Transcriptome Ortholog Alignment Sequence Tools (TOAST) for Phylogenomic Dataset Assembly

Alex Dornburg (alex.dornburg@naturalsciences.org)
NC Museum of Natural Sciences https://orcid.org/0000-0003-0863-2283
Dustin J. Wcisel
North Carolina State University
J. Thomas Howard
North Carolina State University
Jeffrey A. Yoder
North Carolina State University

Software

Keywords: BUSCO ortholog assembly, Cetacean phylogeny, Missing Data Visualization, Transcriptome, Concatenated Alignment

Posted Date: October 21st, 2019

DOI: https://doi.org/10.21203/rs.2.16269/v1

Version of Record: A version of this preprint was published at BMC Evolutionary Biology on March 30th, 2020. See the published version at https://doi.org/10.1186/s12862-020-01603-w.

Abstract

Background Advances in next-generation sequencing technologies have reduced the cost of whole transcriptome analyses, allowing characterization of non-model species at unprecedented levels. The rapid pace of transcriptomic sequencing has driven the public accumulation of a wealth of data for phylogenomic analyses, however lack of tools aimed towards phylogeneticists to efficiently identify orthologous sequences currently hinders effective harnessing of this resource.Results We introduce TOAST, an open source R software package that can utilize the ortholog searches based on the software Benchmarking Universal Single-Copy Orthologs (BUSCO) to assemble multiple sequence alignments of orthologous loci from transcriptomes for any group of organisms. By streamlining search, query, and alignment, TOAST automates the generation of locus and concatenated alignments, and also presents a series of outputs from which users can not only explore missing data patterns across their alignments, but also reassemble alignments based on user-defined acceptable missing data levels for a given research question.Conclusions TOAST provides a comprehensive set of tools for assembly of sequence alignments of orthologs for comparative transcriptomic and phylogenomic studies. This software empowers easy assembly of public and novel sequences for any target database of candidate orthologs, and fills a critically needed niche for tools that enable quantification and testing of the impact of missing data. As open-source software, TOAST is fully customizable for integration into existing or novel custom informatic pipelines for phylogenomic inference.

Background

Advances in sequencing technology have dramatically decreased the cost of transcriptomic sequencing, driving a rapid increase in the representation of non-model organisms in transcriptome public databases. This proliferation of sequence data has already generated tremendous opportunities for studies that span the molecular evolution of gene families [1, 2] to human disease [3, 4]. Nevertheless, publicly available transcriptomic databases remain underutilized in phylogenomic applications. This is unfortunate, as orthologous markers assembled from public transcriptome data have been shown to provide a cost-effective means to resolve some of the most vexing problems in the Tree of Life [5–7]. A major impediment for using public transcriptomic data in phylogenomics has been the lack of ease in implementing bioinformatic tools for ortholog identification. However, software for Benchmarking Universal Single-Copy Orthologs (BUSCO) [8] provides a powerful framework from which to develop much needed tools for aggregating these orthologs for phylogenomic studies.

BUSCO was originally designed to estimate the “completeness” of genome sequences and whole transcriptome datasets by assessing the number of orthologs expected to be present in all species belonging to a selected clade from a list found in the OrthoDB [8]. Performing BUSCO analysis on multiple species results in the annotation of transcripts or genes with universal identifications (IDs) that could be used as a basis for aggregating sequences for later use in phylogenomic analyses. It has been demonstrated that orthologs identified in transcriptomic datasets using BUSCO are sufficient for phylogenetic analyses [9]. However, the absence of easy to use, peer reviewed software tools targeted

towards the phylogenetics community prohibit the widespread adaptation of this approach. Currently, we are only aware of two well-documented pipelines that exist on private websites or github. The first pipeline, QKbusco [10], is a set of python scripts that the user calls sequentially. The second is part of an on-line bioinformatics tutorial that relies heavily on bash loops and user input of command line prompts such as “sed” and “awk” [11]. While both of these pipelines are easy to access, they require a high level of input and computational experience from the user across multiple sequence ‘search & query’ and ‘assembly’ steps. Even for experienced users, such a process is not efficient. Additionally, the resulting sequences still require significant amounts of processing prior to tree inference. Clearly needed is a software package that automates the fetching and alignment of BUSCO-identified orthologs from transcriptomic data in order to empower the community of evolutionary biologists to effectively harness the potential of this growing set of sequence data.

