orthoSNAP: a tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees

Jacob L. Steenwyk[1,*], Dayna C. Goltz[2], Thomas J. Buida III[3], Yuanning Li[1], Xing-Xing Shen[4], & Antonis Rokas[1,*]

[1] Vanderbilt University, Department of Biological Sciences, VU Station B #35-1634, Nashville, TN 37235, United States of America

[2] 2312 Elliston Place #510, Nashville, TN 37203, United States of America

[3] 9 City Place #312, Nashville, TN 37209, United States of America

[4] Ministry of Agriculture Key Lab of Molecular Biology of Crop Pathogens and Insects, Institute of Insect Sciences, Zhejiang University, Hangzhou, China

*Correspondence should be addressed to: jacob.steenwyk@vanderbilt.edu or antonis.rokas@vanderbilt.edu

ORCiDs

J.L. Steenwyk: 0000-0002-8436-595X

T.J. Buida III: 0000-0001-9367-6189

Y. Li: 0000-0002-2206-5804

X. Shen: 0000-0001-5765-1419

A. Rokas: 0000-0002-7248-6551

Running title: Retrieving orthologs from gene family data

Keywords: phylogenetics, phylogenomics, orthology, paralogy, evolutionary genomics, evolutionary biology, molecular evolution, gene family, orthologous group of genes

Abstract

Molecular evolution studies, such as phylogenomic studies and genome-wide surveys of positive selection, often rely on gene families of single-copy orthologs (SC-OGs). In contrast, large gene families with multiple homologs in one or more species—a phenomenon observed among several important families of genes such as transporters and transcription factors—are often ignored because identifying and retrieving SC-OGs nested within them is challenging. To address this issue and increase the number of markers used in molecular evolution studies, we developed orthoSNAP, a software that uses a phylogenetic framework to simultaneously split gene families into SC-OGs and prune species-specific inparalogs. We term SC-OGs identified by orthoSNAP as SNAP-OGs because they are identified using a splitting and pruning procedure. From 46,645 orthologous groups of genes inferred using graph-based clustering of sequence similarity scores across four separate eukaryotic datasets, we identified 6,634 SC-OGs; using orthoSNAP on the remaining 40,011 orthologous groups of genes, we identified an additional 6,630 SNAP-OGs. Comparison of SNAP-OGs and SC-OGs revealed that their phylogenetic information content was similar. orthoSNAP is useful for increasing the number of markers used in molecular evolution data matrices, a critical step for robustly inferring and exploring the tree of life.

Introduction

Molecular evolution studies, such as species tree inference, genome-wide surveys of positive selection, evolutionary rate estimation, measures of gene-gene coevolution, and others typically rely on single-copy orthologs (SC-OGs), a group of homologous genes that originated via speciation and are present in single-copy among species of interest (Rokas et al., 2003; Jeffares et al., 2015; Li et al., 2017; Wu et al., 2017; Dong et al., 2019; Steenwyk et al., 2021c). In contrast, paralogs, homologous genes that originated via duplication and are often members of large gene families, are typically absent from these studies (Fig 1). Gene families of orthologs and paralogs often encode functionally significant proteins such as transcription factors, transporters, and olfactory receptors (Ozcan and Johnston, 1999; Malnic et al., 2004; Wingender et al., 2013; Niimura et al., 2014). The exclusion of SC-OGs from gene families has not only hindered our understanding of their evolution and phylogenetic informativeness but is also artificially reducing the number of gene markers available for molecular evolution studies. Furthermore, as the number of species and / or their evolutionary divergence increases in a data set, the number of SC-OGs decreases (Emms and Kelly, 2018; Thomas et al., 2020); case in point, no SC-OGs were identified in a dataset of 42 plants (Emms and Kelly, 2018). As the number of available genomes across the tree of life continues to increase, our ability to identify SC-OGs present in many taxa will become more challenging.

In light of these issues, several methods have been developed to account for paralogs in specific types of molecular evolution studies—for example, in species tree reconstruction (Smith and Hahn, 2021). Methods such as SpeciesRax, STAG, ASTRAL-PRO, and DISCO can be used to infer a species tree from a set of SC-OGs and gene families composed of orthologs and paralogs (Emms and Kelly, 2018; Zhang et al., 2020; Morel et al., 2021; Willson et al., 2021). Other methods such as PHYLDOG (Boussau et al., 2013) and guenomu (de Oliveira Martins and Posada, 2017) jointly infer the species and gene trees, but require abundant computational resources, which has hindered their use for large datasets. Other software, such as PhyloTreePruner (Kocot et al., 2013), can conduct species-specific inparalog trimming, whereas Agalma (Dunn et al., 2013), as part of a larger automated phylogenomic workflow, can prune gene trees

into maximally inclusive subtrees wherein each taxon is represented by one sequence. Although these methods have expanded the numbers of gene markers used in species tree reconstruction, they were not designed to facilitate the retrieval of as broad a set of SC-OGs as possible for downstream molecular evolution studies such as surveys of positive selection. Furthermore, the phylogenetic information content of these gene families remains unknown.

To address this need, we developed orthoSNAP, a novel tree traversal algorithm that conducts tree splitting and species-specific inparalog pruning to identify SC-OGs nested within larger gene families. We term SC-OGs identified by orthoSNAP as SNAP-OGs because they were retrieved using a splitting and pruning procedure. orthoSNAP takes as input a gene family phylogeny and associated FASTA file and will output individual FASTA files populated with sequences from SNAP-OGs (Fig 2). During tree traversal, tree uncertainty can be accounted for by orthoSNAP via collapsing poorly supported branches. In a set of four eukaryotic datasets that contained 6,634 SC-OGs, we used orthoSNAP to identify an additional 6,630 SNAP-OGs. Using a combination of multivariate statistics and phylogenetic measures, we demonstrate that SNAP-OGs and SC-OGs have similar phylogenetic information content in all four datasets. We also observed similar patterns of support among SNAP-OGs and SC-OGs in a contentious branch in the tree of life. Taken together, these results suggest that orthoSNAP is helpful for expanding the set of gene markers available for molecular evolutionary studies.

