

# aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity

Shigehiro Kuraku<sup>1,\*</sup>, Christian M. Zmasek<sup>2</sup>, Osamu Nishimura<sup>1</sup> and Kazutaka Katoh<sup>3,4,\*</sup>

<sup>1</sup>Genome Resource and Analysis Unit, RIKEN Center for Developmental Biology, Kobe, Hyogo 650-0047, Japan, <sup>2</sup>Program in Bioinformatics and Systems Biology, Sanford-Burnham Medical Research Institute, La Jolla, CA 92037, USA, <sup>3</sup>Immunology Frontier Research Center, Osaka University, Osaka 565-0871, Japan and <sup>4</sup>Computational Biology Research Center, The National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan

Received February 17, 2013; Revised April 11, 2013; Accepted April 18, 2013

## ABSTRACT

We report a new web server, aLeaves (<http://aleaves.cdb.riken.jp/>), for homologue collection from diverse animal genomes. In molecular comparative studies involving multiple species, orthology identification is the basis on which most subsequent biological analyses rely. It can be achieved most accurately by explicit phylogenetic inference. More and more species are subjected to large-scale sequencing, but the resultant resources are scattered in independent project-based, and multi-species, but separate, web sites. This complicates data access and is becoming a serious barrier to the comprehensiveness of molecular phylogenetic analysis. aLeaves, launched to overcome this difficulty, collects sequences similar to an input query sequence from various data sources. The collected sequences can be passed on to the MAFFT sequence alignment server (<http://mafft.cbrc.jp/alignment/server/>), which has been significantly improved in interactivity. This update enables to switch between (i) sequence selection using the Archaeopteryx tree viewer, (ii) multiple sequence alignment and (iii) tree inference. This can be performed as a loop until one reaches a sensible data set, which minimizes redundancy for better visibility and handling in phylogenetic inference while covering relevant taxa. The work flow achieved by the seamless link between aLeaves and MAFFT provides a convenient online platform to address various questions in zoology and evolutionary biology.

## INTRODUCTION

In any cross-species comparison at the molecular level, identification of orthology and paralogy is the basis on which most subsequent analyses rely (1). The most reliable approach for distinguishing orthologues from paralogues is by explicit phylogenetic inference. Some databases host genome-wide sets of molecular phylogenetic trees for individual gene families ('phyloomes') (2–4). However, those existing databases provide only phyloomes for a limited number of species with genome-wide sequence resources, and cannot fully accommodate biologists' daily demands for custom data sets, sometimes including organisms without genome-wide information or sequences identified on their own. To achieve smooth custom analyses from sequence collection to tree inference, several tools have been developed (5–8). However, they require program installation on local systems or/and elaborate database maintenance and allow no handy interface for refining data sets.

Moreover, a large amount of molecular sequence data has been produced, thanks to recent developments in sequencing technologies. In fact, the resultant resources for protein-coding gene sets are not integrated into a single archive, such as GenBank (9) managed by NCBI (10), but are scattered in separate databases including independent project-based web sites. For example, GenBank does not host genome-wide data of many species with sequenced genomes including *Xenopus tropicalis*, *Danio rerio* and *Ciona intestinalis*, all of which are available at Ensembl (<http://www.ensembl.org/>) (2). Conversely, Ensembl does not host many invertebrate species with sequenced genomes that are hosted at EnsemblGenomes (<http://www.ensemblgenomes.org/>), such as those of many insects including *Apis mellifera* (honey bee). Resources for other species that are not in

\*To whom correspondence should be addressed. Tel: +81 78 306 3331; Fax: +81 78 306 3048; Email: shigehiro-kuraku@cdb.riken.jp  
Correspondence may also be addressed to Kazutaka Katoh. Tel: +81 33 599 8684; Fax: +81 33 599 8081; Email: katoh@ifrec.osaka-u.ac.jp

either NCBI or Ensembl are also deposited in project-based individual web sites, such as the site of Joint Genome Institute (<http://www.jgi.doe.gov/>; e.g. for *Capitella teleta*). The heavily scattered data deposition complicates data access and is becoming a serious barrier to large-scale sequence comparison and molecular phylogenetic analyses.

To accommodate the demand for building high-coverage phylogenetic trees by a wide range of biologists including laboratory workers with specific interest in particular molecules, we have established a new web server, aLeaves (èilí:vz; <http://aleaves.cdb.riken.jp/>), which provides a simple and biologist-oriented interface and easy-to-interpret output. Sequences collected by aLeaves can readily be passed on to multiple sequence alignment (MSA) and phylogenetic tree inference on the MAFFT server upgraded to enhance its function through the co-ordination with aLeaves. Here we introduce the seamless work flow achieved by these two web servers.

