Published online 11 May 2022 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* ***W577–W582***
https://doi.org/10.1093/nar/gkac324

PhyloCloud: an online platform for making sense of phylogenomic data

Ziqi Deng, Jorge Botas, Carlos P. Cantalapiedra, Ana Hernández-Plaza, Jordi Burguet-Castell and Jaime Huerta-Cepas*

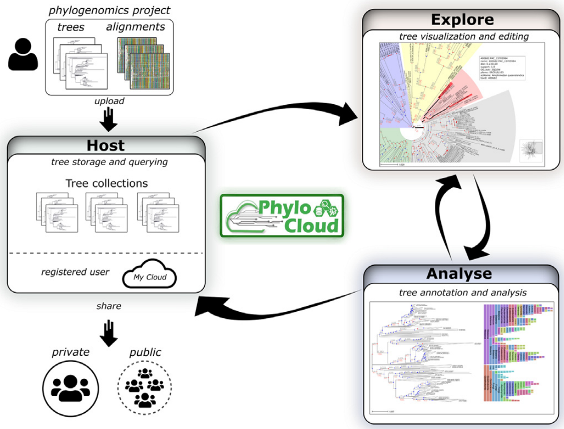
Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) and Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), 28223 Madrid, Spain

Received March 22, 2022; Revised April 18, 2022; Editorial Decision April 19, 2022; Accepted May 03, 2022

ABSTRACT

Phylogenomics data have grown exponentially over the last decades. It is currently common for genome-wide projects to generate hundreds or even thousands of phylogenetic trees and multiple sequence alignments, which may also be very large in size. However, the analysis and interpretation of such data still depends on custom bioinformatic and visualisation workflows that are largely unattainable for non-expert users. Here, we present PhyloCloud, an online platform aimed at hosting, indexing and exploring large phylogenetic tree collections, providing also seamless access to common analyses and operations, such as node annotation, searching, topology editing, automatic tree rooting, orthology detection and more. In addition, PhyloCloud provides quick access to tools that allow users to build their own phylogenies using fast predefined workflows, graphically compare tree topologies, or query taxonomic databases such as NBCI or GTDB. Finally, PhyloCloud offers a novel tree visualisation system based on ETE Toolkit v4.0, which can be used to explore very large trees and enhance them with custom annotations and multiple sequence alignments. The platform allows for sharing tree collections and specific tree views via private links, or make them fully public, serving also as a repository of phylogenomic data. PhyloCloud is available at https://phylocloud.cgmlab.org

GRAPHICAL ABSTRACT



INTRODUCTION

Phylogenomics is a relatively recent field aiming at applying evolutionary analyses at the genomic scale (1). Most common applications include establishing phylogenetic relationships between species based on multiple genes (2), predicting gene function (3), studying the evolutionary history of protein families (4), detecting horizontal gene transfer events (5) and tracing the origin of neo-functionalization events (6). To this end, phylogenomic studies not only rely on the computation of phylogenetic trees, but also on their analysis and interpretation. Extracting meaningful information out of raw phylogenies, usually encoded in Newick or NEXUS formats, involves specialized analyses and operations such as automatic tree rooting, detection of duplication events, calculation of distances between tree nodes or topology comparisons (7). Many of these tasks can only be performed through *ad hoc* bioinformatic solutions, including command line tools (e.g. Newick Utilities (8)) and specialized programming libraries (9–12). In addition, to make sense out of phylogenetic trees, it is usually required that the different tree nodes are annotated with functional, taxonomic or any other type of relevant metadata, which can

[*] To whom correspondence should be addressed. Tel: +34 910679202; Email: j.huerta@csic.es

be used to enrich graphical representations of the results. In fact, graphical exploration of annotated trees is a crucial step in the phylogenetic workflow, often required to identify and interpret evolutionary patterns.

Advanced visualization tools exist on different platforms that assist in the creation of custom layouts and graphical representations. Most notably, iTOL (13) provides numerous graphical options for online interactive visualization of large annotated trees; ETE-Toolkit (12) provides programmatic annotation and visualization of trees using the Python programming language, and ggtree (14) offers custom visualizations for the R programming environment. Additionally, many stand-alone and online tree visualization programs are available that allow for the interactive exploration of trees using different styles and representations (e.g. among the most popular, FigTree, Dendroscope (15), PhyD3 (16), Phylo.io (17)). However, such tools depend on previously annotated trees and tend to require substantial work to adjust custom tree visualization layouts. Most importantly, current phylogenomic studies can produce hundreds of trees along with their corresponding multiple sequence alignments (MSAs), which can also be large, hitting the limits of many of these programs.

