ORTHOSCOPE: An Automatic Web Tool for Phylogenetically Inferring Bilaterian Orthogroups with User-Selected Taxa

ORTHOSCOPE: An Automatic Web Tool for Phylogenetically Inferring Bilaterian Orthogroups with User-Selected Taxa

Jun Inoue*[,1] and Noriyuki Satoh[1]

[1]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan

***Corresponding author:** E-mail: jun.inoue@oist.jp.

Associate editor: Fabia Ursula Battistuzzi

Abstract

Identification of orthologous or paralogous relationships of coding genes is fundamental to all aspects of comparative genomics. For accurate identification of orthologs among deeply diversified bilaterian lineages, precise estimation of gene trees is indispensable, given the complicated histories of genes over millions of years. By estimating gene trees, orthologs can be identified as members of an orthogroup, a set of genes descended from a single gene in the last common ancestor of all the species being considered. In addition to comparisons with a given species tree, purposeful taxonomic sampling increases the accuracy of gene tree estimation and orthogroup identification. Although some major phylogenetic relationships of bilaterians are gradually being unraveled, the scattering of published genomic data among separate web databases is becoming a significant hindrance to identification of orthogroups with appropriate taxonomic sampling. By integrating more than 250 metazoan gene models predicted in genome projects, we developed a web tool called ORTHOSCOPE to identify orthogroups of specific protein-coding genes within major bilaterian lineages. ORTHOSCOPE allows users to employ several sequences of a specific molecule and broadly accepted nodes included in a user-specified species tree as queries and to evaluate the reliability of estimated orthogroups based on topologies and node support values of estimated gene trees. A test analysis using data from 36 bilaterians was accomplished within 140 s. ORTHOSCOPE results can be used to evaluate orthologs identified by other stand-alone programs using genome-scale data. ORTHOSCOPE is freely available at https://www.orthoscope.jp or https://github.com/jun-inoue/orthoscope (last accessed December 28, 2018).

Key words: **ORTHOSCOPE, orthogroup, orthology, gene tree, species tree, bilaterians.**

Introduction

Identifying orthology and paralogy is fundamental to all aspects of molecular biological research, including cross-species comparisons (Fitch 1970). Given that orthologs are genes derived by speciation, they are used to infer gene functions in nonmodel organisms (Gabaldon and Koonin 2013) and phylogenetic analysis of species (Moritz and Hillis 1996). Considering the complicated history of genes that have diverged via speciation and gene gain (duplication) or loss, the most reliable approach for distinguishing orthologs from paralogs is by explicit phylogenetic inference (Gabaldon 2008; Sonnhammer et al. 2014; Kuraku et al. 2016), especially among distantly related groups of bilaterians.

By estimating gene trees, orthologs can be identified as members of an orthogroup (Li et al. 2003; Chen et al. 2006), a set of genes descended from a single gene in the last common ancestor of all the species being considered (Emms and Kelly 2015). However, identifying an orthogroup by estimating gene trees involves large computational costs, especially for genome-scale data sets. To reduce the computational burden of gene tree estimation, stand-alone programs, such as OrthoMCL (Li et al. 2003) and OrthoFinder (Emms and Kelly 2015), compute sequence similarity scores in multiple species comparisons by employing all-versus-all Blast searches. Then the MCL clustering algorithm (Van Dongen 2000) is used for ortholog identification. On another front, some databases such as EnsemblCompara (Vilella et al. 2008) and PhylomeDB (Huerta-Cepas et al. 2014) store and curate genome-scale orthology hypotheses derived from phylogenetic gene trees. These databases, however, cannot accommodate researchers' demands to estimate gene trees using their own sequences and purposeful taxonomic sampling.

The use of a species tree, in addition to a gene alignment, yields better gene trees than methods that only consider gene alignments (Szöllősi et al. 2015). Recently, some major phylogenetic relationships of bilaterians have gradually begun to be unraveled (Dunn et al. 2014). However, scattering of genome resources among databases, such as NCBI (https://www.ncbi.nlm.nih.gov/), Ensembl (http://www.ensembl.org/), and other independent project-based web sites (e.g., OIST Marine Genomics Unit: http://marinegenomics.oist.jp/) prevents appropriate taxonomic sampling to increase the accuracy of phylogenetic estimation (Heath et al. 2008). Kuraku et al. (2013) integrated scattered protein-coding sequences and created a web tool, aLeaves/MAFFT. With this system, ortholog candidates can be collected from selected databases with their 13 classified groups. Thereafter, for purposeful taxon sampling, orthogroup identification should be achieved by estimating a gene tree manually by selecting sequences.

Article

In the course of evolutionary studies of teleost and chordate genes, we constructed databases of genome-scale protein-coding gene sequences, enabling purposeful taxonomic sampling so as to bisect possible long branches. Moreover, we developed an analytical pipeline to identify orthogroups by estimating gene trees and comparing them with their corresponding species trees. This analytical pipeline successfully identified orthologs not only derived from teleost genome duplication (TGD) (Inoue et al. 2015), but it also identified those that contributed to formation of chordate characteristics (Inoue et al. 2017; Inoue and Satoh 2018).

New Approaches

In the process of developing our analytical pipeline to identify orthogroups of major bilaterian lineages, we created a web tool called ORTHOSCOPE. It enables biologists interested in specific molecules to identify orthogroups and to count numbers of orthogroup members in each species/lineage. For this purpose, the database consists of gene models predicted in genome/transcriptome sequencing projects, in an elementary sense. In order to exclude transcript variants of single loci, the database does not incorporate individually reported gene sequences from each species without full genomic or transcriptomic data.

Orthogroup identification using ORTHOSCOPE has the following characteristics: Users can 1) employ several query sequences (fig. 1A) to collect diverse genes derived from ancestral gene/species separation, 2) select from >250 metazoan species with decoded genomes (fig. 2 and supplementary fig. S1, Supplementary Material online) and one of four taxonomic groups (Deuterostomia, Protostomia, Vertebrata, and Actinopterygii) to employ broadly accepted nodes for orthogroup identification, 3) refer to a hypothetical metazoan species tree reconstructed from a literature survey in order to make their own species trees, and 4) evaluate reliability of orthogroups using topologies, node support values, and functions attached to some sequences shown in estimated gene trees.

