


ORTHOSCOPE*: A Phylogenetic Pipeline to Infer Gene Histories from Genome-Wide Data

Jun Inoue *

Center for Earth Surface System Dynamics, Atmosphere and Ocean Research Institute, University of Tokyo, Kashiwa, Japan

*Corresponding author: E-mail: jinoue@g.ecc.u-tokyo.ac.jp.

Associate editor: Rebekah Rogers

Abstract

Comparative genome-scale analyses of protein-coding gene sequences are employed to examine evidence for whole-genome duplication and horizontal gene transfer. For this purpose, an orthogroup should be delineated to infer evolutionary history regarding each gene, and results of all orthogroup analyses need to be integrated to infer a genome-scale history. An orthogroup is a set of genes descended from a single gene in the last common ancestor of all species under consideration. However, such analyses confront several problems: 1) Analytical pipelines to infer all gene histories with methods comparing species and gene trees are not fully developed, and 2) without detailed analyses within orthogroups, evolutionary events of paralogous genes in the same orthogroup cannot be distinguished for genome-wide integration of results derived from multiple orthogroup analyses. Here I present an analytical pipeline, ORTHOSCOPE* (star), to infer evolutionary histories of animal/plant genes from genome-scale data. ORTHOSCOPE* estimates a tree for a specified gene, detects speciation/gene duplication events that occurred at nodes belonging to only one lineage leading to a species of interest, and then integrates results derived from gene trees estimated for all query genes in genome-wide data. Thus, ORTHOSCOPE* can be used to detect species nodes just after whole-genome duplications as a first step of comparative genomic analyses. Moreover, by examining the presence or absence of genes belonging to species lineages with dense taxon sampling available from the ORTHOSCOPE web version, ORTHOSCOPE* can detect genes lost in specific lineages and horizontal gene transfers. This pipeline is available at https://github.com/jun-inoue/ORTHOSCOPE_STAR.

Key words: ORTHOSCOPE*, genome comparisons, orthogroup, orthology, species tree and gene tree, animals and plants.

Introduction

The recent, rapid accumulation of sequenced genomes has made it possible to compare genomic data among distantly related species. As a first step toward comparative studies, genome-wide data regarding protein-coding genes are employed to find evolutionary events such as whole-genome duplication (WGD) and horizontal gene transfers (Futuyma and Kirkpatrick 2017; Nagy et al. 2020).

Orthology is a central concept in evolutionary and comparative genomics and is used to identify corresponding genes in different species (Gabaldon and Koonin 2013; Altenhoff, Glover, et al. 2019). Two genes can have shared ancestry due to any of three phenomena: speciation (orthologs), gene duplication (paralogs), and horizontal gene transfer (xenologs) (Koonin 2005; Fernández et al. 2020). Considering the complicated history of genes that have diverged via speciation and gene acquisition (duplication) or loss, the most reliable approach for distinguishing orthologs from paralogs is by explicit phylogenetic inference (Gabaldon 2008; Sonnhammer et al. 2014). By estimating gene trees, orthologs can be identified as members of an orthogroup (fig. 1), a set of genes descended

from a single gene in the last common ancestor of all species being considered (Emms and Kelly 2019), although paralogous relationships can be included in this set (Trachana et al. 2011; Altenhoff, Levy, et al. 2019).

Orthogroup identification can be achieved by estimating a gene tree and comparing it with a species tree (Fernández et al. 2020). Our web tool, ORTHOSCOPE (Inoue and Satoh 2019), identifies an orthogroup by assigning a node in the gene tree that corresponds to a key node in the species tree (fig. 1), but this tool works only for a specific molecule and does not allow genome-scale analyses. A pioneering tool in this field, OrthoFinder (Emms and Kelly 2019), can identify orthogroups for genome-wide data and can count gene duplications by summarizing gene trees estimated for orthogroups. The history of a genome, however, cannot be inferred by integrating all results obtained with respect to each orthogroup, due to the existence of paralogs within orthogroups (Fernández et al. 2020). In figure 1B, two medaka paralogs, Gene A1 and Gene A2, have different evolutionary histories. The evolutionary history of a genome should be inferred by integrating evolutionary histories estimated gene by gene.