Here, we present PhyloCloud (https://phylocloud.cgmlab.org), an online platform aiming at providing an integrative framework for interactive analysis, annotation, visualization and management of phylogenetic trees, regardless of their number or size. PhyloCloud allows users to upload, organise and share both public and private collections of phylogenetic trees, annotate them using predefined methods, and explore them using built-in visualization layouts. Additionally, the platform provides several handy tools for comparative genomics, such as querying information from the NCBI Taxonomy and GTDB databases (18,19), comparing tree topologies in a graphical manner, and quickly building phylogenetic trees using custom data. Notably, PhyloCloud's tree visualization capabilities are built upon the latest developments of the ETE software (12), which enables fast interactive exploration of huge trees, while providing advanced graphical features and adaptive zooming.

RESULTS

Storing, managing and sharing tree collections

PhyloCloud organizes tree uploads by collections, following a similar approach as the iTOL online viewer (13). Although single trees can also be submitted, the tree upload form allows users to submit up to 1000 trees in Newick or Extended Newick format in batch mode, together with their MSAs in FASTA format. Once loaded, all trees are assigned to a new or existing collection, and their content (i.e. node names) is automatically indexed to enable quick text-based searches. Moreover, trees under a given collection can be easily browsed, expanded and rearranged between collections. By default, users can create collections anonymously, and these can be shared with colleagues through a private link. Optionally, users can register into PhyloCloud for better organizing their collections and fine controlling the sharing options. There is currently no limit on the number of trees that can be assigned to a collection, enabling PhyloCloud as a custom repository of data analogous to project-oriented databases (20,21) As an example, PhyloCloud currently features several public tree collections covering up to 112,563 gene trees from recently published phylogenomic studies (22,23)

Tree annotation and analysis

Various tree annotation and analysis options are provided in PhyloCloud. First, any species or gene tree with NCBI or GTDB v202 species identifiers in their leaf names can be automatically annotated with taxonomic information. This includes not only the assignment of scientific names and full lineage paths to leaf nodes but also the annotation of internal branches with the last common ancestor lineage of their descendants. Such annotations are automatically recognized and can be displayed in the tree explorer panel as vertical color bands at the right side of the tree image, thus guiding the exploration of large trees (Figure 1A). Moreover, based on the species content of each internal node, PhyloCloud can identify duplication and speciation events in gene trees using the species overlap detection algorithm (24) (Figure 1, blue and red nodes), inferring also pairwise orthology relationships between tip nodes. These options are enabled solely by extracting the species identifiers associated with each leaf in the tree, which can be parsed automatically from their names, or manually adjusted using the graphical node editing options.

Link to multiple sequence alignments

One of PhyloCloud's primary features is the availability of linking MSAs to their corresponding trees. By doing so, the interactive MSA panel can be activated alongside with the phylogenetic tree visualization (Figure 1B). By doing so, the dynamic exploration of tree topologies is synchronized with a graphical and also dynamic representation of the sequences associated to each leaf node. When tree nodes are collapsed, one or more representative sequences from their descendants are displayed, enabling smooth interactive visualisation of large datasets.

Tree visualisation, editing and searching

PhyloCloud uses the new visualization framework implemented in ETE 4.0, which allows for the interactive exploration of huge phylogenies based on a context-based adaptive zooming strategy (Figure 2). Moreover, the interactive tree explorer allows users to perform various editing options on specific nodes or the tree as a whole, including topology modifications such as automatic re-rooting, pruning, ladderizing branches, resolving polytomies or converting tree topology into an ultrametric (Figure 2B, C and D). These changes can be either discarded after testing or saved permanently into the database. In addition, specific subtrees can be extracted from particular clades and exported separately in Newick format, including node annotations. Notably, the tree explorer allows for quick searches within large tree topologies, where each search can be associated w a different label and color (Figure 2, red branches in panel A).