Results

SC-OGs and SNAP-OGs have similar information content

To compare SC-OGs and SNAP-OGs, we first independently inferred orthologous groups of genes in three eukaryotic datasets of 24 budding yeasts, 36 filamentous fungi, and 26 mammals (S1 Fig; Table S1). There was variation in the number of SC-OGs and SNAP-OGs in each lineage (S2 Fig; Table S2). Interestingly, the ratio of SNAP-OGs : SC-OGs among budding yeasts, filamentous fungi, and mammals was 0.46, 0.83, and 5.53, respectively, indicating SNAP-OGs can substantially increase the number of

gene markers in certain lineages. The number of SNAP-OGs identified in a gene family with multiple homologs in one or more species also varied (S3 Fig).

Similar taxon occupancy and best fitting models of substitutions were observed among SC-OGs and SNAP-OGs (S4 Fig; Table S3), raising the question of whether SC-OGs and SNAP-OGs have similar information content. To answer this, we calculated nine properties of phylogenetic information content from multiple sequence alignments and phylogenetic trees from SC-OGs and SNAP-OGs (S5 Fig; Table S4) and compared them using multivariate analysis and statistics as well as information theory-based phylogenetic measures. Principal component analysis enabled qualitative comparisons between SC-OGs and SNAP-OGs in reduced dimensional space and revealed a striking similarity (Fig 3, S6 Fig). Multivariate statistics, namely multi-factor analysis of variance, facilitated a quantitative comparison of SC-OGs and SNAP-OGs and revealed no difference between SC-OGs and SNAP-OGs ($p = 0.63$, $F = 0.23$, $df = 1$; Table S5) and no interaction between the nine properties and SC-OGs and SNAP-OGs ($p = 0.16$, $F = 1.46$, $df = 8$). Similarly, multi-factor analysis of variance using an additive model, which assumes each factor is independent and there are no interactions, also revealed no differences between SC-OGs and SNAP-OGs ($p = 0.65$, $F = 0.21$, $df = 1$). Next, we calculated tree certainty, an information theory-based measure of tree congruence from a set of gene trees, and found similar levels of congruence among phylogenetic trees inferred from SC-OGs and SNAP-OGs (Table S6). Taken together, these analyses demonstrate that SC-OGs and SNAP-OGs have similar phylogenetic information content.

SC-OGs and SNAP-OGs have similar patterns of support in a contentious branch in the tree of life

To further compare SC-OGs and SNAP-OGs, we investigated patterns of support in a difficult-to-resolve branch in the tree of life. More specifically, we evaluated the support between two leading hypotheses concerning deep evolutionary relationships among placental mammals: (1) Xenarthra (placental mammals from the Americas) and Afrotheria (placental mammals from Africa) are sister to all other Eutheria (Hallström et al., 2007; Wildman et al., 2007) or (2) Afrotheria are sister to all other Eutheria (Murphy,

2001; Murphy et al., 2001) (Fig 4A). Resolution of this conflict has important implications for understanding the historical biogeography of these organisms. To do so, we first obtained protein-coding gene sequences from six Afrotheria, two Xenarthra, 12 other Eutheria, and eight outgroup taxa from NCBI (S7 Fig; Table S7), which represent all annotated and publicly genome assemblies at the time of this study (Table S8). Using the protein translations of these gene sequences as input to OrthoFinder, we identified 252 SC-OGs shared across taxa; application of orthoSNAP identified an additional 1,428 SNAP-OGs, which represents a greater than five-fold increase in the number of gene markers for this dataset (Table S8). There was variation in the number of SNAP-OGs identified per orthologous group of genes (S8 Fig). The highest number of SNAP-OGs identified in an orthologous group of genes was 10, which was a gene family of olfactory receptors and are known to have expanded in the evolutionary history of placental mammals (Niimura et al., 2014). The best fitting substitution models were similar between SC-OGs and SNAP-OGs (S9 Fig).

Two independent tests examining support between alternative hypotheses of deep evolutionary relationships among placental mammals revealed similar patterns of support between SC-OGs and SNAP-OGs. More specifically, no differences were observed in gene support frequencies—the number of genes that support one of three possible hypotheses at a given branch in a phylogeny—without or with accounting for single-gene tree uncertainty by collapsing branches with low support values ($p = 0.26$ and $p = 0.05$, respectively; Fischer's exact test with Benjamini-Hochberg multi-test correction; Fig 4B; Table S9). We next conducted a second test of single-gene support for hypothesis one or hypothesis two by measuring gene-wise log-likelihood scores (ΔGLS), which is the difference in the log-likelihood score of a single gene when constrained to the topologies of the two hypotheses. In this case, positive ΔGLS are reflective of greater support for hypothesis one and negative ΔGLS are reflective of greater support for hypothesis two. No difference was observed in the distribution of ΔGLS values ($p = 0.39$; Wilcoxon rank sum test). Examination of patterns of support in a contentious branch in the tree of life using two

independent tests revealed SC-OGs and SNAP-OGs are similar and further supports the observation that they contain similar phylogenetic information.

In summary, 46,645 orthologous groups of genes across four datasets contained 6,634 SC-OGs; application of orthoSNAP identified an additional 6,630 SNAP-OGs, doubling the number of gene markers. Comprehensive comparison of the phylogenetic information content among SC-OGs and SNAP-OGs revealed no differences in phylogenetic information content. Strikingly, this observation held true when conducting hypothesis testing in a difficult-to-resolve branch in the tree of life. These findings suggest that SNAP-OGs may be useful for exploring patterns of molecular evolution ranging from genome-wide surveys of positive selection, phylogenomics, gene-gene coevolution analysis, and others.

Discussion

Molecular evolution studies typically rely on SC-OGs. Recently developed methods can integrate gene families of orthologs and paralogs into species tree inference but are not designed to broadly facilitate the retrieval of gene markers for molecular evolution analyses. Furthermore, the phylogenetic information content of gene families of orthologs and paralogs remains unknown. This observation underscores the need for algorithms that can identify SC-OGs nested within larger gene families, which can be in turn be incorporated into diverse molecular evolution analyses, and a comprehensive assessment of their phylogenetic properties.

