Published on Sep 28, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-z08x-v798

NovelTree: Highly parallelized phylogenomic inference

We want to find and use evolutionary innovations to solve present-day problems. We developed NovelTree, an efficient phylogenomic workflow that will empower us to decode the evolutionary traces of these innovations across the tree of life.

Contributors (A-Z)

Feridun Mert Celebi, Seemay Chou, Jonathan A. Eisen, Megan L. Hochstrasser, Elizabeth A. McDaniel, Erin McGeever, Austin H. Patton, Taylor Reiter, Dennis A. Sun, Ryan York

Version 2 · Mar 31, 2025

Purpose

One of our central research goals is to identify evolution's greatest innovations. The space we're searching (i.e., the entire tree of life) is vast and remains mostly uncharted. Luckily, given continued expansions in genome sequencing and computing power, we're entering an era in which we can start reading the record of life more comprehensively than ever before. Key to this effort is the application of phylogenetic methods to identify conserved genes, map their evolutionary histories, reconstruct species relationships, and pinpoint when and where evolution has innovated the most.

The methods available to us today are largely optimized for more specific, narrow evolutionary scales of interest. We created the NovelTree workflow to efficiently perform phylogenetic comparisons at any evolutionary scale. Our goal was to generate a single framework that could be applied again and again across the tree of life in a quick and cost-effective manner. Starting with protein sequence data, NovelTree outputs gene families, gene family trees, species trees, and genome-wide evolutionary dynamics. NovelTree does all this in NextFlow, allowing it to run in parallel and with substantial cost savings compared to other methods.

We hope NovelTree will be of use to anybody employing large-scale phylogenetic inference in their research.

- This pub is part of the **platform effort**, "Genetics: Decoding evolutionary drivers across biology." Visit the platform narrative for more background and context.
- You can find the **NovelTree workflow** in this GitHub repository.
- The **code** for the TSAR analyses presented herein is available in this GitHub repository.
  - An **R script** containing functions used to summarize and visualize NovelTree's outputs is available here.
  - An **R markdown script** that walks through how to use these functions and produce visualizations is available here.
  - A **bash script** that downloads all workflow outputs associated with this pub from Zenodo and runs the above R markdown script to produce a user-friendly, interactive HTML is available here.
- The **original protein sequence data** prior to pre-processing, **filtered protein sequences**, and all **outputs** from NovelTree are available here.

The context

Life is a record of outcomes molded by a single process: evolution. The record is vast. At least several million species exist today (potentially trillions, depending on who you

ask **[1]**). Over the past 3.7 billion years, many, many more have evolved, lived for a time, and gone extinct. During this time, evolution has generated an expansive "parts list" in the form of proteins. The design space for these parts is virtually infinite **[2][3]**. They've been combined and recombined, over and over, to form a toolkit of options for addressing life's myriad challenges — a living record of engineering possibilities, as diverse as life itself. Contained within this record are countless novel technologies just waiting to be discovered, many of which we believe may be transformative.

Lucky for us, we can read the record of life. As evolution tinkers, it leaves behind decodable signals in the form of gene sequence changes. These changes may adjust the structure and function of proteins and, in turn, affect biological processes. Over time, useful changes arise, are modified, and subsequent generations inherit them. By comparing these changes across species, we reconstruct their histories and, in doing so, decode the record.

The first step in reading life's record is determining which sets of gene sequences are related: we must find "gene families." Gene families are groups of genes found in multiple species that are derived from a common ancestor gene through duplications, and often have related functions. Once identified, we can compare gene families to query how evolution has employed certain sequences over time. Gene families can vary extensively across species, and the evolutionary histories of individual genes often do not perfectly match that of the species that possess them. Sometimes gene families grow in size when species gain new versions of genes. At other times they may contract, or be lost altogether, such as when species no longer "need" a gene. And in some cases, genes may originate from scratch, generating unforeseen novelty. By accounting for these patterns, we can develop hypotheses about the utility of gene families over evolutionary time and shine a light on the key innovations contained within the vastness of life **[4][5][6]**.

It's an extremely exciting time to be reading the record of life. Genome sequences are constantly expanding in their taxonomic variety and number (e.g., **[7][8]**). So too are computational resources for handling them **[9][10]**. Given these trends, it may soon be possible to conceive of mapping gene family histories at the scale of the tree of life, allowing us to navigate the full universe of natural engineering experiments conducted by evolution.

The problem

Unfortunately, it's not easy to comprehensively identify gene families and to reconstruct their evolutionary histories. There are both technical and computational scaling challenges.

Traditional approaches are constrained to using small sets of the most broadly conserved genes, known as housekeeping genes **[9]**. Newer methods can identify families across entire sets of genomes **[11]**, but often only work if we make unrealistic (and potentially detrimental) assumptions **[12]**. For example, many can only handle single-copy gene families **[13]** despite nearly every gene family being multi-copy in the genome of at least one species **[12][14][15]** and the clear relevance of gene copy number to a number of biological innovations **[4][5]**.

Furthermore, even if accurately identified, tracing the evolutionary history of all gene families across the full tree of life requires a scale of computational analysis that is extremely challenging, if not impossible, with existing tools. A major bottleneck in these sorts of analyses is the extent to which the analysis of individual gene families can be done simultaneously. Application of existing methods to a handful, even several hundreds of gene families is doable — but what if we're interested in studying all ~10,000 gene families in animals **[16]**, for instance? In practice, such analyses are largely unviable, for even modest numbers of species. For instance, just the initial steps of a conservative analysis of ~500 gene families could require hours to days of compute time. This requirement would need to be reduced by at least 10–100× for an average user to gain traction on this problem, let alone analyze 10,000+ gene families.

Current phylogenomic approaches fall short for a variety of reasons. Many are limited in scope to a specific biological or technical problem, making generalization difficult **[17][18]**. Some can be generalized, but are likely too complex or computationally expensive to make broad adoption likely. Furthermore, the methodologies can be highly heterogeneous, leading to vastly different outcomes depending on the approach taken **[19]**.

We sought a Swiss Army knife solution for phylogenomic analyses at a tree-of-life scale. We reasoned that any such approach must meet several requirements. First, it must be able to comprehensively identify gene family histories across any set of species (including multi-copy families). Second, it must be flexible enough to accommodate a broad array of phylogenomic methods — tools that not only identify,

but help to retrace and interrogate the evolutionary history of all gene families. Finally, it must be computationally efficient, integrated, and parallelizable so we can run it again and again.

Our solution, in brief

To conduct phylogenomic studies of gene family evolution at scale, we developed NovelTree. The workflow takes proteomes (i.e. one protein amino acid sequence for each gene in a genome) from diverse organisms and infers orthology, gene family trees, species trees, and gene family evolutionary dynamics (Figure 1). We designed this workflow to let us more easily and efficiently address a broad range of comparative questions, bringing us much closer to our ultimate goal of uncovering evolutionary novelties (Table 1). As with other workflows developed here at Arcadia **[20]**, we chose to implement NovelTree in Nextflow.

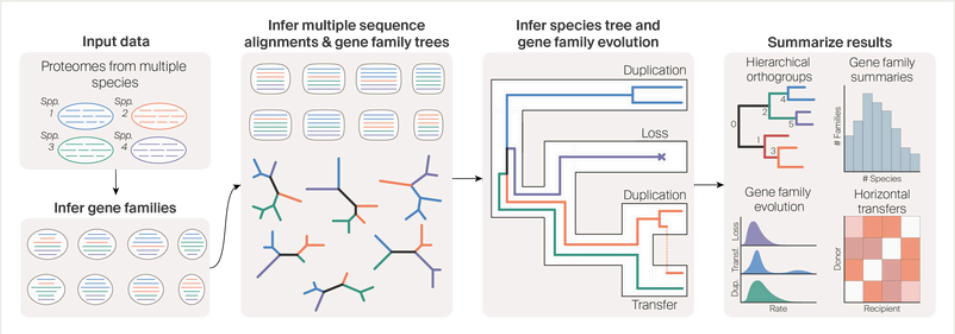

Figure 1

Conceptual outline of NovelTree.

Beginning with a set of species' proteomes, the workflow carries out everything from gene family inference, all the way through phylogenetic inference (i.e., multiple sequence alignment and gene family tree inference) and the inference gene family evolutionary dynamics such as gene duplication, transfer, and loss. We provide resources to facilitate meaningful summarization of all results, and intend to integrate some of these into future versions of the workflow.

Briefly, NovelTree's major contributions are the following:

1. Efficient, parallelized implementation of the core steps in phylogenomic analysis of whole-genome protein sequence data for a set of species using sensible defaults and appropriate methods for general use.
2. Flexibility at nearly all stages of analysis, including:
    - Methodological alternatives at multiple stages of the workflow (e.g., multiple sequence alignment and gene family or species tree inference).
    - Ability to modify behavior of software at all stages of the workflow via user-specification of parameters for software implemented in each module.
3. Improvement to/extension of OrthoFinder's gene family inference procedure to (optionally) assess the performance of different MCL inflation parameters before inferring gene families with the optimal parameter value.
4. Implementation of improved methods for inference of multiple sequence alignments from data sets that are heterogeneous in quality, contain highly divergent species, or fragmentary protein sequences.
5. Extension of the scope of existing phylogenomic workflows — users can actually infer gene family evolutionary history and dynamics under explicit models of gene duplication, transfer, and loss.
6. Modularity (see Next steps for expanded discussion).