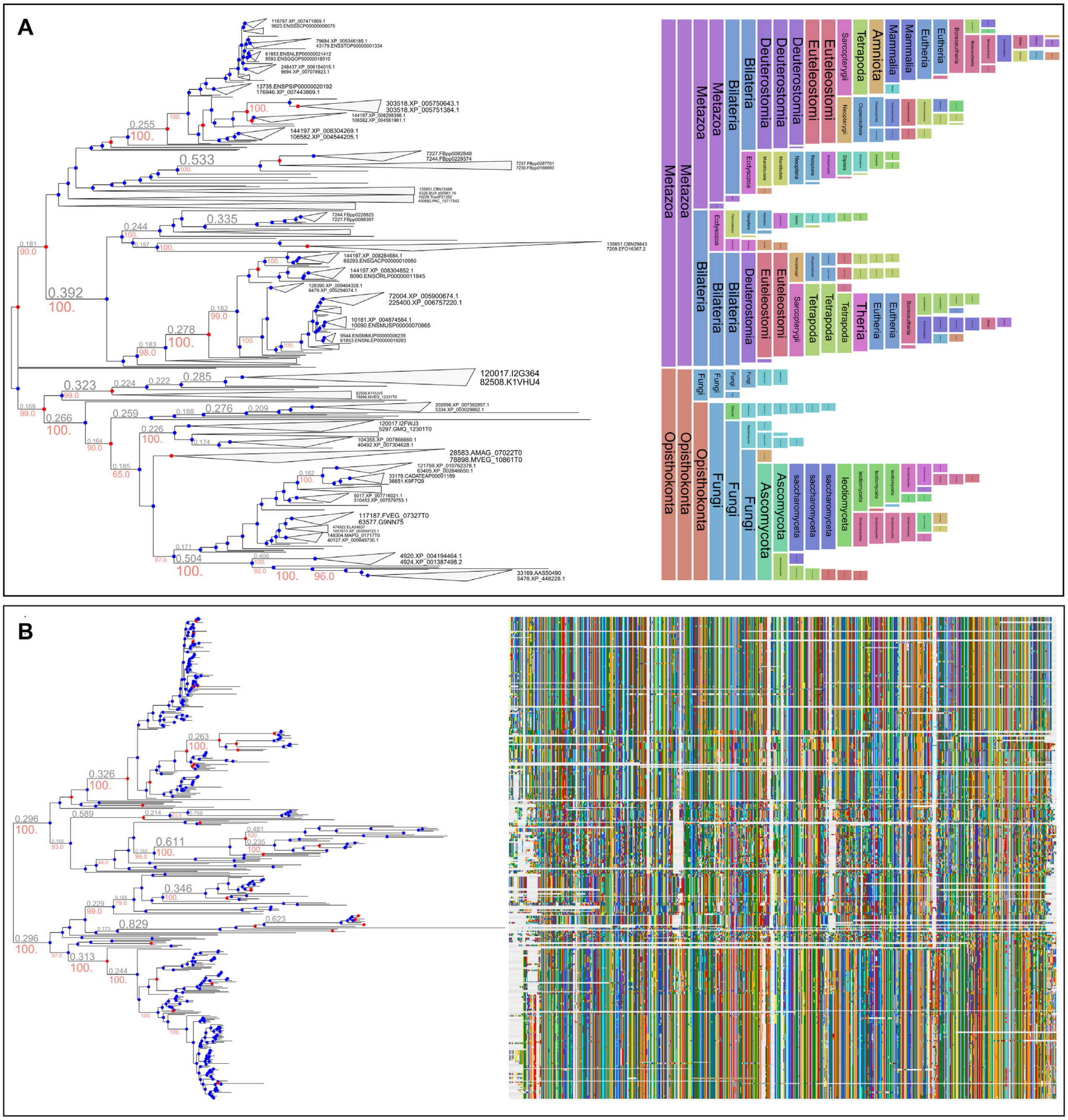
