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

alex dornburg
  NC Museum of Natural Sciences

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

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.

Software, a detailed manual, and example data files are available through github carolinafishes.github.io

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 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 [9]. However, the absence of easy to use, peer reviewed software tools targeted towards the phylogenetics community prohibit the widespread adaptation of this approach. In addition to the original python scripts used to harvest BUSCO orthologs for phylogenetic studies [9], we are aware of only two other 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 these three 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 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.

Implementation

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 cDNA sequence fasta files from species within a specified taxonomy group. These fasta files include everything in the National Center for Biotechnology Information (NCBI) nucleotide database which includes all transcriptome shotgun assembly sequence database (TSA) sequences for each species. These sequences 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 and users are able to specify an acceptable degree of fragmentation to retain based on the expected length of target loci. 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. It is important to note that TOAST is not limited to only analyzing public data as 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 here, we envision the primary utility of this option will be to empower streamlined integration of public and novel sequence data.

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 [14] 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 as well as the distribution of fragmented loci. 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 both in UNIX environments as well as windows. TOAST additionally has the ability to compute missing data across a directory of fasta files for any set of loci (e.g., anchored hybrid enrichment [19–21], ultraconserved elements [22–24] etc), 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 and also explore potential limitations and pitfalls of BUSCO orthologs for phylogenetics by analyzing three datasets that illustrate different classes of phylogenetic problems. First we test the utility of TOAST for resolving the earliest divergences within delphinoids, a clade that experienced a geologically recent pulse of diversification [25–27]. Second, we test whether strongly supported resolution of the sister lineage to teleost fishes is achievable using BUSCO orthologs. This dataset spans an evolutionary timescale on par with some of the earliest divergences within major groups of living jawed vertebrates. Finally, we assess the effect of tissue specific expression patterns on the representation of BUSCO orthologs between two closely related species of camelids. This third dataset provides expectations of heterogeneity of data coverage and also demonstrates the ability of

TOAST to link recovered loci to the gene ontology (GO) database to assess how the estimated functional roles of loci impact expression patterns.

Resolving the early divergences of delphids

We accessed nucleotide and TSA sequences for each species of Cetacea (whales and dolphins; NCBI ID = 9721) and 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]. Using TOAST, we create a set of files for each species that includes gene IDs that match specific BUSCO IDs from the laurasiatheria_odb9 dataset. The representation of ortholog identification within a species varied from complete coverage (6253 loci) to less than 14 loci. However, 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, Supplemental materials**).

Using this threshold, we constructed a concatenated alignment of 1,473,632 nucleotides for representatives of the major delphinid lineages Monodontidae (*Delphinapterus leucas* and *Monodon monoceros*), Phocoenidae (*Neophocaena asiaeorientalis*), Delphinidae (*Lagenorhynchus obliquidens*, *Tursiops truncatus*, *Orcinus orca*), and Lipotidae (*Lipotes vexillifer*) using *Physeter catodon* as an outgroup. Maximum likelihood based phylogenetic tree searches were conducted using this data in conjunction with best fit nucleotide substitution models and partitions chosen using Bayesian Information Criteria (BIC) in IQ-TREE [15]. Confidence in the topological inference was assessed using 1000 bootstrap replicates (**Figure 1B**). This resulting tree provides strongly supported topological resolution for the evolutionary relationships of major delphinid lineages, supporting previous hypotheses of a sister group relationship between Delphinidae and a clade comprised of Monodontidae + Phocoenidae [28–30] Our analyses also strongly support an early divergence of *Lipotes* prior to the radiation of delphioids (**Figure 1B**). This result is consistent with previous phylogenetic analyses that suggested *Lipotes* to be one of several deeply divergent lineages of river dolphins that invaded freshwater in the Miocene [31–33]. During this period, sea level rise led to the creation of shallow seaways that provided new estuarine habitats for cetaceans on several continents. Subsequent lowering sea levels is thought to have isolated these lineages from their marine relatives, resulting in the independent evolution of modern river dolphins [32]. Our phylogenetic results are consistent with other topological inferences that have been used as a basis to hypothesize that river dolphins such as *Lipotes* are vestiges of previously diverse marine cetacean clades that were replaced by the radiation of delphinoids [32], and simultaneously demonstrate the potential utility of TOAST for generating phylogenetic datasets for recent radiations.

