# PhyKIT: A Multitool for Phylogenomics

Jacob L. Steenwyk,[1,13] Gemma I. Martínez-Redondo,[2] Thomas J. Buida III,[3] Emile Gluck-Thaler,[4] Xing-Xing Shen,[5] Toni Gabaldón,[6,7,8,9,12] Antonis Rokas,[10,11,12] and Rosa Fernández[2,12]

[1]Howards Hughes Medical Institute and the Department of Molecular and Cell Biology, University of California, Berkeley, California
[2]Metazoa Phylogenomics Lab, Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain
[3]Independent Researcher, Nashville, Tennessee
[4]Department of Plant Pathology, University of Wisconsin-Madison, Madison, Wisconsin
[5]College of Agriculture and Biotechnology, Centre for Evolutionary & Organismal Biology, Zhejiang University, Hangzhou, China
[6]Barcelona Supercomputing Centre (BSC-CNS), Plaça Eusebi Güell, Barcelona, Spain
[7]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain
[8]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
[9]CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain
[10]Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee
[11]Evolutionary Studies Initiative, Vanderbilt University, Nashville, Tennessee
[12]These authors contributed equally to this work.
[13]Corresponding author: *jlsteenwyk@berkeley.edu*

Published in the Bioinformatics section

Multiple sequence alignments and phylogenetic trees are rich in biological information and are fundamental to research in biology. PhyKIT is a tool for processing and analyzing the information content of multiple sequence alignments and phylogenetic trees. Here, we describe how to use PhyKIT for diverse analyses, including (i) constructing a phylogenomic supermatrix, (ii) detecting errors in orthology inference, (iii) quantifying biases in phylogenomic data sets, (iv) identifying radiation events or lack of resolution using gene support frequencies, and (v) conducting evolution-based screens to facilitate gene function prediction. Several PhyKIT functions that streamline multiple sequence alignment and phylogenetic processing—such as renaming FASTA entries or tree tips—are also discussed. These protocols demonstrate how simple command-line operations in the unified framework of PhyKIT facilitate diverse phylogenomic data analysis and processing, from supermatrix construction and diagnosis to gaining clues about gene function. © 2024 The Author(s). Current Protocols published by Wiley Periodicals LLC.

**Basic Protocol 1:** Installing PhyKIT and syntax for usage
**Basic Protocol 2:** Constructing a phylogenomic supermatrix
**Basic Protocol 3:** Detecting anomalies in orthology relationships
**Basic Protocol 4:** Quantifying biases in phylogenomic data matrices and related measures
**Basic Protocol 5:** Identifying polytomies
**Basic Protocol 6:** Assessing gene-gene coevolution as a genetic screen

Keywords: comparative genomics • genomics • phylogeny • phylogenomics • sequence alignments • software tools

CURRENT PROTOCOLS
A Wiley Brand

*Current Protocols* e70016, Volume 4
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpz1.70016

**How to cite this article:**
Steenwyk, J. L., Martínez-Redondo, G. I., Buida, T. J., Gluck-Thaler, E., Shen, X.-X., Gabaldón, T., Rokas, A., & Fernández, R. (2024). PhyKIT: A multitool for phylogenomics. *Current Protocols*, *4,* e70016. doi: 10.1002/cpz1.70016

## INTRODUCTION

Phylogenomics combines principles and methodologies in phylogenetics, genomics, and bioinformatics to understand the evolutionary relationships and functions of genome-scale nucleotide and amino acid sequences (Delsuc et al., 2005; Eisen, 1998). Phylogenomic analyses can also help uncover the tempo and modes of evolution across lineages, facilitate accurate identification of taxa, and serve as the backbone for downstream comparative studies (Green et al., 2014; Jarvis et al., 2014; Sierra-Patev et al., 2023; Steenwyk et al., 2023b; Steenwyk et al., 2024; Thornton & DeSalle, 2000).

Two data types are fundamental to phylogenomic analysis: multiple sequence alignments, wherein putative site-wise homologies are represented as columns, and phylogenetic trees, which are diagrams representing evolutionary relationships with branch-length information frequently represented as time or evolutionary rate. Efficient processing and analysis of these data types is key for phylogenomics.