Results and Discussion

Interface and Analytical Pipeline

ORTHOSCOPE can be accessed via a web browser (fig. 1A). To start an analysis, ORTHOSCOPE requires a set of sequences consisting only of coding (DNA) or amino acid sequences of protein-coding genes. When sequences (FASTA format) and a species tree (NEWICK format) are provided by the user (fig. 3A), ORTHOSCOPE estimates a gene tree and an orthogroup within several minutes (e.g., 57 s in a case study of deuterostome Brachyury) without the need for user input. Users can modify the species tree in reference to a hypothesis that can be obtained from the ORTHOSCOPE front page.

Before starting an analysis, the user needs to select one of the four "Focal groups" of species to identify orthologs with a focal gene in a specific lineage (fig. 1A). The user can set parameters in "Sequence collection" for the BlastP search (fig. 3B). A threshold (Aligned site rate) in "Alignment" (fig. 3C) is used to remove extremely short sequences when such sequences prevent estimation of data matrices for phylogenetic analysis (fig. 3D). Parameters in "Tree search" are used for gene tree estimation (fig. 3E–G). Taxonomic sampling is determined by selecting species in "Genome taxon sampling" (fig. 1A). In order to count orthologs, ORTHOSCOPE employs a genome-scale protein-coding gene database (coding and amino acid sequence data sets) constructed for each species using only the longest sequence when transcript variants exist for single locus. If a species targeted by a query sequence is not present in the ORTHOSCOPE database, the user needs to add the species name to his species tree.

When the analysis starts (fig. 3A), ORTHOSCOPE first collects amino acid sequences of ortholog candidates by performing a BlastP search (fig. 3B) against selected protein sequence databases. Corresponding coding sequences are also selected from the database. The collected sequences (fig. 3C) are aligned using MAFFT (Katoh and Standley 2013). The resultant multiple sequence alignment is trimmed by removing poorly aligned regions using trimAl (Capella-Gutierrez et al. 2009) with the option "gappyout." Corresponding coding sequences are forced onto the amino acid alignment using PAL2NAL (Suyama et al. 2006) to generate nucleotide alignments for subsequent comparative analysis.

To achieve faster analysis speed than is possible with the maximum likelihood method, phylogenetic analyses (fig. 3E) employ the neighbor joining (NJ) method (Saitou and Nei 1987) implemented in APE in R (Popescu et al. 2012) for DNA alignments and FastME (Lefort et al. 2015) for amino acid alignments. For analyses of DNA alignments, the most parameter-rich model in the program, the TN 93 model (Tamura and Nei 1993), is applied with a gamma-distributed rate for site heterogeneity (Yang 1994). For analyses of amino acid alignments, a widely used substitution model for nuclear gene analysis, the WAG model (Whelan and Goldman 2001), is applied with the gamma model. To evaluate robustness of internal branches, 100 bootstrap replications are calculated for each data set.

Resultant gene trees (fig. 3E), however, often have weakly supported nodes. In such cases, one can revise these ambiguous nodes in comparison with a specific species tree. For this purpose, ORTHOSCOPE conducts rearrangement/reconciliation analysis using a method implemented in NOTUNG (Chen et al. 2000) for the NJ gene tree (fig. 3E) in comparison with the uploaded species tree (fig. 3F). As a first step, NOTUNG rearranges weakly supported nodes of the gene tree, to minimize duplication and extinction of genes, using parsimony with equal weights and the threshold parameter for bootstrap support values of nodes (fig. 1A). Then, the rearranged gene tree is reconciled with the species tree. Finally, an orthogroup is identified (fig. 3G).

Orthogroup of ORTHOSCOPE

Orthogroups are defined as sets of genes descended from single genes in the last common ancestor of all the species being considered (Emms and Kelly 2015). In gene trees, ancestral states of genes are single at speciation nodes (fig. 3G).