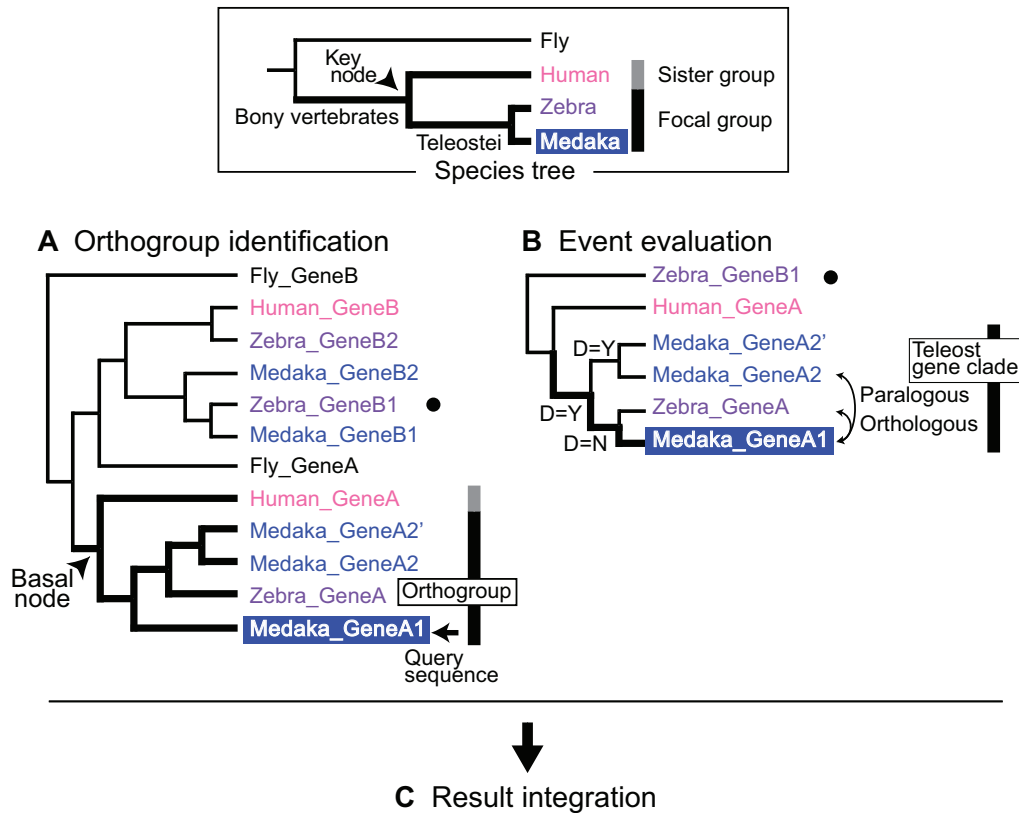


Fig. 1. Schematic overview of ORTHOSCOPE* analysis. (A) Orthogroup identification. Thick branches connect orthogroup members. (B) Event evaluation. Thick branches indicate lineages leading to the query sequence. Two paralogs have different histories: in comparison with Medaka_GeneA1, Medaka_GeneA2 experienced an additional duplication event. In the species tree column, the vertical bar separated by black and gray segments (boundary for the orthogroup) indicates species whose genes can be included in orthogroups. Black segments denote focal groups of species and gray segments denote their sister groups. The key node is used for orthogroup identification by finding their basal nodes in gene trees. Black circles indicate the rooting sequence used in (B). D = N: speciation; D = Y: gene duplication.

New Approach

While developing the ORTHOSCOPE web version, I created a new analytical pipeline called ORTHOSCOPE* (star) (last accessed July 18, 2021). The ORTHOSCOPE web version infers an evolutionary history of one orthogroup derived from a specific query gene sequence. In contrast, to facilitate future genome comparative studies, ORTHOSCOPE* version 1 accommodates genome-wide data of protein-coding genes. To infer evolutionary events of an entire genome, evolutionary events of all genes should be considered by integrating results derived from orthogroups delineated gene by gene, and by calculating the percentage of gene trees retaining traces of focal events. For this purpose, ORTHOSCOPE* 1) focuses on speciation/gene duplication events that occurred at nodes belonging to only one lineage leading to a focal gene in a gene tree (fig. 1B) and 2) chooses accurately estimated orthogroups/gene nodes with a criterion based upon bootstrap support. In addition, purposeful taxonomic sampling can be accomplished by using more than 550 animal/plant gene models available from ORTHOSCOPE web version (<https://github.com/jun-inoue/orthoscope>). Thus, in addition to the accuracy of the gene model for focal species, purposeful or denser taxonomic sampling also enables users to examine the presence or absence of genes in species lineages, and to

identify genes derived from horizontal gene transfers. Moreover, ORTHOSCOPE* produces visualized gene trees and orthogroups, enabling users to check their results in biological contexts.

Results and Discussion

ORTHOSCOPE* analysis comprises three steps (fig. 1).

Orthogroup identification: ORTHOSCOPE* estimates a gene tree using all BLAST-hit sequences. In the estimated gene tree, the analytical pipeline delineates an orthogroup by finding the basal node corresponding to the user-defined key-node in the species tree (fig. 1A).

Event evaluation: Using sequences of rooting and orthogroup members, the analytical pipeline estimates a more accurate gene tree for the orthogroup (fig. 1B). Then, for nodes leading to the query sequence, it evaluates their status as speciation events or gene duplications.

Result integration: The analytical pipeline integrates results from orthogroups delineated for all gene sequences used as queries (fig. 1C).

Three Analysis Modes

In the following, the analytical procedure is described with an example (Case Study 1 data, see below) downloaded from