Resolving the sister lineage to teleosts

We used transcriptomes of representative species of ray finned fishes from a recent study by Hughes et. al [34] to resolve the sister lineage to teleost fishes. Species selected span all major clades of non-teleost fishes, including representative bichirs (*Erpetoichthys calabaricus*, *Polypterus bichir*, *Polypterus endlicheri*), paddlefish (*Polyodon spathula*), gar (*Atractosteus spatula*), and bowfin (*Amia calva*). Teleosts were selected to include major divergences in the earliest diverging teleost lineages elopomorpha (*Megalops cyprinoides*, *Gymnothorax reevesii*, *Conger cinereus*) and osteoglossomorpha (*Osteoglossum bicirrhosum*, *Mormyrus tapirus*, *Pantodon buchholzi*, *Papyrocranus afer*), as well as several euteleost species (*Engraulis encrasicolus*, *Lepidogalaxias salamandroides*, *Synodus intermedius*, *Porichthys notatus*). Using the Actinopterygii_odb9 dataset (database of single-copy 4584 orthologs that are present in at least 90% of species) from the OrthoDB website [9, 13], we then used Using TOAST to harvest gene IDs that had a match to specific BUSCO IDs (4529 loci). Quantification of missing data patterns revealed that almost taxa had greater than 70% representation. However, high levels of missing data was localized to three species (*Gymnothorax reevesii*: 81% missing, *Conger cinereus*: 84 % missing, and *Porichthys notatus*: 94% missing **Figure 2A**). We constructed a concatenated alignment of 24,225,167 nucleotides for all taxa excluding these three species with high missing data values. We conducted a maximum likelihood analyses using best fit models and partitions selected with BIC in IQ-TREE with 1000 bootstrap replicates (**Figure 2B**). The resulting tree provides strongly supported topological resolution for every node, supporting previous hypotheses of a sister group relationship between teleosts and Holostei, a clade comprised of gars and bowfin.

For over 150 years, gars and bowfin have been repeatedly hypothesized to form a natural group: Holostei. Originally defined to include a combination of bichirs, gar, and bowfin [35], a revision by Huxley [36] redefined the clade to comprise only bowfin and gar. However, anatomical investigations in the 20th century have repeatedly challenged the validity of Holostei, often proposing instead that bowfin share a closer affinity to teleosts than gar, though recent work by Grande [37] again revived Holostei as sister to teleosts. In sharp contrast to decades of debate among morphologists [38–41], molecular phylogenetic analyses have been nearly unanimous in resolving a strongly supported holostei in studies using mtDNA genomes [42–44], nuclear exons [42, 45–47], as well as transcriptomic and genomic analyses [34, 48]. Such congruence is surprising across years of independent study, and our results add yet another line of evidence strongly supporting the monophyly of Holostei (**Figure 2**).

The resolution of Holostei as sister to teleosts has important implications for biomedical research using teleost model organisms. Linking genomic work in models such as zebrafish, sticklebacks, or medaka back to humans is often challenged by factors that include the loss of ohnologs following the early rounds of vertebrate genome duplication [49, 50], duplicates of loci found in humans as a consequence of the teleost genome duplication [51], and generally rapid rates of molecular evolution in target loci of many teleosts [52, 53]. However, sequencing of the genome of spotted gar (*Lepisosteus oculatus*) has revealed surprising similarity between this holostean genome and genomic features of both teleosts and tetrapods [48]. For example, analysis of the major histocompatibility (MHC) genes in gar revealed synteny to human loci as well as identification of other major groups of immune receptors thought to be teleost specific [48]. Similarly, the gar genome contains sequence information that can link hidden orthology of