PhyKIT is a software package with multiple functions for high-throughput exploration of multiple sequence alignments and phylogenetic trees in phylogenomic datasets (Fig. 1 and Table 1; Steenwyk et al., 2021). This article introduces the basics of processing phylogenomic datasets with PhyKIT, followed by more complex analyses ranging from diagnosing errors and biases in phylogenomic data matrices to identifying radiation events and conducting evolution-based screens to facilitate the prediction of gene function. Although this article covers frequently used PhyKIT functions, explanations for all 47 utilities are available in the online documentation (*https://jlsteenwyk.com/PhyKIT*). While these protocols demonstrate how PhyKIT functions democratize the processing and analysis of phylogenomic data matrices, they also underscore how complex phylogenomic analyses can be easily done using a unified toolkit.

### Operation System Requirements

Access to a machine with Unix, Linux, Apple OS X, or Windows operating system is required.

### Conventions

In this article, PhyKIT usage are depicted as if working in the Unix environment. Unix commands are in Courier font, with the `$` character indicating the command line. Comments, set off by the # character, are used to describe the command being executed and associated output.

### Background Knowledge

Previous experience with the Unix command line is assumed. Familiarity with FASTA and Newick file formats—used for multiple sequence alignments and phylogenetic trees, respectively—is also required. FASTA and Newick file formats are the main inputs and outputs of PhyKIT. Files not in these formats can be converted to FASTA or Newick format using software such as the sibling toolkit BioKIT (Steenwyk et al., 2022a) or other utilities such as biopython (Cock et al., 2009) and web portals such as phylogeny.fr (*http://phylogeny.lirmm.fr/phylo_cgi/data_converter.cgi*).