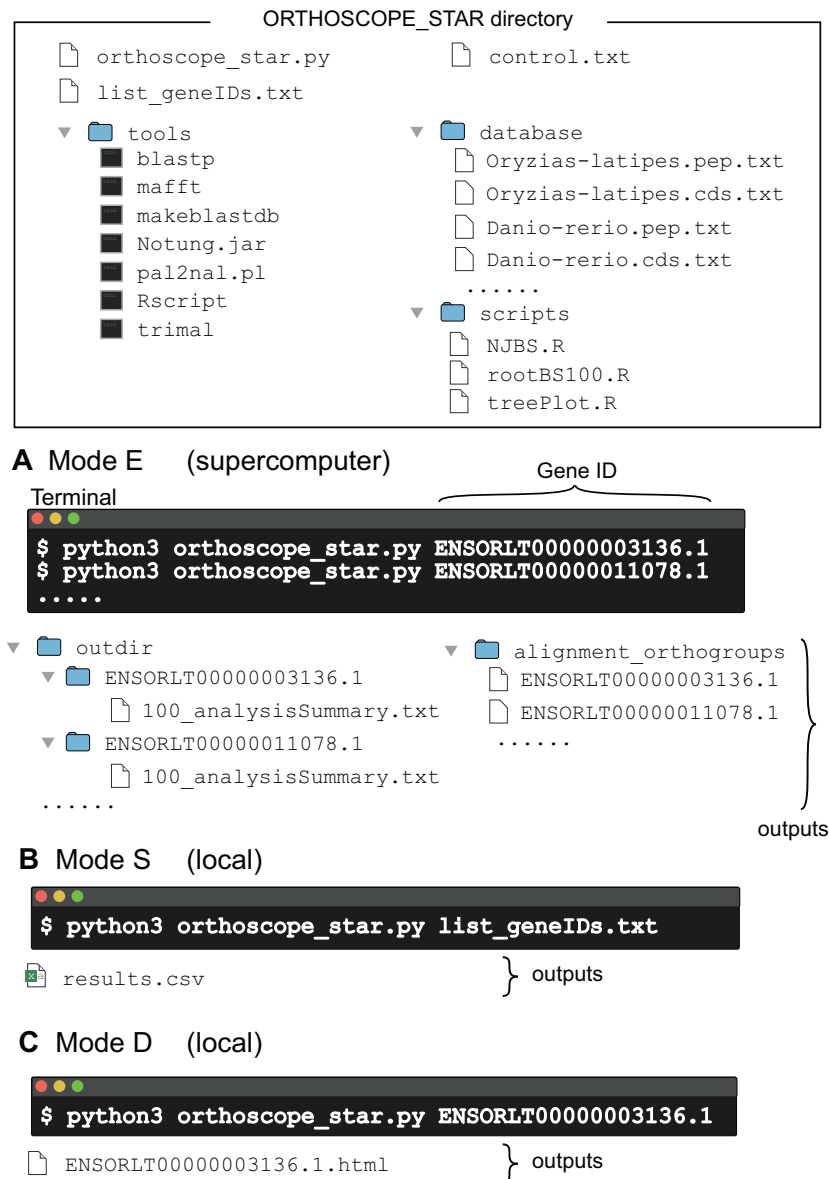


Fig. 2. Three modes of ORTHOSCOPE* analysis. The downloaded directory contains files of the main script (orthoscope_star.py), parameters (control.txt), and gene ID list (list_geneIDs.txt), and directories of dependencies (tools), nucleotide/amino acid fasta files (database), and additional scripts for R (scripts). (A) Mode E analysis. Those analyses produce two directories containing results of each query analysis: estimated gene trees, orthogroup members, etc. (outdir) and DNA alignment of orthogroup members and outgroup (alignment_orthogroups). Mode E analyses are designed to be conducted using supercomputers. (B) Mode S analysis. By using outputs (outdir) from Mode E analyses, Mode S analysis produces a file (results.csv) summarizing results of all query analyses assigned in the list_geneIDs.txt file. This analysis can be done using local computers. (C) Mode D analysis. By using an output (outdir) from Mode E analysis, Mode D analysis produces an html file showing estimated gene trees (fig. 3).

github: https://github.com/jun-inoue/ORTHOSCOPE_STAR. It is designed to work on Linux or Unix (Mac) systems (fig. 2).

(A) Mode E: The “Mode E” option estimates a gene tree. All ORTHOSCOPE* options can be selected in the control.txt file (supplementary fig. S1, Supplementary Material online). With the “TaxonSampling” option, taxonomic sampling is determined by describing species names and corresponding colors used in drawing trees (fig. 3). To the right of the species name/color, file names of amino acid, and coding sequences are described. Those files should be saved in the “database” directory (fig. 2). A gene model (amino acid and coding sequence data sets) should be constructed for each species.

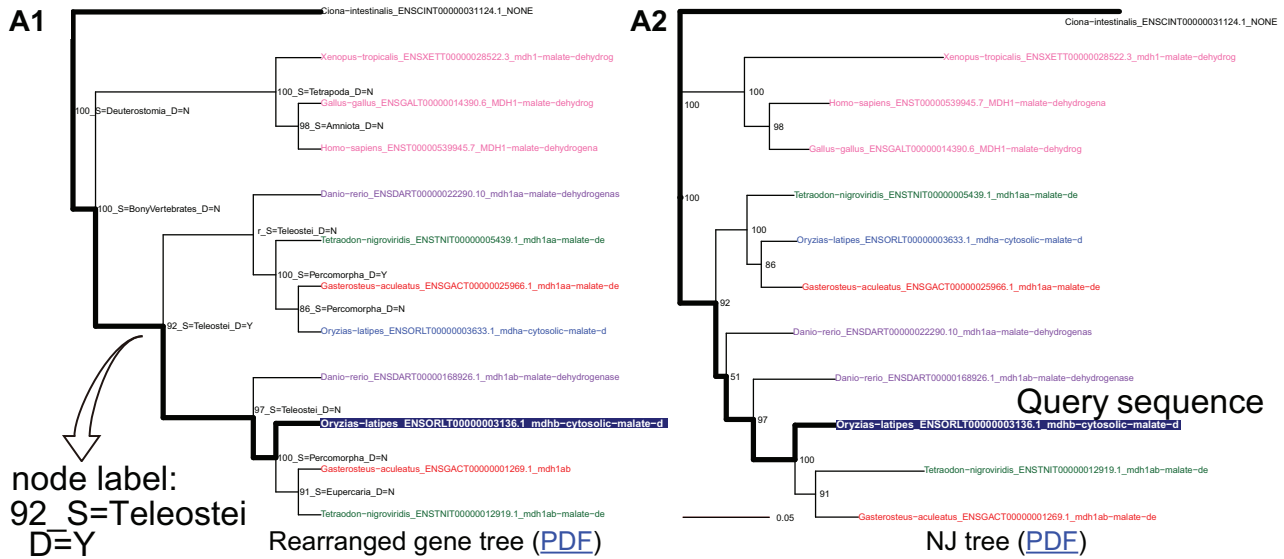
Gene models for more than 550 animal species can be downloaded from the ORTHOSCOPE website (see the ORTHOSCOPE* instruction page). Users wishing to run their own data sets should replace the database directory with that containing input fasta files using name lines, following examples and explanations shown in supplementary figure S2, Supplementary Material online. With the “SpeciesTree” option, the species tree is described as the Newick format, including node names.

ORTHOSCOPE* starts an analysis when receiving a query gene ID as a parameter for the query sequence (as shown in the Terminal in fig. 2A). This gene ID should be included in

ORTHOSCOPE STAR

Query sequence: Oryzias-latipes_ENSORLT00000003136.1
 _mdhb-cytosolic-malate-d
[Summary](#)
 BS value of orthogroup-basal node (1st tree): 98