teleost and human microRNAs as well as conserved non-coding elements (CNEs) that appear to be highly informative for understanding the "fin to limb" transition [48]. Further investigation of gar CNEs also linked numerous human disease-associated haplotypes to locations within the zebrafish genome, providing new opportunities for functional experiments [48]. The biomedical importance of the genome of a single gar in combination with the recognition of holostean monophyly raises the question of what discoveries await discovery in not only the genomes of more gar, but also the bowfin genome. Given that both the fossil record and molecular clock studies point to an ancient divergence of gar and bowfin several hundred million years ago [45, 46], investigation of the bowfin genome is sure to illuminate not only biomedically relevant research, but also fundamental aspects of vertebrate genome biology.

Assessing tissue-level expression data in camelids

Transcriptome sequences were downloaded from NCBI's TSA for two camel species (Cambac:*Camelus bactrianus* and Camdro:*Camelus dromedarius*) and eight matching tissues (brain: GAES01|GADT01, heart: GAET01|GADU01, hypothalamus: GAEU01|GADV01, kidney: GAEV01|GADW01, liver: GAEW01|GADX01, lung: GAEX01|GADY01, muscle: GAEY01|GADZ01, skin: GAFA01|GAEA01, and testis: GAEZ01|GAEB01). Utilizing these transcriptome datasets, we demonstrate TOAST's ability to compare patterns data coverage within the same species and between species and/or tissue types and compare these patterns to expectations of predicted function.

TOAST builds upon the R packages GSEABase and GOstats [54] to assign the orthologs within a given BUSCO database into Gene Ontology Terms (GoTerm) categories. For the camelids, we used TOAST to organize the Laurasiatheria_odb9 database [55, 56], and display GoTerm assignation and overlap (GoTerms) using the dependent R package UpSetR [57] (Figure 3A). Our results demonstrate that the BUSCO loci for this subset of mammals are largely involved in developmental processes and signalling responses (**Figure 3A**). We demonstrate different degrees of missing data between species and tissue samples, with generally high numbers of fragmented loci in the testis and and very little coverage of loci within liver samples of both species (**Figure 3B**). Additionally, there appear to be minor differences in the relative percent of GoTerms sampled in brain tissues with the GoTerm "Development" to compared to other tissue types (**Figure 3C**). Depending on user goals, TOAST offers users the ability to remove fragmented loci based on a length threshold. In this case, our results demonstrate the heterogeneity of BUSCO coverage that can occur between the same tissue types of closely related species. Although it is premature to interpret biological significance from missing data patterns of this dataset, our results demonstrate the potential for BUSCO coverage limitations stemming from tissue and lineage specific patterns of expression.

Conclusions

TOAST provides a comprehensive set of tools for assembly of sequence alignments of orthologs from public and/or local transcriptomic datasets, enabling streamlined assembly of sequence datasets for any target database of candidate orthologs. TOAST not only allows for effective capture of public data, but

can also be used to integrate novel sequencing with existing (public or private) transcriptomic data. As BUSCO datasets now capturing lineages that span vertebrates to arthropods, fungi, and bacteria, TOAST offers a versatile framework for incorporating transcriptomic resources and BUSCO orthologs into phylogenetic studies. Currently, TOAST is designed for working with transcriptomic data. However, intrepid users could also gather predicted genes from genomic assemblies for compiling BUSCO orthologs in TOAST. Our results reveal the potential utility of these loci for phylogenetic problems of spanning different temporal scales. However, we also demonstrate the potential for heterogeneous sequence coverage between species and tissue types necessitating assessment of missing data patterns.

Output from TOAST facilitates visual and quantitative assessment of missing data patterns that can be integrated with existing approaches to quantify matrix decisiveness [58] or phylogenetic information content [59]. 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 [58]. Using interactive TOAST functions, users can determine acceptable thresholds for minimum representation in the sequence data matrix and readily subsample their data along preset criteria. 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

Lists the following:

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

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.