Example testable questions	Example tests	Outputs used
Can we predict how some organism obtains its food?	Phylogenetic profiling	☒ Gene families
Which species do or do not have my favorite gene?	Presence/absence of gene in focal gene family	☒ Gene families
What is the likely function of some uncharacterized protein(s) in my favorite species?	Functional annotation transfer among orthologs/closely related genes	☒ Gene families
		☒ Gene family trees
How are species related to each other, and how far back in time did they diverge?	Species tree inference and divergence time estimation	☒ Multiple sequence alignments
		☒ Species trees
When did my favorite gene family originate?	Mapping gene family evolution to species tree	☒ Gene families
		☒ Gene family trees
		☒ Species trees
What was the protein sequence in the ancestor of my favorite species?	Ancestral sequence reconstruction	☒ Multiple sequence alignments
		☒ Gene family trees
In which organisms has my favorite gene evolved most rapidly?	Comparing rates of protein-wide amino acid substitution across species	☒ Multiple sequence alignments
		☒ Gene family trees
		☒ Species trees
Are different parts of a gene's sequence more	Comparing local rates of amino acid substitution	☒

Example testable questions	Example tests	Outputs used
or less highly conserved?	within protein	Multiple sequence alignments
		⊠ Gene family trees
When organisms move to different environments, do certain gene families expand or contract?	Correlate rates of duplication/ transfer/loss with environment of species and their common ancestors	⊠ Gene family trees
		⊠ Species trees
		−⊠+ Gene family evolutionary dynamics (duplication/transfer/ loss)

Table 1

We also list example approaches to test corresponding hypotheses and the data (NovelTree outputs) used to do so.

We've written this pub with two intended audiences in mind. We wrote the introductory sections for a general audience with a less extensive background in evolutionary biology and phylogenetics. We've intentionally omitted most of the granular methodological details from these sections (Study system and Application). If that's exactly what you're looking for, we've got just the thing for you! You'll find all relevant details about how we generated and prepared our data set in the Methods: Data management section, and fine-grained details about the software implemented within NovelTree in the Methods: Workflow implementation section.

TRY IT: You can find the **NovelTree workflow** in this GitHub repository (DOI: 10.5281/zenodo.8387630).

NovelTree: An applied walkthrough

In the following section, we show what NovelTree can do by applying it to an example use case — analyzing an interesting group of eukaryotes referred to as the TSAR supergroup. We've written this section for a general scientific audience without expertise in phylogenomics. We don't go into great detail as we describe these results because we plan to share a more in-depth "Result" pub in the near future.

We wrote the Methods section for an expert/technical audience. That's where you'll get a deep dive into the methodology of data curation (Data management) and NovelTree's inner workings (Workflow implementation).

Study system

To demonstrate NovelTree's utility for phylogenomic and comparative evolutionary inference, we applied the workflow to a dataset of 36 TSAR eukaryotes (*Telonemia*, *Stramenopila*, *Alveolata*, and *Rhizaria*). Collectively known under the moniker TSAR, these four lineages comprise an understudied collection of four eukaryotic supergroups (Table 2, **[21]**) that diverged from their most-recent common ancestor approximately 1.4 billion years ago **[22]**. Perhaps most importantly, diversity within TSAR is immense, yet proportionally very little is known about them — estimated species diversity across all four groups exceeds 120,000, yet the vast majority (> 90%) are only known to science from morphological data (Table 2). That said, availability of genetic sequence data within the group is increasing at an accelerating rate, in turn providing increased resolution of their evolutionary relationships to one another and improved understanding of how their immense ecological and morphological diversity came to be. For an in-depth review of the biology and current state of scientific knowledge of SAR specifically, see **[23]**.

Supergroup

Supergroup	Major subclades	Approx. # species	# sampled species	N
Telonemia **[24]**	Telonema	Two described, likely several genera **[25]**	UniProt proteomes: 0 EukProt genomes: 0 EukProt single-cell Genomes: 0 EukProt transcriptomes: 2	**Te** H pr
Stramenopila **[26]**	Gyrista	> 100,000 **[27]**	UniProt proteomes: 7 EukProt genomes: 1 EukProt single-cell Genomes: 1 EukProt transcriptomes: 3	**G** Pr ph (C pa (P
	Opalozoa (Bigyra)			**O** Pr he sc
	Sagenista (Bigyra)			**Sa** he
Alveolata **[28]**	Apicomplexa	> 10,000 **[28]**	UniProt proteomes: 10 EukProt genomes: 0 EukProt single-cell genomes: 0 EukProt transcriptomes: 4	**A** O

Supergroup	Major subclades	Approx. # species	# sampled species	N
	Dinoflagellata			**D** M pl ot er pr pa
	Ciliophora			**C** Pr
Rhizaria **[29]**	Cercozoa	> 10,000 **[23]**	UniProt proteomes: 2 EukProt genomes: 1 EukProt single-cell genomes: 0 EukProt transcriptomes: 5	**C** he sc pl
	Retaria			**R** Pl Fo Ra Fo ac m he sc pa m Ra zo m sc he

Table 2

***Telonemia*, *Stramenopila*, *Alveolata*, and *Rhizaria*: TSAR) studied herein.**

For instance, *Rhizaria*, a group that lacks a defining characteristic, was only recently recognized and proposed as a eukaryotic supergroup on the basis of genetic data **[30]**. This clade contains immense biological diversity. Most are free-living heterotrophs, but there exist numerous parasitic or photosynthetic species. On the other hand, alveolates are proportionately better studied, being among the first

microbes to become known to science **[23]**. Alveolates are composed of mostly parasitic, single-celled organisms such as the photosynthetic and toxin-producing dinoflagellates that cause “red tides,” and parasitic apicomplexans such as *Plasmodium,* which causes malaria. Stramenopiles are thought to be the most diverse among the four groups. Stramenopiles range from autotrophic to heterotrophic, parasitic to endocommensal, and single-celled to multicellular, as in the case of large brown algae such as kelp. *Telonemia* is the least diverse, and also the least well understood group of the bunch, having only been formally recognized as recently as 2006 **[24]**.

Given the immense array of life-history diversity that has originated, and in many cases repeatedly and independently in the TSAR eukaryotes, we hypothesize they harbor a wealth of transformative evolutionary innovations that are yet to be discovered. Our efforts presented here represent an initial effort to enable their discovery.

Application

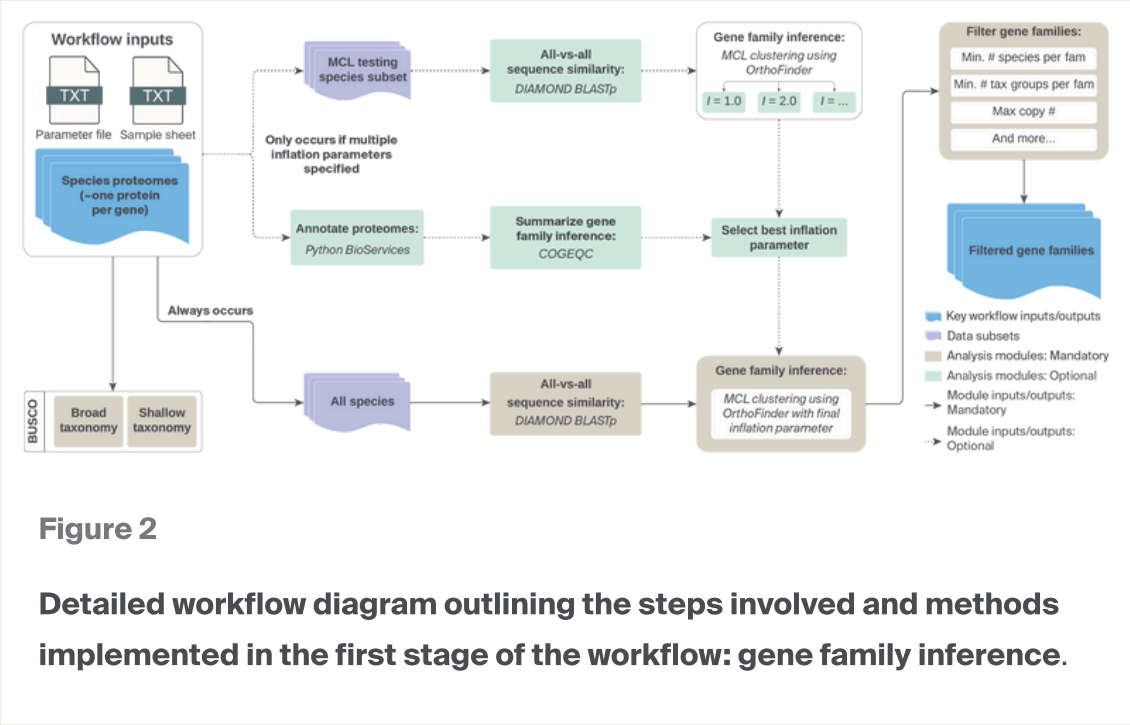

Figure 2

Detailed workflow diagram outlining the steps involved and methods implemented in the first stage of the workflow: gene family inference.