2nd tree: Speciation/duplication events in the query sequence lineage



1st tree: Orthogroup

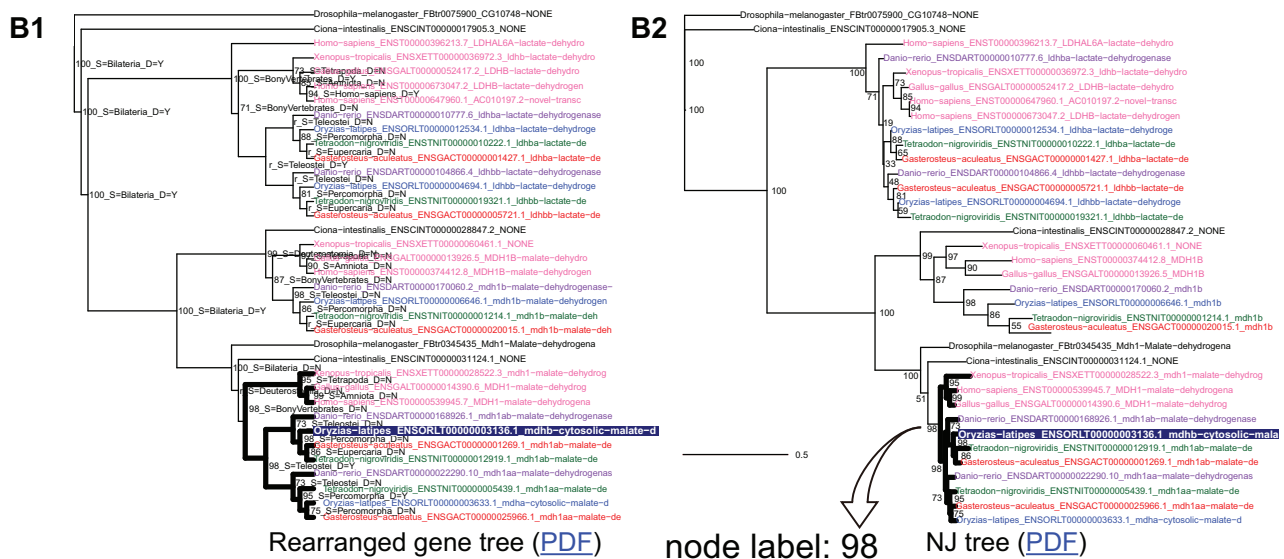


Fig. 3. Example of visualized ORTHOSCOPE* results. “ENSORLT00000003136.1” was used as the query gene ID (highlighted with dark blue background) for Mode D analysis. (A1) A rearranged gene tree of orthogroup members. Node labels indicate bootstrap values, gene node names, and assigned statuses, duplication (D = Y) or speciation (D = N). (A2) A gene tree of orthogroup members. (B1) A rearranged gene tree of all BLAST hits. (B2) A gene tree of all BLAST hits.

name lines of gene model files (supplementary fig. S2, Supplementary Material online) constructed for the query species. The query species is defined using the “QuerySpecies” option in the control.txt file (supplementary fig. S1, Supplementary Material online). A BLAST search is first conducted using the query sequence against all data from

user-defined species. Based on multiple alignments using all BLAST-hit sequences, ORTHOSCOPE* estimates a gene tree. In the estimated gene tree (fig. 1A), the analytical pipeline delineates an orthogroup by finding a basal node corresponding to the user-defined key node using the “KeyNode” option. The defined key node is used throughout all query analyses.

To rigorously infer the evolutionary history of the query sequence, the analytical pipeline estimates a gene tree (fig. 1B) using sequences of orthogroup members and rooting selected from the first gene tree (fig. 1A). Results are saved in the “100_analysisSummary.txt” file with respect to each query sequence in the “outdir” directory (fig. 2A). The “BSthreshold” parameter, the threshold to bootstrap support values of gene node, is used to evaluate the accuracy of resultant orthogroup/gene nodes. Alignments of orthogroup members are saved in the “alignment_orthogroups” directory (fig. 2) to check the quality of alignments. For genome scale analyses, ORTHOSCOPE* should be run with the Mode E option by employing multiple query sequences separately.

(B) Mode S: After conducting analyses with the Mode E option, Mode S analysis (fig. 2B) integrates results from multiple query sequences. The analysis starts when the program receives a gene ID list (list_geneIDs.txt) as a parameter in your Terminal. In the downloaded example, results derived from 121 query sequences are integrated in the results.csv file (supplementary table S2, Supplementary Material online). Each row contains: the query gene ID (QueryGeneID), the gene ID of a species with a gene function to represent functions of orthogroup members (SpeciesWithGeneFunction), the bootstrap value of the orthogroup basal node (BS_of_orthogroupBasalNode), numbers of BLAST hits (BHnum_*) and orthogroup members (OGnum_*) of each species, and then, bootstrap values (BS_of_*_monophyly), duplication statuses (dupStatus_*), and sister gene nodes (Sister_of_*) of nodes leading to the query sequence (see supplementary fig. S3, Supplementary Material online for details).

(C) Mode D: Mode D analysis visualizes the resultant orthogroup and gene trees. The analysis starts when the analytical pipeline receives a gene ID as a parameter (fig. 2C). Example output for “ENSORLT00000003136.1” is shown in figure 3.

Three Case Studies

Here, the utility of ORTHOSCOPE* is demonstrated using case studies with three data sets, with taxonomic sampling relative to teleosts, actinopterygians, and deuterostomes.

Case Study 1: Evaluation of Gene Node Status

By comparing manually estimated gene trees (Sato et al. 2009), I tested whether ORTHOSCOPE* can properly assign gene node status (speciation or gene duplication). Teleost-specific genome duplication (TGD) occurred at the beginning of teleost evolution (fig. 4) (Braasch and Postlethwait 2012). Therefore, traces of TGD can be found in some gene trees. In such cases, the basal node of the teleost-gene clade is evaluated as a gene duplication ($D = Y$ in fig. 1B). In contrast, if the paired gene derived from the TGD was lost just after the event, the status of the basal node is evaluated as speciation ($D = N$).

To evaluate traces of TGD remaining in teleost gene trees, Sato et al. (2009) estimated gene trees manually for 130 human query sequences using data from four teleost species

(fig. 4A). To identify orthogroups, Sato et al. (2009) added data from tetrapods and invertebrates, although they did not mention the use of orthogroups at that time. For gene trees estimated using human query sequences, they evaluated status, gene duplication ($D = Y$; 3R in Sato et al. [2009]) or speciation ($D = N$; 1:1 in Sato et al. [2009]), for basal nodes of teleost-gene clades (supplementary table S1, Supplementary Material online).

In this case study, ORTHOSCOPE* analyses were conducted for 121 medaka gene sequences (supplementary table S2, Supplementary Material online) included in the gene trees estimated in Sato et al. (2009). To exclude ambiguously estimated orthogroups/gene nodes, the BSthreshold was set at 70. When the “BonyVertebrates” species node was used as the key node, 95 orthogroups were identified (supplementary table S2, Supplementary Material online) as fulfilling the BS threshold. Among these 95 orthogroups, 78 have basal nodes of teleost-gene clades fulfilling the BS threshold. For these 78 orthogroups, ORTHOSCOPE* identifies duplication ($D = Y$) for 38 orthogroups and speciation ($D = N$) for the remaining 40. For the same 78 orthogroups, Sato et al. evaluated $D = Y$ for 43 orthogroups and $D = N$ for 35. Between these two studies, different conclusions were drawn for nine orthogroups: 1) For one orthogroup, Sato et al. (2009) used ambiguous gene topology. 2) For the remaining eight orthogroups, ORTHOSCOPE* used ambiguous tree topologies. Considering that Sato et al. (2009) employed only one representative for four teleost lineages (fig. 4A), ORTHOSCOPE* can evaluate the status of gene nodes more accurately once denser taxonomic sampling is employed.