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

BIC - Bayesian Information Criterion

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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. Lemmon AR, Emme SA, Lemmon EM. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Systematic Biology. 2012;61:727–44. doi:10.1093/sysbio/sys049.

20. Dornburg A, Townsend JP, Brooks W, Spriggs E, Eytan RI, Moore JA, et al. New insights on the sister lineage of percomorph fishes with an anchored hybrid enrichment dataset. Molecular Phylogenetics and Evolution. 2017;110:27–38. doi:10.1016/j.ympev.2017.02.017.

21. Karin BR, Gamble T, Jackman TR. Optimizing Phylogenomics with Rapidly Evolving Long Exons: Comparison with Anchored Hybrid Enrichment and Ultraconserved Elements. Mol Biol Evol. 2019. doi:10.1093/molbev/msz263.

22. Faircloth BC. Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods in Ecology and Evolution. 2017;8:1103–12. doi:10.1111/2041-210x.12754.

23. McCormack J, Tsai WLE, Faircloth BC. Sequence capture of ultraconserved elements from bird museum specimens. doi:10.1101/020271.

24. Chakrabarty P, Faircloth BC, Alda F, Ludt WB, Mcmahan CD, Near TJ, et al. Phylogenomic Systematics of Ostariophysan Fishes: Ultraconserved Elements Support the Surprising Non-Monophyly of Characiformes. Syst Biol. 2017;66:881–95.

25. Boessenecker RW, Fraser D, Churchill M, Geisler JH. A toothless dwarf dolphin (Odontoceti: Xenorophidae) points to explosive feeding diversification of modern whales (Neoceti). Proc Biol Sci. 2017;284. doi:10.1098/rspb.2017.0531.

26. Moura AE, Nielsen SCA, Vilstrup JT, Moreno-Mayar JV, Gilbert MTP, Gray HWI, et al. Recent diversification of a marine genus (Tursiops spp.) tracks habitat preference and environmental change. Syst Biol. 2013;62:865–77.

27. Steeman ME, Hebsgaard MB, Fordyce RE, Ho SYW, Rabosky DL, Nielsen R, et al. Radiation of extant cetaceans driven by restructuring of the oceans. Syst Biol. 2009;58:573–85.

28. 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.

29. 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.

30. 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.

31. Yan J, Zhou K, Yang G. Molecular phylogenetics of "river dolphins" and the baiji mitochondrial genome. Molecular Phylogenetics and Evolution. 2005;37:743–50. doi:10.1016/j.ympev.2005.06.009.

32. Hamilton H, Caballero S, Collins AG, Brownell RL. Evolution of river dolphins. Proceedings of the Royal Society of London. Series B: Biological Sciences. 2001;268:549–56. doi:10.1098/rspb.2000.1385.

33. Muizon C de, de Muizon C, Lambert O, Bianucci G. River Dolphins, Evolution. Encyclopedia of Marine Mammals. 2018;:829–35. doi:10.1016/b978-0-12-804327-1.00221-1.

34. Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc Natl Acad Sci U S A. 2018;115:6249–54.

35. Müller J. Über den Bau und die Grenzen der Ganoiden und über das natürliche System der Fische. 1846.

36. Huxley TH. Preliminary Essay Upon the Systematic Arrangement of the Fishes of the Devonian Epoch. 1861.

37. Grande L. An Empirical Synthetic Pattern Study of Gars (Lepisosteiformes) and Closely Related Species, Based Mostly on Skeletal Anatomy: The Resurrection of Holostei. 2010.

38. Arratia G. The sister-group of Teleostei: consensus and disagreements. Journal of Vertebrate Paleontology. 2001;21:767–73. doi:10.1671/0272-4634(2001)021[0767:tsgotc]2.0.co;2.

39. Gosline WA. Fishes Interrelationships of Fishes P. H. Greenwood R. S. Miles Colin Patterson. BioScience. 1974;24:464–464. doi:10.2307/1296857.