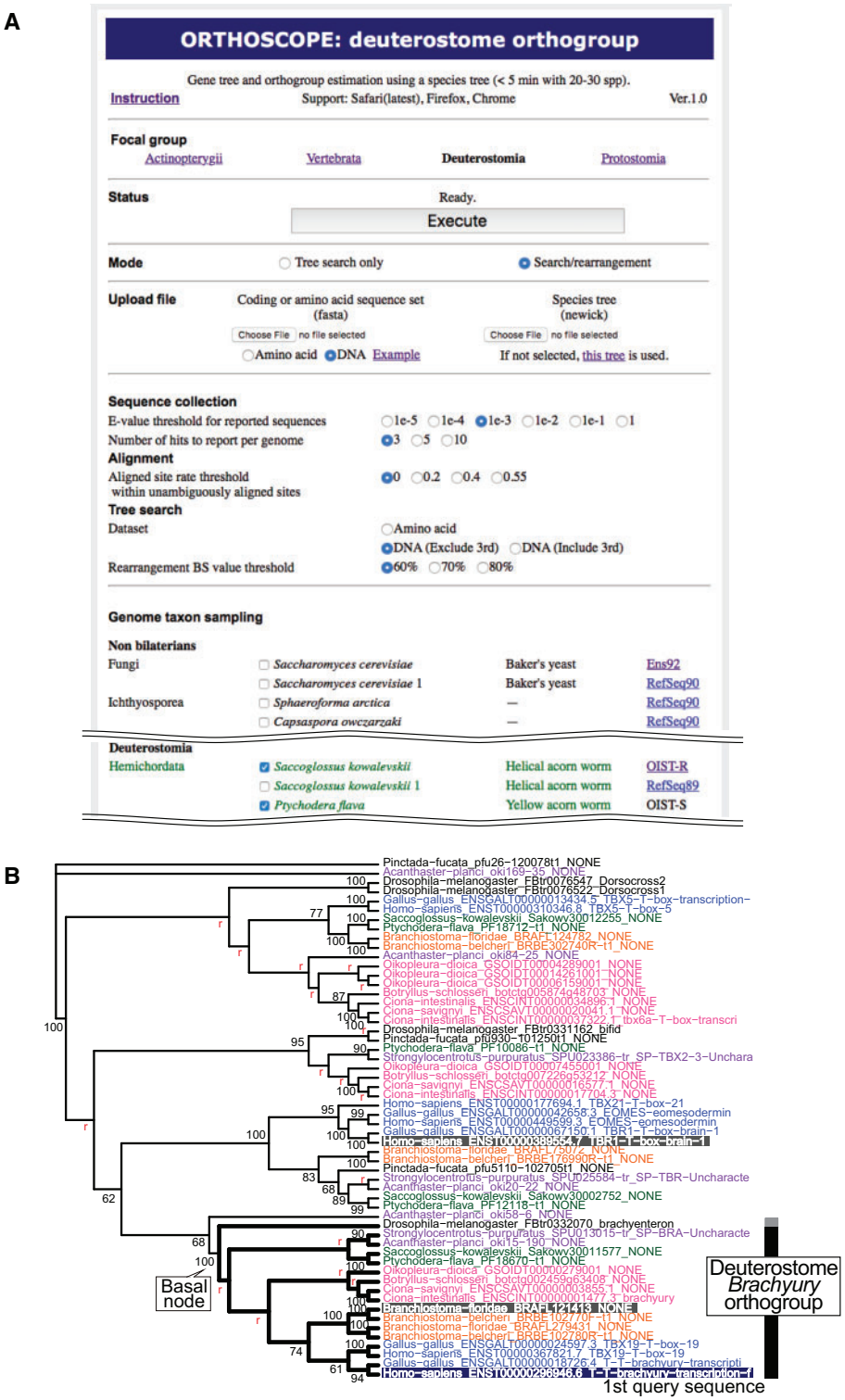

Fig. 1. An overview of the interface of the ORTHOSCOPE web server. (*A*) The front page. Ortholog identification is conducted by selecting one of four focal groups of species, Actinopterygii, Vertebrata, Deuterostomia, or Protostomia. The user can select species for orthogroup/tree estimation. (*B*) The resultant tree of *Brachyury* gene analysis using the focal group Deuterostomia (supplementary fig. S2*A*2, Supplementary Material online). White letters on a navy blue back ground (*Homo sapiens Brachyury* gene sequence) indicate the first query (ENST00000296946.6 from Ensembl) and those on a gray back ground indicate others (BRAFL121413 from JGI and ENST00000389554.7). The smallest bilaterian clade, including the first query sequence, is identified as the orthogroup (connected by thick branches). The orthogroup is shown with a vertical bar consisting of black segment (focal group: deuterostome genes) and gray (its sister group: a protostome gene) segment. The basal node denotes the basal split of the orthogroup. Nodes marked with an "r" were rearranged using NOTUNG during comparisons with the species tree, because they had lower bootstrap support values than the user-defined threshold (60%).

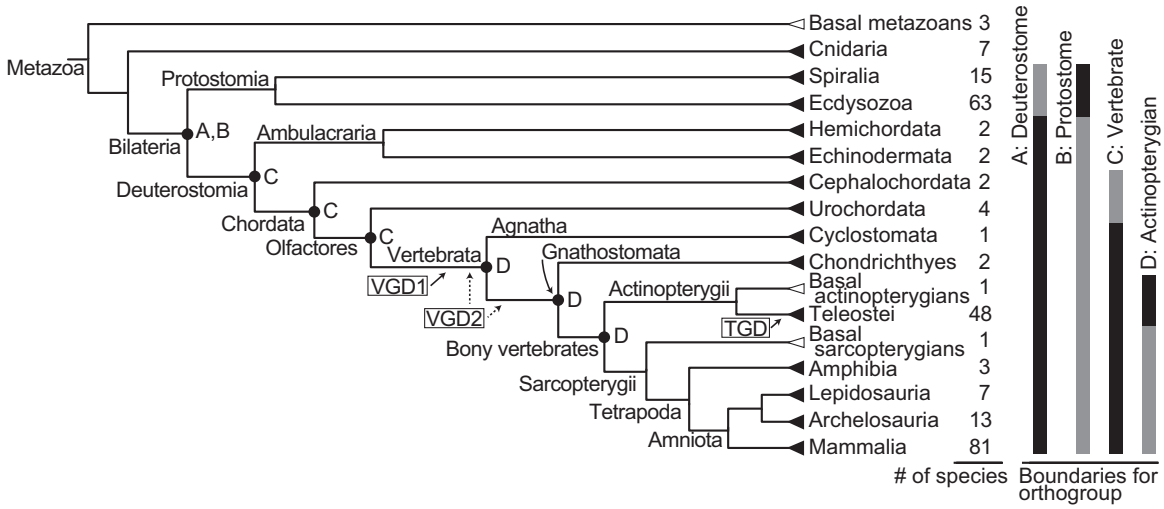

Fig. 2. Phylogenetic relationships of bilaterian lineages and the number of species included in the ORTHOSCOPE database. With respect to each focal group (*A*–*D*) of species, an orthogroup is identified in an estimated gene tree by finding the broadly accepted node (marked with black circle and alphabet of focal group of species): Basal splits of Bilateria (Dunn et al. 2014), Deuterostomia (Satoh 2016), Chordata/Olfactores (Satoh 2016), Vertebrata, and bony vertebrates (Meyer and Zardoya 2003). Those broadly accepted nodes are appropriate for corresponding nodes as orthogroup basal nodes because these nodes are insulated from the influence of whole genome duplications when identifying corresponding nodes in gene trees. Phylogenetic positions of whole genome duplications (VGD, vertebrate genome duplication; TGD, teleost genome duplication) follow Braasch and Postlethwait (2012). Whether the second vertebrate genome duplication (VGD2) occurred before or after divergence of jawless fish remains controversial. Black segments denote focal groups of species and gray segments denote their sister groups. Those species groups are used for orthogroup identification by finding their basal nodes (key nodes) in gene trees. Triangles indicate species groups in which monophyly is supported (black) or unsupported (white). For details, see supplementary figure S1, Supplementary Material online.