The ever-growing availability of publicly available transcriptome and genome assemblies presents immense opportunity for comparative genomic and phylogenetic studies across broad swaths of biodiversity. However, these data vary extensively in their quality. It is often quite useful to obtain estimates of proteome quality and

completeness for each species within a data set, as such information can provide additional context for observed patterns of gene family evolution. Consequently, NovelTree returns estimates of proteome quality and completeness as estimated using BUSCO **[31]** (Figure 2). These analyses provide a high-level summary of how "complete" a proteome is with respect to a reference set of evolutionarily conserved genes, as well as how many of these are fragmented, duplicated, or missing. Results of these analyses for the TSAR eukaryotes we're studying are plotted in Figure 3, A. Although we observe substantial heterogeneity in proteome completeness, the proteomes of most species are more than 75% complete; alveolates were the most complete overall, whereas stramenopiles and rhizarians had a greater number of missing or fragmented BUSCOs.

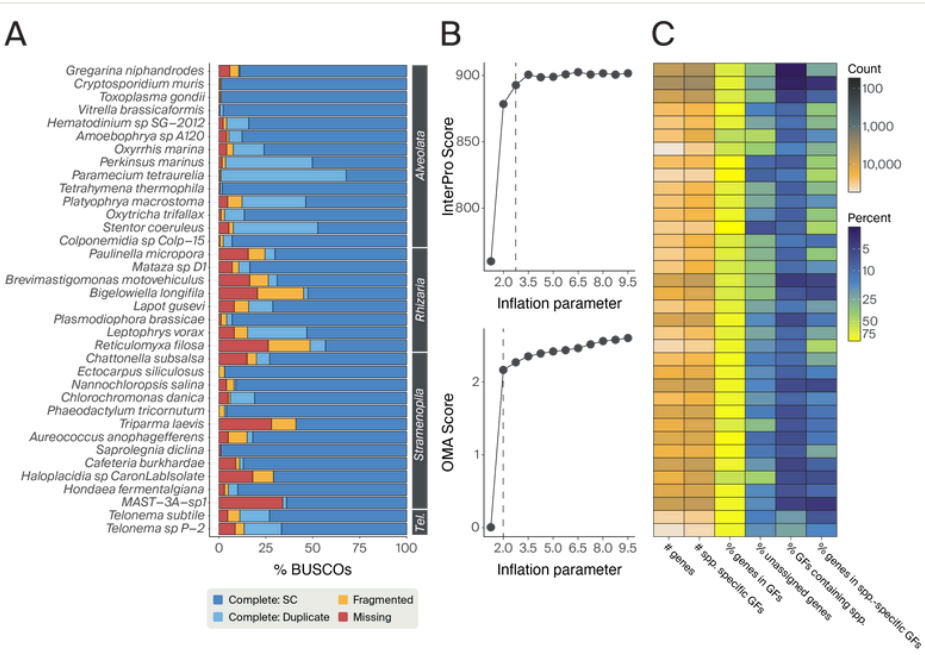

Figure 3

NovelTree produces BUSCO proteome quality summaries and high-level overview of results after gene family inference.

(A) BUSCO scores for each species obtained using the shallow taxonomic databases.

(B) Summary statistics from COGEQC for interpro protein domains and OMA orthology, summarizing for multiple values of MCL clustering inflation parameters how well annotations are clustered within inferred gene families, penalizing against dispersion of these annotations among inferred gene families. Greater values are better, and the selected values for each summary stat are indicated by the dashed vertical line, selected using the “elbow method” — greater values are better.

(C) Heatmaps showing the count (warmer colors) or percent (cooler colors) of different gene family (GF) count/species membership statistics, each indicated at the bottom of each column.

See our interactive walkthrough to learn how to produce plots like these.

Next, we focused on the issue of gene family inference. To do this, we must first cluster each species' proteins into homologous (evolutionarily "equivalent") groups or gene families. In a given gene family, each protein is assumed to have descended from the same ancestral gene copy through speciation or gene duplication events. To accomplish this, we used a custom approach to cluster protein sequences into gene families given their sequence similarity (Figure 2). Specifically, we optimized the inference of gene families by assessing the impact of a critical clustering parameter on the distribution of biologically informative protein annotations within and among gene families, favoring parameters that produced largely homogenous compositions (see Workflow implementation for details). The resulting gene families represented a close-to-optimal tradeoff between performance and cluster granularity (Figure 3, B–C). Additionally, we implemented a number of ways to filter gene families on the basis of gene copy number, number of species, and more (Figure 2). For a detailed description of these filters and the parameter values used for our data set, jump to Gene family filtering and Non-default workflow specifications.

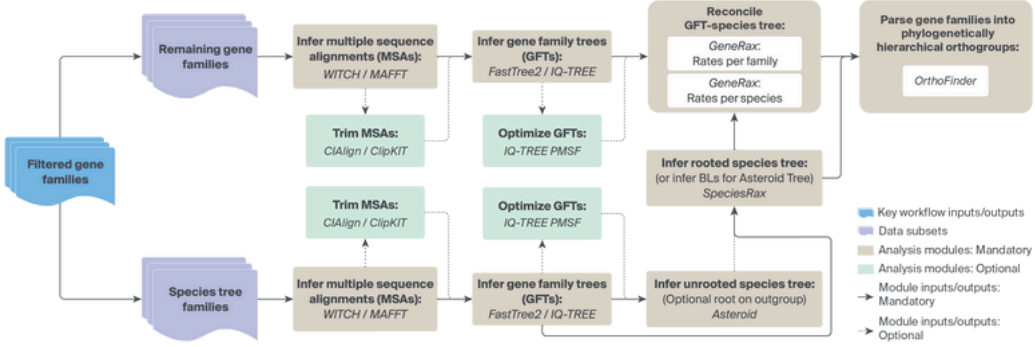

Figure 4

Detailed workflow diagram outlining the steps involved and methods implemented in the second stage of the workflow: phylogenetic inference, and inference of gene family evolution.

Using our filtered gene families, we proceeded with phylogenetic analysis of each family, beginning first with the inference of multiple sequence alignments (MSAs), followed by the inference of gene family trees using these MSAs (Figure 4). We used

these gene family trees to infer a species tree depicting the history of evolutionary divergence and relationships among our study species (Figure 5). In the future, we can use these data to (for instance) conduct ancestral sequence reconstruction, investigate rates of molecular evolution, explicitly test for and identify cases of protein sequence convergence, and more.

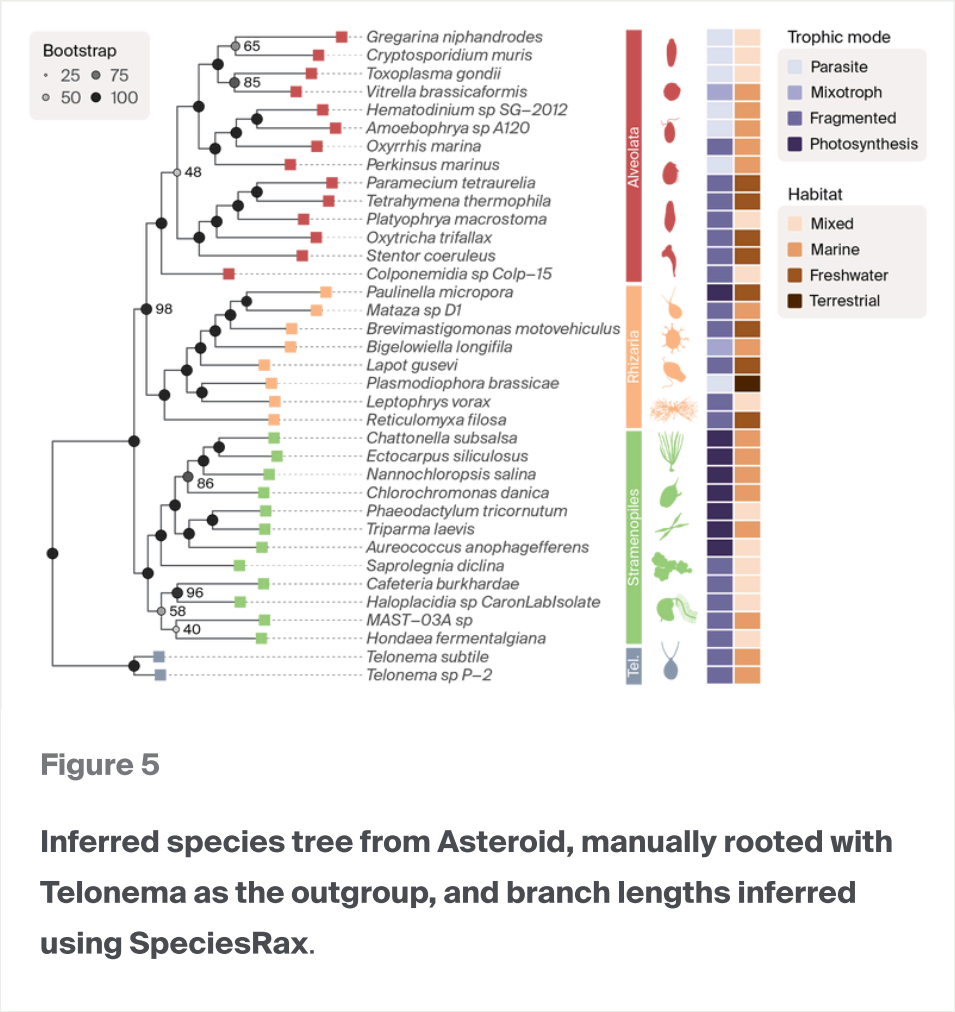

Figure 5

Inferred species tree from Asteroid, manually rooted with Telonema as the outgroup, and branch lengths inferred using SpeciesRax.

