene probably also had an intron in this region and intron removal has left the highly varied nucleotides in this region (Fig. 4B). However, this second highly varied region of bacterial CapI has a relatively higher GT content (3.8%) and no associated downstream AG-rich region (Fig. 4 and Supplementary File 2). Therefore, whether the presumably existed intron at the second highly varied region was removed using the same mechanism mentioned earlier is not clear.

In contrast to coding regions of bacterial CapI that have low GT content in the second highly varied region (Fig. 4B), those of eukaryotic UGE have high AG content in this second highly varied region (Fig. 5C). This AG-rich region has an AG content of 9.9%, about 90% higher than the average (5.2%) in all the 10 aligned sequences (see Supplementary File 3 for nucleotides of the 10 sequences). Moreover, an associated GT-poor region is present at about 50 bp upstream (Fig. 5B). This region has a GT content of 4.2%, about 26% lower than the average (5.7%) of all the 10 aligned sequences. This coexistence of GT-poor and AG-rich regions suggests that the nucleotides of UGEs at sites 644–784 may also come from exonized intron sequences. And, the exonization probably has been realized mainly through base deletion, because nucleotide sequences containing less GT and/or AG at these two regions are generally shorter than the other sequences, eg, SaUGE, SrUGE, AcUGE, BdUGE, and MbUGE (Fig. 5B and C).

In correspondence to the first highly varied region in coding regions of bacterial CapI (Fig. 4A), there is also a highly varied region in those of eukaryotic UGE (Fig. 5A). What is more, coding regions of SaUGE and CsUGE have a phase 1 intron in this region, same like that of NvHTG1. However, this region has a GT content of 4.8%, only slightly lower than the average (5.7%). Besides, no associated AG-rich region is found downstream. Therefore, how the intron in this region of eukaryotic UGE gene was removed remains for future exploration.

In searching for bacterial HTGs harbored in eukaryotes, we have used 1e-50 and 0.1 as the cutoff *E* value for our primary and secondary Blastp searches, respectively. This approach seems to yield certain false discovery rate. In our case, among the 29 eukaryotic proteins with an AI value of over 45, only 9 are verified to be HTGs by phylogenetic analyses (Fig. 1). Therefore, to validate a horizontal gene transfer event, multiple measures including calculation of AI value, phylogenetic analysis, and gene structure comparison must be employed.[21] Meanwhile, to reduce the number of potential HTGs for later phylogenetic and gene structure analyses, the AI cutoff value can be raised to 100.

In the evolution of eukaryotic genes, both intron gain and intron loss have been found to occur in specific organisms, and a number of mechanisms have been proposed to explain how intron gain and loss happen. For example, intron gain has been found to occur through intron retrotransposition in fungi,[40] through intronization of exonic sequences in fruit fly,[41] and through double-strand break repair in *Daphnia pulex*.[42] On the other hand, intron loss was found to happen through genomic deletion in seven eukaryotic genomes[43] and through reverse transcription of mRNA followed by homologous recombination in human.[44] Here, we present two examples of intron loss through mutation at splice sites, which is in relation with genomic deletion. First, the presumably existed intron in bacterial *CapI* gene was finally removed mainly through deleting its 5′-splice site (Fig. 4A). Second, one of the introns in eukaryotic UGE gene was finally removed mainly through deleting both 5′- and 3′-splice sites (Fig. 5B and C). However, this mechanism would probably only work when the intron is quite short so that the left nucleotides after removal of its 5′- and/or 3′-splice sites have less chances to contain a stop codon. Therefore, a long intron may first be reduced to a short length, say, less than 100 bp, through other mechanisms. Then, mutation at splice sites may be used as the final step to remove the intron. Once the splice site(s) are successfully removed, the remaining nucleotides may then be used to encode amino acids and thus become exonized.

Upon completion of the above analyses, we also made additional attempts to identify more such horizontal gene transfer events that could have left *intron-like sequences* in current bacteria. However, our Blastp searches using all protein sequences of three other bacterial strains against eukaryotic protein databases failed in returning such a new event. Here, we consider our present work as of genomic archeology, which involves discovery of nucleotide sequences left by ancestral bacterial strains during the evolution of their genomes. It seems that such sequences are not present in high number. Therefore, in this report, we could only present this single

example of possible existence of intron in bacterial *CapI* gene. We anticipate more discoveries of such sequences in various genomes of bacteria in the future.

Conclusions

In this study, we identified an intron in HTG harbored in sea anemone. This intron has high sequence identity with bases encoding amino acids in current bacterial *CapI* genes. 5′-Splice site of this intron is located at a GT-poor region followed by an AG-rich region in *CapI* gene. These data led us to conclude that an intron could have existed in ancestral bacterial *CapI* gene and mutation at 5′-splice site had been employed to remove this intron, which had left exonized nucleotides in current *CapI* gene. This is the first report providing the result of sequence analysis to suggest possible existence of introns in ancestral bacterial protein-coding genes. The methodology employed in this study may be used to identify more such evidence that would aid in settlement of the dispute between *introns-early* and *introns-late* theories.