### COLLECTING SEQUENCES ON THE ALEAVES SERVER

The aLeaves server accepts a protein sequence query to run a BLASTP search based on NCBI BLAST (11) over pre-compiled databases covering multiple species (Figures 1 and 2). Users can set (or unset) low complexity filtering, the threshold for search hits and the number of sequences to retrieve (Figure 1). The databases include genome project-based datasets covering ~100 metazoan species (Table 1), as well as sequences from the widely used public database GenBank. The latter category contains biologically annotated sequences submitted by individual researchers and serves as an indispensable information source in analyses focusing on particular targets. As explained above, the taxonomic coverage of aLeaves is wider than that of Ensembl or NCBI Genome, and none of other currently available tools can explore such a comprehensive list of diverse metazoan species' genomes online in a single search. We do not perform any gene prediction on genome assemblies, but instead adopt data sets that already exist as protein sequences. The list of databases available at aLeaves will be enriched by incorporating emerging information from more species with frequent updates. Updates of the aLeaves server will be announced in the external blog site linked from the 'History' page. The 'Help' page of aLeaves serves step-by-step tutorial of its function as well as frequently asked questions with answers for them.

After the BLASTP search, the specified numbers of sequences are retrieved in form of a multi-fasta file, sorted by E-value, and downloadable from the results page. The output of collected sequences has species identifiers ('Species ID'; for example, 'HOMSA' for *Homo sapiens*) for sequences derived from genome-wide databases and category identifiers ('Category ID'; for example, 'ART' for arthropods) for sequences derived from the GenBank database (the lists of the identifiers are available at <http://aleaves.cdb.riken.jp/aleaves/species.html>). The results page also provides a link to the sequence clustering

functions on the MAFFT server (12) where subsequent analyses are performed.

### ALIGNMENT AND TREE INFERENCE ON THE MAFFT SERVER

The sequences transferred from the aLeaves server are first subjected to all-to-all pairwise comparison implemented in the MAFFT program (13,14). Several different methods, a rapid one based on the number of shared *k*-mers (15) and more rigorous ones based on local or global pairwise alignment scores (16–18), can be selected in this step, according to the data size. The pairwise distances are used for initial clustering by UPGMA or minimal linkage. The resulting tree is intended to roughly visualize the tree-like relationship for even large numbers of sequences and to provide visual guidance for the next steps, and is displayed with a Java applet version of the Archaeopteryx phylogenetic tree visualization and analysis tool (19,20). By using the new features described below, users can select any subset of the collected sequences. Then, an MSA of the selected sequences is calculated by MAFFT and subjected to phylogenetic tree inference with the neighbour-joining (NJ) method (21).

One highlight of the new features of the MAFFT server is the communication between Archaeopteryx and the sequence data set refinement process in the web browser for user-friendly sequence selection (Figure 3). A useful practice in evolutionary analysis is the detailed (re)analyses of the phylogenetic relationships of the sequences making up one or more subtree of a larger, previously inferred but relatively inaccurate, tree. One advantage of this approach is that such an analysis is usually more accurate because moderately divergent sequences can be aligned more precisely than larger sets of too divergent sequences—inaccurate MSAs are one of the main sources of errors in phylogenetic analysis. Furthermore, the removal of redundant uninteresting sequences allows for analysis with more accurate but more time-consuming approaches, and simplifies visual analysis. While manual selection of sequences and realignment is possible, it is labour intensive and error prone. Therefore, we modified Archaeopteryx in such a way that one can select or unselect, via its graphical user interface, individual sequences as well as complete sets of sequences included in particular subtrees (by clicking on their common root node) for reanalysis. This improvement should dramatically ease the manual process to select or unselect sequences from a preliminary data set often containing truncated, redundant or phylogenetically too distant sequences. To further facilitate visual tree analysis, species identifiers and category identifiers given by aLeaves are automatically recognized and used for species-specific colouring of sequence labels in trees visualized in Archaeopteryx (Figure 3). This function allows users to get a rough overview of the molecular phylogeny and in this initial step, delete distantly related sequences that can complicate and slow down the subsequent steps.

Two additional tools, CD-HIT (22) and MaxAlign (23), have been adopted in the MAFFT service to help the data

## A

### aLeaves - [èilf: vz]

first step to build zoologically informative phylogenetic trees

Top Database Species History Help About Links

**What is 'aLeaves'?**

aLeaves allows you to collect homologs of the sequence of your interest. Those sequences can be passed on to the web server for sophisticated multiple alignment, and then to tree building. This initial process one can perform here is just like collecting [leaves](#) (terminals of trees) from diverse animals. This is why we designated this tool 'aLeaves'.