Case Study 2: Calculating the Fraction of Duplicated Genes at Species Nodes

By counting numbers of gene duplications, ORTHOSCOPE* can be used to estimate phylogenetic positions of WGD events. For this purpose, an ORTHOSCOPE* analysis was conducted using actinopterygian data (fig. 4B). To bisect possible long branches of teleost gene lineages, two species were selected to represent each of the three major teleost lineages (Otophysi, Protacanthopterygii, and Percomorpha) with known phylogenetic relationships (Nelson et al. 2016). In order to count the number of gene duplication events before the TGD, data from two nonteleost actinopterygians were included. When employing all 19,699 medaka gene sequences as queries (supplementary table S3, Supplementary Material online), 11,539 query analyses produced orthogroups fulfilling the 70% BS criterion for their basal node (BS_of_orthogroupBasalNode).

Traces of TGD. Among these 11,539 orthogroups, 6,269 orthogroups contained monophyletic teleost-gene clades supported by $>70\%$ BS values (BS_of_Teleostei_monophyly in supplementary table S3, Supplementary Material online). Among them, for the status of the teleost gene node (dupStatus_Teleostei in supplementary table S3, Supplementary Material online), 2,062 orthogroups were evaluated as duplications ($D = Y$) and the remaining 4,207

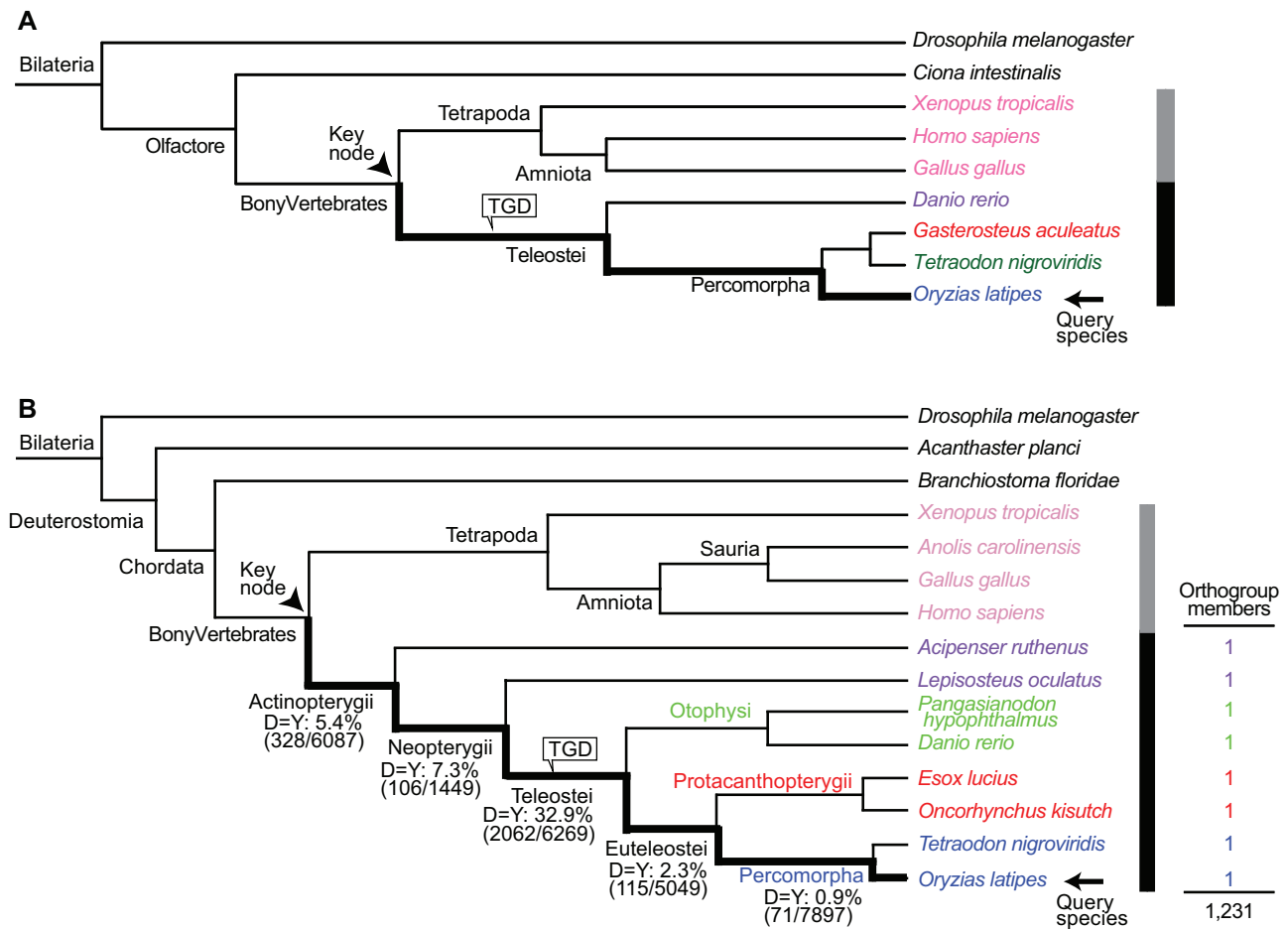


Fig. 4. Species trees employed in Case Studies 1 and 2. (A) Case Study 1. (B) Case Study 2. Fractions of genes showing traces of gene duplication events are shown at nodes leading to the query species. Numbers under species node names indicate fractions of gene duplication traces that occurred just before divergence at each node. Those fractions are obtained by dividing the numbers of duplication status ($D = Y$) by those of all observed monophyletic nodes fulfilling the bootstrap criterion. Thus, those fractions indicate percentages of genes retaining counterparts of duplicated genes in ancestral genomes. At the basal node of the Actinopterygii, the fraction is calculated by removing 3,133 orthogroups, including nodes assigned as duplications ($D = Y$) due to unusual placements of *Acipenser* and/or *Lepisosteus* gene sequences. Those were identified as sistergroups of the teleost-gene clade shown in the Sister_of_Teleostei column (supplementary table S3, Supplementary Material online). On the right side, the pattern of orthogroup members (sharing one member among all actinopterygian species) and the number of orthogroups having such a member pattern are shown. For vertical bars separated by black and gray segments, see fig. 1. TGD, teleost genome duplication.