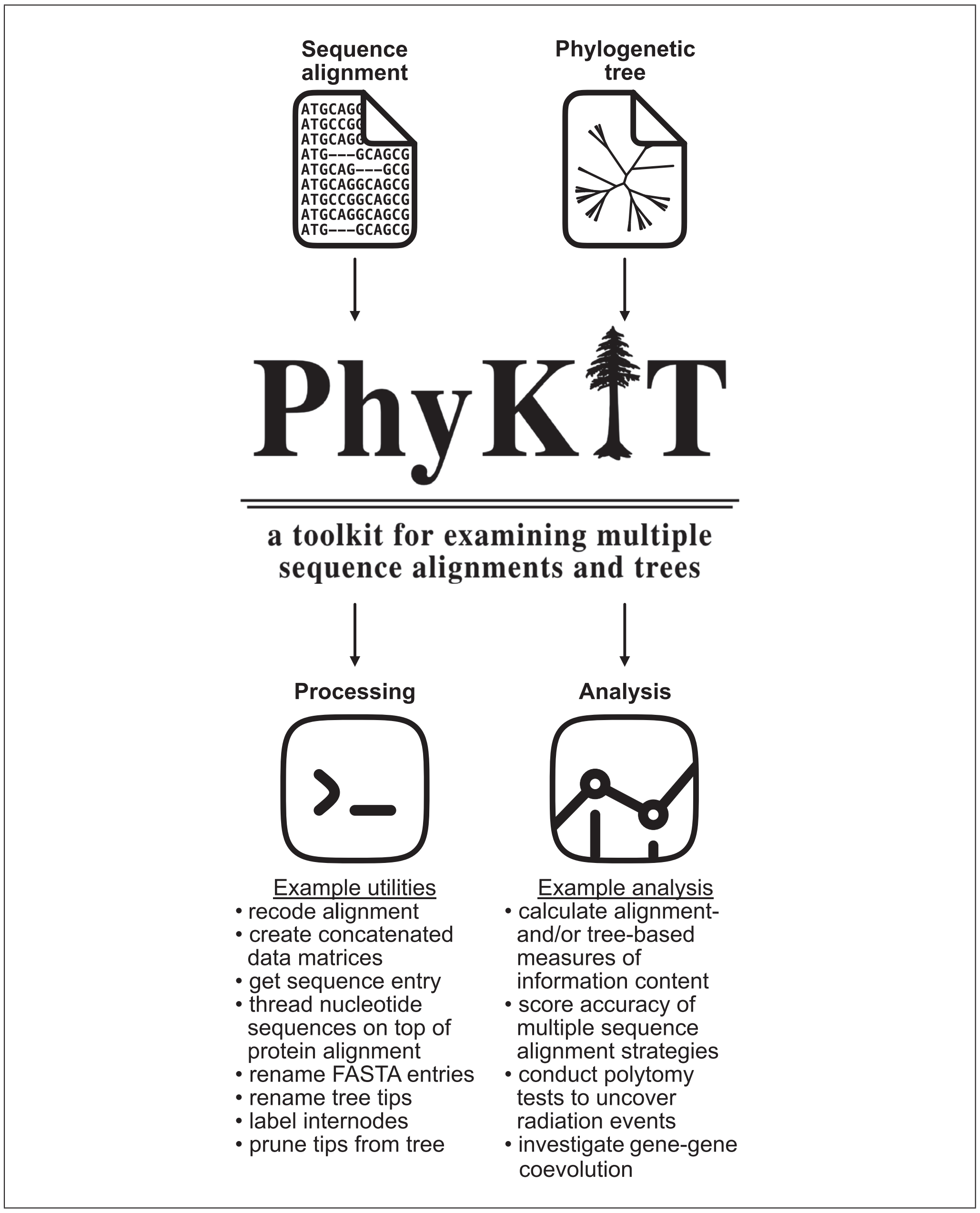

Sequence alignment
ATGCAGG
ATGCCGG
ATGCAGG
ATG---GCAGCG
ATGCAG---GCG
ATGCAGGCAGCG
ATGCCGGCAGCG
ATGCAGGCAGCG
ATG---GCAGCG
Phylogenetic tree
PhyKIT
a toolkit for examining multiple sequence alignments and trees
Processing
Analysis
Example utilities
• recode alignment
• create concatenated data matrices
• get sequence entry
• thread nucleotide sequences on top of protein alignment
• rename FASTA entries
• rename tree tips
• label internodes
• prune tips from tree
Example analysis
• calculate alignment- and/or tree-based measures of information content
• score accuracy of multiple sequence alignment strategies
• conduct polytomy tests to uncover radiation events
• investigate gene-gene coevolution

**Figure 1** PhyKIT is a multitool with diverse functions for processing and analyzing phylogenomic data. For example, processing can include recoding alignment, creating concatenated supermatrices, renaming FASTA entries, and more. Exemplary analysis functions include calculating the information content of multiple sequence alignments and phylogenetic trees, and quantifying gene-gene coevolution. Overall, PhyKIT offers diverse features under a unified framework for facilitating and streamlining phylogenomic data processing and analysis.

## INSTALLING PhyKIT AND SYNTAX FOR USAGE

*BASIC PROTOCOL 1*

PhyKIT is freely available under the MIT license via GitHub (*https://github.com/JLSteenwyk/PhyKIT*) and is distributed through multiple repositories, including Python Package Index (PyPI; *https://pypi.org/project/phykit/*) and the Anaconda Cloud (*https://anaconda.org/bioconda/phykit*). As of writing, the latest PhyKIT release is version 1.19.2. The installed version can easily be updated to the most recent release following the protocol for each distribution platform.

Perhaps the easiest way to install PhyKIT is to use package and environment manager programs such as PIP or Conda.