To address this need, we developed orthoSNAP, a tree splitting and pruning algorithm that identifies SNAP-OGs, which refers to SC-OGs nested within larger gene families wherein species specific inparalogs have also been pruned. Comprehensive examination of the phylogenetic information content of SNAP-OGs and SC-OGs from four empirical datasets of diverse eukaryotic species revealed striking similarities. In certain datasets, SNAP-OGs were five times more prevalent than SC-OGs indicating SNAP-OGs can substantially increase the size of molecular evolution datasets. We note that our results

are qualitatively similar to those reported recently by Smith et al. (Smith et al., 2021), which retrieved SC-OGs nested within larger families from 26 primates and examined their performance in gene tree and species tree inference. Three noteworthy differences are that we also conduct species-specific inparalog trimming, provide a user-friendly command-line software for SNAP-OG identification, and evaluated the phylogenetic information content of SNAP-OGs and SC-OGs across four diverse datasets. We also note that our algorithm can account for diverse types of paralogy—outparalogs, inparalogs, and species-specific inparalogs—whereas other software like PhyloTreePruner, which conducts species-specific inparalog trimming (Kocot et al., 2013), and Agalma, which identifies single-copy outparalogs and inparalogs (Dunn et al., 2013), can account for some, but not all, types of paralogs. Our results, together with other studies, demonstrate the utility of SC-OGs that are nested within larger families (van der Heijden et al., 2007; Dunn et al., 2013; Smith et al., 2021; Willson et al., 2021).

Next, we discuss some practical considerations when using orthoSNAP. In the present study, we inferred orthology information using OrthoFinder (Emms and Kelly, 2019), but several other approaches can be used upstream of orthoSNAP. For example, other graph-based algorithms such as OrthoMCL (Li et al., 2003) or sequence similarity-based algorithms such as orthofisher (Steenwyk and Rokas, 2021), can be used to infer gene families. Similarly, sequence similarity search algorithms like BLAST+ (Camacho et al., 2009), USEARCH (Edgar, 2010), and HMMER (Eddy, 2011), can be used to retrieve homologous sets of sequences that are used as input for orthoSNAP.

We suggest employing “best practices” when inferring groups of putatively orthologous genes, including SNAP-OGs. Specifically, orthology information can be further scrutinized using phylogenetic methods. Orthology inference errors may occur upstream of orthoSNAP; for example, SNAP-OGs may be susceptible to erroneous inference of orthology during upstream clustering of putatively orthologous genes. One method to identify putatively spurious orthology inference is by identifying long terminal branches (Shen et al., 2018). Terminal branches of outlier length can be identified using the

"spurious_sequence" function in PhyKIT (Steenwyk et al., 2021b). Other tools, such as PhyloFisher, UPhO, and other orthology inference pipelines employ similar strategies to refine orthology inference (Yang and Smith, 2014; Ballesteros and Hormiga, 2016; Tice et al., 2021).

Taken together, we suggest that orthoSNAP is useful for retrieving single-copy orthologous groups of genes from gene family data and that the identified SNAP-OGs have similar phylogenetic information content compared to SC-OGs. In combination with other phylogenomic toolkits, orthoSNAP may be helpful for reconstructing the tree of life and expanding our understanding of the tempo and mode of evolution therein.

Methods

orthoSNAP availability and documentation

orthoSNAP is a command-line software written in the Python programming language (https://www.python.org/) and requires Biopython (Cock et al., 2009) and NumPy (Van Der Walt et al., 2011). orthoSNAP is available under the MIT license from GitHub (https://github.com/JLSteenwyk/orthosnap), PyPi (https://pypi.org/project/orthosnap), and the Anaconda cloud (https://anaconda.org/JLSteenwyk/orthosnap). Documentation describes the orthoSNAP algorithm, parameters, and provides user tutorials (https://jlsteenwyk.com/orthosnap).

orthoSNAP algorithm description and usage

We next describe how orthoSNAP identifies SNAP-OGs. orthoSNAP requires two files as input: one is a FASTA file that contains two or more homologous sequences in one or more species and the other the corresponding gene family phylogeny in Newick format. In both the FASTA and Newick file, users must follow a naming scheme—wherein taxon identifiers and gene sequences identifiers are separated by a vertical bar (also known as a pipe character or "|")—which allows orthoSNAP to determine which sequences were encoded in the genome of each taxon. After initiating orthoSNAP, the gene family

phylogeny is first mid-point rooted and then SNAP-OGs are identified using a tree-traversal algorithm. To do so, orthoSNAP will loop through the internal branches in the gene family phylogeny and evaluate the number of distinct taxa identifiers among children terminal branches. If the number of unique taxa identifiers is greater than or equal to the taxon occupancy threshold (default: 50% of total taxa in the inputted phylogeny; users can specify an integer threshold), then all children branches and termini are examined further; otherwise, orthoSNAP will examine the next internal branch. Next, orthoSNAP will collapse branches with low support (default: 80, which is motivated by using ultrafast bootstrap approximations (Hoang et al., 2018) to evaluate bipartition support; users can specify an integer threshold) and conduct species-specific inparalog trimming wherein the longest sequence is maintained, a practice common in transcriptomics. Species-specific inparalogs are defined as sequences from the same taxon that are sister to one another or belong to the same polytomy (Kocot et al., 2013). The resulting set of taxa and sequences are examined to determine if one taxon is represented by one sequence and ensure these sequences have not yet been assigned to a SNAP-OG. If so, they are considered a SNAP-OG; if not, orthoSNAP will examine the next internal branch.

The orthoSNAP algorithm is also described using the following pseudocode:

FOR internal branch in midpoint rooted gene family phylogeny:

> IF taxon occupancy among children termini is greater than or equal to taxon occupancy threshold;

>> Collapse poorly supported bipartitions and trim species-specific inparalogs;

>> IF each taxon among the trimmed set of taxa is represented by only one sequence and no sequences being examined have been assigned to a SNAP-OG yet;

>>> Sequences represent a SNAP-OG and are outputted to a FASTA file

>> ELSE

>>> examine next internal branch

> ELSE

>> examine next internal branch

ENDFOR

To enhance the user experience, arguments or default values are printed to the standard output, a progress bar informs the user of how of the analysis has been completed, and the number of SNAP-OGs identified as well as the names of the outputted FASTA files are printed to the standard output.

Development practices and design principles to ensure long-term software stability

Archival instabilities among software threatens the reproducibility of bioinformatics research (Mangul et al., 2019). To ensure long-term stability of orthoSNAP, we implemented previously established rigorous development practices and design principles (Steenwyk et al., 2020, 2021b, 2021a; Steenwyk and Rokas, 2021). For example, orthoSNAP features a refactored codebase, which facilitates debugging, testing, and future development. We also implemented a continuous integration pipeline to automatically build, package, and install orthoSNAP across Python versions 3.8, 3.8, and 3.9. The continuous integration pipeline also conducts 29 unit and integration tests, which span 95.92% of the codebase and ensure faithful function of orthoSNAP.