40. Olsen PE. The skull and pectoral girdle of the parasemionotid fishWatsonulus eugnathoidesfrom the Early Triassic Sakamena Group of Madagascar, with comments on the relationships of the holostean fishes. Journal of Vertebrate Paleontology. 1984;4:481–99. doi:10.1080/02724634.1984.10012024.

41. Gardiner BG, Grande L, Bemis W. A Comprehensive Phylogenetic Study of Amiid Fishes (Amiidae) Based on Comparative Skeletal Anatomy. An Empirical Search for Interconnected Patterns of Natural History. Copeia. 1999;1999:240. doi:10.2307/1447417.

42. Dornburg A, Townsend JP, Friedman M, Near TJ. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. BMC Evol Biol. 2014;14:169.

43. Inoue JG, Miya M, Tsukamoto K, Nishida M. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the "ancient fish." Molecular Phylogenetics and Evolution. 2003;26:110–20. doi:10.1016/s1055-7903(02)00331-7.

44. Miya M, Nishida M. The mitogenomic contributions to molecular phylogenetics and evolution of fishes: a 15-year retrospect. Ichthyological Research. 2015;62:29–71. doi:10.1007/s10228-014-0440-9.

45. Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, et al. Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci U S A. 2012;109:13698–703.

46. Near TJ, Dornburg A, Tokita M, Suzuki D, Brandley MC, Friedman M. Boom and bust: ancient and recent diversification in bichirs (Polypteridae: Actinopterygii), a relictual lineage of ray-finned fishes. Evolution. 2014;68:1014–26.

47. Betancur-R R, Wiley EO, Arratia G, Acero A, Bailly N, Miya M, et al. Phylogenetic classification of bony fishes. BMC Evol Biol. 2017;17:162.

48. Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat Genet. 2016;48:427–37.

49. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 2005;3:e314.

50. Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H. Evidence of en bloc duplication in vertebrate genomes. Nat Genet. 2002;31:100–5.

51. Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res. 2003;13:382–90.

52. Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, et al. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. BMC Genomics. 2016;17:368.

53. Pasquier J, Braasch I, Batzel P, Cabau C, Montfort J, Nguyen T, et al. Evolution of gene expression after whole-genome duplication: New insights from the spotted gar genome. J Exp Zool B Mol Dev Evol. 2017;328:709–21.

54. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23:257–8.

55. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.

56. The Gene Ontology Consortium, The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research. 2019;47:D330–8. doi:10.1093/nar/gky1055.

57. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33:2938–40. doi:10.1093/bioinformatics/btx364.

58. Sanderson MJ, McMahon MM, Steel M. Terraces in phylogenetic tree space. Science. 2011;333:448–50.

59. Dornburg A, Fisk JN, Tamagnan J, Townsend JP. PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. BMC Evol Biol. 2016;16:262.