We report bootstrap topological support values as node circles, with the size and shade proportional to the percent of bootstrapped phylogenies supporting each bipartition. We only annotate bipartitions (“branching points”) with support < 100. We include example images of each group and species’ trophic style and habitat to the right of the phylogeny.

Species trees have a great number of potential utilities, particularly when paired with the species’ phenotypes or other functionally or ecologically relevant metadata. Within such a framework, researchers can, for instance, formally test hypotheses of environmentally mediated convergent or divergent phenotypic evolution, infer the timing, order, or number of origins of traits, and much more. The species tree inferred

here, rooted with *Telonemia* and inferred using a total of 2,018 gene families, supports a hypothesis in which Stramenopiles are sister to a clade composed of Alveolates and Rhizarians, each of which are reciprocally monophyletic. In a later pub, we will conduct a deeper dive into this species tree, the distribution of support for alternative topologies across gene families, and more.

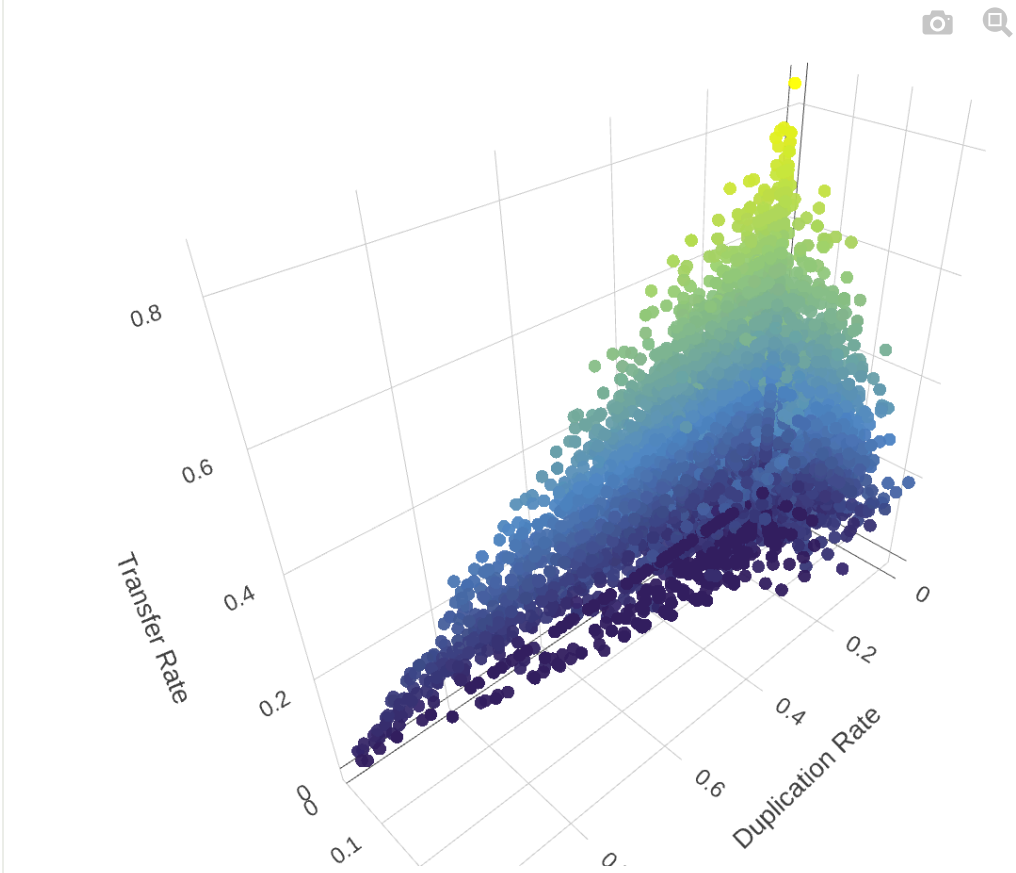

Figure 6

Interactive 3D plot of the inferred rates of gene family duplication, transfer, and loss as inferred under GeneRax's per-family model.

Click, drag, and scroll to rotate, pan, and zoom in the plot!

Each point is a gene family, and the gene family composition is indicated for each when hovered over. Gene families are colored according to their verticality, which is defined as the proportion of all gene family evolutionary events that are vertical in nature (i.e., excluding transfers).

This figure is intended to highlight the exploratory potential of these data. For instance, we see that rates of gene loss are typically exceeded by both rates of duplication and transfer, a pattern that is consistent with previous studies.

View this interactive figure in a new tab.

One of our primary goals is to investigate how gene family evolution can contribute to biological innovation. Thus, we paired our inferred gene family and species trees, enabling us to infer rates of gene duplication, transfer, and loss for each species, and for each gene family (Figure 6, Figure 7, Figure 8). We have barely begun to scratch the surface of these data, but even a cursory investigation makes the potential clear. For example, when looking at rates of gene loss across all gene families for each branch of the species tree (Figure 7, B), we observe some of the greatest rates of gene loss at the base of the alveolates. As another example, we observe an unusually high number of transfer events between *Lapot gusevi* (*Rhizaria*) and *Platyophrya macrosoma* (*Stramenopila*) (Figure 8). In contrast, we see two marked reductions in the rate of horizontal transfer occur within *Alveolata* (specifically Apicomplexa and Ciliophora), a pattern that is perhaps driven by the elevated rates of gene loss within this group (Figure 7, B).

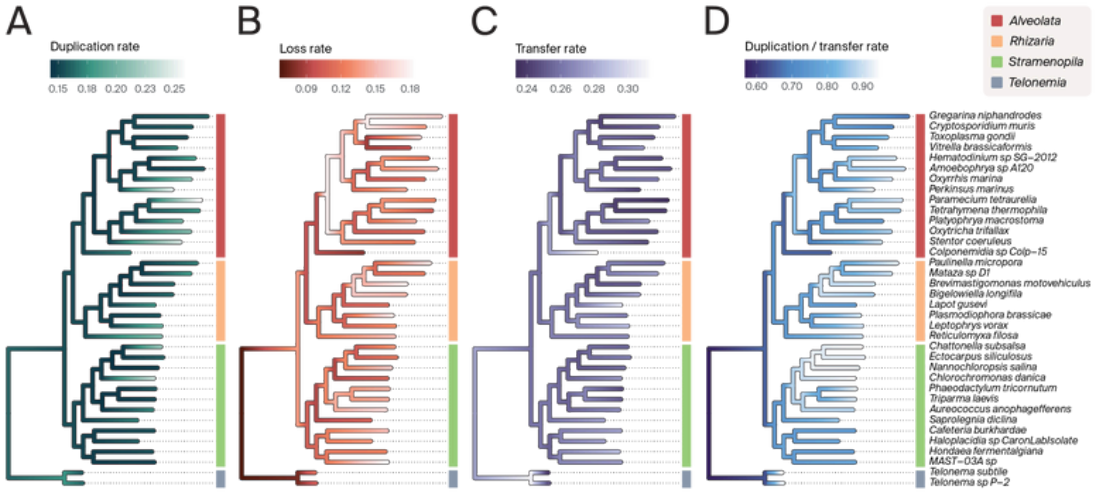

Figure 7

Rates of gene family duplication, transfer, loss, and the ratio of duplication to loss rate, as summarized as the average across all gene families, for each branch in the species tree.

Finally, knowing whether two protein sequences are direct orthologs of one another is useful for numerous scientific efforts. Direct orthologs are not only members of the same gene family, but they are descended from the same ancestral gene copy

following speciation (as opposed to paralogs, which originated from duplication events). For instance, orthologs have long been used to transfer or propagate functional annotations between proteins — that is, annotating one protein using the annotation of a closely related protein — a practice motivated by the so-called "ortholog conjecture" **[32]**. Although we've recently learned that transference of protein annotations in this way benefits from the inclusion of both paralogs and orthologs, orthologs do still tend to be more functionally similar than paralogs **[32]**. For this reason, and because ortholog inference is standard procedure in related phylogenomic workflows, we used OrthoFinder once more to parse each gene family into phylogenetically hierarchical orthogroups by comparing the inferred gene family trees with the species tree. These hierarchical orthogroups indicate the set of orthologous protein sequences that derive from the same ancestral gene copy following speciation for each internal node within the species tree. In other words, in the final step of the workflow, we return a set of orthologous proteins at each node in the tree, enabling researchers to more easily zoom in on the closely related protein sequences possessed by specific lineages within the tree as needed. We have hosted all hierarchical orthogroups, as well as all other workflow outputs in a Zenodo data repository (DOI: 10.5281/zenodo.8237421).

> **SHOW ME THE DATA**: The **original protein sequence data** prior to pre-processing, **filtered protein sequences**, and all **outputs** from NovelTree are available here.