For this reason, the basal node of orthogroup should be a speciation node, when finding it in a gene tree. However, considering the presence of duplication nodes and weak resolution of gene tree nodes, identification of the orthogroup basal node is difficult without *a priori* information about species relationships and phylogenetic positions of genome duplication events related to the node (fig. 2).

As a corresponding node of the orthogroup basal node (fig. 3*G*), ORTHOSCOPE uses a key node (fig. 3*F*), one of the broadly accepted nodes of a species tree (fig. 2). From a given species tree (fig. 3*F*), ORTHOSCOPE identifies focal and sister groups for two species lineages separated at a key node. Accordingly, an orthogroup identified by ORTHOSCOPE (fig. 3*G*) contains genes not only of the focal group of species, but also of its sister group species. Therefore, when comparing genes within a focal group of genes, some relationships are paralogous. However, when comparing genes between a focal group of genes and its sister group, all relationships are orthologous.

In the Deuterostome Brachyury analysis, ORTHOSCOPE identifies deuterostomes as the focal group and protostomes as their sister group (fig. 3*F*). In this case, the separation between deuterostomes and protostomes is used as the key node. By finding the corresponding node of this key node from the rearranged gene tree (fig. 3*G*), ORTHOSCOPE identifies an orthogroup, a bilaterian gene clade including the first query sequence. The bootstrap value of the basal node can be used to evaluate the accuracy of orthogroup identification.

Case Studies

We demonstrate the utility of ORTHOSCOPE using case studies with four focal groups of species. In each case, to show novelty in ORTHOSCOPE, resultant orthogroups were compared with those estimated using two pioneering tools in this field, OrthoFinder (ver. 2.2.6) and aLeaves (last access date: June 24, 2018). Although these two programs also facilitate ortholog estimation, their scopes are different from that of ORTHOSCOPE: 1) OrthoFinder estimates orthogroups for all protein-coding genes at one time using user-specified data sets; and 2) aLeaves collects as many ortholog candidates as possible for a particular molecule using their database including individually reported gene sequences from each species without full genomic/transcriptomic data.

Deuterostome Brachyury

ORTHOSCOPE can identify orthologs of a gene that creates morphological novelty in deuterostomes (fig. 2). The *Brachyury* gene encodes a member of the T-box transcription factor family and is crucial for notochord formation in chordates (Satoh 2016). Using *Brachyury* gene sequences of *Homo sapiens* and *Branchiostoma floridae* (Florida lancelet) as queries (fig. 1*B*), ORTHOSCOPE identified orthologs from all five deuterostome lineages (table 1*A*; fig. 4*A*, and supplementary fig. S2*A*, Supplementary Material online). As suggested in Inoue et al. (2017), two copies of the *Brachyury* ortholog were identified in each of two cephalochordate species. We confirmed that one of the three queries, *H. sapiens* TBR1 (ENST00000389554.7 in Ensembl), is placed outside the

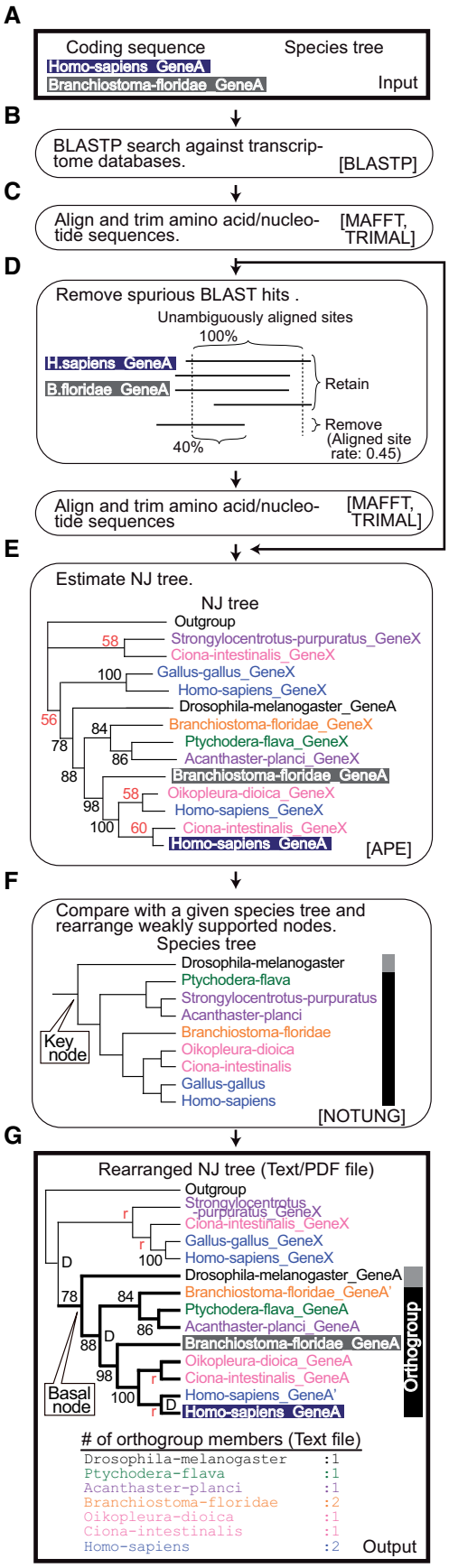

Fig. 3. An overview of the ORTHOSCOPE analytical pipeline of orthogroup identification. By uploading query sequences and a species tree (input), after an ORTHOSCOPE analysis, the estimated gene tree and candidate ortholog sequences are downloaded as text/PDF files (output). The species tree (F) consists of a focal group (black segment) and its sister group (gray segment). In the rearranged gene tree (G), the orthogroup consists of a focal group of genes (black segment) and its sister group (gray segment). Nodes marked with "D" are duplication nodes whereas those with no mark are speciation nodes. Refer to the main text for details of each procedure.

vertebrate *Brachyury* orthogroup because the orthogroup was identified based solely on the first query, *H. sapiens* Brachyury.