Dataset generation

To generate a dataset for identifying SNAP-OGs and comparing them to SC-OGs, we first identified putative groups of orthologous genes across four empirical datasets. To do so, we first downloaded proteomes for each dataset, which were obtained from publicly available repositories on NCBI (S1 and S7 Fig; Table S1 and S7) or figshare (Shen et al., 2018). Each dataset varied in its sampling of sequence diversity and in the evolutionary divergence of the sampled taxa. The dataset of 24 budding yeasts spans approximately 275 million years of evolution (Shen et al., 2018); the dataset of 36 filamentous fungi spans approximately 94 million years of evolution (Steenwyk et al., 2019); the dataset of 26 mammals spans approximately 160 million years of evolution (Tarver et al., 2016); and the dataset of 28 placental mammals—which was used to study the contentious deep evolutionary relationships among placental

mammals—concerns an ancient divergence that occurred approximately 160 million years ago (Luo et al., 2011). Putatively orthologous groups of genes were identified using OrthoFinder, v2.3.8 (Emms and Kelly, 2019), with default parameters, which resulted in 46,645 orthologous groups of genes with at least 50% taxon occupancy (Table S8).

To infer the evolutionary history of each orthologous group of genes, we first individually aligned and trimmed each group of sequences using MAFFT, v7.402 (Katoh and Standley, 2013), with the "auto" parameter and ClipKIT, v1.1.3 (Steenwyk et al., 2020), with the "smart-gap" parameter, respectively. Thereafter, we inferred the best-fitting substitution model using Bayesian information criterion and evolutionary history of each orthologous group of genes using IQ-TREE2, v2.0.6 (Minh et al., 2020). Bipartition support was examined using 1,000 ultrafast bootstrap approximations (Hoang et al., 2018).

To identify SNAP-OGs, the FASTA file and associated phylogenetic tree for each gene family with multiple homologs in one or more species was used as input for orthoSNAP, v0.0.1 (this study). Across 40,011 gene families with multiple homologs in one or more species in all datasets, we identified 6,630 SNAP-OGs with at least 50% taxon occupancy (S2 Fig; Table S8). Unaligned sequences of SNAP-OGs were then individually aligned and trimmed using the same strategy as described above. To determine gene families that were SC-OGs, we identified orthologous groups of genes with at least 50% taxon occupancy and each taxon was represented by only one sequence—6,634 orthologous groups of genes were SC-OGs.

Measuring and comparing information content among SC-OGs and SNAP-OGs

To compare the information content of SC-OGs and SNAP-OGs, we calculated nine properties of multiple sequence alignments and phylogenetic trees associated with robust phylogenetic signal in the budding yeasts, filamentous fungi, and mammalian datasets (Table S4). More specifically, we calculated information content from phylogenetic trees such as measures of tree certainty (average bootstrap

support), accuracy (Robinson-Foulds distance (Robinson and Foulds, 1981)), signal-to-noise ratios (treeness (Phillips and Penny, 2003)), and violation of clock-like evolution (degree of violation of a molecular clock or DVMC (Liu et al., 2017)). Information content was also measured among multiple sequence alignments by examining alignment length and the number of parsimony-informative sites, which are associated with robust and accurate inferences of evolutionary histories (Shen et al., 2016) as well as biases in sequence composition (RCV (Phillips and Penny, 2003)). Lastly, information content was also evaluated using metrics that consider characteristics of phylogenetic trees and multiple sequence alignments such as the degree of saturation, which refers to multiple substitutions in multiple sequence alignments that underestimate the distance between two taxa (Philippe et al., 2011), and treeness / RCV, a measure of signal-to-noise ratios in phylogenetic trees and sequence composition biases (Phillips and Penny, 2003). For tree accuracy, phylogenetic trees were compared to species trees reported in previous studies (Tarver et al., 2016; Shen et al., 2018; Steenwyk et al., 2019). All properties were calculated using functions in PhyKIT, v1.1.2 (Steenwyk et al., 2021b). The function used to calculate each metric and additional information are described in Table S4.

Principal component analysis across the nine properties that summarize phylogenetic information content was used to qualitatively compare SC-OGs and SNAP-OGs in reduced dimensional space. Principal component analysis, visualization, and determination of property contribution to each principal component was conducted using factoextra, v1.0.7 (Kassambara and Mundt, 2017), and FactoMineR, v2.4 (Lê et al., 2008), in the R, v4.0.2 (https://cran.r-project.org/), programming environment. Statistical analysis using a multi-factor ANOVA was used to quantitatively compare SC-OGs and SNAP-OGs using the res.aov() function in R.

Information theory-based approaches were used to evaluate incongruence among SC-OGs and SNAP-OGs phylogenetic trees. More specifically, we calculated tree certainty and tree certainty-all (Salichos and Rokas, 2013; Salichos et al., 2014; Kobert et al., 2016), which are conceptually similar to entropy

values and are derived from examining support among a set of gene trees and the two most supported topologies or all topologies that occur with a frequency of ≥5%, respectively. More simply, tree certainty values range from 0 to 1 in which low values are indicative of low congruence among gene trees and high values are indicative of high congruence among gene trees. Tree certainty and tree certainty-all values were calculated using RAxML, v8.2.10 (Stamatakis, 2014).

To examine patterns of support in a contentious branch concerning deep evolutionary relationships among placental mammals, we calculated gene support frequencies and ΔGLS. Gene support frequencies were calculated using the "polytomy_test" function in PhyKIT, v1.1.2 (Steenwyk et al., 2021b). To account for uncertainty in gene tree topology, we also examined patterns of gene support frequencies after collapsing bipartitions with less than 75 ultrafast bootstrap approximation support using the "collapse" function in PhyKIT. To calculate ΔGLS, partition log-likelihoods were calculated using the "wpl" parameter in IQ-TREE2 (Minh et al., 2020), which required as input a phylogeny in Newick format that represented either hypothesis one or hypothesis two (Fig 4A) and a concatenated alignment of SC-OGs and SNAP-OGs with partition information. Thereafter, gene-wise log-likelihood scores associated with hypothesis two were subtracted from gene-wise log-likelihood scores associated with hypothesis one. The resulting score is ΔGLS wherein values greater than zero support hypothesis one and values less than zero support hypothesis two.