Here we present ‘Transcriptome Ortholog Alignment Sequence Tools (TOAST)’, a versatile and efficient R package for aggregating single-copy orthologs from either public and/or local transcriptomic resources of targeted organisms and aligning those sequences for subsequent phylogenomic studies. For a given clade, nucleotide sequences from the NCBI database [12] or within a directory on the user’s hard drive can be accessed and assigned a universal ortholog ID from OrthoDB using BUSCO [13]. From this annotated data, we can correctly orient (e.g. 5’ to 3’ direction of transcripts) and align sequences to each other using existing alignment methods [14]. Using returned alignments, TOAST facilitates visualization of missing data patterns, and options are available for generating additional alignments based on user specified levels of data matrix completeness, including customizable concatenated datasets with a corresponding partition file that can be fed directly into the phylogenetic software IQtree [15] for partition+model selection and tree inference.

Methods

TOAST was designed to be run locally, e.g., on a laptop or desktop with modest capabilities. Most of TOAST’s functionality (data fetching and alignment) is currently designed for UNIX systems to accommodate the UNIX reliancy of BUSCO and it’s dependencies (Linux/Mac) [16]. As BUSCO utilizes parallel processing, advanced users may speed up the BUSCO step by utilizing a computer cluster to perform this analysis across more cores and then moving the result files to a local machine. TOAST begins by downloading fasta files from species within a specified taxonomy group. These fasta files include the nucleotide and transcriptome shotgun assembly sequence database (TSA) sequences for each species from the National Center for Biotechnology Information (NCBI) database, and are stored in taxonomically informed file name, *genus_species*.fasta, within the designated folder.

Using internal functions, TOAST employs BUSCO v3.0.2 [9, 13], along with HMMER v3.1b2 [17] and NCBI BLAST+ [18] to find orthologs within the selected OrthoDB database [9, 13]. TOAST will input all of the fasta files from the specified directory, run each through BUSCO, and write the results in a new directory. TOAST next parses the information from the full table of ortholog matches within the results folder for each organism. Both complete and fragmented BUSCO IDs are retained. In the case of duplicated results,

the best scoring sequence is reported. In the event duplicated sequences have the same score, the first sequence encountered is reported. Sequences from each organism are binned into fasta files based on BUSCO IDs.

The reported fasta files contain the best BUSCO nucleotide sequence for each organism. However, the direction of the DNA sequence may be reversed for some species. Therefore, TOAST uses MAFFT to both align the sequences, and assure that all sequences are oriented in the same direction. Individual alignment files are written to relaxed phylip format, and are ready for use with phylogenetic software such as IQ-TREE [15]. These alignments can be concatenated into a single alignment using the TOAST function “SuperAlign”. In addition, the location of sequence partitions within this supermatrix can be written using the TOAST function “PartitionTable” in the nexus format read by IQ-TREE [15].

As given loci will vary with regard to their representation across target taxa, TOAST users have a series of options from which to visualize missing data patterns. These include missing data patterns across all loci, as well as missing data patterns for user defined hierarchical levels (i.e., taxonomy). These functions work on any platform and TOAST additionally has the ability to compute missing data across a directory of fasta files for any loci, filling a critical software need for phylogeneticists. Based on the user defined criterion of acceptable levels of missing information, TOAST can omit taxa from the original alignment, realign each locus, and provide a new concatenated matrix and associated partition block file that defines the location of each locus. These files can be directly read into IQ-TREE [15] for inference.

Results And Discussion

We demonstrate the utility of our program by first analyzing the nucleotide and TSA sequences for each species of Cetacea (whales and dolphins; NCBI ID = 9721). Using the laurasiatheria_odb9 dataset (database of single-copy laurasiatherian orthologs that are present in at least 90% of species) from the OrthoDB website [9, 13], we used TOAST to create a set of files for each species that includes the gene IDs that had a match to specific BUSCO IDs from the laurasiatheria_odb9 dataset. The representation of sequences of an ortholog within a species varied from complete coverage to less than 10 loci.