**Start the search (powered by NCBI Blast)**

Enter your "query" **peptide sequence in the fasta format** (example):

```
>zebrafish bmp1b
MFFASLLVLMILLPQASSGHOEGPSOHTGKLDLSLEPSLAHTIQNLLLTRLGLQSHNPSTKAQVPOYL
LDLYRFHTDOOYHLIEDPEFSYPSKVGOGANTVRFHHTDSPTPSLPEEQKTTVDGIHIGFNLSISPSEES
VYSAEIIRLLHEGSSGGSHVASLYLSNHOPSSKPIILLHSRQLTRDRKSAQLWETFFLDREVFQNLKSTSGS
LSFILDVLPDSNSSLTPKQRHLRVRKSTLQDQPTWERORPLLYTSHDGRSEPFVNLKRRKSSRSRSRW
TRFKDGRGOGSDWNERRKRNRRAAKLRLSRARCRRHPLYDFKDVGVNKKWI IAPSGYDAFFCLGECR
FLIDRNRGSGHWTIYVQVQVAVDFDQVPTLQALFLDPECFVNLKRRKSSRSRSRW
```

Or upload a file:

**Select database (one or more)** (number of sequences in parentheses):

- 1. Human - Refseq (26,077) [detail]
- 2. Human - Ensembl (104,785) [detail]
- 3\*. Non-human eutherians - Ensembl (743,219) [detail]
- 4. Non-eutherian mammals - Ensembl (83,640) [detail]
- 5. Non-mammalian vertebrates - Ensembl (377,477) [detail]
- 6. Cartilaginous fishes and cyclostomes (33,508) [detail]
- 7\*. All vertebrate entries except mammals in NCBI Protein (906,275) [detail]
- 8. Invertebrate deuterostomes (106,784) [detail]
- 9. Arthropods (408,642) [detail]
- 10. Nematodes (111,266) [detail]
- 11. Other protostomes (182,289) [detail]
- 12. Non-bilaterian metazoans (cnidarians, placozoon & poriferan) (342,161) [detail]
- 13\*. All metazoan entries except vertebrates in NCBI Protein (2,331,372) [detail]

(caution: selecting the database marked with \* will largely slow down the search)

Number of homologs to collect:

Threshold E-value:

Set low complexity filter or not [help of NCBI Blast]:  no  yes

## B

### aLeaves - [èilf: vz]

first step to build zoologically informative phylogenetic trees

Top Database Species History Help About Links

Search finished!

**Results**

Collected sequences in multifasta format: [Open or download](#)  
(File size = 0.41 Mbyte)

List of the collected sequences with database links: [Open or download](#)

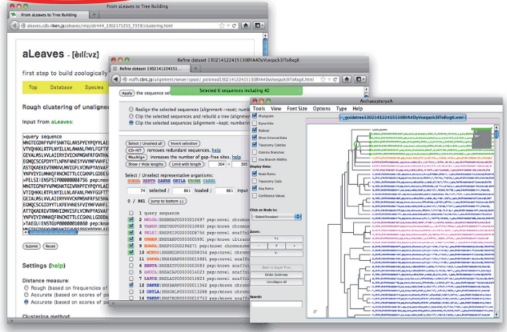
Raw result of the BLAST search: [Open or download](#)

(Caution: The results will be stored under the above links for only 7 days)

Redo the search

**Proceed to tree building (powered by MAFFT server)**

Click here! (open in a new window)