When the same amino acid databases were used, OrthoFinder produced exactly the same orthogroup as that estimated by ORTHOSCOPE (table 1A and supplementary fig. S2*A*, Supplementary Material online), identifying *Brachyury* orthologs in every deuterostome lineage. Moreover, orthogroup members identified by ORTHOSCOPE were also the same as those estimated based on sequences collected by aLeaves (supplementary fig. S2*A4*, Supplementary Material online), except for hemichordates, which were not included in the aLeaves database. The main difference between the results of ORTHOSCOPE and aLeaves lies in the number of species with identified orthologs from vertebrates and protostomes due to the limitation of purposeful taxonomic sampling.

Protostome Brachyury

ORTHOSCOPE can also evaluate the presence or absence of orthologs in morphologically and genetically diverse protostomes (fig. 2). A *Brachyury* ortholog has not been identified in the *C. elegans* (nematode worm) genome (Hejnol and Martin-Duran 2015; Inoue et al. 2017). In order to confirm whether this lack of a *Brachyury* ortholog is shared among other nematodes, an ORTHOSCOPE analysis was conducted using protostome *Brachyury* gene sequences and a *C. elegans mab-9* sequence (T27A1.6 in WormBase: https://www.wormbase.org), which is related to Brachyury (Woollard and Hodgkin 2000) as queries (supplementary fig. S2*B1–B3*, Supplementary Material online). The resultant tree confirmed that no *Brachyury* ortholog is found in 11 nematode species (table 1B, fig. 4*B*, and supplementary fig. S2*B*, Supplementary Material online). In addition, no *Brachyury* ortholog was found in platyhelminth genomes, as reported previously (Martin-Duran and Romero 2011; Hejnol and Martin-Duran 2015).

To evaluate results indicating the absence of *Brachyury* orthologs in the nematode and platyhelminth genomes, we estimated the protostome *Brachyury* orthogroup using OrthoFinder and aLeaves. OrthoFinder identified *Brachyury* orthologs from nematodes and platyhelminths (table 1B and supplementary table S1B, Supplementary Material online), conflicting with results from ORTHOSCOPE. Divergent protostome sequences and analyses without the broadly accepted node, the basal split of bilaterians, may prevent OrthoFinder from delineating the protostome Brachyury orthogroup. On the other hand, the resultant tree based on sequences collected by aLeaves identified no *Brachyury*

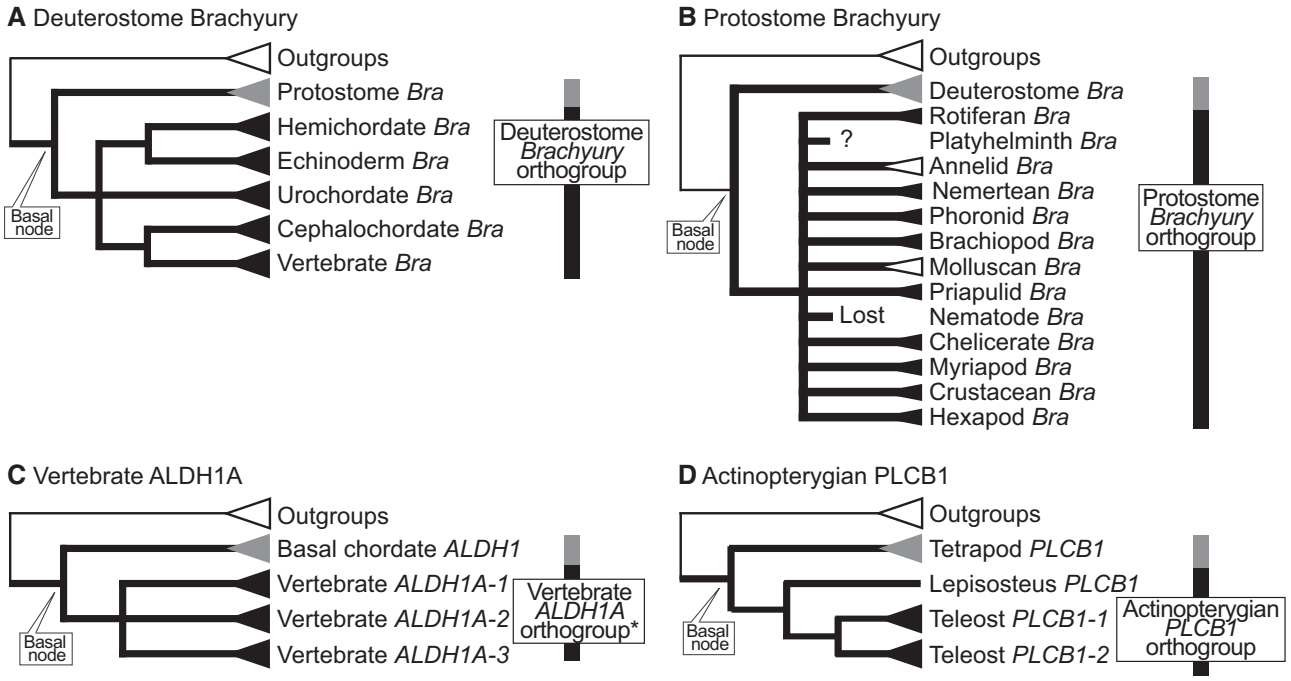

Fig. 4. Schematic of estimated gene trees using ORTHOSCOPE (supplementary fig. S2*A*–*D*, Supplementary Material online). (*A*) Deuterostome *Brachyury* gene tree. (*B*) Protostome *Brachyury* gene tree. (*C*) Vertebrate *ALDH1A* gene tree. An asterisk indicates that the orthogroup was not supported by the 60% bootstrap value criterion for the basal node of orthogroup (basal chordate vs. vertebrate lineages). (*D*) Actinopterygian *PLCB1* gene tree. Orthogroups are shown with black (focal group of genes) and gray (sister group of genes) segments.

ortholog in either lineage and supported the ORTHOSCOPE results (supplementary fig. S2*B4*, Supplementary Material online).