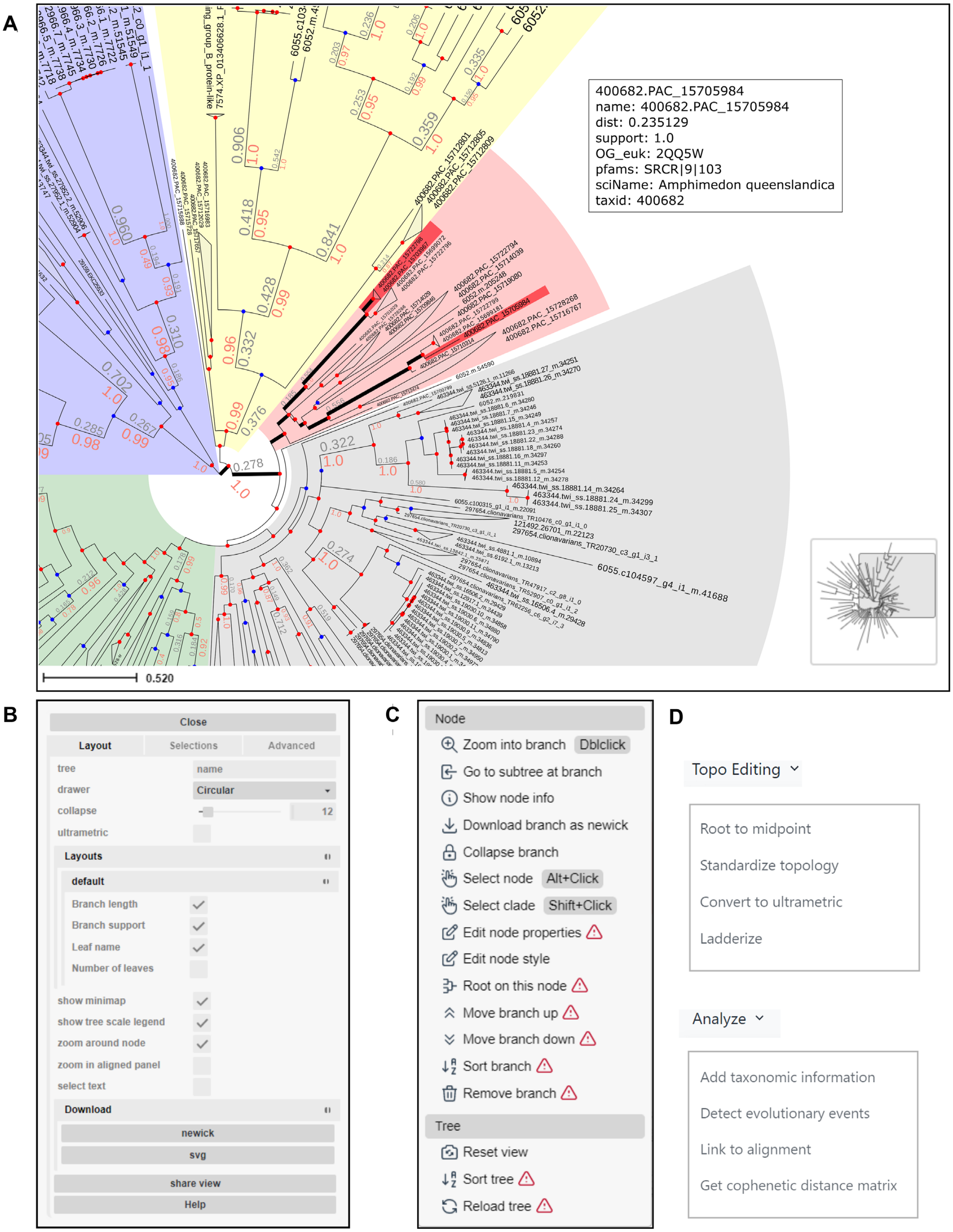