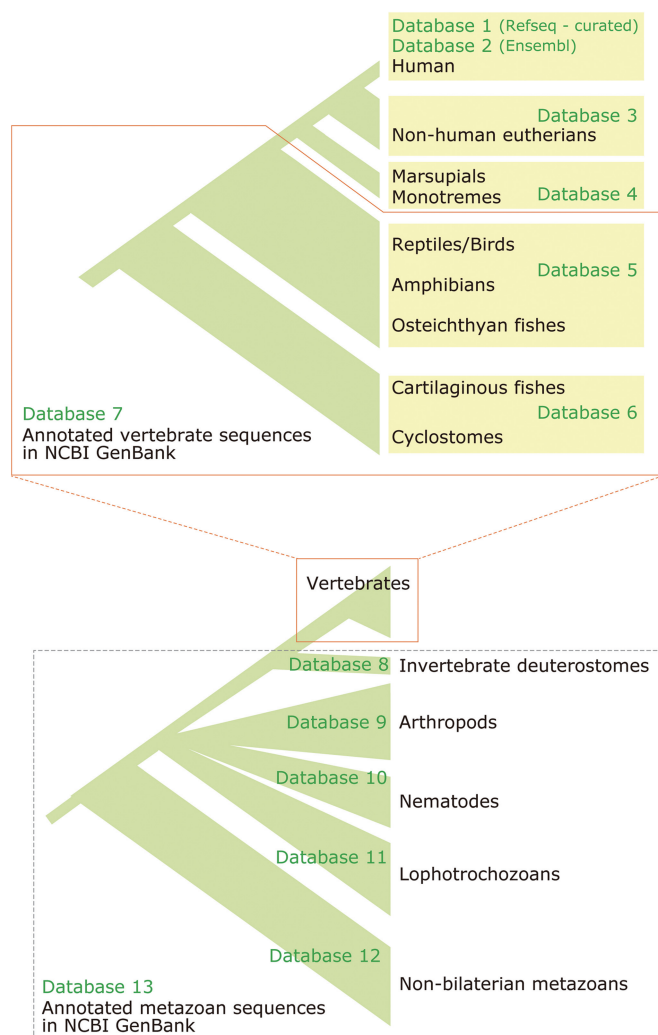


**Figure 1.** Overview of the interface of the aLeaves web server. (A) The 'Top' page of the aLeaves server, containing the search interface. (B) The results page shown when a search and sequence collection is completed. The 'Proceed to tree building' section (red oval) provides a gateway to the MAFFT server for the rest of the process from data set refinement to molecular phylogenetic tree inference.

refinement process. CD-HIT is used to exclude redundant sequences. MaxAlign is used to exclude short amino acid sequences that are inconvenient for phylogenetic analyses. Even genome-based databases have such sequences mainly because of incomplete sequencing or gene misidentification. Users can check the result of these automatic processes and decide whether they accept the result in the next cycle. Whenever preferable, users can also secondarily add sequences of their choice to the data set using the 'New sequences' input form in the MAFFT web page. Refinement of sequence data set is a seemingly simple but virtually tedious process in molecular phylogenetic analyses. Retrieval of similar sequences from various databases usually results in a bulky data set, which can obscure the target of the analysis. In contrast, if one reduces too many sequences, the resultant data set may miss a crucial subset of molecular phylogeny, causing artefacts, such as 'hidden paralogy' (24). The MAFFT server with the enhancement of user-controlled interactive function for sequence selection provides a solution to

minimize the risk of handling data sets that tend to be unnecessarily huge initially.

The refined sequence set can then be aligned, and the aligned sequences can be subjected to molecular phylogenetic tree inference with the NJ method (21). For distance calculation in this step, several measures can be selected: the Poisson correction (25), maximum-likelihood (ML) estimation with the JTT model (26) and ML estimation with the WAG model (27). We modified the MOLPHY package (28) to consider rate heterogeneity across sites with the discrete  $\Gamma$  model (29) and use it in the ML distance estimation. Bootstrap analysis (30,31) is also available. To show a relationship among the selected sequences, at present, this service supports the NJ method, but does not support the ML method because of too much load on the server. After checking the NJ tree, users can download the MSA in the multi-fasta format and use it for tree inference with the ML or other methods outside the server. The sequences are also downloadable to apply different MSA methods other than MAFFT.



**Figure 2.** Phylogenetic coverage of the compiled databases available at the aLeaves web server. Numbering of the databases (Database 1–13) corresponds to that in the aLeaves server (<http://aleaves.cdb.riken.jp/aleaves/database.html>).

At any step in the process, one can go back to the data set refinement page with a single click and delete sequences through the Archaeopteryx applet. While a user switches between these different steps, previous versions of the data sets are stored and shown in the ‘History’ section of the HTML page, and a particular version of interest can be easily restored on demand.