Figures
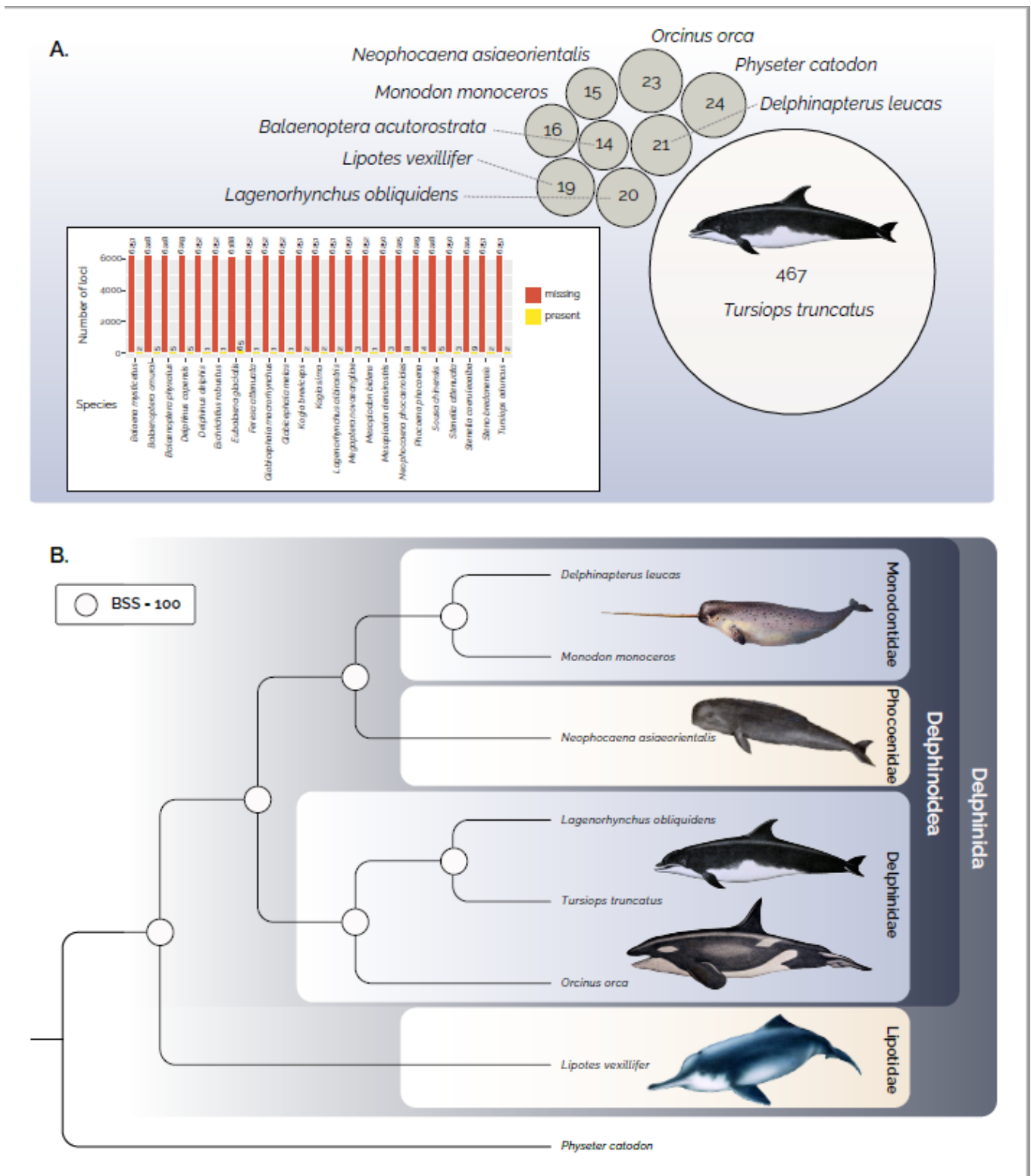
Figure 1

Visualization of missing data patterns and estimation of delphinid phylogeny enabled by TOAST. A. Circle pack plot showing the number of missing loci within taxa that contain at least 1000 loci, contrasted with a barplot of missing (red) versus present (yellow) data patterns within taxa that do not contain at least 1000 loci (insert). B. 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.