Further exploration of missing data patterns demonstrates higher levels of missing data within toothed whales. Using an arbitrary threshold of including only cetaceans with at least 1000 of the over 6000 possible orthologs revealed that most missing data was localized in the dolphin *Tursiops* (**Figure 1A**), and that most taxa not meeting this threshold contained very few loci (**Figure 1A**). Further visualizations possible with TOAST demonstrate that this threshold of minimally containing 1000 loci would remove the majority of the missing data the concatenated alignment (**Figure 1B**). TOAST also enables users to utilize interactive plots to explore missing data at different hierarchical levels (in this case taxonomy), which would demonstrate the distribution of missing data is most heavy in dolphins (**Supplemental materials**). Using this threshold, we constructed a concatenated alignment of dolphins + two outgroups and conducted a maximum likelihood analyses based on this data + best fit models and partitions in IQ-TREE (**Figure 2**). This tree provides strongly supported topological resolution for the evolutionary

relationships of major delphinoid lineages, supporting previous hypotheses of a sister group relationship between Delphinidae and a clade comprised of Monodontidae + Phocoenidae [19–21] while simultaneously demonstrating the utility of TOAST for generating phylogenetic datasets.

It is important to note that TOAST is not limited to generating alignments of public data. TOAST can also be used to fetch and align orthologs from local transcriptomes or combine local searches with online fetching. While we use public data for illustrative purposes, we envision the primary utility of this option to empower easy integration of public and novel sequence data for any target set of orthologs. Further, the missing data visualization functions are designed to work with any delimited file of data presence/absence such as behavioural, phenotypic, gene expression, etc studies. As missing elements within a matrix are a common feature of large datasets, TOAST provides a useful set of tools for visually scrutinizing data and considering the potential for biases, such as tree terracing in phylogenetic inference[22], that could result as a consequence of over-looking this potential axis of error.

Conclusions

TOAST provides a comprehensive set of tools for assembly of sequence alignments of orthologs from public or local transcriptomic datasets. For comparative genomic and phylogenomic studies, TOAST enables streamlined assembly of sequence datasets for any target database of candidate orthologs. Output from TOAST facilitates visual and quantitative assessment of missing data patterns that can be integrated with existing approaches to quantify matrix decisiveness [22] or phylogenetic information content [23]. Using these interactive TOAST functions, users can determine acceptable thresholds for minimum representation in the sequence data matrix and readily subsample their data along preset criteria. TOAST not only allows for effective capture of public data, but can also be used to integrate novel sequencing with existing (public or private) gene alignments. As such TOAST empowers phylogeneticists to effectively harness the potential of transcriptomic data as well as investigate the impact of missing data patterns on inferences, filling two important niches of high utility for resolution of a genomic Tree of Life.

Availability And Requirements

Project name: TOAST

Project home page: https://github.com/carolinafishes/TOAST

Operating system(s): UNIX (Mac and Linux)

Programming language: R

Other requirements: R 3.6 or higher, Python 3 or higher, BUSCO, HMMer, BLAST, and Mafft installed

License: e.g. GNU GPL 3

Any restrictions to use by non-academics: license needed

Abbreviations

TOAST - Transcriptome Ortholog Alignment Sequence Tools

BUSCO - Benchmarking Universal Single-Copy Orthologs

IDs - Identifications

TSA - Transcriptome shotgun assembly sequence database

NCBI - National Center for Biotechnology Information

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Software with an additional manual and example data used in the text is hosted on github: https://github.com/carolinafishes/TOAST and an online tutorial is hosted here https://carolinafishes.github.io/software/TOAST/

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported, in part, by grants from the National Science Foundation (IOS-1755242 to AD and IOS-1755330 to JAY).

Authors' contributions

DJW and AD designed the software. JTH, DJW and AD wrote the software. JTH, DJW, JAY and AD wrote the manuscript.

Acknowledgements

We thank E. Ferraro, K. Carlson, and E. Parker for comments on earlier versions of this manuscript as well as help with software documentation and tutorials. We additionally would like to thank participants of the Physalia comparative methods workshop in Berlin for beta-testing and providing useful feedback on earlier versions of this software.