Acknowledgment

We are thankful to the National Center for Biotechnology Information (www.ncbi.nlm.nlh.gov) for free access to various sequence databases.

Author Contributions

Performed the overall analysis and wrote the article: YW. Did the Blast searches and phylogenetic analyses: X-FT, A-KL, T-LL, LS. Proposed and conceived the work: Z-XS, XG, QY, K-PC. All the authors discussed and put forward views on the results and approved the final article.

Supplementary Material

Supplementary File 1. Multiple sequence alignment of NvHTG1, BdHTG1, 100 bacetial CapI proteins, and 100 eukaryotic UGE proteins. XP_001618246.1 and XP_006683402.1 are the accession numbers of NvHTG1 and BdHTG1, respectively. Accession numbers beginning with *WP* are of bacterial proteins. Those beginning with *XP* or *NP* are of eukaryotic proteins. Total length of the alignment is 795. The phylogenetic tree shown in Figure 2 was constructed using the aligned amino acids from sites 219 to 585.

Supplementary File 2. Multiple sequence alignment of genomic DNAs from NvHTG1, BdHTG1, and 10 bacterial *CapI* genes. All aligned sequences are of genomic DNAs coding for NvHTG1 (HTG harbored in *Nematostella vectensis*, a species of sea anemones), BdHTG1 (HTG harbored in *Batrachochytrium dendrobatidis*, a species of chytrids), and capsule biosynthesis protein CapI from *To* (*Tenacibaculum ovolyticum*, a strain of CFB group bacteria), *Np* (*Nostoc punctiforme*, a strain of cyanobacteria), *Nc* (*Neptuniibacter caesariensis*, a strain of g-proteobacteria), *Cp* (*Chlorobaculum parvum*, a strain of green sulfur bacteria), *Bb* (*Bacillus bombysepticus*, a strain of firmicutes), *Fp* (*Franconibacter pulveris*, a strain of enterobacteria), *Md* (*Methyloversatilis discipulorum*, a strain of b-proteobacteria), *Pc* (*Pelobacter carbinolicus*, a strain of d-proteobacteria), *Cu* (*Campylobacter ureolyticus*, a strain of e-proteobacteria), and *Ma* (*Methanococcus aeolicus*, a strain of euryarchaeotes, which is of archaebacteria). Each sequence has its own start and stop codons. Lowercase letters of NvHTG1 aligned at sites 218–257 are of a phase 1 intron of 40 bp. Two highly varied regions each of which does not have identical bases among 50 aligned sites can easily be seen at sites 147–220 and at sites 758–831.

Supplementary File 3. Multiple sequence alignment of exonic DNAs from NvHTG1 and nine eukaryotic UGE genes. All aligned sequences are of exonic DNAs coding for NvHTG1 (one HTG harbored in *Nematostella vectensis*, a species of sea anemones) and eukaryotic UGE from *Sa* (*Sphaeroforma arctica*, a species of unclassified eukaryotes), *Sr* (*Salpingoeca rosetta*, a species of choanoflagellates), *Ac* (*Acanthamoeba castellanii*, a species of unclassified eukaryotes), *Bd* (*Batrachochytrium dendrobatidis*, a species of chytrids), *Mb* (*Monosiga brevicollis*, a species of choanoflagellates), *Cr* (*Chlamydomonas reinhardtii*, a species of green algae), *Cs* (*Camelina sativa*, the false flax, a species of eudicots), *Mt* (*Medicago truncatula*, the barrel medic, a species of eudicots), and *Mn* (*Morus notabilis*, a species of eudicots). Coding sequence of NvHTG1 is added here for comparing the aligned nucleotides with Figure 4. Locations of various phase introns are marked using small red triangles. The two highly varied regions each of which does not have identical bases among aligned 50 sites can easily be seen at sites 192–241 and at sites 726–784.

REFERENCES

1. Doolittle WF. Genes in pieces: were they ever together? *Nature*. 1978;272:581–2.
2. Gilbert W. The exon theory of genes. *Cold Spring Harb Symp Quant Biol*. 1987;52:901–5.
3. Roy SW. Recent evidence for the exon theory of genes. *Genetica*. 2003;118:251–66.
4. Doolittle WF, Stoltzfus A. Molecular evolution: genes-in-pieces revisited. *Nature*. 1993;361:403.
5. Logsdon JM Jr. The recent origins of spliceosomal introns revisited. *Curr Opin Genet Dev*. 1998;8:637–48.
6. Penny D, Hoeppner MP, Poole AM, Jeffares DC. An overview of the introns-first theory. *J Mol Evol*. 2009;69:527–40.
7. Rodríguez-Trelles F, Tarrío R, Ayala FJ. Origins and evolution of spliceosomal introns. *Annu Rev Genet*. 2006;40:47–76.
8. Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. *Biol Direct*. 2012;7:11.
9. 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;13:1512–7.
10. Csuros M. Likely scenarios of intron evolution. *Comparat Gen Lecture Notes Comput Sci*. 2005;3678:47–60.
11. Niu DK, Hou WR, Li SW. mRNA-mediated intron losses: evidence from extraordinarily large exons. *Mol Biol Evol*. 2005;22:1475–81.
12. Carmel L, Wolf YI, Rogozin IB, Koonin EV. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res*. 2007;17:1034–44.
13. 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.
14. Yang YF, Zhu T, Niu DK. Association of intron loss with high mutation rate in *Arabidopsis*: implications for genome size evolution. *Genome Biol Evol*. 2013;5:723–33.
15. Jackson S, Cannone J, Lee J, Gutell R, Woodson S. Distribution of rRNA introns in the three-dimensional structure of the ribosome. *J Mol Biol*. 2002;323:35–52.