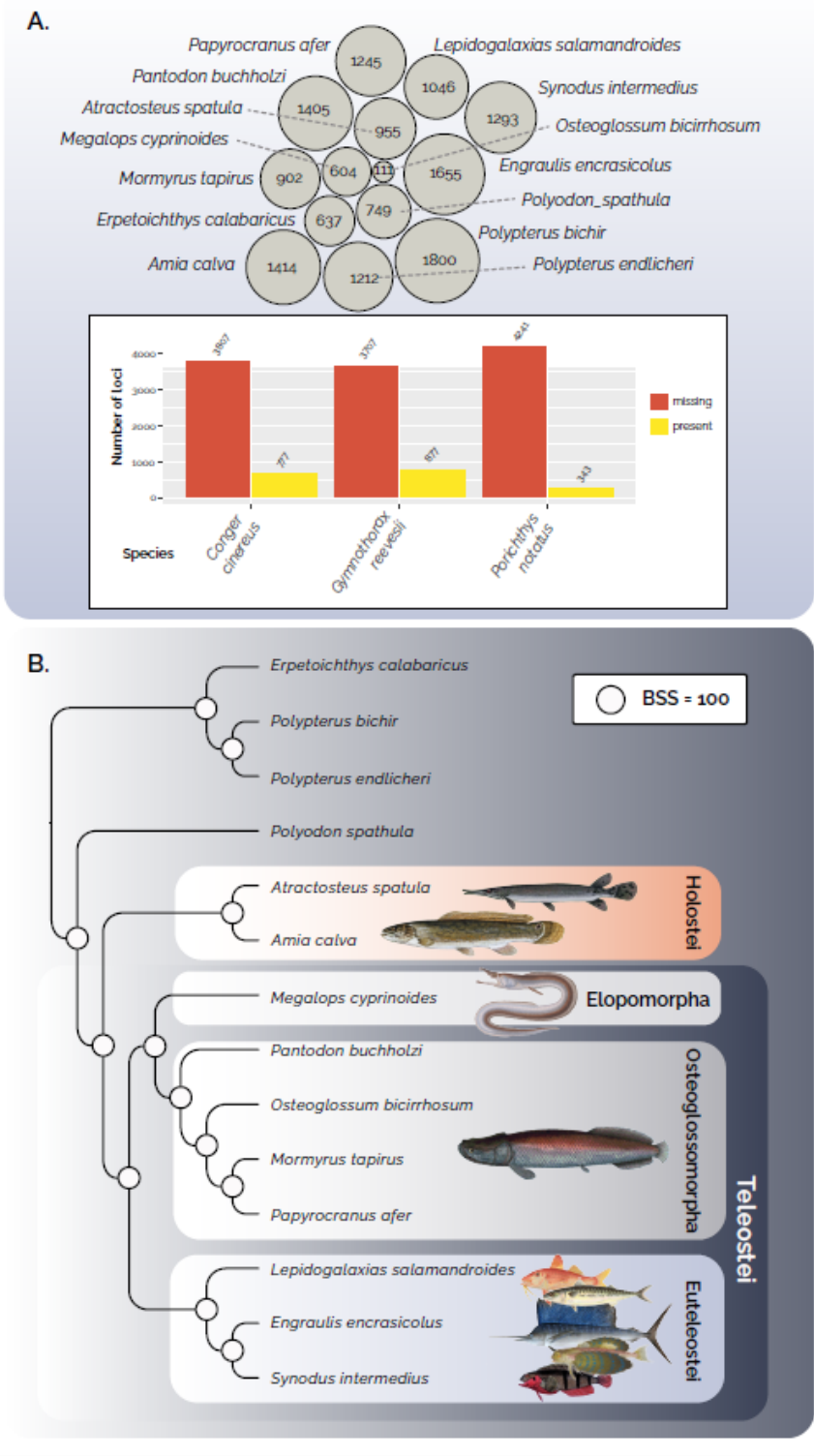
Figure 2

Visualization of missing data patterns and estimation of early ray-finned fish phylogeny enabled by TOAST. A. Circle pack plot showing the number of missing loci within taxa that contain at least 1000 loci, contrasted with a barplot of missing (red) versus present (yellow) data patterns within taxa that do not