## EXAMPLES

### Insect wingless

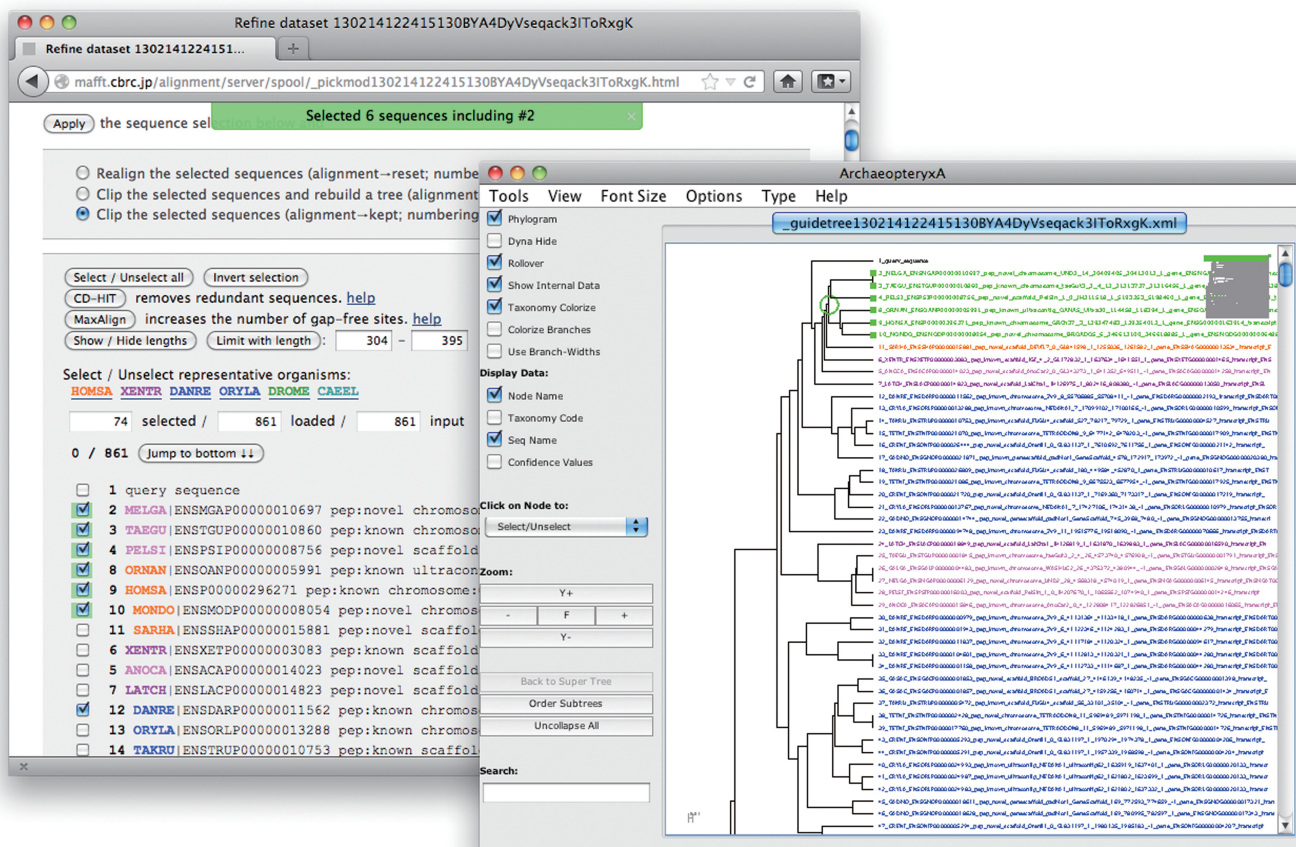
Wingless (Wg) is the invertebrate orthologue of vertebrate Wnt1 genes (32,33). Using the amino acid sequence of the *Drosophila melanogaster* Wg gene (AAF52501 in NCBI Protein), a search was performed at aLeaves with the default parameter settings in all the databases covering invertebrates (Database 8–13), as well as Database 1 and 5 covering vertebrates as key phylogenetic landmarks. The search resulted in a data set with 1000 sequences, which included not only vertebrate Wnt1 and invertebrate Wingless but also phylogenetically distant sequences, such as Wnt4 (<http://aleaves.cdb.riken.jp/aleaves/sample/wg/>). This data set was then subjected to clustering with 6-mer and the ‘UPGMA’ method, which allowed to visualize a preliminary tree in <15s with only a few clicks. Many identical Wg sequences are redundantly deposited mainly in GenBank by individual researchers because of long-standing interest in this gene. In this particular situation, the application of CD-HIT effectively deleted more than half of the initial sequences. Further refinement of the data set was demonstrated by identifying a subtree including all arthropod Wg genes. Because some sequences with truncated N- or C-ends limited the number of gap-free sites in the alignment, they were excluded from the data set. Finally, 36 arthropod sequences, including chelicerates (spiders and tick) as outgroup, were retained in the data set harbouring 276 gap-free sites ([http://mafft.cbrc.jp/alignment/server/spool/aleaves\\_example\\_wingless.html](http://mafft.cbrc.jp/alignment/server/spool/aleaves_example_wingless.html)). The resultant tree therein did not indicate any gene