References

1. Carmona SJ, Teichmann SA, Ferreira L, Macaulay IC, Stubbington MJT, Cvejic A, et al. Single-cell transcriptome analysis of fish immune cells provides insight into the evolution of vertebrate immune cell types. Genome Res. 2017;27:451–61. doi:10.1101/gr.207704.116.
2. McConnell SC, Hernandez KM, Wcisel DJ, Kettleborough RN, Stemple DL, Yoder JA, et al. Alternative haplotypes of antigen processing genes in zebrafish diverged early in vertebrate evolution. Proc Natl Acad Sci U S A. 2016;113:E5014–23. doi:10.1073/pnas.1607602113.
3. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489:391–9. doi:10.1038/nature11405.
4. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357:eaan2507. doi:10.1126/science.aan2507.
5. Hawkins JA, Kaczmarek ME, Müller MA, Drosten C, Press WH, Sawyer SL. A metaanalysis of bat phylogenetics and positive selection based on genomes and transcriptomes from 18 species. Proc Natl Acad Sci U S A. 2019;116:11351–60. doi:10.1073/pnas.1814995116.
6. Lindgren AR, Anderson FE. Assessing the utility of transcriptome data for inferring phylogenetic relationships among coleoid cephalopods. Mol Phylogenet Evol. 2018;118:330–42. doi:10.1016/j.ympev.2017.10.004.
7. Saunders GW, Jackson C, Salomaki ED. Phylogenetic analyses of transcriptome data resolve familial assignments for genera of the red-algal Acrochaetiales-Palmariales Complex (Nemaliophycidae). Mol Phylogenet Evol. 2018;119:151–9. doi:10.1016/j.ympev.2017.11.002.

8. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 2013;41 Database issue:D358–65. doi:10.1093/nar/gks1116.
9. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017. doi:10.1093/molbev/msx319.
10. Moscou M. QKbusco (initial release). 2018. https://github.com/matthewmoscou/QKbusco.
11. Severin A, Chudalayandi S, Masonbrink RE, Sayadi M, Seetharam AS. Building maximum likelihood phylogenetic tree using BUSCO genes. 2019. https://isugenomics.github.io/bioinformatics-workbook//dataAnalysis/phylogenetics/reconstructing-species-phylogenetic-tree-with-busco-genes-using-maximum-liklihood-method.html.
12. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35 Database issue:D61–5. doi:10.1093/nar/gkl842.
13. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2. doi:10.1093/bioinformatics/btv351.
14. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 2002;30:3059–66. doi:10.1093/nar/gkf436.
15. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74. doi:10.1093/molbev/msu300.
16. Winter DJ. rentrez: An R package for the NCBI eUtils API. 2017. https://peerj.com/preprints/3179.pdf.
17. Eddy SR and the HMMER development team. HMMER, version 3.1b2. 2015. http://hmmer.org/.
18. Camacho C. BLAST+ Release Notes (Mar 12, 2013 release; Apr 1, 2019 update). 2019. https://www.ncbi.nlm.nih.gov/books/NBK131777/.
19. Dornburg A, Brandley MC, McGowen MR, Near TJ. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). Mol Biol Evol. 2012;29:721–36. doi:10.1093/molbev/msr228.
20. McGowen MR. Toward the resolution of an explosive radiation—A multilocus phylogeny of oceanic dolphins (Delphinidae). Molecular Phylogenetics and Evolution. 2011;60:345–57. doi:10.1016/j.ympev.2011.05.003.
21. Gatesy J, Geisler JH, Chang J, Buell C, Berta A, Meredith RW, et al. A phylogenetic blueprint for a modern whale. Mol Phylogenet Evol. 2013;66:479–506. doi:10.1016/j.ympev.2012.10.012.
22. Sanderson MJ, McMahon MM, Steel M. Terraces in phylogenetic tree space. Science. 2011;333:448–50. doi:10.1126/science.1206357.

23. Dornburg A, Fisk JN, Tamagnan J, Townsend JP. PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. BMC Evol Biol. 2016;16:262. doi:10.1186/s12862-016-0837-3.