Vertebrate ALDH1A

From a transcriptome assembly, ORTHOSCOPE can identify orthologs of genes that experienced ancient whole genome duplications. A comparative genomic study suggested that 20–30% of duplicate genes (Makino and McLysaght 2010) derived from vertebrate genome duplications (VGDs) are still retained in the human genome, even after several hundred million years (fig. 2). Their duplicates, called ohnologs, complicate identification of vertebrate orthologs (Kuraku et al. 2016).

The vertebrate *ALDH1A* (*retinaldehyde dehydrogenase 1 A*) gene is thought to have been foundational for the emergence of vertebrates (Duester 2008). In vertebrates, the *ALDH1A* gene encodes cytosolic enzymes capable of metabolizing all-*trans*-retinaldehyde to retinoic acid, a molecular signal that guides vertebrate development and adipogenesis (Holmes 2015). In order to identify *ALDH1A* orthologs of *Tylototriton wenxianensis* (wenxian knobby newt), ORTHOSCOPE analysis was conducted. At first, using a Blast search, five candidate sequences similar to the *H. sapiens ALDH1A1* gene sequence (ENST00000297785.7) were selected from the *T. wenxianensis* transcriptome assembly (GESS00000000 in NCBI). Then an ORTHOSCOPE analysis was conducted using these five sequences as queries. As a result, four out of the five sequences were identified as members of the vertebrate *ALDH1A* orthogroup (table 1C, fig. 4C, and supplementary fig. S2*C*, Supplementary Material online). A phylogenetic analysis focusing on orthogroup members (supplementary fig. S2*C3*, Supplementary Material online) indicated that the *T. wenxianensis* sequences distributed among these *ALDH1A* gene lineages were duplicated during VGD events. Although the analysis did not provide strong support for relationships among *ALDH1A1-3* genes of *T. wenxianensis*, orthology can be identified by means of conserved synteny. In fact, a syntenic analysis (Canestro et al. 2009) suggests a closer relationship between *ALDH1A-1* and *ALDH1A-2* gene lineages and the loss of the *ALDH1A-3* gene lineage counterpart just after VGDs.

We compared the ORTHOSCOPE result with that of OrthoFinder analysis using the same *T. wenxianensis* transcriptome assembly. Under taxonomic sampling comprising only vertebrates (table 1C), OrthoFinder identified the same four orthologs found by ORTHOSCOPE (supplementary fig. S2*C2*, Supplementary Material online). However, the OrthoFinder analysis also identified additional sequences, including *T. wenxianensis* (supplementary table S1C, Supplementary Material online, GESS01039398.1) with an extremely short sequence (324 bp) compared with the others (1,539–1,722 bp). In order to determine their phylogenetic positions, by including this additional *T. wenxianensis* sequence as one of queries (supplementary fig. S2*C4*, Supplementary Material online), an ORTHOSCOPE analysis was conducted with the top five Blast hits. As a result, although these additional sequences were included in the vertebrate *ALDH1A* gene lineage, except for a short sequence (777 bp) of *H. sapiens* (ENST00000546840.2), the sequence of *T. wenxianensis* was not grouped with the other *T. wenxianensis* sequences within the same gene lineage. The sequence

Table 1. Taxon Samplings and Estimated Numbers of Orthogroup Members.

	No. of orthogroup members	
Taxon sampling	ORTHOSCOPE	OrthoFinder
A. Deuterostome Brachyury		Bilaterians[a]
Protostomia		
Spiralia		
Pinctada fucata	0	0
Ecdysozoa		
Drosophila melanogaster	1	1
Deuterostomia		
Hemichordata		
Saccoglossus kowalevskii	1	1
Ptychodera flava	1	1
Echinodermata		
Strongylocentrotus purpuratus	1	1
Acanthaster planci	1	1
Cephalochordata		
Branchiostoma floridae	2	2
B. belcheri	2	2
Urochordata		
Oikopleura dioica	1	1
Botryllus schlosseri	1	1
Ciona savignyi	1	1
C. intestinalis	1	1
Vertebrata		
Gallus gallus	2	2
Homo sapiens	2	2
B. Protostome Brachyury		Bilaterians[a]
Deuterostomia		
Gallus gallus	2	17
Homo sapiens	2	16
Protostomia		
Spiralia		
Rotifera		
Adineta vaga	3	28
Platyhelminthes		
Schistosoma mansoni	0	6
Annelida		
Capitella teleta	1	8
Helobdella robusta	1	18
Nemertea		
Notospermus geniculatus	1	9
Phoronida		
Phoronis australis	1	6
Brachiopoda		
Lingula anatina	1	7
Cephalopoda		
Octopus bimaculoides	1	10
Gastropoda		
Lottia gigantea	1	11
Biomphalaria glabrata	1	9
Aplysia californica	1	12
Bivalvia		
Crassostrea virginica	1	13
Crassostrea gigas	1	11
Mizuhopecten yessoensis	1	12
Pinctada fucata	0	7
Ecdysozoa		
Priapulida		
Priapulus caudatus	1	5
Nematoda		
Trichinella spiralis	0	4
Strongyloides ratti	0	4
Onchocerca volvulus	0	4
Loa loa	0	9

(continued)