**Table 1** A Complete List of PhyKIT Functions as of Version 1.19.0

| | Function name | Function alias(es) | Description |
|---|---|---|---|
| **Alignment-based functions** | alignment_ length | aln_len; al | Calculates alignment length. |
| | alignment_ length_ no_gaps | aln_len_no_ gaps; alng | Calculates alignment length excluding sites with gaps. |
| | alignment_ recoding | aln_recoding; recode | Recodes alignments using reduced character states. |
| | column_ score | cs | Calculates column score, an accuracy metric for a multiple alignment relative to a reference alignment. |
| | compositional _bias_ per_site | comp_bias_per_ site; cbps | Calculates compositional bias per site in an alignment. Site-wise $\chi^2$ tests are conducted in an alignment to detect compositional biases. |
| | create_ concatenation_ matrix | create_concat; cc | Create a concatenated alignment file. This function is used to help in the construction of multi-locus data matrices. |
| | evolutionary_ rate_per_site | evo_rate_per_ site; erps | Estimates the evolutionary rate per site. Values may range from 0 (slow evolving; no diversity at the given site) to 1 (fast evolving; all characters appear only once). |
| | faidx | get_entry; ge | Extracts sequence entry from FASTA file. |
| | gc_content | gc | Calculates GC content of a FASTA file. |
| | pairwise_ identity | pairwise_ id; pi | Calculates the average pairwise identity among sequences. Pairwise identities can be used as proxies for the evolutionary rate of sequences. |
| | parsimony_ informative_sites | pis | Calculates the number and percentage of parsimony informative sites in an alignment. |
| | relative_ composition_ variability | rel_comp_ var; rcv | Calculates RCV (relative composition variability) for an alignment. Lower RCV values are thought to be desirable because they represent a lower composition bias in an alignment. |
| | relative_ composition_ variability_taxon | rel_comp_ var_taxon; rcvt | Calculates RCVT (relative composition variability, taxon) for an alignment, representing the relative composition variability metric for individual taxa. Lower RCVT values are more desirable because they indicate a lower composition bias for a given taxon in an alignment. |
| | rename_fasta_ entries | rename_fasta | Renames entries in a FASTA file. Note: The input FASTA file does not need to be a multiple sequence alignment. |
| | sum_of_pairs_ score | sops; sop | Calculates sum-of-pairs score, an accuracy metric for a multiple alignment relative to a reference alignment. |
| | thread_dna | pal2nal; p2n | Threads DNA sequence onto a protein alignment to create a codon-based alignment. |
| | variable_sites | vs | Calculates the number and percentage of variable sites in an alignment. |

(*Continued*)

**Table 1** A Complete List of PhyKIT Functions as of Version 1.19.0, *continued*

| | Function name | Function alias(es) | Description |
|---|---|---|---|
| **Tree-based functions** | bipartition_ support_stats | bss | Calculates summary statistics for bipartition support. |
| | branch_length_ multiplier | blm | Multiplies branch lengths in a phylogeny by a given factor. This can help modify reference trees when conducting simulations or other analyses. |
| | collapse_ branches | collapse; cb | Collapses branches on a phylogeny according to bipartition support. Bipartitions will be collapsed if they are less than the user-specified value. |
| | covarying_ evolutionary_ rates | cover | Quantify the degree of coevolution between two single-gene trees. |
| | degree_of_ violation_of_a_ molecular_clock | dvmc | Calculates degree of violation of a molecular clock (or DVMC) in a phylogeny. Lower DVMC values are thought to be desirable because they are indicative of a lower degree of violation in the molecular clock assumption. |
| | evolutionary_ rate | evo_rate | Calculates a tree-based estimation of the evolutionary rate of a gene. |
| | hidden_ paralogy_ check | clan_check | Scans tree for evidence of hidden paralogy. Specifically, this method will examine if a set of well-known monophyletic taxa are, in fact, exclusively monophyletic. |
| | internal_ branch_ stats | ibs | Calculates summary statistics for internal branch lengths in a phylogeny. Internal branch lengths can be useful for phylogeny diagnostics. |
| | internode_ labeler | il | Appends numerical identifiers to bipartitions in place of support values. This is helpful for pointing to specific internodes in supplementary files or otherwise. |
| | last_common_ ancestor_subtree | lca_subtree | Obtains subtree from a phylogeny by getting the last common ancestor from a list of taxa. |
| | long_branch_ score | lb_score; lbs | Calculates long branch scores in a phylogeny. Lower long branch scores are thought to be desirable because they are indicative of taxa or trees that likely do not have issues with long branch attraction. |
| | monophyly_ check | is_monophyletic | Can be used to determine if a set of taxa are exclusively monophyletic. If other taxa are in the same clade, the lineage will not be considered exclusively monophyletic. |
| | nearest_ neighbor_ interchange | nni | Generates all nearest-neighbor interchange moves for a binary rooted tree. |

(*Continued*)