**Table 1.** Sources of genome-wide protein data sets available at aLeaves but not available at Ensemble-based or NCBI-based sites

Species name	English common name	Phylum	Database ID at aLeaves <sup>a</sup>	URL
<i>Callorhynchus milii</i>	Elephant shark (or ghost shark)	Chordata	6	<a href="http://people.inf.ethz.ch/cdessimo/Cmilii/">http://people.inf.ethz.ch/cdessimo/Cmilii/</a>
<i>Oikopleura dioica</i>		Chordata	8	<a href="http://www.genoscope.cns.fr/externe/GenomeBrowser/Oikopleura/">http://www.genoscope.cns.fr/externe/GenomeBrowser/Oikopleura/</a>
<i>Branchiostoma floridae</i>	Amphioxus	Chordata	8	<a href="http://genome.jgi-psf.org/Brafl1/">http://genome.jgi-psf.org/Brafl1/</a>
<i>Meloidogyne incognita</i>	Phytoparasitic root-knot nematode	Nematoda	10	<a href="http://www.inra.fr/meloidogyne_incognita/">http://www.inra.fr/meloidogyne_incognita/</a>
<i>Schistosoma japonicum</i>	Parasitic flatworm	Platyhelminthes	11	<a href="http://lifecenter.sgst.cn/schistosoma/cn/schistosomaCnIndexPage.do">http://lifecenter.sgst.cn/schistosoma/cn/schistosomaCnIndexPage.do</a>
<i>Capitella teleta</i>	Polychaete worm	Annelida	11	<a href="http://genome.jgi-psf.org/Capca1/">http://genome.jgi-psf.org/Capca1/</a>
<i>Helobdella robusta</i>	Leech	Annelida	11	<a href="http://genome.jgi-psf.org/Helro1/">http://genome.jgi-psf.org/Helro1/</a>
<i>Pinctada fucata</i>	Pearl oyster	Mollusca	11	<a href="http://marinegenomics.oist.jp/">http://marinegenomics.oist.jp/</a>
<i>Crassostrea gigas</i>	Pacific oyster	Mollusca	11	<a href="http://gigadb.org/pacific_oyster/">http://gigadb.org/pacific_oyster/</a>
<i>Acropora digitifera</i>	Okinawan staghorn coral	Cnidaria	12	<a href="http://marinegenomics.oist.jp/">http://marinegenomics.oist.jp/</a>

These species are available at aLeaves but not available at ‘Ensembl’, ‘EnsemblGenomes Metazoa’ or NCBI Genome (as of April 8, 2013). The complete list of species available at aLeaves is found at <http://aleaves.cdb.riken.jp/aleaves/species.html>.

<sup>a</sup>The detail of the aLeaves databases is found in Figure 2 (also see <http://aleaves.cdb.riken.jp/aleaves/database.html>).



**Figure 3.** Sequence data set refinement at the MAFFT web server through Archaeopteryx. Shown as inset is a view of the Archaeopteryx applet, in which a single node containing six sequences is selected (highlighted in bright green with parent node marked by a circle). The parental web browser window shows an HTML page with a list of sequences in the present data set, in which the six sequences selected in Archaeopteryx are newly selected with ticks on the left. The colouring of the different sequences indicates their taxonomic categorization (detailed in the 'Species' page of the aLeaves server).

duplication inside the arthropod lineage, and its tree topology was largely consistent with arthropod species phylogeny proposed with larger data sets (34).

### Vertebrate Hox14

The second example is the molecular phylogeny of vertebrate Hox14 genes (35). Hox genes are an intensively studied group of metazoan regulatory genes (36), and above all, Hox14 marks a unique phylogenetic feature that this group of genes has not been identified in the genomes of any traditional laboratory vertebrate species (37,38). As a query, the amino acid sequence of *Neoceratodus forsteri* (Australian lungfish) HoxA14 gene (CBY85303 in NCBI Protein) was used, and the search at the aLeaves server was performed in the all databases (Database 1–13) with the default setting. The resultant data set contained previously documented Hox14 members and other Hox genes as well as non-Hox genes, such as *Cdx* (<http://aleaves.cdb.riken.jp/aleaves/sample/hox14/>). After a clustering of the sequences in the initial data set, a subtree containing the previously documented Hox14 members was identified in the preliminary tree shown in Archaeopteryx. No sequence other

than the previously documented Hox14 members was identified inside this subtree, in the existing protein-coding gene sets available for all the genomes of the species covered by aLeaves except one species, namely coelacanth *Latimeria* ([http://mafft.cbrc.jp/alignment/server/spool/aleaves\\_example\\_hox14.html](http://mafft.cbrc.jp/alignment/server/spool/aleaves_example_hox14.html)). The tree topology and the non-identification of any Hox14 member in the genomes of most vertebrates were consistent with the results of a previous study (37). The unified platform based on genome resources of dozens of species at aLeaves facilitates the identification of homologues of users' particular interest.