Table 1. Continued

Taxon sampling	No. of orthogroup members	
	ORTHOSCOPE	OrthoFinder
Brugia malayi	0	4
Pristionchus pacificus	0	6
Caenorhabditis japonica	0	13
C. brenneri	0	9
C. remanei	0	14
C. briggsae	0	13
C. elegans	0	8
Chelicerata		
Limulus polyphemus	1	32
Stegodyphus mimosarum	0	12
Myriapoda		
Strigamia maritima	2	6
Crustacea		
Daphnia pulex	1	7
Hexapoda		
Nasonia vitripennis	1	5
Bombyx mori	1	11
Drosophila melanogaster	1	8
C. Vertebrate ALDH1A		Vertebrates[a]
Urochordata		
Ciona savignyi	1	–
C. intestinalis	1	–
Vertebrata		
Chondrichthyes		
Callorhinchus milii	3	3
Rhincodon typus	3	3
Actinopterygii		
Lepisosteus oculatus	3	3
Danio rerio	2	2
Salmo salar	5	5
Oncorhynchus mykiss	3	5
Tetraodon nigroviridis	1	2
Oreochromis niloticus	2	2
Oryzias latipes	1	1
Sarcopterygii		
Amphibia		
Xenopus tropicalis	3	3
Tylototriton wenxianensis[b]	4[c]	5
Lepidosauria		
Anolis carolinensis	3	3
Testudines		
Pelodiscus sinensis	3	3
Aves		
Gallus gallus	3	3
Mammalia		
Bos taurus	3	3
Mus musculus	4	4
Homo sapiens	3	4
D. Actinopterygian PLCB1		Actinops[a]
Chondrichthyes		
Callorhinchus milii	1	–
Rhincodon typus	1	–
Sarcopterygii		
Gallus gallus	1	–
Homo sapiens	1	–
Actinopterygii		
Neopterygii		
Lepisosteidae		
Lepisosteus oculatus	1	1
Teleostei		
Osteoglossomorpha		
Scleropages formosus	2	2

(continued)

Table 1. Continued

Taxon sampling	No. of orthogroup members	
	ORTHOSCOPE	OrthoFinder
Paramormyrops kingsleyae	2	2
Otomorpha		
Astyanax mexicanus	2	2
Danio rerio	0	0
Cyprinus carpio	2	3
Protacanthopterygii		
Esox lucius	2	2
Coregonus lavaretus[b]	2[c]	2
Salmo salar	3	4
Acanthomorphata		
Gadus morhua	2	2
Takifugu rubripes	2	2
Oreochromis niloticus	2	2
Oryzias latipes	1	1

[a]Taxon sampling (supplementary fig. S2, Supplementary Material online).
[b]Databases constructed from NCBI transcriptome shotgun assembly (TSA).
[c]Numbers manually counted.

alignment and the resultant gene tree produced by ORTHOSCOPE highlighted an ambiguity in the assembly of this extremely short sequence.

Actinopterygian Phospholipase C Beta 1

ORTHOSCOPE can also identify orthologs from genes that experienced TGD (in fig. 2). In order to identify *Phospholipase C beta 1* (*PLCB1*) orthologs of *Coregonus lavaretus* (common whitefish) from a transcriptome assembly (GESS00000000), an ORTHOSCOPE analysis was conducted using three ortholog candidates of the *C. lavaretus PLCB1* gene as queries. The resultant tree showed that two of the three candidate sequences were found in the actinopterygian *PLCB1* orthogroup and placed in two gene lineages, teleost *PLCB1-1* and *-2* (table 1D, fig. 4*D*, and supplementary fig. S2*D*, Supplementary Material online). Teleost gene lineages *PLCB1-1* and *-2* are thought to have been derived from TGD, according to phylogenetic and synteny analyses (figs. S27 and S64 in Sato et al. 2009, respectively). Moreover, the ORTHOSCOPE analysis identified duplicated genes in the lineage leading to *Cyprinus carpio* (common carp) and *Salmo salar* (Atlantic salmon)/*C. lavaretus* (in the teleost *PLCB1-2* gene lineage [supplementary fig. S2*D*3, Supplementary Material online]). They may have been derived from the carp genome duplication or the salmonid genome duplication, respectively (supplementary fig. S1*D*, Supplementary Material online).

From the same transcriptome assembly, OrthoFinder identified the same orthologs of *C. lavaretus* under a taxonomic sampling comprising only actinopterygians (table 1D and supplementary fig. S2*D*2, Supplementary Material online). For other orthogroup members, however, two additional sequences, *C. carpio* (XP018928569.1) and *S. salar* (XP014066076.1), were included in the orthogroup (supplementary table S1D, Supplementary Material online). When the top five sequences of the Blast search were employed (supplementary fig. S2*D*4, Supplementary Material online), an ORTHOSCOPE analysis included the *C. carpio* sequence in the actinopterygian *PLCB1* gene lineage as an orthogroup member. Again, in comparison with the three *C. lavaretus* sequences (3,597–3,768 bp), the short length found in this sequence (1,497 bp) may have prevented its inclusion among the top three Blast hits. On another front, the *S. salar* sequence was not included in the bony vertebrate *PLCB1* gene lineage, probably due to its long branch (supplementary fig. S2*D*4, Supplementary Material online). In the alignment produce by ORTHOSCOPE, a highly diversified region was found in this long *S. salar* sequence (6,328 bp). A possible mis-assembly made ortholog identification of this sequence difficult.

Conclusions

ORTHOSCOPE, a fully automatic web pipeline, successfully identified orthologs in the present four example analyses, consistent with manual identifications in prior research. As shown in the present study, ORTHOSCOPE can be used to evaluate orthologs identified in genome-scale analyses by other programs. ORTHOSCOPE users can evaluate reliability of orthogroups using estimated gene trees in light of user knowledge of species/gene evolutionary histories, even when the same orthogroups were identified among different programs. In addition to inferring gene function from model to nonmodel organisms (but see Gabaldon and Koonin 2013), orthogroups identified by ORTHOSCOPE can be applied to evolutionary studies of gene regulatory networks (Marti-Solans et al. 2016) and local synteny (Inoue et al. 2017) including nonmodel organisms. Moreover, with regard to genes derived from VGDs or TGD, ORTHOSCOPE can evaluate phylogenetic markers in vertebrates or teleosts by identifying the presence or absence of ohnologs, which

complicate phylogenetic analyses. We will include newly published genome-wide protein-coding sequences from various metazoan species and expand focal groups in ORTHOSCOPE (e.g., spiralians and urochordates) in response to user requests.