orthogroups were considered speciation events ($D = N$). This means that 32.9% (2,062/6,269) of medaka genes produced gene trees retaining TGD traces. Thus, 32.9% of genes in the ancestral genome at this node are considered to retain counterparts derived from TGD. This fraction is comparable to that estimated in my previous study (Inoue et al. 2015). In that study, the fraction of TGD traces was estimated using gene model data from nine teleosts. That analysis produced a smaller fraction, 18% (1,237/6,892), than the present study, because of removing orthogroups counted more than once due to the presence of TGD-derived medaka paralogs and by using human gene sequences as queries.

Some gene trees also retain traces of gene duplications that occurred in other species nodes. In the same manner as the above TGD analysis, fractions of gene duplications that occurred just before species nodes leading to query species were calculated (fig. 4B). As a result, among species nodes from the base of the Actinopterygii to the Percomorpha,

ORTHOSCOPE* produced much smaller fractions of gene duplications (0.9–7.3%) than Teleostei (32.9%). This indicates that ORTHOSCOPE* can be used as a first step for inferring phylogenetic positions of WGDs, by comparing calculated fractions along lineages leading to query species, although there is no criterion for the fraction of duplications needed to identify WGD traces. Therefore, occurrences of WGD should be confirmed by comparing genomic positions of WGD-derived paralogs, as shown in Inoue et al. (2015).

To illustrate the novelty of ORTHOSCOPE*, numbers of gene duplication events were compared with those estimated using a pioneering analytical pipeline in this field, OrthoFinder (Emms and Kelly 2019; last access date May 3, 2021), with special reference to nodes leading to *Oryzias latipes*. Except for ORTHOSCOPE*, OrthoFinder is the only other tool that can count numbers of gene duplication events by delineating orthogroups. Based on the same nucleotide databases (-d option), OrthoFinder analysis was conducted based on a

multiple sequence alignment (-M masa option) using the given species tree (-s option) (supplementary fig. S4A, Supplementary Material online). As a result, OrthoFinder also produced a larger number of gene duplication events for the teleost node (814) than for other nodes (21–294). OrthoFinder, however, does not estimate percentages of duplication events in ancestral genomes, and cannot exclude ambiguous events due to the lack of support at gene nodes.

Sistergroup Evaluation. ORTHOSCOPE* also evaluates sistergroup hypotheses in the species tree. Results obtained for the Case Study 2 data set can be used to evaluate three sistergroup hypotheses for the Percomorpha (fig. 4B): (A) Protacanthopterygii, (B) Otophysi, (C) Protacanthopterygii + Otophysi. Among 6,269 orthogroups having teleost-gene clades supported by >70% bootstrap values, 4,752 orthogroups showed a Percomorph gene clade with >70% bootstrap probability (BS_of_Percomorpha_monophyly in supplementary table S3, Supplementary Material online). Of these, 2,850 orthogroups supported one of the three sistergroup hypothesis (Sister_of_Percomorpha) with >70% bootstrap support (BS_with_Percomorpha). As expected, the number of orthogroups supporting the Protacanthopterygii hypothesis (2,520) was much larger than the number of remaining orthogroups (Otophysi, 66; Protacanthopterygii, 264).

When the same nucleotide databases were used without a given species tree option (-s option), OrthoFinder also identified Protacanthopterygii as the sistergroup of the Percomorpha (supplementary fig. S4B, Supplementary Material online) based on a concatenated multiple sequence alignment of single-copy genes (-M msa). OrthoFinder, however, cannot compare alternative sistergroup hypotheses among all orthogroup analyses.

Environmental DNA Marker Selection. ORTHOSCOPE* can offer a set of orthology-confirmed gene markers for environmental DNA (eDNA) analyses. The recently developed eDNA analysis has been used to estimate the distribution of aquatic vertebrates using mitochondrial DNA (mtDNA) as a genetic marker (Wang et al. 2021). However, mtDNA markers have certain drawbacks, such as low-resolution species identification due to low sequence variability. Although some studies reported the availability of nuclear DNA markers for eDNA analyses (Jo et al. 2020), candidates of nuclear environmental DNA of macro-organisms were not selected from genome-wide data.

For further progress in the eDNA analysis of aquatic vertebrates, a greater number of reliable, orthology-confirmed nuclear gene markers is required. They are desired to be 1:1 single-copy genes that have lost one of a pair after TGD, but before teleost diversification. We found 1,231 genes belonging to 1:1 orthogroups between four tetrapods and six teleosts by excluding cases of reciprocal gene lineage loss between teleost lineages (fig. 4B). A script for extracting these gene markers from the “results.csv” file is available from the instruction page.

Case Study 3: Presence or Absence of Genes

ORTHOSCOPE* can evaluate presence or absence of genes within a species lineage. By using the web version of ORTHOSCOPE, Inoue et al. (2019) confirmed the presence of cellulose synthase (CesA) orthologs in all sequenced tunicate genomes, but its absence in other metazoan genomes. This indicated that the prokaryotic CesA gene was horizontally transferred into the genome of a tunicate ancestor from a bacterium. Tunicates are the only metazoans that can synthesize cellulose, a biological function associated with bacteria and plants, but not animals. Are there any other genes transferred to the genome of a tunicate ancestor from bacteria?

Based on data comprising the 49 metazoans used in Inoue et al. (2019), an ORTHOSCOPE* analysis was conducted employing 16,671 *Ciona intestinalis* sequences as queries (fig. 5). As a result, the CesA gene was the only gene with orthologs in all sequenced tunicate genomes, but absent in other metazoan genomes (e.g., BHnum_ *Drosophila-melanogaster* in supplementary table S4, Supplementary Material online). This indicates that no additional genes were transferred to ancestral tunicates from bacteria to form the cellulose synthesis system in tunicates.