**Table 1** A Complete List of PhyKIT Functions as of Version 1.19.0, *continued*

| | Function name | Function alias(es) | Description |
|---|---|---|---|
| | patristic_ distances | pd | Calculates summary statistics for patristic distances—i.e., all tip-to-tip distances—in a phylogeny. |
| | polytomy_ test | polyt_test; polyt; ptt | Conducts a polytomy test for three clades in a phylogeny. Tests for polytomy (multifurcation) can be used to identify putative radiations as well as to identify well-supported alternative topologies. |
| | print_tree | print; pt | Prints ASCII tree of input phylogeny. |
| | prune_tree | prune | Prunes tips from a phylogeny. |
| | rename_tree | rename_tips | Renames tips in a phylogeny. |
| | robinson_ foulds_ distance | rf_distance; rf_dist; rf | Calculates Robinson-Foulds distance between two trees. Low Robinson-Foulds distances reflect greater similarity between two phylogenies. This function prints out two values, the plain Robinson-Foulds value, and the normalized Robinson-Foulds value, which are separated by a tab. |
| | root_tree | root; rt | Roots phylogeny using user-specified taxa. |
| | spurious_ sequence | spurious_ seq; ss | Identifies potentially spurious sequences and reports tips in the phylogeny that could possibly be removed from the associated multiple sequence alignment. PhyKIT does this by identifying and reporting long terminal branches, defined as branches that are $\geq 20 \times$ the median length of all branches. |
| | terminal_ branch_ stats | tbs | Calculates summary statistics for terminal branch lengths in a phylogeny. |
| | tip_labels | tree_labels; labels; tl | Prints the tip labels (or names) in a phylogeny. |
| | tip_to_tip_ distance | t2t_dist; t2t | Calculates distance between two tips (or leaves) in a phylogeny. |
| | tip_to_tip_ node_distance | t2t_node_dist; t2t_nd | Calculates distance between two tips (or leaves) in a phylogeny, measured by the number of nodes between one tip and another. |
| | total_tree_ length | tree_len | Calculates total tree length, which is a sum of all branches. |
| | treeness | tness | Calculates treeness statistic for a phylogeny. Higher treeness values are thought to be desirable because they represent a higher signal-to-noise ratio. |
| **Alignment- and tree-based functions** | saturation | sat | Calculates saturation for a given tree and alignment, defined as sequences in multiple sequence alignments that have undergone numerous substitutions such that the distances between taxa are underestimated. |
| | treeness_ over_rcv | toverr; tor | Calculates treeness/RCV for a given alignment and tree. Higher treeness/RCV values are thought to be desirable because they reflect a high signal-to-noise ratio and are least susceptible to composition bias. |

## Installing PhyKIT

### *Install using PIP (Preferred Installer Program)*

```
# install
$ pip install phykit
```

### *Install using Conda*

```
# install
$ conda install bioconda::phykit
```

### *Install from source*

```
# download
$ git clone https://github.com/JLSteenwyk/PhyKIT.git
$ cd PhyKIT/
# install
$ make install
```

#### *Install in a virtual environment*

It may also be useful to install PhyKIT in a separate virtual environment. These isolated environments help separate variable dependencies required by different software, overcoming conflicting dependencies between software packages (e.g., one software may require an older version of NumPy; Harris et al., 2020). PhyKIT is engineered to have relatively few dependencies—Biopython, NumPy, SciPy, and Cython (Behnel et al., 2011; Cock et al., 2009; Harris et al., 2020; Virtanen et al., 2020)—to help ensure ease of compatibility with other software and long-term stability. One approach to managing virtual environments is to have one environment per large project or substantial portion of a large project. Virtual environments may be stored in the same directory as the project. Ultimately, users must find a system that works best for them. The following are examples of how to install PhyKIT using a virtual environment.

#### *Install in a virtual environment using PIP*

```
# create a virtual environment
$ python -m venv venv_phykit
# activate the virtual environment
$ source venv_phykit/bin/activate
# install
$ pip install PhyKIT
```

After using software in your virtual environment, you may wish to deactivate (or exit) the virtual environment.

```
# deactivate virtual environment
$ deactivate
```

#### *Install in a virtual environment using Conda*

```
# create a virtual environment
$ conda create -n venv_phykit
# activate the virtual environment
$ conda activate venv_phykit
# install
$ conda install -n venv_phykit bioconda::phykit
# deactivate environment when you are done using PhyKIT
$ conda deactivate
```

PIP and Conda both provide easy ways to automatically check for new software releases and install them if available.