Materials and Methods

The server runs on the Linux operating system and an Apache HTTP Server provides web services. Python scripts process all data and requests from users. All these resources have been extensively used and are well supported.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Asuka Arimoto for advice on the analytical pipeline and manuscript, and for providing customized Ruby scripts. Mutsumi Nishida, Yukuto Sato, and Robert Sinclair helped greatly with discussions about early versions of the analytical pipeline. We thank all members of the Marine Genomics Unit, especially Yuuri Yasuoka and Takeshi Takeuchi, for discussions on an earlier version of the manuscript/pipeline, and two anonymous reviewers for helpful comments on the manuscript. We thank Shinobu Kinjo and Atsushi Kawai in the Information Service Section of the Okinawa Institute of Science and Technology Graduate University (OIST) for technical support in putting ORTHOSCOPE online, and Steven D. Aird for editing the manuscript. This work was supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (C) (15K07172) to J.I. and (B) (16H04824) to N.S.

References

Braasch I, Postlethwait J. 2012. Polyploidy in fish and the teleost genome duplication. In: Soltis PS, Soltis DE, editors. Polyploidy and genome evolution. Berlin (Germany): Springer. p. 341–383.

Canestro C, Catchen JM, Rodriguez-Mari A, Yokoi H, Postlethwait JH. 2009. Consequences of lineage-specific gene loss on functional evolution of surviving paralogs: aLDH1A and retinoic acid signaling in vertebrate genomes. *PLoS Genet.* 5(5): e1000496.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15): 1972–1973.

Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34(Database issue): D363–D368.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol.* 7(3–4): 429–447.

Duester G. 2008. Retinoic acid synthesis and signaling during early organogenesis. *Cell* 134(6): 921–931.

Dunn CW, Giribet G, Edgecombe GD, Hejnol A. 2014. Animal phylogeny and its evolutionary implications. *Annu Rev Ecol Evol Syst.* 45(1): 371–395.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool.* 19(2): 99–113.

Gabaldon T. 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.* 9:235.

Gabaldon T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet.* 14(5): 360–366.

Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. *J Syst Evol.* 46:239–257.

Hejnol A, Martin-Duran DJM. 2015. Getting to the bottom of anal evolution. *Zool Anz.* 256:61–74.

Holmes RS. 2015. Comparative and evolutionary studies of vertebrate ALDH1A-like genes and proteins. *Chem Biol Interact.* 234:4–11.

Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Marcet-Houben M, Gabaldon T. 2014. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* 42(Database issue): D897–D902.

Inoue J, Sato Y, Sinclair R, Tsukamoto K, Nishida M. 2015. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proc Natl Acad Sci U S A.* 112(48): 14918–14923.

Inoue J, Satoh N. 2018. Deuterostome genomics: lineage-specific protein expansions that enabled chordate muscle evolution. *Mol Biol Evol.* 35(4): 914–924.

Inoue J, Yasuoka Y, Takahashi H, Satoh N. 2017. The chordate ancestor possessed a single copy of the *Brachyury* gene for notochord acquisition. *Zoological Lett.* 3:4.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4): 772–780.

Kuraku S, Feiner N, Keeley SD, Hara Y. 2016. Incorporating tree-thinking and evolutionary time scale into developmental biology. *Dev Growth Differ.* 58(1): 131–142.

Kuraku S, Zmasek CM, Nishimura O, Katoh K. 2013. aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. *Nucleic Acids Res.* 41(Web Server issue): W22–W28.

Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol.* 32(10): 2798–2800.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13(9): 2178–2189.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A.* 107(20): 9270–9274.

Marti-Solans J, Belyaeva OV, Torres-Aguila NP, Kedishvili NY, Albalat R, Canestro C. 2016. Coelimination and survival in gene network evolution: dismantling the RA-signaling in a chordate. *Mol Biol Evol.* 33:2401–2416.

Martin-Duran JM, Romero R. 2011. Evolutionary implications of morphogenesis and molecular patterning of the blind gut in the planarian *Schmidtea polychroa*. *Dev Biol.* 352(1): 164–176.

Meyer A, Zardoya R. 2003. Recent advances in the (molecular) phylogeny of vertebrates. *Annu Rev Ecol Evol Syst.* 34(1): 311–338.

Moritz C, Hillis DM. 1996. Molecular systematics: context and controversies. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. Sunderland (MA): Sinauer Associates. p. 1–13.

Popescu AA, Huber KT, Paradis E. 2012. ape 3.0: new tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* 28(11): 1536–1537.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4(4): 406–425.

Sato Y, Hashiguchi Y, Nishida M. 2009. Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after teleost-specific genome duplication. BMC Evolutionary Biology 9:127.

Satoh N. 2016. Chordate origins and evolution: the molecular evolutionary road to vertebrates. Boston: Elsevier.

Sonnhammer EL, Gabaldon T, Sousa da Silva AW, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas PD, Dessimoz C, Quest for

Orthologs. 2014. Big data and other challenges in the quest for orthologs. *Bioinformatics* 30(21): 2993–2998.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34(Web Server issue): W609–W612.
Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. *Syst Biol.* 64(1): e42–e62.
Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Mol Biol Evol.* 10(3): 512–526.
Van Dongen S. 2000. Graph clustering by flow simulation. Utrecht (The Netherlands): University of Utrecht.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2008. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19(2): 327–335.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol.* 18(5): 691–699.
Woollard A, Hodgkin J. 2000. The *Caenorhabditis elegans* fate-determining gene *mab-9* encodes a T-box protein required to pattern the posterior hindgut. *Genes Dev.* 14(5): 596–603.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate method. *J Mol Evol.* 39(3): 306–314.