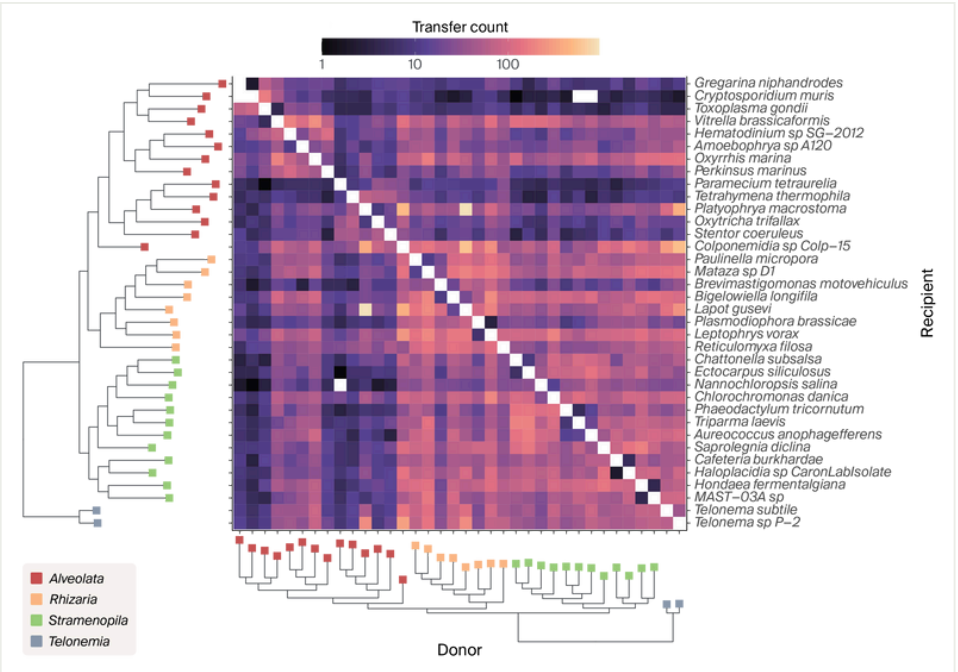

Figure 8

Heatmap depicting the number of horizontal transfer events among pairs of species, summed across all gene families as inferred under GeneRax's per-species model.

The inferred species tree is plotted along the x- and y-axes, with species names corresponding to the species tree (rows of cells) along the right. The matrix is not symmetrical — recipient species are plotted along the y-axes, and donor species are plotted along the x-axis.

Check out an interactive version of this plot to investigate specific details.

In sum, NovelTree is a scalable, generalizable, and integrated approach for empowering high-throughput evolutionary biology. Since it packages the major components of phylogenomic analysis in a straightforward-to-implement and cost-effective workflow, NovelTree makes it possible to identify a taxonomic group of interest and quickly start dissecting its evolutionary history. Importantly, the results presented here are just scratching the surface of the signals that can be extracted by NovelTree. Phylogenomic libraries such as these contain myriad interesting patterns, each ripe for downstream investigation of the tempo and mode of evolution. It is a certainty that there are many to be mined from the TSAR library (and will be the basis

of future pubs). Given the flexibility of NovelTree, we look forward to exploring what the tree of life has to offer.

> The **code** we used for all the **TSAR analyses** is available in this GitHub repository (DOI: 10.5281/zenodo.8384568).

Output summary

Below, we have embedded an interactive HTML document that walks through how users may summarize and visualize workflow outputs using the R scripts we provide in this associated GitHub repository. For easier viewing, you can download this HTML doc from here or open in a new browser tab.

1 Application of `NovelTree` to TSAR Eukaryotes.

Code

NovelTree Results: TSAR Eukaryotes

Austin H. Patton

2023-08-24

1 Application of `NovelTree` to TSAR Eukaryotes.

Here, we summarize the multitude of results produced by the `NovelTree` workflow. Because results come from numerous distinct software, we have provided a complementary script containing helper functions to summarize these outputs.

1.1 Preparation

We will begin by creating a variable that specifies where all workflow outputs are stored

TRY IT: You can find the **NovelTree workflow** in this GitHub repository (DOI: 10.5281/zenodo.8387630).

Next steps

As suggested in "Our solution" and as highlighted throughout our description of the workflow, NovelTree is highly modular by design, and users have a unique degree of control over their analyses as compared to other phylogenomic workflows, being able to modify nearly all parameters for all included software. Because it is implemented in Nextflow, it is entirely possible to build upon and modify the workflow. We intend to do exactly that, but also enthusiastically welcome community engagement with and contributions to NovelTree. We would love for NovelTree to serve as a highly fine-tuneable and generally applicable toolkit for standard and specialized phylogenetic analysis alike.

NovelTree is ripe for continued development, including:

- Alternative methods/approaches to gene family inference, relaxing the need to use functional annotations
- Generalization to handle different types of inputs (e.g., nucleotide sequences of protein sequences)
- Expanded protein/gene family filtering capabilities such as identifying and removing contaminated sequences (e.g., by removing spuriously long branches)
- Automated tailoring of phylogenetic inference to individual gene families (e.g., IQ-TREE vs. FastTree2 based on family size, improved model specification, etc.)
- Building in additional functionalities/types of analyses (e.g., inference of molecular evolution)
- Evolutionarily informed/tree-guided propagation of functional annotations to gene families or constituent proteins
- Automated generation of workflow output summaries, akin to what is done in the walkthrough associated with this pub
- And more!

Additionally, we intend to build and release pubs describing “recipes” to easily and appropriately conduct different combinations of analyses implemented within the workflow. These recipes will enable scientists to easily, and with limited familiarity with Nextflow, be able to run analyses that are more tailored to their needs. For instance, they’ll let users toggle specific modules on or off, or select a range of parameters suitable for specific use cases, all without requiring the user to make each modification themselves. Ultimately, we’ll curate these recipes into a “phylogenomic cookbook” that enables any scientist to apply robust phylogenetic methodology to a broad range of biological systems and data sets.

Is there anything you would like to see implemented in NovelTree? Please let us know!

Methods

While the rest of this pub is for a more general audience, here’s where experts can find all the technical details.

Data management

Curation

Leveraging publicly available data, we curated a data set of whole-proteomes from 36 species within TSAR, including two from *Telonemia*, 12 from *Stramenopila*, 14 from *Alveolata*, and eight from *Rhizaria* (Table 2). We selected these representative species to strike a balance between taxonomic completeness and data set quality. Specifically, we sought to exclude species with particularly poor BUSCO completeness (as reported for each proteome in both source databases described below), or unusually high protein counts prior to filtering (i.e. > 20,000–40,000). The latter issue presented most often for transcriptome-derived proteomes.

We obtained protein sets from either UniProt reference proteomes or from EukProt (version 3) **[33][34]**. We favored UniProt reference proteomes over EukProt entries where both existed due to their additional rigorous curation. In particular, their reduction down to one protein per-gene makes them particularly amenable to gene family evolutionary analysis, as the inclusion of alternative isoforms (i.e. in

transcriptomes) can confound estimates of gene duplication, transfer, or loss. Proteomes obtained from EukProt include either those originating from whole-genome assemblies, single-cell genome assemblies, or transcriptome data. We used transcriptome-derived proteomes only as necessary for taxonomic completeness and subsequently filtered them more conservatively prior to phylogenomic inference (see the next subsection for an extended description).

Data preprocessing

At this point, we had a protein set for each species that was either a UniProt reference proteome, a EukProt genome-derived proteome, or a EukProt transcriptome-derived proteome. To improve uniformity across datasets, we subsequently filtered these datasets more conservatively to reduce protein redundancy and to mitigate any impact of data source rates of gene family evolution across species. Because transcriptome-derived proteomes have a greater likelihood of retaining multiple transcripts per-gene, we filtered these data more extensively than the former two sources.

Starting from the original, unmodified protein sequences for each species, we first excluded any protein sequence shorter than 50 amino acids (AA) using seqkit (version 2.3.1) **[35]**. Next, for transcriptome-derived proteomes, we corresponded each protein ID to gene ID using information contained within the sequence headers, using bioawk (version 1.0) to record the length of each protein sequence, and retained the single longest transcript per-gene using seqinr (version 4.2_23) **[36]** as implemented in R (version 4.2.2) **[37]**. Next, we used CD-HIT (version 4.8.1) **[38]** to further reduce protein redundancy in the proteomes of each species such that we retained a single representative protein sequence for each cluster. As stated earlier, we sought to filter transcriptome-derived proteomes more conservatively than genome-derived proteomes, treating UniProt reference proteomes as our “standard” — the filtering here reflects this. Thus, we reduced transcriptome-derived protein data sets using a sequence similarity threshold of 0.90 (90%), and reduced genome-derived sequences (excluding UniProt reference proteomes) using a threshold of 0.95. Last, we used seqinr once more to rename protein sequences into a format that is consistent across all samples, independent of their source. Specifically, we renamed proteins according to the convention, `Genus_species:ProteinID`.

Workflow implementation

Caveats

We are actively developing NovelTree. Consequently, aspects of implementation, features, and behaviors of the workflow are subject to change — we intend to openly and clearly document these changes as development progresses. For this reason, we welcome and encourage user feedback, as we would like the workflow to be as broadly useful for other scientists as possible. Users of the workflow are encouraged to publicly post any feature requests or report any bugs they encounter on NovelTree's GitHub issues page.