Figure 1. Alignment and taxonomic annotation panels. (**A**) A phylogenetic tree shown along with NCBI taxonomic labels inferred for each internal branch, corresponding to the last common ancestor rank and lineage of its descendants. Speciation (dots in blue) and duplication (dots in red) events were automatically detected and annotated on internal nodes. (**B**) A phylogenetic tree is visualized along with the condensed representation of the MSA used to reconstruct the phylogeny. All three panels (tree, alignments and taxonomic annotations) can be shown together and are synchronized and dynamically adjusted to the zoom level and browsing region of the tree image.

This provides a quick view on the distribution of specific features.

Quick phylogenetic reconstruction

Although phylogenetic reconstruction is not the main focus of PhyloCloud, the platform provides access to a quick click-and-go phylogenetic reconstruction workflow based on those available from the command line tools included in ETE v3.2.1. To ensure relatively fast results, the available workflows offer two aligners (MAFFT (25) and Clustal Ω (26)), several MSA trimming options (trimAl 2.0 (27)), and FastTree for approximate likelihood tree inference (28). Input data, which can be either nucleotide or amino acids sequences, are expected in FASTA format. Results include MSA files (also in FASTA format) and phylogenetic tree



400682.PAC_15705984
name: 400682.PAC_15705984
dist: 0.235129
support: 1.0
OG_euk: 2QQ5W
pfams: SRCR|9|103
sciName: Amphimedon queenslandica
taxid: 400682

tree: name
drawer: Circular
collapse: 12
ultrametric: □
Branch length: ✓
Branch support: ✓
Leaf name: ✓
Number of leaves: □
show minimap: ✓
show tree scale legend: ✓
zoom around node: ✓
zoom in aligned panel: □
select text: □

Figure 2. Overview of the PhyloCloud tree explorer interface. (**A**) A phylogenetic tree from the Spongilla featured collection, visualized in circular layout. Some clades were shaded with different background colors. Support values (red) and branch lengths (grey) are displayed on top of branches and collapsed nodes are shown as a triangle summarizing the length of underlying branches. Hits from two searches are highlighted with thicker black lineage branches and bright red tips. The annotations of one of the hits are shown on the top right corner. The minimap (bottom right) facilitates navigation. (**B**) The control panel allows users to customise visualization layout and features, and to perform text-based searches. (**C**) The node editor panel provides access to node-specific actions, such as creating subtrees, collapsing, pruning, rooting and more. (**D**) Drop-down menus showing the topological and analytic actions which can be performed on the current tree in PhyloCloud.

files (Newick format), which can be quickly inspected in place, or saved into a PhyloCloud collection for further sharing and analysis. Phylogenetic results obtained from more advanced workflows like the ones available in other platforms such as NGPhylogeny (29) can also be loaded into PhyloCloud.

Comparing tree topologies

PhyloCloud provides online access to the tree comparison capabilities implemented in the ETE Toolkit library, enabling the possibility of highlighting the topological differences between two trees of medium size. Besides providing general metrics such Robinson-Foulds distance and the percentage of identical branches found in both trees, PhyloCloud allows users to explore, side-by-side, which branches in one tree are different from the other. To this end, the visualization panel is synchronized between both trees allowing users to quickly identify which is the closest match of a given branch, as well as the euclidean distance to it. The algorithm used for comparing tree topologies is based on minimizing the overall Euclidean distance between all branches in both trees.

Querying taxonomic databases

Querying taxonomic databases is a common task in phylogenetics and other evolutionary analyses. PhyloCloud provides a convenient interface to retrieve full lineage information and subtrees from the NCBI and GTDB taxonomy databases. Thus, users can query with either NCBI or GTDB taxonomic identifiers and obtain a pruned and fully annotated tree including all the descendants of the queried clade.

CONCLUSIONS

PhyloCloud is an online platform that combines analytic tools, utilities and advanced visualization options relevant for a wide range of phylogenetic and phylogenomic studies. PhyloCloud focuses on simplicity and aims at allowing the use of commonly used workflows by non-expert users. The platform is, by design, intended for massive datasets, providing novel solutions for the exploration of large phylogenetic trees and multiple sequence alignments, serving also a global repository for hosting and sharing public phylogenomics datasets. We expect future improvements to provide new analytic workflows, annotation layouts, and sharing capabilities.

FUNDING

[2020-218584], Chan Zuckerberg Initiative DAF an advised fund of Silicon Valley Community Foundation (to J.B., J.BC, and J.HC); [PGC2018-098073-A-I00 MCIU/AEI/FEDER, UE], National Programme for Fostering Excellence in Scientific and Technical Research (to J.HC) [SEV-2016-0672-18-2:PRE2018-084075] FPI-Severo Ochoa predoctoral fellowship (to Z.D); [SEV-2016-0672 (2017–2021)] Severo Ochoa Centres of Excellence Programme from the State Research Agency (AEI) of Spain (to CP.C); [PTA2019-017593-I/AEI/10.13039/501100011033], State Research Agency (AEI) of Spain, Subprograma Personal Técnico de Apoyo (to A.HP)

Conflict of interest statement. None declared.