16. Kjems J, Garrett R. Novel splicing mechanism for the ribosomal RNA intron in the archaebacterium *Desulfurococcus mobilis*. *Cell*. 1988;54:693–703.
17. Marck C, Grosjean H. Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. *RNA*. 2003;9:1516–31.
18. Salman V, Amann R, Shub DA, Schulz-Vogt HN. Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. *Proc Natl Acad Sci U S A*. 2012;109:4203–8.
19. Syvanen M. Cross-species gene transfer; implications for a new theory of evolution. *J Theor Biol*. 1985;112:333–43.
20. Sprague GF Jr. Genetic exchange between kingdoms. *Curr Opin Genet Dev*. 1991;1:530–3.
21. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*. 2001;55:709–42.
22. Gladyshev EA, Meselson M, Arkhipova IR. Massive horizontal gene transfer in bdelloid rotifers. *Science*. 2008;320:1210–3.
23. Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*. 2008;9:605–18.
24. Richards TA, Soanes DM, Foster PG, Leonard G, Thomton CR, Talbot NJ. Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. *Plant Cell*. 2009;21:1897–911.
25. Zhu B, Lou MM, Xie GL, et al. Horizontal gene transfer in silkworm, *Bombyx mori*. *BMC Genomics*. 2011;12:248.
26. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–7.
27. 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;28:2731–9.
28. Nicholas KB, Nicholas HB Jr, Deerfield DW II. GeneDoc: analysis and visualization of genetic variation. *Embnet News*. 1997;4:14.
29. Koonin EV. Evolution of genome architecture. *Int J Biochem Cell Biol*. 2009; 41:298–306.
30. Hurst LD, McVean GT. Molecular evolution: a difficult phase for introns-early. *Curr Biol*. 1996;6(5):533–6.
31. Fedorov A, Suboch G, Bujakov M, Fedorova L. Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res*. 1992;20(10):2553–7.
32. de Souza SJ, Long MY, Klein RJ, Roy S, Lin S, Gilbert W. Toward a resolution of the introns earlyylate debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc Natl Acad Sci U S A*. 1998;95:5094–9.
33. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. Evidence of splice signal migration from exon to intron during intron evolution. *Curr Biol*. 2003; 13:2170–4.
34. Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet*. 2006;7(3):211–21.
35. Stoltzfus A, Spencer DF, Zuker M, Logsdon JMJ, Doolittle WF. Testing the exon theory of genes: the evidence from protein structure. *Science*. 1994; 265(5169):202–7.
36. Rzhetsky A, Ayala FJ, Hsu LC, Chang C, Yoshida A. Exon/intron structure of aldehyde dehydrogenase genes supports the "introns-late" theory. *Proc Natl Acad Sci U S A*. 1997;94(13):6820–5.
37. Cho G, Doolittle RF. Intron distribution in ancient paralogs supports random insertion and not random loss. *J Mol Evol*. 1997;44(6):573–84.
38. Yoshihama M, Nakao A, Nguyen HD, Kenmochi N. Analysis of ribosomal protein gene structures: implications for intron evolution. *PLoS Genet*. 2006;2(3):e25.
39. 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.
40. Torriani SF, Stukenbrock EH, Brunner PC, McDonald BA, Croll D. Evidence for extensive recent intron transposition in closely related fungi. *Curr Biol*. 2011;21:2017–22.
41. Farlow A, Meduri E, Dolezal M, Hua L, Schlotterer C. Nonsense-mediated decay enables intron gain in *Drosophila*. *PLoS Genet*. 2010;6(1):e1000819.
42. Li W, Tucker AE, Sung W, Thomas WK, Lynch M. Extensive, recent intron gains in *Daphnia* populations. *Science*. 2009;326(5957):1260–2.
43. Roy SW, Gilbert W. The pattern of intron loss. *Proc Natl Acad Sci U S A*. 2005; 102:713–8.
44. Bernstein LB, Mount SM, Weiner AM. Pseudogenes for human small nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites. *Cell*. 1983;32:461–72.