Data Availability

Results and data are available from figshare (doi: 10.6084/m9.figshare.16875904) – the link will become accessible upon publication.

Acknowledgements

We thank the Rokas lab for helpful discussion and feedback. J.L.S. and A.R. were funded by the Howard Hughes Medical Institute through the James H. Gilliam Fellowships for Advanced Study program.

Research in A.R.'s lab is supported by grants from the National Science Foundation (DEB-1442113 and DEB-2110404), the National Institutes of Health/National Institute of Allergy and Infectious Diseases (R56 AI146096), and the Burroughs Wellcome Fund.

Conflict of Interest

A.R. is a scientific consultant for LifeMine Therapeutics, Inc.

References

Ballesteros, J. A., and Hormiga, G. (2016). A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology. *Mol. Biol. Evol.* 33, 2117–2134. doi:10.1093/molbev/msw069.

Boussau, B., Szollosi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Res.* 23, 323–330. doi:10.1101/gr.141978.112.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi:10.1093/bioinformatics/btp163.

de Oliveira Martins, L., and Posada, D. (2017). "Species Tree Estimation from Genome-Wide Data with guenomu," in, 461–478. doi:10.1007/978-1-4939-6622-6_18.

Dong, Y., Chen, S., Cheng, S., Zhou, W., Ma, Q., Chen, Z., et al. (2019). Natural selection and repeated patterns of molecular evolution following allopatric divergence. *Elife* 8. doi:10.7554/eLife.45199.

Doyle, V. P., Young, R. E., Naylor, G. J. P., and Brown, J. M. (2015). Can We Identify Genes with Increased Phylogenetic Reliability? *Syst. Biol.* 64, 824–837. doi:10.1093/sysbio/syv041.

Dunn, C. W., Howison, M., and Zapata, F. (2013). Agalma: an automated phylogenomics workflow. *BMC Bioinformatics* 14, 330. doi:10.1186/1471-2105-14-330.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.* 7, e1002195. doi:10.1371/journal.pcbi.1002195.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi:10.1093/bioinformatics/btq461.

Emms, D. M., and Kelly, S. (2018). STAG: Species Tree Inference from All Genes. *bioRxiv*, 267914. doi:10.1101/267914.

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative

genomics. *Genome Biol.* 20, 238. doi:10.1186/s13059-019-1832-y.

Hallström, B. M., Kullberg, M., Nilsson, M. A., and Janke, A. (2007). Phylogenomic Data Analyses Provide Evidence that Xenarthra and Afrotheria Are Sister Groups. *Mol. Biol. Evol.* 24, 2059–2068. doi:10.1093/molbev/msm136.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* 35, 518–522. doi:10.1093/molbev/msx281.

Jeffares, D. C., Tomiczek, B., Sojo, V., and dos Reis, M. (2015). "A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome," in, 65–90. doi:10.1007/978-1-4939-1438-8_4.

Kassambara, A., and Mundt, F. (2017). factoextra. *R Packag. v. 1.0.5*.

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. doi:10.1093/molbev/mst010.

Kobert, K., Salichos, L., Rokas, A., and Stamatakis, A. (2016). Computing the Internode Certainty and Related Measures from Partial Gene Trees. *Mol. Biol. Evol.* 33, 1606–1617. doi:10.1093/molbev/msw040.

Kocot, K. M., Citarella, M. R., Moroz, L. L., and Halanych, K. M. (2013). PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol. Bioinforma.* 2013, 429–435. doi:10.4137/EBO.S12813.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR□: An R Package for Multivariate Analysis. *J. Stat. Softw.* 25, 1–18. doi:10.18637/jss.v025.i01.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503.

Li, Z., De La Torre, A. R., Sterck, L., Cánovas, F. M., Avila, C., Merino, I., et al. (2017). Single-Copy Genes as Molecular Markers for Phylogenomic Studies in Seed Plants. *Genome Biol. Evol.* 9, 1130–

1147. doi:10.1093/gbe/evx070.

Luo, Z.-X., Yuan, C.-X., Meng, Q.-J., and Ji, Q. (2011). A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature* 476, 442–445. doi:10.1038/nature10291.

Malnic, B., Godfrey, P. A., and Buck, L. B. (2004). The human olfactory receptor gene family. *Proc. Natl. Acad. Sci.* 101, 2584–2589. doi:10.1073/pnas.0307882100.

Mangul, S., Martin, L. S., Eskin, E., and Blekhman, R. (2019). Improving the usability and archival stability of bioinformatics software. *Genome Biol.* 20, 47. doi:10.1186/s13059-019-1649-8.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37, 1530–1534. doi:10.1093/molbev/msaa015.

Morel, B., Schade, P., Lutteropp, S., Williams, T. A., Szöllősi, G. J., and Stamatakis, A. (2021). SpeciesRax: A tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *bioRxiv*, 2021.03.29.437460. doi:10.1101/2021.03.29.437460.

Murphy, W. J. (2001). Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics. *Science (80-.).* 294, 2348–2351. doi:10.1126/science.1067179.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature* 409, 614–618. doi:10.1038/35054550.

Niimura, Y., Matsui, A., and Touhara, K. (2014). Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome Res.* 24, 1485–1496. doi:10.1101/gr.169532.113.

Ozcan, S., and Johnston, M. (1999). Function and regulation of yeast hexose transporters. *Microbiol. Mol. Biol. Rev.* 63, 554–69. doi:10.1128/MMBR.63.3.554-569.1999.

Phillips, M. J., and Penny, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* 28, 171–185. doi:10.1016/S1055-7903(03)00057-5.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving

incongruence in molecular phylogenies. *Nature* 425, 798–804. doi:10.1038/nature02053.

Salichos, L., and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 327–331. doi:10.1038/nature12130.

Salichos, L., Stamatakis, A., and Rokas, A. (2014). Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Mol. Biol. Evol.* 31, 1261–1271. doi:10.1093/molbev/msu061.

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., et al. (2018). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* 175, 1533-1545.e20. doi:10.1016/j.cell.2018.10.023.