Second, NovelTree is not a substitute for due diligence in experimental design. Prior to running the workflow, the user should have conducted sufficient data curation such that the species and their proteomes are well-suited to the user's goals.

For instance, take a scenario in which a researcher would like to ask whether opsin gene family contractions or expansions are associated with transitions to a pelagic life history in fishes (e.g., **[39]**). In such a scenario, the user should make a concerted effort to steer clear of acquisition bias in their experimental design. This would mean avoiding phylogenetic pseudo-replication wherein a single sampled evolutionary transition to a pelagic life history renders all sampled pelagic species evolutionarily non-independent **[40]**. This could confound the researcher's ability to meaningfully recover evolutionary associations.

Furthermore, where only transcriptomes are available for certain species, the user must be sure to assess whether the source tissues are appropriate. Under the previous example, assembled transcriptomes should include data obtained from eye tissue, otherwise it is likely that these species will be absent from analyses of gene family evolutionary dynamics. Conversely, transcriptomes derived exclusively from eye tissue will include a non-random sample of gene families, confounding the user's ability to assess whether rates of opsin gene family evolution in these species are exceptional with respect to genome-wide patterns. In general, transcriptomes derived from whole-organism samples, or at the very least multiple tissue sources, are preferable.

Last, we emphasize that the user must sufficiently clean their data set prior to analysis. Others have written extensively about what this means exactly — for example, see **[19]**.

Our approach to data cleaning is described in the Data preprocessing section. As discussed in more detail within that section, it is critical that the user make efforts to filter their data such that as close to one protein per gene family is retained so as not to confound estimates of the rate of gene family evolution. In later updates to the pub, we will provide additional guidelines/recommendations with respect to experimental design, including scenarios where such acquisition bias may be unavoidable.

Inputs

NovelTree takes two files as input:

1. A sample sheet:
    - This comma-separated text file provides metadata for each sample/species as well as a filepath/s3 URI to each input file, which is an uncompressed, filtered proteome/amino acid FASTA file.
    - A detailed description of each parameter is available in the NovelTree GitHub repository.
    - The properly formatted sample sheet we used to run the analyses described in this pub may be found here.
2. A parameter file:
    - This JSON-formatted text file simplifies user interactions with the command-line interface, and defines mandatory user-defined parameters.
    - Here, the user may specify:
        - Which tools they would like to use for multiple sequence alignment, alignment cleaning, gene family inference, and species tree inference.
        - The values of the MCL inflation parameter they would like to assess for use in gene family inference.
        - Any preferred filtering of gene families prior to phylogenetic analysis, such as requiring a minimum number of species, maximum mean per-species copy number, outgroup taxa, and more.
        - Optional or module-specific parameters, which users can also define within the module configuration file.

- The properly formatted parameter file used to run the analyses described herein may be found here, and a detailed description of parameter options are described on the NovelTree GitHub README.

After cloning the NovelTree repository and navigating to that directory, the user simply calls the workflow using the following command:

```
nextflow run main.nf -profile docker -params-file nextflow_parameters.json
```

Note that both Docker and Nextflow must be installed and configured prior to running the workflow. Additionally, the bulk of our testing of the workflow was conducted on machines using the Ubuntu v20.04 operating system.

Proteome quality summaries

Upon initiation of the workflow, the first analysis NovelTree conducts is estimating BUSCO scores for each species using user-specified lineage data sets. Up to two reference BUSCO databases may be used for each species; these data sets are specified within the sample sheet under the `shallow_db` and `broad_db` fields and correspond to the databases that are taxonomically closest (e.g., `alveolata_odb10`, or `stramenopiles_odb10`) to each species, or the most taxonomically inclusive of all species (e.g., `eukaryota_odb10`).

Importantly, these BUSCO analyses are optional and may be skipped by entering `NA` into one or both fields. We chose to allow the user to specify multiple taxonomic scales because use of a single database for all species may be misleading when studying a set of species that are highly divergent, as in our case. This choice is consistent with the recommendation of BUSCO's developers to use the specific lineages when possible. Inspection of our own results further supports this — Figure 3, A plots BUSCO quality summaries for each species using the shallow taxonomic-scale data sets, but you can see results from both analyses here.

Gene family inference: MCL clustering optimization

Unless the user provides a single value of the inflation parameter when running NovelTree, the first stage in the pipeline's gene family inference procedure is to determine which value of the MCL inflation parameter provides the most sensible results with respect to the functional composition of inferred gene families. This is

accomplished using a taxonomically representative subset of species (to reduce computational burden) including several species that use UniProt protein accessions as their protein IDs. The workflow uses these UniProt protein accessions to obtain functional annotations to facilitate optimization of the MCL inflation parameter. If the user provides only one parameter value, the workflow skips the optimization steps and goes directly to the Gene family inference: Post-optimization stage of the workflow. No species with UniProt protein accessions are required under this scenario.

Using this subset of species, NovelTree implements the following steps:

1. **UniProt proteome annotation**:
    - We use the BioServices **[41]** Python package.
    - The user can specify up to 16 different sets of protein annotations to download from UniProt given each species' UniProt protein accessions. These include protein IDs for external reference databases (e.g., OrthoDB **[42]**, EggNOG **[43]**), annotations of any known post-translational modifications, and more.
    - We outlined these details in more depth on NovelTree's usage documentation.
2. **All-v-All DIAMOND BLASTp [44]** of protein sequence similarity among and within all sampled species.
3. **Gene family inference** using OrthoFinder's implementation of MCL clustering with an array of (user-specified) inflation parameters:
    - OrthoFinder's implementation of MCL clustering **[45]** normalizes BLAST similarity scores to account for sequence length divergence between species.
    - Sequence similarity (both bit scores and e-values) is pathologically confounded by sequence length divergence, and thus by evolutionary divergence between species.
4. **Gene family inference quality assessment for tested inflation parameters**:
    - COGEQC **[46]** scores inflation parameters on the distribution of functional annotations within and among inferred gene families using a summary statistic:

      $$Score = \frac{Homogeneity}{Dispersal}$$

    - This score favors inflation parameters that produce gene families composed of largely homogenous functional annotations, and penalizes those for which

annotations are dispersed across multiple gene families.

- Currently, we use InterPro domains as our functional annotations, and OMA orthology IDs as our known orthology information (obtained in Step 1). In future versions, this may change or even be user-specified.

5. **Inflation parameter selection (Figure 3, B)**
    - Using the elbow method, for each score, we identify the inflection point at which increasing the inflation parameter leads to diminishing improvements.
    - We take the mean of these two inflation parameters and use this for downstream gene family inference for the complete data set.

Gene family inference: Post-optimization

The second stage of NovelTree's gene family inference procedure largely mirrors that of the first, but consists of fewer steps and uses all species.

In brief, these stages are:

1. **All-v-All DIAMOND BLASTp [44]** of protein sequence similarity among and within all sampled species.
2. **Gene family inference** using the inflation parameter selected during the optimization stage.
    - We provide summary scripts in an associated GitHub repository (with a walkthrough of their application to the workflow results presented herein) to summarize and visualize results of inferred gene families.
    - This includes information pertaining to the taxonomic composition and gene families, per-species gene family statistics, and more (Figure 3, C).

Gene family filtering

Following gene family inference using the complete data set, NovelTree filters gene families (according to the user's specification) into a set of gene families to use for species tree inference, and another set for all other phylogenetic analysis.

These filters include:

1. The minimum number of proteins a gene family must contain:
    - *Default = 4*
    - Increasing this parameter has a large influence on the number of retained gene families, as the distribution of gene family size (number of proteins) is highly right-skewed.
    - Families with fewer than four sequences inherently lack phylogenetic resolution, limiting their utility; we recommend against their inclusion.
2. The minimum number of species per gene family:
    - *Default = 2*
    - *Max = Number of species*
    - This default is somewhat arbitrary, but is intended to retain only gene families that are informative of evolutionary relationships among species.
    - Increasing this will lend priority to more deeply conserved gene families.
3. The minimum number of taxonomic groups per family:
    - *Default = 1*
    - *Max = Number of taxonomic groups in sample sheet*
    - Lower values enable the discovery of lineage-specific gene families (such as those originating through gene family births).
    - Larger values led priority to families that are conserved across greater numbers of the user-defined taxonomic groups.
4. The maximum mean copy number per species:
    - *Default = 10*
    - This default is chosen as a conservative upper limit to prevent the inclusion of exceptionally large, computationally intensive gene families.
    - Choice of this parameter will depend on the needs of the user, data set size, and computational resources available to them.
    - Species absent from a gene family are not included in the estimate of this value.
5. The maximum mean copy number per species for use in species tree estimation:
    - *Default = 5*

- *Max = Parameter specification for maximum mean copy number per species*
- This parameter is set lower than the one described above so as to ensure/prioritize that the species tree is inferred in advance of the GeneRax modules.
- Increasing this parameter will increase the number of gene family trees that Asteroid and SpeciesRax use by including larger and more complex gene families.

6. The minimum proportion of species that a gene family must contain for use in species tree estimation:
    - *Default = 0.75*
    - This default is conservative. Asteroid is robust to the presence of missing species in gene families, and SpeciesRax implements an approach to limit their impact.
    - Increasing this parameter will decrease the number of gene families NovelTree uses in species tree inference, although each gene family will be more "complete."
    - Decreasing the parameter will include a larger number of gene families at the expense of gene family "completeness" across species.