### Vertebrate CTCF

CCCTC-binding factor (CTCF) is a Zinc finger containing transcription factor regulating chromatin compartments by marking particular genomic regions as insulators (39). A search at the aLeaves server was performed using the human CTCF amino acid sequence (AAB07788 in NCBI Protein) with the default parameter setting in the Database 1, 4, 5 and 7 covering major vertebrate lineages and Database 8 containing their potential orthologues of invertebrate deuterostomes as outgroup.

The search resulted in 1000 sequences as requested (<http://aleaves.cdb.riken.jp/aleaves/sample/ctcf/>), but the data set contained many sequences that seemed phylogenetically distant from CTCF.

The entire set of the collected sequences was subjected to clustering as in the above examples. To create a data set focused on vertebrate CTCF genes, we first excluded many mammalian Zinc finger protein (ZNF) sequences apparently distantly related from CTCF, by identifying a subtree consisting almost completely exclusively of abundant mammalian sequences including ZNF484 and ZNF180. The identification of the uninteresting subtree, which was recognized as a paraphyletic group with initial rooting, was facilitated by using the ‘Root/Reroot’ function of Archaeopteryx, and this operation deleted >650 sequences out of the initial 1000. To further reduce the number of sequences in the data set, CD-HIT was applied to retain 285 sequences. Clustering was performed again with ‘accurate (global)’ distance calculation and the ‘minimum linkage’ method to more carefully assess the relationships among the sequences in the data set. The resultant tree in Archaeopteryx allowed the identification of the monophyletic group consisting of vertebrate CTCF genes and their close relative, CTCF-like (or BORIS), rooted by their immediate outgroup (invertebrate deuterostomes). The data set of this subtree containing 58 sequences was subjected to MSA, but the resulting MSA did not contain any gap-free sites. As applying MaxAlign turned out to delete the invertebrate outgroup, which is necessary for rooting, we thus decided not to rely on the automated process but to exclude individual sequences in the data set that are reducing gap-free sites with careful inspection. As demonstrated here, depending on the nature of data sets and aims, users can choose more sensible strategies among diverse implemented functions to refine data sets.

For this example, the 30 sequences remaining after the final careful refinement were aligned again, and using the 405 gap-free sites, an NJ tree was inferred ([http://mafft.cbrc.jp/alignment/server/spool/aleaves\\_example\\_ctcf.html](http://mafft.cbrc.jp/alignment/server/spool/aleaves_example_ctcf.html)). The resultant tree topology was largely consistent with those in previous studies (40) and Ensembl Tree view ([http://www.ensembl.org/Homo\\_sapiens/Gene/Compare\\_Tree?g=ENSG00000102974](http://www.ensembl.org/Homo_sapiens/Gene/Compare_Tree?g=ENSG00000102974)), and the data set derived from aLeaves covered some additional species. On demand, some of the sequences can further be deleted, and conversely, additional sequences determined by one’s own can easily be integrated into the MSA for more customized tree based on amino acid sites of the user’s choice. The major difference between these results was the phylogenetic timing of the gene duplication between CTCF and CTCF-like (BORIS). This question can be further scrutinized by downloading the data set file and applying other phylogenetic tree inference methods on it.

## CONCLUSION

The two web servers, aLeaves and MAFFT, provide a unique online platform to explore metazoan gene family trees on demand. It is expected to handle demands on two

extreme ends—automated processing of a large sequence dataset from diverse genomes and highly interactive analysis with manual inspections in molecular phylogenetics—that are impossible to reconcile with other existing tools. The framework of the aLeaves server is planned to expand in the future, which includes addition of more organisms inside as well as outside Metazoa.

## ACKNOWLEDGEMENTS

The authors thank Naoyuki Iwabe, Yuichiro Hara, Nathalie Feiner and Miyuki Noro for testing earlier versions of aLeaves and valuable comments. The authors also thank Hiroshi Suga, Kei-ichi Kuma and Go Sasaki for providing a part of the tree inference programs.