Smith, M. L., and Hahn, M. W. (2021). New Approaches for Inferring Phylogenies in the Presence of Paralogs. *Trends Genet.* 37, 174–187. doi:10.1016/j.tig.2020.08.012.

Smith, M. L., Vanderpool, D., and Hahn, M. W. (2021). Using all gene families vastly expands data available for phylogenomic inference in primates. *bioRxiv*, 2021.09.22.461252. doi:10.1101/2021.09.22.461252.

Song, S., Liu, L., Edwards, S. V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci.* 109, 14942–14947. doi:10.1073/pnas.1211733109.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Steenwyk, J. L., Buida, T. J., Gonçalves, C., Goltz, D. C., Morales, G., Mead, M. E., et al. (2021a). BioKIT: a versatile toolkit for processing and analyzing diverse types of sequence data. *bioRxiv*, 2021.10.02.462868. doi:10.1101/2021.10.02.462868.

Steenwyk, J. L., Buida, T. J., Labella, A. L., Li, Y., Shen, X.-X., and Rokas, A. (2021b). PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics*. doi:10.1093/bioinformatics/btab096.

Steenwyk, J. L., Buida, T. J., Li, Y., Shen, X.-X., and Rokas, A. (2020). ClipKIT: A multiple sequence

alignment trimming software for accurate phylogenomic inference. *PLOS Biol.* 18, e3001007. doi:10.1371/journal.pbio.3001007.

Steenwyk, J. L., Phillips, M. A., Yang, F., Date, S. S., Graham, T. R., Berman, J., et al. (2021c). A gene coevolution network provides insight into eukaryotic cellular and genomic structure and function. *bioRxiv*, 2021.07.09.451830. doi:10.1101/2021.07.09.451830.

Steenwyk, J. L., and Rokas, A. (2021). orthofisher: a broadly applicable tool for automated gene identification and retrieval. *G3 Genes|Genomes|Genetics* 11. doi:10.1093/g3journal/jkab250.

Steenwyk, J. L., Shen, X.-X., Lind, A. L., Goldman, G. H., and Rokas, A. (2019). A Robust Phylogenomic Time Tree for Biotechnologically and Medically Important Fungi in the Genera Aspergillus and Penicillium. *MBio* 10. doi:10.1128/mBio.00925-19.

Tarver, J. E., dos Reis, M., Mirarab, S., Moran, R. J., Parker, S., O'Reilly, J. E., et al. (2016). The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. *Genome Biol. Evol.* 8, 330–344. doi:10.1093/gbe/evv261.

Thomas, G. W. C., Dohmen, E., Hughes, D. S. T., Murali, S. C., Poelchau, M., Glastad, K., et al. (2020). Gene content evolution in the arthropods. *Genome Biol.* 21, 15. doi:10.1186/s13059-019-1925-7.

Tice, A. K., Žihala, D., Pánek, T., Jones, R. E., Salomaki, E. D., Nenarokov, S., et al. (2021). PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. *PLOS Biol.* 19, e3001365. doi:10.1371/journal.pbio.3001365.

van der Heijden, R. T., Snel, B., van Noort, V., and Huynen, M. A. (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 8, 83. doi:10.1186/1471-2105-8-83.

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* 13, 22–30. doi:10.1109/MCSE.2011.37.

Wildman, D. E., Uddin, M., Opazo, J. C., Liu, G., Lefort, V., Guindon, S., et al. (2007). Genomics, biogeography, and the diversification of placental mammals. *Proc. Natl. Acad. Sci.* 104, 14395–14400. doi:10.1073/pnas.0704342104.

Willson, J., Roddur, M. S., Liu, B., Zaharias, P., and Warnow, T. (2021). DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. *Syst. Biol.* doi:10.1093/sysbio/syab070.

Wingender, E., Schoeps, T., and Dönitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* 41, D165–D170. doi:10.1093/nar/gks1123.

Wu, J., Yonezawa, T., and Kishino, H. (2017). Rates of Molecular Evolution Suggest Natural History of Life History Traits and a Post-K-Pg Nocturnal Bottleneck of Placentals. *Curr. Biol.* 27, 3025-3033.e5. doi:10.1016/j.cub.2017.08.043.

Yang, Y., and Smith, S. A. (2014). Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Mol. Biol. Evol.* 31, 3081–3092. doi:10.1093/molbev/msu245.

Zhang, C., Scornavacca, C., Molloy, E. K., and Mirarab, S. (2020). ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Mol. Biol. Evol.* 37, 3292–3307. doi:10.1093/molbev/msaa139.