Multiple sequence alignment

After filtering gene families, NovelTree proceeds into its third stage — multiple sequence alignment and alignment cleaning. We implement two multiple sequence aligners, and two methods for alignment cleaning (Figure 4).

Multiple sequence alignment:

1. **WITCH [47]** (default)
    - We have chosen WITCH to be the default alignment method, as it is ideally suited for types of data sets typical of analysis with NovelTree.
    - These data sets are often composed of many gene families that contain substantial sequence length heterogeneity, whether due to the inclusion of fragmentary sequences, or to substantial sequence length divergence.

- From our experience, use of WITCH leads to dramatic improvement of multiple sequence alignment quality, particularly for large, complex gene families.

2. **MAFFT [48]**
    - When MAFFT is selected, NovelTree implements the L-INS-i iterative algorithm by default.

Alignment cleaning:

1. **None (default)**
    - By default, no additional processing of multiple sequence alignments occurs. This is because the WITCH (the default aligner) module implements its own alignment cleaning to remove gappy sites and sequences that are subsequently shorter than some user-specified threshold (*default = 20 AA*).
2. **ClAlign (optional) [49]**
    - Lightweight, efficient, and highly customizable alignment trimmer to remove both difficult-to-align positions and sequences from alignments.
3. **ClipKit (optional) [50]**
    - Lightweight, efficient alignment trimmer with several optional user-specified alignment trimming “modes” to remove difficult-to-align positions from alignments.
    - By default, the pipeline uses “smart-gap” — the recommended method.

Gene family tree inference

Using the (cleaned) multiple sequence alignments inferred during the previous stage from the filtered sets of gene families, NovelTree then proceeds to infer gene family trees for each. As with all software implemented herein, the parameters for each method are fully customizable by the user (Figure 4).

The workflow implements two methods:

1. **FastTree2 [51]** (default, for expediency)
    - We applied this method to our data set of TSAR eukaryotes.

2. **IQ-TREE2 [52]** (for accuracy)

Optionally, a researcher may use the gene family trees inferred at this stage to parameterize and infer trees under the more complex posterior mean site frequency amino acid mixture models **[53]**. When chosen, this method currently applies to all gene families, but in future versions we intend to increase the flexibility of this approach, such as by applying it only to those gene families destined for species tree inference.

Species tree inference

Using the gene family trees inferred from the families destined for species tree inference, we implement two approaches to infer a species tree — rooted and unrooted (Figure 4).

1. **Asteroid [54]**
    - Asteroid infers an unrooted species tree without branch lengths, is robust to the presence of missing data, and can use both single- and multi-copy gene family trees.
    - Following the procedure implemented in its original publication, we decompose multi-copy gene family trees into their corresponding set of possible single-copy trees using DISCO **[55]** and use these for species tree inference. In later versions, we intend to make this optional.
    - If outgroup species are included in the data set, the user may specify these in the parameter file and use them to root the inferred species tree using phytools (version 1.5.1) **[56]**.
2. **SpeciesRax [57]**
    - If no outgroups are specified, SpeciesRax uses both single- and multi-copy gene family trees to infer a rooted species tree with branch lengths under a model of gene family duplication, transfer, and loss (DTL).
    - If outgroups are provided, the workflow uses SpeciesRax to infer branch lengths for the rooted species tree inferred using Asteroid.

Gene family evolutionary dynamics

Now, with gene family trees and a rooted species tree in hand, NovelTree proceeds to use GeneRax **[58]** to quantify rates of gene duplication, transfer, and loss for each gene family, and for each species and branch in the species tree (Figure 4). Currently, GeneRax is the only model we've implemented — in future versions we hope to implement alternative approaches, such as ALE **[59]**, CAFE **[60]**, or the parsimony-based COMPARE **[61]**.

This is accomplished by formally reconciling each inferred gene family tree with the species tree under a model wherein the likelihood of the reconciled gene tree is a function of either per-family or per-species rates of duplication, transfer and loss. Under the per-family model, it's assumed that all species or branches in the species tree share the same rates. To expedite these analyses, NovelTree analyzes gene families in parallel, separately under each model. GeneRax produces a wealth of information, the results of which are suitable for numerous downstream applications. We provide a set of functions to summarize these outputs into user-friendly formats, and additionally provide a walkthrough for how to use them, including example visualizations.

Briefly, the results of GeneRax enable the user to:

1. Improve gene family tree topologies using maximum likelihood under the model of DTL by reconciling them with the rooted species tree.
    - The workflow saves all reconciled gene family trees and returns them as output.
    - Users may visualize gene tree/species tree reconciliations using software such as ThirdKind **[62]** (as depicted within the fourth panel of Figure 1).
2. Infer rates of gene duplication, transfer, and loss for each gene family.
    - Under the per-family model, it's assumed that all species or branches in the species tree share the same rates (see Figure 6)
    - Under the per-species model, each species or branch in the species tree is assumed to have its own rates (see Figure 7)
        - These rates may still be summarized at the level of gene family (e.g., using the mean, median, or some measure of variance among branches).
        - Alternatively, per-branch rates may be summarized by taxonomic group.

3. Infer the count of gene family duplications, transfers, and losses for each gene family and each species.
    - NovelTree obtains event counts for each species from the reconciliations and thus from both the per-family and per-species models.
    - The workflow records the donor and recipient species for each transfer event and the user can easily summarize them (Figure 8) for exploratory or quantitative analysis using scripts we provide along with this pub.

Phylogenetically hierarchical orthogroups

In the last stage of NovelTree, the workflow reconciles gene family trees from GeneRax's per-species rate model and parses them into phylogenetically hierarchical orthogroups using OrthoFinder **[17]** (Figure 4). These hierarchical orthogroups (HOGs) are defined for each node in the species tree; the proteins contained in each HOG include all orthologs and in-paralogs that are derived from the same ancestral gene copy at the time of speciation. The output from this procedure is identical to what OrthoFinder outputs, except that NovelTree replaces OrthoFinder's default "Gene_Trees" directory with "GeneRax_Reconciled_GFTs," which contains the gene family trees used to infer HOGs.

Non-default workflow specifications for TSAR analysis

Whereas NovelTree tests MCL inflation parameters (*I*) 1.5–5.0 in intervals of 0.5 by default, we sought to increase the range of values assessed as we had no a priori knowledge regarding the value at which we expected performance to plateau. Consequently, we tested *I* = 1.25–9.5 in intervals of 0.75. Additionally, because we were analyzing a modest number of species here, we chose to relax our gene family filtering thresholds, allowing for a mean copy number per species of 50 (*default = 10*), and included gene families in species tree estimation if they had at most a mean per species copy number of 20 (*default = 5*). This led to the retention of even the largest gene families that met our other filtering criteria. Lastly, the literature strongly supports a hypothesis in which *Telonemia* is sister to *Stramenopila*, *Alveolata*, and *Rhizaria*. Accordingly, we thus manually rooted the Asteroid species tree using the two species

of *Telonema*, using SpeciesRax to infer branch lengths from all gene family trees rather than directly inferring a rooted species tree using SpeciesRax.

Additional methods

We used ChatGPT to write some code and clean up other code.

References

1. Locey KJ, Lennon JT. (2016). Scaling laws predict global microbial diversity. https://doi.org/10.1073/pnas.1521291113
2. MAYNARD SMITH J. (1970). Natural Selection and the Concept of a Protein Space. https://doi.org/10.1038/225563a0
3. Romero PA, Arnold FH. (2009). Exploring protein fitness landscapes by directed evolution. https://doi.org/10.1038/nrm2805
4. Crombach A, Hogeweg P. (2008). Evolution of Evolvability in Gene Regulatory Networks. https://doi.org/10.1371/journal.pcbi.1000112
5. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. (2007). Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. https://doi.org/10.1126/science.1136678
6. Fernández R, Gabaldón T. (2020). Gene gain and loss across the metazoan tree of life. https://doi.org/10.1038/s41559-019-1069-x
7. Hotaling S, Kelley JL, Frandsen PB. (2021). Toward a genome sequence for every animal: Where are we now? https://doi.org/10.1073/pnas.2109019118
8. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. (2015). Insights from 20 years of bacterial genome sequencing. https://doi.org/10.1007/s10142-015-0433-4

9 Kapli P, Yang Z, Telford MJ. (2020). Phylogenetic tree building in the genomic age. https://doi.org/10.1038/s41576-020-0233-0

10 Scornavacca C, Delsuc D, Galtier N. (2020). Phylogenetics in the Genomic Era. https://inria.hal.science/PGE/

11 Smith ML, Hahn MW. (2021). New Approaches for Inferring Phylogenies in the Presence of Paralogs. https://doi.org/10.1016/j.tig.2020.08.012

12 Smith ML, Vanderpool D, Hahn MW. (2022). Using all Gene Families Vastly Expands Data Available for Phylogenomic Inference. https://doi.org/10.1093/molbev/msac112

13 Yan Z, Smith ML, Du P, Hahn MW, Nakhleh L. (2021). Species Tree Inference Methods Intended to Deal with Incomplete Lineage Sorting Are Robust to the Presence of Paralogs. https://doi.org/10.1093/sysbio/syab056