contain at least 1000 loci (insert). B. Maximum likelihood phylogeny of ray-finned fish lineages inferred from TOAST harvested BUSCO loci using IQTree. Circles at nodes represent bootstrap support (BSS) values of 100. Fish images modified from public domain illustrations.
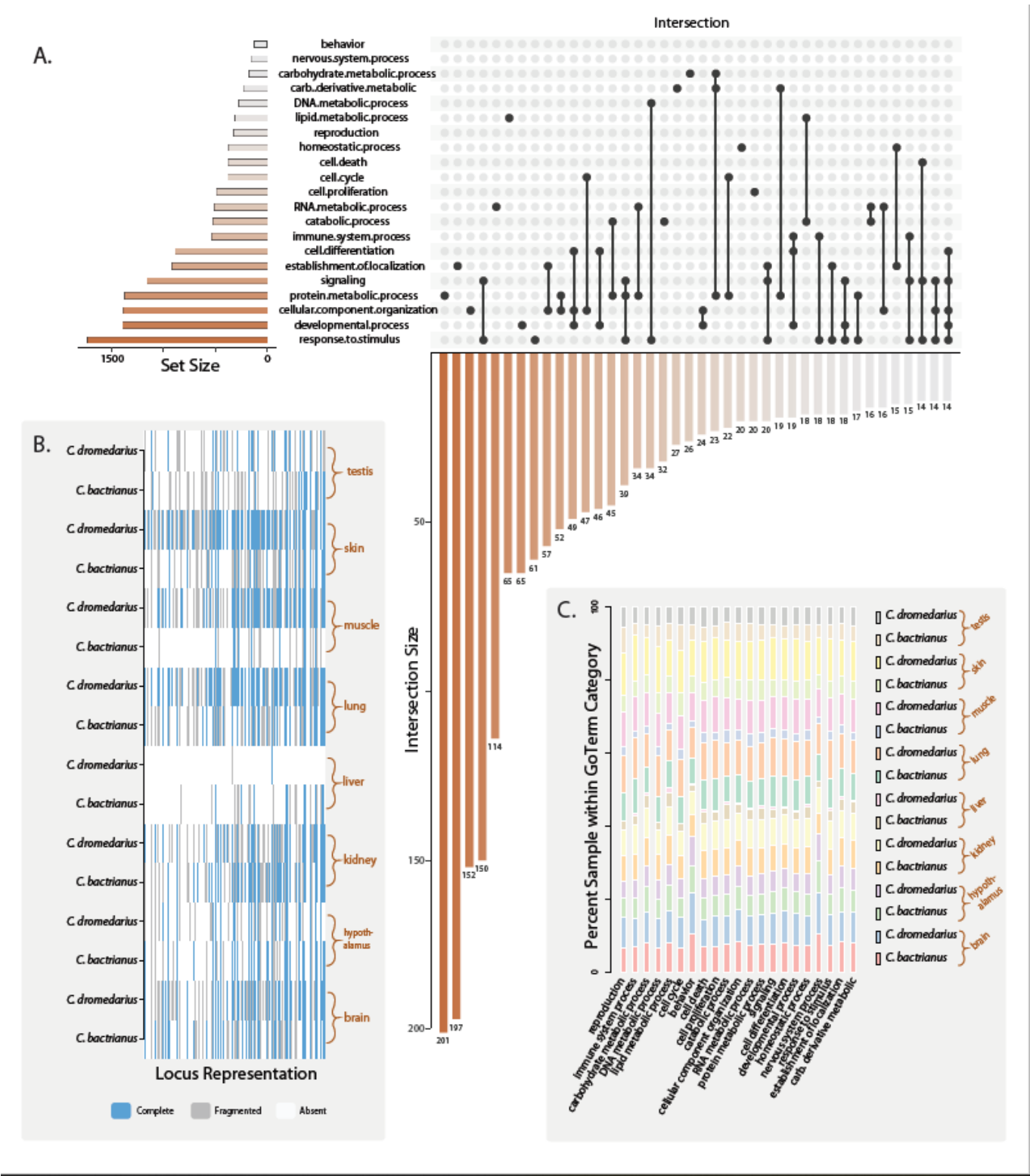
Figure 3

Visualizationing the representation and intersection of GoTerms between tissues of Camelus dromedarius and C. bactianus using TOAST. A. Upset plot showing the intersection of GoTerms across all

loci. Set size represents the number of loci found within each specific category. Intersection size (lower histogram) represents the number of loci within each intersection column. B. Occupancy matrix of complete (blue), fragmented (gray), and missing (white) loci between species and specific tissues. C. Relative percent of GoTerms sampled between tissues and species.