*Update installation using PIP*

```
# The “-U” is short for “--upgrade”
$ pip install phykit -U
```

*Update installation using Conda*

```
# This is the same command as to install
$ conda install -n venv_phykit bioconda::phykit
```

*Activate environment before using PhyKIT*

If PhyKIT has been installed in a virtual environment, it must be activated when using PhyKIT.

```
# If installed using PIP, activate the environment
$ source venv_phykit/bin/activate
# If installed using Conda, activate the environment
$ conda activate venv_phykit
```

**PhyKIT Help Menu and PhyKIT Functions**

The PhyKIT Help Menu is accessed using the following command:

```
$ phykit -h
```

If PhyKIT has been successfully installed, this command should return the list of available functions. If installation was unsuccessful, an error message may appear, and users are encouraged to revisit the installation instructions or contact the developers (*https://github.com/JLSteenwyk or https://jlsteenwyk.com/contact.html*).

PhyKIT functions are organized based on the input file type. Specifically, some functions take multiple sequence alignments as input, others take as input phylogenetic trees, and some functions take both. See Table 1 for a complete list of PhyKIT functions.

PhyKIT functions that take multiple sequence alignment files as input include:

- alignment_length: calculates alignment length
- gc_content: calculate GC content of a nucleotide FASTA entries or entries thereof
- pairwise_identity: calculates average pairwise identify among sequences in an alignment file—a proxy for evolutionary rate
- relative_composition_variability: calculates relative composition variability in an alignment

A complete list of alignment-based functions, including detailed explanations of each, is available here: *https://jlsteenwyk.com/PhyKIT/usage/index.html#alignment-based-functions*.

PhyKIT functions that take phylogenetic trees as input include:

- bipartition_support_stats: calculates summary statistics for bipartition support
- degree_of_violation_of_a_molecular_clock: reports the degree of violation of the molecular clock assumption that sequences evolve at a relatively constant rate over time and among different organisms
- evolutionary_rate: reports a tree-based estimation of evolutionary rate for a gene
- prune_tree: prune taxa from a phylogeny

A complete list of tree-based functions, including detailed explanations of each, is available here: *https://jlsteenwyk.com/PhyKIT/usage/index.html#tree-based-functions*.

PhyKIT functions that take multiple sequence alignments and phylogenetic trees as input include:

- saturation: calculates saturation by examining the slope of patristic distances and uncorrected distances;
- treeness_over_rcv: calculates treeness divided by relative composition variability (rcv), treeness, and rcv; and

A complete list of alignment- and tree-based functions, including detailed explanations of each, is available here: *https://jlsteenwyk.com/PhyKIT/usage/index.html#alignment-and-tree-based-functions*.

The help message has abbreviated descriptions of each function. Users are encouraged to read the corresponding section in the documentation (*https://jlsteenwyk.com/PhyKIT*) for more details about each function. Alternatively, users can print out the help message for specific functions. For example, to see the help message for the "alignment_length" function, which provides a much more detailed description of the function and its utility, requires executing the following command:

```
# Call the help message of a specific function
$ phykit alignment_length -h
```

Throughout the manuscript, we also provide links to relevant documentation of each PhyKIT function, allowing easy access to additional details and explanations of PhyKIT functionality.

## PhyKIT Syntax

Although this will be expanded upon in later sections, the syntax of using PhyKIT functions is that users first call PhyKIT, then specify the function to be executed, and next specify the arguments required for the function.

```
# Description of PhyKIT syntax
$ phykit <command> <arguments> [optional arguments]
# Note, optional arguments will always have square brackets around them
```

For example, the following command demonstrates how to calculate the length of a multiple sequence alignment:

```
# Calculate alignment length
$ phykit alignment_length input.fa
```

### *Function aliases*

Every PhyKIT function also has aliases, which are abbreviations of function names. For example, the "alignment_length" function can also be called using "aln_len" or "al", as exemplified in the following commands:

```
# Calculate alignment length using the aln_len alias
$ phykit aln_len input.fa
# Calculate alignment length using the al alias
$ phykit al input.fa
```

### *Shorthand syntax*

To accelerate executing PhyKIT functions, an alternative, shorthand syntax is also available. Specifically, PhyKIT and associated functions can be called by combining the prefix "pk_" (shorthand for phykit) with the function name or alias. For example, the shorthand syntax can be used to calculate the length of an alignment using the following command:

```
# Calculate alignment length using the full function name
$ pk_alignment_length input.fa
# Calculate alignment length using the aln_len alias
$ pk_aln_len input.fa
# Calculate alignment length using the al alias
$ pk_al input.fa
```

A benefit of the shorthand syntax is the ability to use command-line completion (or tab completion). For example, executing command-line completion after typing `pk_` will display all PhyKIT commands and their aliases.