14 Demuth JP, Bie TD, Stajich JE, Cristianini N, Hahn MW. (2006). The Evolution of Mammalian Gene Families. https://doi.org/10.1371/journal.pone.0000085

15 Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, Anstead CA, Ayoub NA, Batterham P, Bellair M, Binford GJ, Chao H, Chen YH, Childers C, Dinh H, Doddapaneni HV, Duan JJ, Dugan S, Esposito LA, Friedrich M, Garb J, Gasser RB, Goodisman MAD, Gundersen-Rindal DE, Han Y, Handler AM, Hatakeyama M, Hering L, Hunter WB, Ioannidis P, Jayaseelan JC, Kalra D, Khila A, Korhonen PK, Lee CE, Lee SL, Li Y, Lindsey ARI, Mayer G, McGregor AP, McKenna DD, Misof B, Munidasa M, Munoz-Torres M, Muzny DM, Niehuis O, Osuji-Lacy N, Palli SR, Panfilio KA, Pechmann M, Perry T, Peters RS, Poynton HC, Prpic N-M, Qu J, Rotenberg D, Schal C, Schoville SD, Scully ED, Skinner E, Sloan DB, Stouthamer R, Strand MR, Szucsich NU, Wijeratne A, Young ND, Zattara EE, Benoit JB, Zdobnov EM, Pfrender ME, Hackett KJ, Werren JH, Worley KC, Gibbs RA, Chipman AD, Waterhouse RM, Bornberg-Bauer E, Hahn MW, Richards S. (2020). Gene content evolution in the arthropods. https://doi.org/10.1186/s13059-019-1925-7

16 Pan D, Zhang L. (2009). An Atlas of the Speed of Copy Number Changes in Animal Gene Families and Its Implications. https://doi.org/10.1371/journal.pone.0007342

17 Emms DM, Kelly S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. https://doi.org/10.1186/s13059-019-1832-y

18 Morel B, Kozlov AM, Stamatakis A. (2018). ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. https://doi.org/10.1093/bioinformatics/bty839

19 Lozano-Fernandez J. (2022). A Practical Guide to Design and Assess a Phylogenomic Study. https://doi.org/10.1093/gbe/evac129

20 Celebi FM, McDaniel EA, Reiter T. (2024). Creating reproducible workflows for complex computational pipelines. https://doi.org/10.57844/ARCADIA-CC5J-A519

21 Burki F, Roger AJ, Brown MW, Simpson AGB. (2020). The New Tree of Eukaryotes. https://doi.org/10.1016/j.tree.2019.08.008

22 Strassert JFH, Irisarri I, Williams TA, Burki F. (2021). A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. https://doi.org/10.1038/s41467-021-22044-z

23 Grattepanche J, Walker LM, Ott BM, Paim Pinto DL, Delwiche CF, Lane CE, Katz LA. (2018). Microbial Diversity in the Eukaryotic SAR Clade: Illuminating the Darkness Between Morphology and Molecular Data. https://doi.org/10.1002/bies.201700198

24 Shalchian-Tabrizi K, Eikrem W, Klaveness D, Vaulot D, Minge MA, Le Gall F, Romari K, Throndsen J, Botnen A, Massana R, Thomsen HA, Jakobsen KS. (2006). Telonemia, a new protist phylum with affinity to chromist lineages. https://doi.org/10.1098/rspb.2006.3515

25 Strassert JFH, Jamy M, Mylnikov AP, Tikhonenkov DV, Burki F. (2019). New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life. https://doi.org/10.1093/molbev/msz012

26 Derelle R, López-García P, Timpano H, Moreira D. (2016). A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). https://doi.org/10.1093/molbev/msw168

27 Schaechter, M. (2009). Encyclopedia of Microbiology. https://books.google.com/books?id=TvbbzwEACAAJ

28 Tikhonenkov DV, Strassert JFH, Janouškovec J, Mylnikov AP, Aleoshin VV, Burki F, Keeling PJ. (2020). Predatory colponemids are the sister group to all other alveolates. https://doi.org/10.1016/j.ympev.2020.106839

29 Burki F, Keeling PJ. (2014). Rhizaria. https://doi.org/10.1016/j.cub.2013.12.025

30 Cavalier-Smith T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. https://doi.org/10.1099/00207713-52-2-297

31 Manni M, Berkeley MR, Seppey M, Zdobnov EM. (2021). BUSCO: Assessing Genomic Data Quality and Beyond. https://doi.org/10.1002/cpz1.323

32 Stamboulian M, Guerrero RF, Hahn MW, Radivojac P. (2020). The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. https://doi.org/10.1093/bioinformatics/btaa468

33 Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, Wideman JG, Burki F, de Vargas C. (2020). EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. https://doi.org/10.1101/2020.06.30.180687

34 Richter D, Berney C, Strassert J, Poh Y-P, Herman EK, Muñoz-Gómez SA, Wideman JG, Burki F, de Vargas C. (2022). EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. https://doi.org/10.6084/M9.FIGSHARE.12417881.V3

35 Shen W, Le S, Li Y, Hu F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. https://doi.org/10.1371/journal.pone.0163962

36 Charif D, Lobry JR. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. https://doi.org/10.1007/978-3-540-35306-5_10

37 R Core Team (2021). R: A language and environment for statistical computing. https://www.R-project.org/

38 Fu L, Niu B, Zhu Z, Wu S, Li W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. https://doi.org/10.1093/bioinformatics/bts565

39 Cortesi F, Musilová Z, Stieb SM, Hart NS, Siebeck UE, Malmstrøm M, Tørresen OK, Jentoft S, Cheney KL, Marshall NJ, Carleton KL, Salzburger W. (2014). Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes. https://doi.org/10.1073/pnas.1417803112

40 Uyeda JC, Zenil-Ferguson R, Pennell MW. (2018). Rethinking phylogenetic comparative methods. https://doi.org/10.1093/sysbio/syy031

41 Cokelaer T, Pultz D, Harder LM, Serra-Musach J, Saez-Rodriguez J. (2013). BioServices: a common Python package to access biological Web Services programmatically. https://doi.org/10.1093/bioinformatics/btt547

42 Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM. (2022). OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. https://doi.org/10.1093/nar/gkac998

43 Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. (2018). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated

orthology resource based on 5090 organisms and 2502 viruses. https://doi.org/10.1093/nar/gky1085

44 Buchfink B, Reuter K, Drost H-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. https://doi.org/10.1038/s41592-021-01101-x

45 Emms DM, Kelly S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. https://doi.org/10.1186/s13059-015-0721-2

46 Almeida-Silva F, Van de Peer Y. (2023). Assessing the quality of comparative genomics data and results with the*cogeqc*R/Bioconductor package. https://doi.org/10.1101/2023.04.14.536860

47 Shen C, Park M, Warnow T. (2022). WITCH: Improved Multiple Sequence Alignment Through Weighted Consensus Hidden Markov Model Alignment. https://doi.org/10.1089/cmb.2021.0585

48 Katoh K, Standley DM. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. https://doi.org/10.1093/molbev/mst010

49 Tumescheit C, Firth AE, Brown K. (2022). CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. https://doi.org/10.7717/peerj.12983

50 Steenwyk JL, Buida TJ, Li Y, Shen X-X, Rokas A. (2020). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. https://doi.org/10.1371/journal.pbio.3001007

51 Price MN, Dehal PS, Arkin AP. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. https://doi.org/10.1371/journal.pone.0009490

52 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. https://doi.org/10.1093/molbev/msaa015

53 Wang H-C, Minh BQ, Susko E, Roger AJ. (2017). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. https://doi.org/10.1093/sysbio/syx068

54 Morel B, Williams TA, Stamatakis A. (2022). Asteroid: a new algorithm to infer species trees from gene trees under high proportions of missing data. https://doi.org/10.1093/bioinformatics/btac832

55 Willson J, Roddur MS, Liu B, Zaharias P, Warnow T. (2021). DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. https://doi.org/10.1093/sysbio/syab070

56 Revell LJ. (2011). phytools: an R package for phylogenetic comparative biology (and other things). https://doi.org/10.1111/j.2041-210x.2011.00169.x

57 Morel B, Schade P, Lutteropp S, Williams TA, Szöllősi GJ, Stamatakis A. (2022). SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. https://doi.org/10.1093/molbev/msab365

58 Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. (2020). GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. https://doi.org/10.1093/molbev/msaa141

59 Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. (2013). Efficient Exploration of the Space of Reconciled Gene Trees. https://doi.org/10.1093/sysbio/syt054

60 Mendes FK, Vanderpool D, Fulton B, Hahn MW. (2020). CAFE 5 models variation in evolutionary rates among gene families. https://doi.org/10.1093/bioinformatics/btaa1022

61 Nagy LG, Ohm RA, Kovács GM, Floudas D, Riley R, Gácser A, Sipiczki M, Davis JM, Doty SL, de Hoog GS, Lang BF, Spatafora JW, Martin FM, Grigoriev IV, Hibbett DS. (2014). Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. https://doi.org/10.1038/ncomms5471

62 Penel S, Menet H, Tricou T, Daubin V, Tannier E. (2022). Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation. https://doi.org/10.1093/bioinformatics/btac062