## FUNDING

Center for Developmental Biology (CDB), RIKEN (to S.K.); Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science and Technology, Japan (to K.K.); National Institutes of Health NIMG [R01GM101457 to C.M.Z.]. Funding for open access charge: Center for Developmental Biology, RIKEN, Japan (the first author’s affiliation).

*Conflict of interest statement.* None declared.

## REFERENCES

- Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
- Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
- Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L.P., Denisov, I., Kormes, D., Marcet-Houben, M. and Gabaldon, T. (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.*, **39**, D556–D560.
- Takeda, J., Yamasaki, C., Murakami, K., Nagai, Y., Sera, M., Hara, Y., Obi, N., Habara, T., Gojobori, T. and Imanishi, T. (2013) H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. *Nucleic Acids Res.*, **41**, D915–D919.
- Brinkman, F.S., Wan, I., Hancock, R.E., Rose, A.M. and Jones, S.J. (2001) PhyloBLAST: facilitating phylogenetic analysis of BLAST results. *Bioinformatics*, **17**, 385–387.
- Dereeper, A., Audic, S., Claverie, J.M. and Blanc, G. (2010) BLAST-EXPLORER helps you building datasets for phylogenetic analysis. *BMC Evol. Biol.*, **10**, 8.
- Frickey, T. and Lupas, A.N. (2004) PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res.*, **32**, 5231–5238.
- Juliusdottir, T., Pettersson, F. and Copley, R.R. (2008) POPE—a tool to aid high-throughput phylogenetic analysis. *Bioinformatics*, **24**, 2778–2779.
- Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
- Coordinators, N.R. (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and

- PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Katoh, K. and Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.*, **9**, 286–298.
  13. Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
  14. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
  15. Higgins, D.G. and Sharp, P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.
  16. Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
  17. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
  18. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
  19. Han, M.V. and Zmasek, C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.
  20. Zmasek, C.M. and Eddy, S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
  21. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
  22. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
  23. Gouveia-Oliveira, R., Sackett, P.W. and Pedersen, A.G. (2007) MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics*, **8**, 312.
  24. Kuraku, S. (2010) Palaeophylogenomics of the vertebrate ancestor—impact of hidden paralogy on hagfish and lamprey gene phylogeny. *Integr. Comp. Biol.*, **50**, 124–129.
  25. Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *J. Theor. Biol.*, **8**, 357–366.
  26. Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.
  27. Whelan, S. and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.
  28. Adachi, J. and Hasegawa, M. (1996) *Computer Science Monographs*, Vol. 28. Institute of Statistical Mathematics, Tokyo.
  29. Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.
  30. Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.
  31. Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
  32. Rijsewijk, F., Schuermann, M., Wagenaar, E., Parren, P., Weigel, D. and Nusse, R. (1987) The Drosophila homolog of the mouse mammary oncogene int-1 is identical to the segment polarity gene wingless. *Cell*, **50**, 649–657.
  33. Wainwright, B.J., Scambler, P.J., Stanier, P., Watson, E.K., Bell, G., Wicking, C., Estivill, X., Courtney, M., Boue, A., Pedersen, P.S. *et al.* (1988) Isolation of a human gene with protein sequence similarity to human and murine int-1 and the Drosophila segment polarity mutant wingless. *EMBO J.*, **7**, 1743–1748.
  34. Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H. and Telford, M.J. (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc. Biol. Sci.*, **278**, 298–306.
  35. Powers, T.P. and Amemiya, C.T. (2004) Evidence for a Hox14 paralog group in vertebrates. *Curr. Biol.*, **14**, R183–R184.
  36. Kuraku, S. and Meyer, A. (2009) The evolution and maintenance of Hox gene clusters in vertebrates and the teleost-specific genome duplication. *Int. J. Dev. Biol.*, **53**, 765–773.
  37. Feiner, N., Ericsson, R., Meyer, A. and Kuraku, S. (2011) Revisiting the origin of the vertebrate Hox14 by including its relict sarcopterygian members. *J. Exp. Zool. B Mol. Dev. Evol.*, **316**, 515–525.
  38. Kuraku, S., Takio, Y., Tamura, K., Aono, H., Meyer, A. and Kuratani, S. (2008) Noncanonical role of Hox14 revealed by its expression patterns in lamprey and shark. *Proc. Natl Acad. Sci. USA*, **105**, 6679–6683.
  39. Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Goncalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P. and Odom, D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335–348.
  40. Hore, T.A., Deakin, J.E. and Marshall Graves, J.A. (2008) The evolution of epigenetic regulators CTCF and BORIS/CTCF1 in amniotes. *PLoS Genet.*, **4**, e1000169.