REFERENCES

1. Eisen,J.A. and Fraser,C.M. (2003) Phylogenomics: intersection of evolution and genomics. *Science*, **300**, 1706–1707.
2. Misof,B., Liu,S., Meusemann,K., Peters,R.S., Donath,A., Mayer,C., Frandsen,P.B., Ware,J., Flouri,T., Beutel,R.G. *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, **346**, 763–767.
3. Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.
4. Liebeskind,B.J., Hillis,D.M. and Zakon,H.H. (2015) Convergence of ion channel genome content in early animal evolution. *Proc. Natl. Acad. Sci. USA*, **112**, E846–E51.
5. Marcet-Houben,M. and Gabaldón,T. (2010) Acquisition of prokaryotic genes by fungal genomes. *Trends Genet.*, **26**, 5–8.
6. Higo,A., Kawashima,T., Borg,M., Zhao,M., López-Vidriero,I., Sakayama,H., Montgomery,S.A., Sekimoto,H., Hackenberg,D., Shimamura,M. *et al.* (2018) Transcription factor DUO1 generated by neo-functionalization is associated with evolution of sperm differentiation in plants. *Nat. Commun.*, **9**, 5283.
7. Posada,D. (2016) Phylogenomics for systematic biology. *Syst. Biol.*, **65**, 353–356.
8. Junier,T. and Zdobnov,E.M. (2010) The newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*, **26**, 1669–1670.
9. Talevich,E., Invergo,B.M., Cock,P.J.A. and Chapman,B.A. (2012) Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in biopython. *BMC Bioinf.*, **13**, 209.
10. Sukumaran,J. and Holder,M.T. (2010) DendroPy: a python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.
11. Paradis,E. and Schliep,K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.
12. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.
13. Letunic,I. and Bork,P. (2021) Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.
14. Yu,G., Smith,D.K., Zhu,H., Guan,Y. and Lam,T.T.-Y. (2017) Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*, **8**, 28–36.
15. Huson,D.H. and Scornavacca,C. (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.*, **61**, 1061–1067.
16. Kreft,L., Botzki,A., Coppens,F., Vandepoele,K. and Van Bel,M. (2017) PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*, **33**, 2946–2947.
17. Robinson,O., Dylus,D. and Dessimoz,C. (2016) Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. *Mol. Biol. Evol.*, **33**, 2163–2166.
18. Schoch,C.L., Ciufo,S., Domrachev,M., Hotton,C.L., Kannan,S., Khovanskaya,R., Leipe,D., Mcveigh,R., O'Neill,K., Robbertse,B. *et al.* (2020) NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**, baaa062.
19. Parks,D.H., Chuvochina,M., Rinke,C., Mussig,A.J., Chaumeil,P.-A. and Hugenholtz,P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**, D785–D794.
20. Fuentes,D., Molina,M., Chorostecki,U., Capella-Gutiérrez,S., Marcet-Houben,M. and Gabaldón,T. (2022) PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Res.*, **50**, D1062–D1068.

21. Vos,R.A., Balhoff,J.P., Caravas,J.A., Holder,M.T., Lapp,H., Maddison,W.P., Midford,P.E., Priyam,A., Sukumaran,J., Xia,X. *et al.* (2012) NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Syst. Biol.*, **61**, 675–689.
22. Musser,J.M., Schippers,K.J., Nickel,M., Mizzon,G., Kohn,A.B., Pape,C., Ronchi,P., Papadopoulos,N., Tarashansky,A.J., Hammel,J.U. *et al.* (2021) Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. *Science*, **374**, 717–723.
23. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernández-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
24. Huerta-Cepas,J., Dopazo,H., Dopazo,J. and Gabaldón,T. (2007) The human phylome. *Genome Biol.*, **8**, R109.
25. 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.
26. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, **7**, 539.
27. Capella-Gutiérrez,S., Silla-Martínez,J.M. and Gabaldón,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
28. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
29. Lemoine,F., Correia,D., Lefort,V., Doppelt-Azeroual,O., Mareuil,F., Cohen-Boulakia,S. and Gascuel,O. (2019) NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Res.*, **47**, W260–W265.