Figures

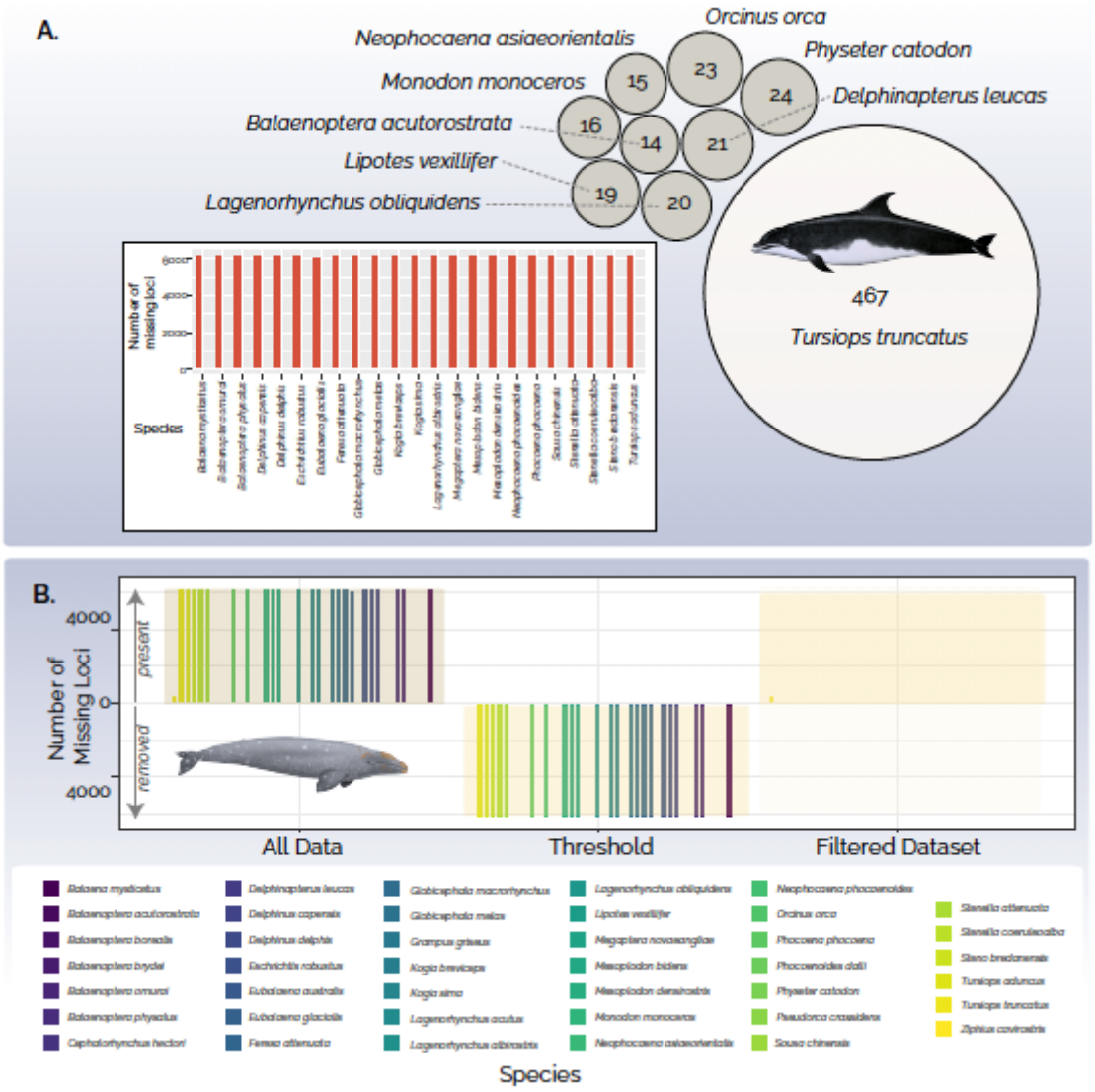

Figure 1

Visualization of missing locus patterns across cetaceans enabled by TOAST. A. Circle pack plot showing missing loci within taxa that contain at least 1000 loci, contrasted with a barplot of missing data levels within taxa that do not contain at least 1000 loci (insert). B. Depiction of missing data by taxon in all available data (left), the amount of missing data lost under a threshold criterion of taxa containing at least 1000 loci (middle), and (right) the number of missing loci within remaining by taxa following filtration by a threshold of acceptable missing data. Taxon colors correspond with taxon color codes.