Figures and Figure Legends

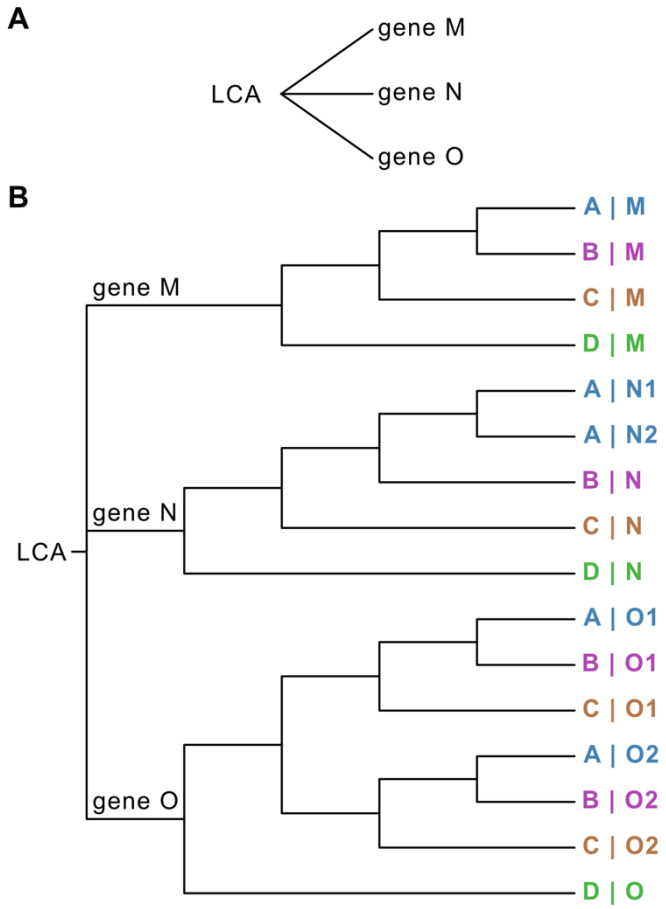

Figure 1. Cartoon depiction of three classes of paralogs: outparalogs, inparalogs, and coorthologs. (A) Paralogs refer to related genes that have originated via gene duplication, such as genes M, N, and O. (B) Outparalogs and inparalogs refer to paralogs that are related to one another via a duplication event that took place prior to or after a speciation event, respectively. With respect to the speciation event that led to the split of taxa A, B, and C from D, genes M, N, and O are outparalogs because they arose prior to the speciation event; genes O1 and O2 in taxa A, B, and C are inparalogs because they arose after the speciation event. Species-specific inparalogs are paralogous genes observed only in one taxon in a dataset, such as gene N1 and N2 in taxon A. Species-specific inparalogs N1 and N2 in taxon A are also coorthologs of gene N in taxa B, C, and D; the same is true for inparalogs O1 and O2 from taxon A, which are coorthologs of gene O from taxon D.

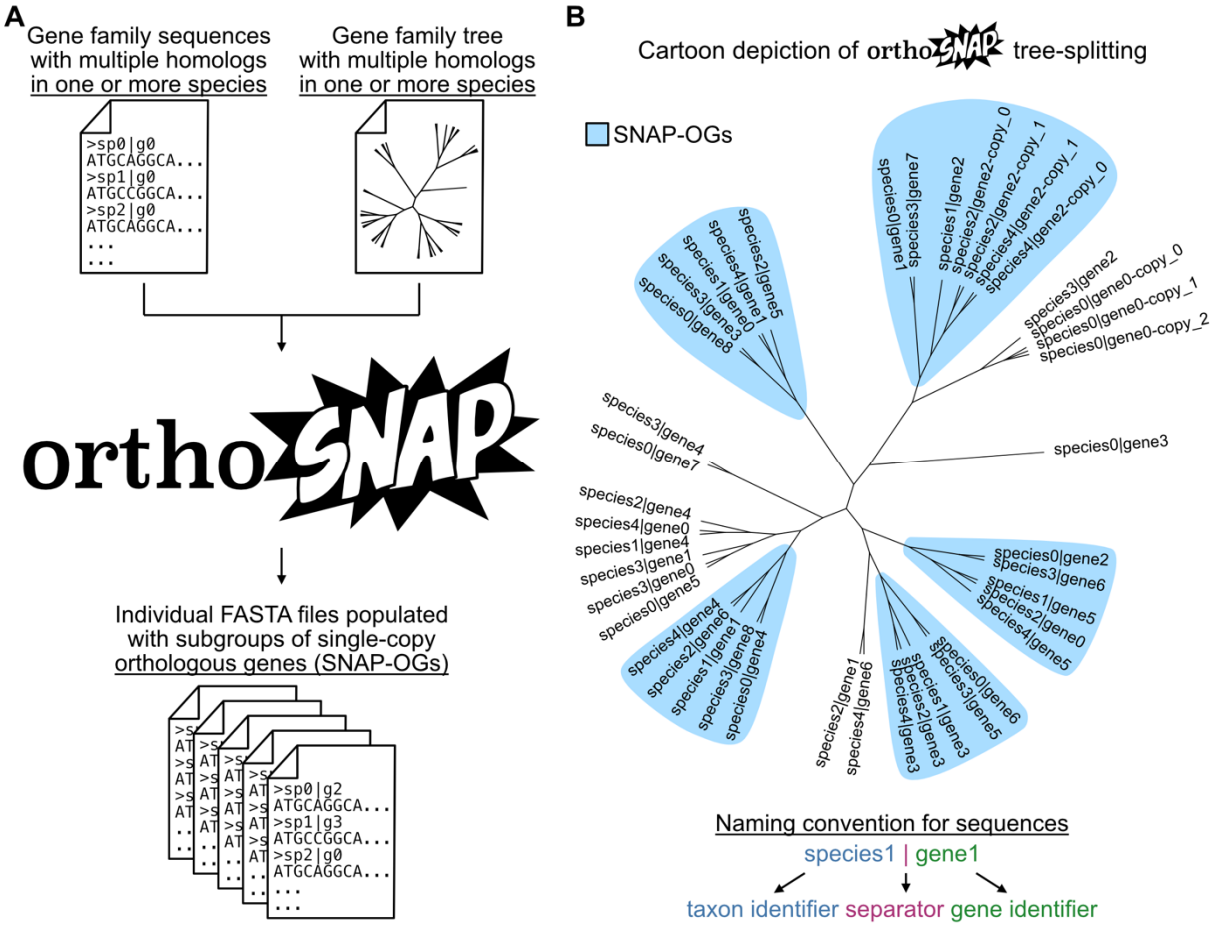

Figure 2. Cartoon depiction of orthoSNAP workflow. (A) orthoSNAP takes as input two files: a FASTA file of a gene family with multiple homologs observed in one or more species and the associated gene family tree. The outputted file(s) will be individual FASTA files of SNAP-OGs. (B) A cartoon phylogenetic tree that depicts the evolutionary history of a gene family and five SNAP-OGs therein. While identifying SNAP-OGs, orthoSNAP also identifies and prunes species-specific inparalogs (e.g., species2|gene2-copy_0 and species2|gene2-copy_1), retaining only the inparalog with the longest sequence, a practice common in transcriptomics. Note, orthoSNAP requires that sequence naming schemes must be the same in both sequences and follow the convention in which a taxon identifier and gene identifier are separated by pipe (or vertical bar; "|") character.