In addition, lineage-specific gene losses were compared among chordate lineages using ORTHOSCOPE*. In this case study, the bootstrap criterion was set to 60% due to the use of the bilaterian species node as the key node. As a result, 739 orthogroups fulfilled the BS value criterion (supplementary table S4, Supplementary Material online). When detecting lineage-specific lost genes in nonurochordate chordates (e.g., OGnum_ *Homo-sapiens* [number of orthogroup member] in supplementary table S4, Supplementary Material online), gene losses (85) that occurred in the vertebrate ancestor were more numerous than in the cephalochordate ancestor (24, such as ARSG and JPT2 of SpeciesWithGeneFunction in supplementary table S4, Supplementary Material online). Given that gene numbers in common vertebrate genomes (~26,500) are similar to those of cephalochordate genomes (~30,400) (Sato 2016), this result supports the hypothesis that vertebrate ancestors lost genes shared among bilaterians after increasing gene numbers via vertebrate-specific genome duplication events.

Conclusion

ORTHOSCOPE* Version 1 infers gene duplication events that occurred at ancestral species nodes and evaluates presence or absence of genes among species lineages. Although case studies were for actinopterygians and chordates, ORTHOSCOPE* analysis can be applied to protostome, cnidarian, and plant gene models available from the web version. In addition, orthogroups identified by ORTHOSCOPE* can be evaluated by ORTHOSCOPE web version. ORTHOSCOPE* has several limitations: 1) Mode E analyses for genome-scale data are premised on parallel analyses using a supercomputer; 2) users should employ a fully bifurcated species tree, although it can include ambiguous relationships, except for key nodes; and 3) some phylogenetic relationships of genes cannot be resolved due to low sequence variability at the population level or

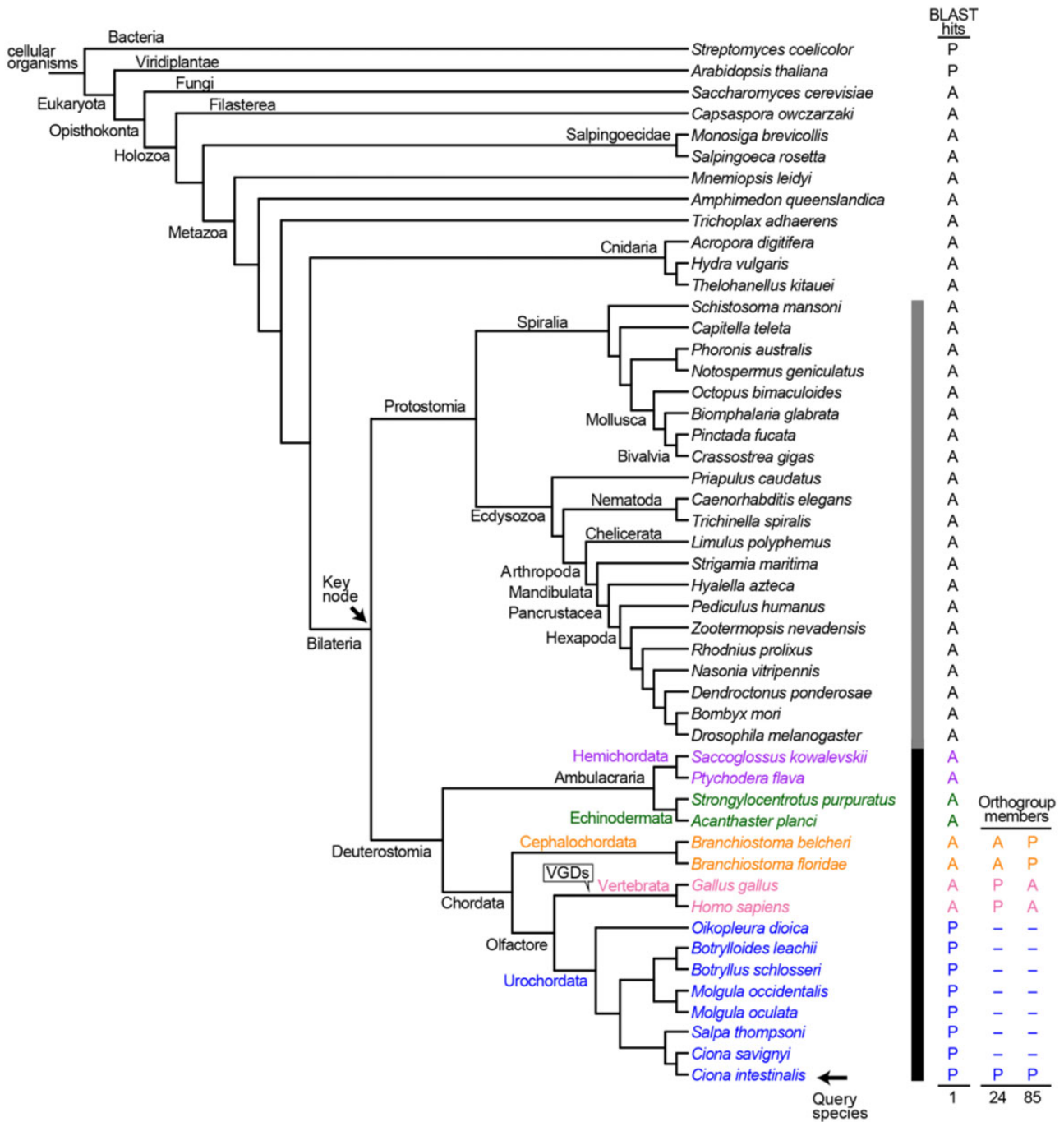


Fig. 5. Species trees employed in Case Study 3. Presence or absence of BLAST hits/orthogroup members and numbers of gene analyses having such patterns are shown at the right side of the tree. For “BLAST hits,” P means presence of BLAST hits and A means absence of such hits in the gene model of target species. For “Orthogroup members,” P means presence of orthogroup members and A means absence of such members in each estimated gene tree. A dash means no-evaluation for P or A in the analysis. VDGs, vertebrate genome duplications.

saturated sequence substitutions at the ancient divergence level.

Materials and Methods

ORTHOSCOPE* is written in Python. It requires standalone versions of BLASTP and MAKEBLASTDB in BLAST+ (Camacho et al. 2009), MAFFT (Katoh and Standley 2013), TRIMAL (Capella-Gutierrez et al. 2009), PAL2NAL (Suyama et al. 2006), APE (Popescu et al. 2012) and RSCRIPT in R (Ihaka

and Gentleman 1996) with R itself, and NOTUNG (Chen et al. 2000). These freely available applications must be installed separately. Parameters of options (below) applied in those dependencies are consistently used through three case studies.