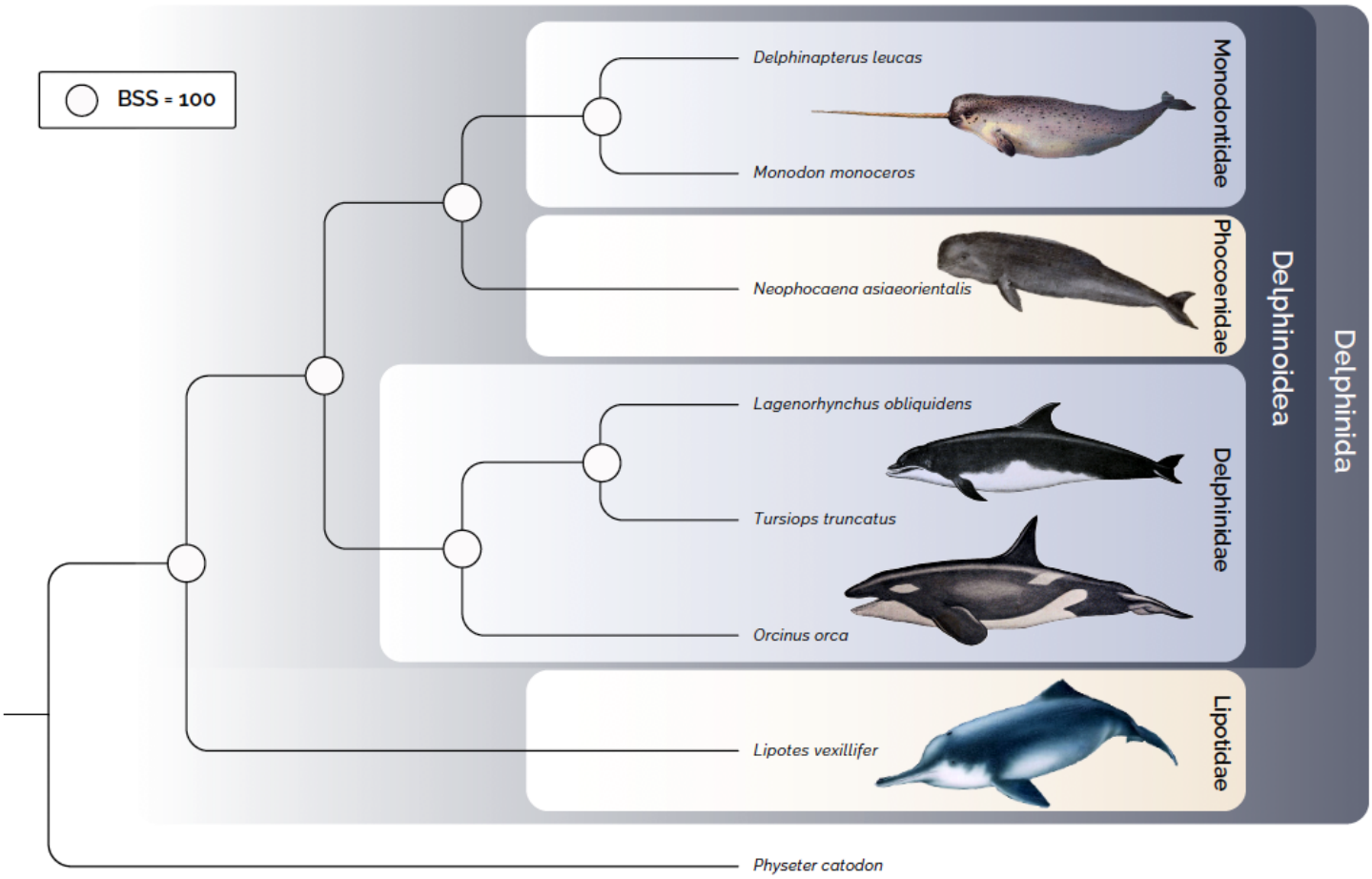

Figure 2

Maximum likelihood phylogeny of delphinid lineages inferred from TOAST harvested BUSCO loci using IQTree. Circles at nodes represent bootstrap support (BSS) values of 100. Delphid images modified from public domain illustrations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TOASTUserManual2019.pdf