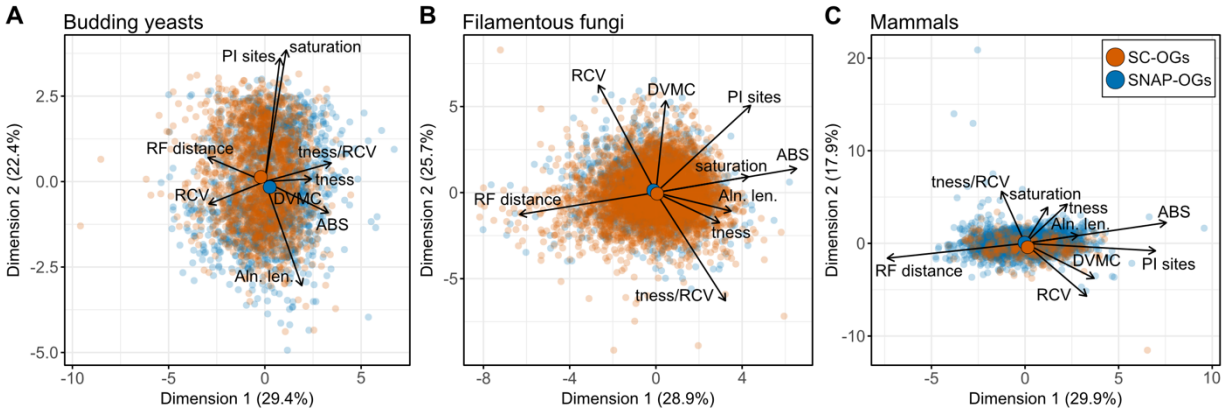

Figure 3. SC-OGs and SNAP-OGs have similar phylogenetic information content. To evaluate similarities and differences between SC-OGs (orange dots) and SNAP-OGs (blue dots), we examined each gene's phylogenetic information content by measuring nine properties of multiple-sequence alignments and phylogenetic trees. We performed these analyses on 12,764 gene families from three datasets—24 budding yeasts (1,668 SC-OGs and 1,392 SNAP-OGs) (A), 36 filamentous fungi (4,393 SC-OGs and 2,035 SNAP-OGs) (B), and 26 mammals (321 SC-OGs and 1,775 SNAP-OGs) (C). Principal component analysis revealed striking similarities between SC-OGs and SNAP-OGs in all three datasets. For example, the centroid (i.e., the mean across all metrics and genes) for SC-OGs and SNAP-OGs, which is depicted as an opaque and larger dot, are very close to one another in reduced dimensional space. Supporting this observation, multi-factor analysis of variance with interaction effects of the 6,630 SNAP-OGs and 6,634 SC-OGs revealed no difference between SC-OGs and SNAP-OGs ($p = 0.63$, $F = 0.23$, $df = 1$) and no interaction between the nine properties and SC-OGs and SNAP-OGs ($p = 0.16$, $F = 1.46$, $df = 8$). Multi-factor analysis of variance using an additive model yielded similar results wherein SC-OGs and SNAP-OGs do not differ ($p = 0.65$, $F = 0.21$, $df = 1$). There are also very few outliers of individual SC-OGs and SNAP-OGs, which are represented as translucent dots, in all three panels. For example, SNAP-OGs outliers at the top of panel C are driven by high treeness/RCV values, which is associated with a high signal-to-noise ratio and/or low composition bias (Phillips and Penny, 2003); SNAP-OG outliers at the right of panel C are driven by high average bootstrap support values, which is associated with greater tree certainty (Salichos and Rokas, 2013); and the single SC-OG outlier observed in the bottom right of

panel C is driven by a SC-OG with a high degree of violation of a molecular clock (Song et al., 2012), which is associated with lower tree certainty (Doyle et al., 2015). Multiple-sequence alignment and phylogenetic tree properties used in principal component analysis and abbreviations thereof are as follows: average bootstrap support (ABS), degree of violation of the molecular clock (DVMC), relative composition variability, Robinson-Foulds distance (RF distance), alignment length (Aln. len.), the number of parsimony informative sites (PI sites), saturation, treeness (tness), and treeness/RCV (tness/RCV).

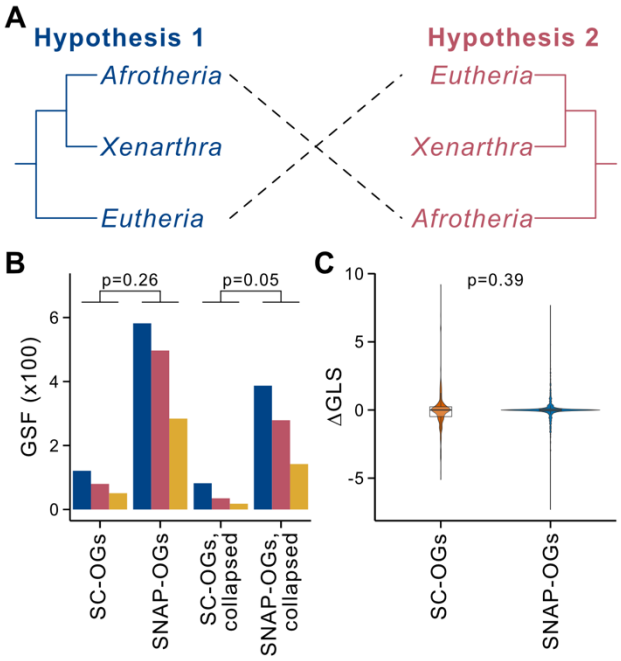

Figure 4. SC-OGs and SNAP-OGs display similar patterns of support in a contentious branch concerning deep evolutionary relationships among placental mammals. (A) Two leading hypotheses for the evolutionary relationships among Eutheria, which have implications for the evolution and biogeography of the clade, are that Afrotheria and Xenarthra are sister to all other Eutheria (hypothesis one; blue) and that Afrotheria are sister to all other Eutheria (hypothesis two; pink). (B) Comparison of gene support frequency (GSF) values for hypotheses one, hypothesis two, as well as a third hypothesis (Xenarthra as sister to all other Eutheria represented in yellow) among 252 SC-OGs and 1,428 SNAP-OGs using an α level of 0.01 revealed no differences in support ($p = 0.26$, Fischer's exact test with Benjamini-Hochberg multi-test correction). Comparison after accounting for gene tree uncertainty by collapsing bipartitions with lower than 75 ultrafast bootstrap approximation support (SC-OGs collapsed vs. SNAP-OGs collapsed) also revealed no differences ($p = 0.05$; Fischer's exact test with Benjamini-Hochberg multi-test correction). (C) Examination of the distribution of gene-wise log-likelihood scores (ΔGLS) revealed no difference between SNAP-OGs and SC-OGs ($p = 0.39$; Wilcoxon rank sum test). ΔGLS values greater than zero are indicative of genes with greater support for hypothesis one; values less than zero are indicative of genes with greater support for hypothesis two.