Mode E: In ORTHOSCOPE* analyses, BLAST searches are conducted by constructing amino acid databases. By using an amino acid file defined in the control.txt file, ORTHOSCOPE* automatically creates an amino acid database for each species

by MAKEBLASTDB with `-dbtype prot` option (for more detail, see <https://www.ncbi.nlm.nih.gov/books>). Against those amino acid databases created for each species, protein-coding gene sequences are used as queries for BLASTP searches. The resulting BLAST top hits are screened according to a user-defined number (`Number_of_hits_to_report_per_genome` parameter) using an E-value cutoff (`BLAST_Evalue_threshold_for_reported_sequences`).

Primary sequences of proteins obtained from BLASTP searches are aligned using MAFFT with the default settings. Multiple sequence alignments are trimmed by removing poorly aligned regions using TRIMAL with the option, “gappayout.” Corresponding coding sequences are forced onto the amino acid alignment using PAL2NAL with the default settings to generate nucleotide alignments for later comparative analyses. Each gene sequence is checked and removed from the alignment as a spurious BLAST hit if the sequence is shorter than a user-defined value (`ShortSequence_threshold`) for the length of the query sequence in unambiguously aligned sites.

Phylogenetic analyses are conducted using the NJ method aligned with bootstrap analysis based upon 100 replicates using the software package, APE, in R. Analyses are conducted using a user-defined data set (excluding third codon positions) with the TN93 model (Tamura and Nei 1993).

Resulting gene trees, however, often have some weakly supported nodes. In such cases, one needs to revise ambiguous nodes in comparison with the species tree. For this purpose, ORTHOSCOPE* conducts rearrangement/reconciliation analysis using a method implemented in NOTUNG. As a first step, with the `-rearrange` option, NOTUNG rearranges weakly supported nodes of the gene tree to minimize duplication and extinction of genes, using parsimony with equal weights. To save more than one tree rooted on the highest scoring edges with feasible reconciliations, `-maxtree` option is set as five. A user-defined value (`BSthreshold` option, `-threshold` option in NOTUNG) is used as the threshold for bootstrap support of nodes. Then the rearranged tree is reconciled with the species tree.

To select reliable orthogroups, the first NJ trees derived from rearrangement/reconciliation analyses are filtered using the user-defined bootstrap value for orthogroup basal nodes. To estimate more accurate gene trees, orthogroup members and rooting sequences are selected from rearranged first gene trees and realigned to conduct second NJ analyses. The resulting second NJ gene trees are subjected to rearrangement/reconciliation analyses for species/gene duplication identification. The result of each query analysis is saved in a text file.

Mode S: Using results from multiple query sequences by Mode E analyses, ORTHOSCOPE* produces the results.csv file.

Mode D: Using the result file, ORTHOSCOPE* plots gene trees using APE, in R.

Supplementary Material

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

Acknowledgments

The author thanks Noriyuki Satoh, Atsuo Nishino, and Seiko Yoshikawa for advice on the early version of manuscript, Steven D. Aird for English language editing, and three anonymous reviewers for helpful comments on the manuscript. Cluster computing resources were provided by OIST and the Atmosphere and Ocean Research Institute (AORI), the University of Tokyo. This work was conducted in part under the FSI project Ocean DNA: Constructing “Bio-map” of Marine Organisms using DNA Sequence Analyses from the University of Tokyo, and supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (C) (18K06396) and (A) (21H04922).

References

- Altenhoff AM, Glover NM, Dessimoz C. 2019. Inferring orthology and paralogy. *Methods Mol Biol.* 1910:149–175.
- Altenhoff AM, Levy J, Zarowiecki M, Tomiczek B, Warwick Vesztrocy A, Dalquen DA, Muller S, Telford MJ, Glover NM, Dylus D, et al. 2019. OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.* 29(7):1152–1163.
- 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.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
- 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 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.
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20(1):238.
- Fernández R, Gabaldon T, Dessimoz C. 2020. Orthology: definitions, prediction, and impact on species phylogeny inference. In: Celine S, Frédéric D, Nicolas G, editors. *Phylogenetics in the genomic era*. p. 2.4:1–2.4:14.
- Futuyma DJ, Kirkpatrick M. 2017. *Evolution*. Sunderland (MA): Sinauer Associates, Inc.
- 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.
- Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. *J Comp Graph Stat.* 5(3):299–314.
- Inoue J, Nakashima K, Satoh N. 2019. ORTHOSCOPE analysis reveals the presence of the cellulose synthase gene in all tunicate genomes but not in other animal genomes. *Genes (Basel)* 10: 294.
- 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. 2019. ORTHOSCOPE: an automatic web tool for phylogenetically inferring bilaterian orthogroups with user-selected taxa. *Mol Biol Evol.* 36(3):621–631.
- Jo T, Arimoto M, Murakami H, Masuda R, Minamoto T. 2020. Estimating shedding and decay rates of environmental nuclear DNA with relation to water temperature and biomass. *Environ DNA* 2(2):140–151.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* 39:309–338.

- Nagy LG, Merenyi Z, Hegedus B, Balint B. 2020. Novel phylogenetic methods are needed for understanding gene function in the era of megascale genome sequencing. *Nucleic Acids Res.* 48(5):2209–2219.
- Nelson JS, Grande T, Wilson MVH. 2016. *Fishes of the world*. Hoboken (NJ): John Wiley & Sons.
- 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.
- 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 Evol Biol.* 9:127.
- Satoh N. 2016. *Chordate origins and evolution: the molecular evolutionary road to vertebrates*. Boston (MA): Elsevier.
- Sonnhammer EL, Gabaldon T, Sousa da Silva AW, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas PD, Dessimoz C, Quest for Orthologs Consortium. 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.
- 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.
- Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P. 2011. Orthology prediction methods: a quality assessment using curated protein families. *Bioessays* 33(10):769–780.
- Wang S, Yan Z, Hanfling B, Zheng X, Wang P, Fan J, Li J. 2021. Methodology of fish eDNA and its applications in ecology and environment. *Sci Total Environ.* 755(Pt 2):142622.