Throughout the remainder of the manuscript, other software tools that facilitate phylogenomic workflows are mentioned. Although aspects of these tools are covered, users are encouraged to read the respective installation instructions and documentation for more information.

### *Summary statistics and the verbose option*

Numerous functions report summary statistics. Summary statistics include mean, median, 25th percentile, 75th percentile, minimum, maximum, standard deviation, and variance. For these functions, verbose options (evoked using `-v/--verbose`) allow every value that generates the underlying values to be reported instead.

### *Requesting new functions*

Protocols described herein highlight how PhyKIT serves as a multitool for phylogenomic data processing and analysis. Moreover, PhyKIT is actively undergoing development to serve the research community better, resulting in new functions and additional utilities to existing functions. Thus, users are encouraged to read the PhyKIT documentation for a complete list of functions (*https://jlsteenwyk.com/PhyKIT*).

If PhyKIT lacks a function that users would want to have, we welcome requests and recommend either contacting the developers via email (*https://jlsteenwyk.com/contact.html*) or opening a GitHub issue (*https://github.com/JLSteenwyk/PhyKIT/issues*) with a feature request.

***BASIC PROTOCOL 2***

## CONSTRUCTING A PHYLOGENOMIC SUPERMATRIX

Concatenation and the multispecies coalescence are two popular methods for species tree inference (Gatesy et al., 2017; Philippe et al., 2017; Steenwyk et al., 2023a). The concatenation approach requires generating multiple sequence alignments among orthologous loci and combining them into a supermatrix, which is then used to reconstruct evolutionary relationships using a maximum-likelihood or Bayesian framework (Kapli et al., 2020; Philippe et al., 2017; Steenwyk et al., 2023a). In contrast, in a popular coalescence-based approach ("two-step" coalescence), first single-locus phylogenies are inferred and then the resulting set of trees is used to reconstruct organismal history using, for example, quartet graph construction (Han & Molloy, 2023; Mirarab et al., 2014).

### *Data acquisition*

Assembled genomes are publicly available from online databases such as GenBank and RefSeq from the National Center for Biotechnology Information (*https://www.ncbi.nlm.nih.gov*) and the Universal Protein Resource (UniProt; *https://www.uniprot.org*). Genomes and transcriptomes can also be assembled de novo using sequencing data available from the Sequence Read Archive (SRA) hosted by the National Center for Biotechnology Information (*https://www.ncbi.nlm.nih.gov/sra*) or the European Nucleotide Archive (*https://www.ebi.ac.uk/ena*) hosted by the European Bioinformatics Institute. Data can also be downloaded from dedicated databases such as MolluscDB and MATEdb/2 (Caurcel et al., 2021; Fernández et al., 2022; Martínez-Redondo et al., 2024), among others. Detailed instructions for genome assembly, quality control, and orthology inference are beyond the scope of this manuscript; instead, we cite relevant protocols and briefly describe key steps (Coombe et al., 2023; Manni et al., 2021; Raghavan et al., 2022; Zhao et al., 2023).

### *Orthology inference*

Next, orthologs are predicted, typically from proteome sequences. *De novo* orthology can be inferred using, for example, OrthoFinder (Emms & Kelly, 2019). The resulting output can be parsed for single-copy genes. Additional single-copy orthologs nested within larger multi-copy gene families can be identified using OrthoSNAP (Steenwyk et al., 2022b). Although methods have been developed to reconstruct evolutionary relationships from gene families with paralogous genes, we focus on phylogenomics using traditional single-copy orthologs and single-copy orthologs identified by OrthoSNAP (Steenwyk et al., 2022b; Zhang et al., 2020). Alternatively, single-copy orthologs can be identified from predetermined single-copy orthologs using tools like BUSCO and orthofisher (Steenwyk & Rokas, 2021b; Waterhouse et al., 2018).

### *Multiple sequence alignment and trimming*

Next, single-copy orthologs are aligned and trimmed. During alignment, site-wise homologies are identified across multiple sequences (Kapli et al., 2020; Steenwyk et al., 2023a). Software packages that infer site-wise homologies include MAFFT, MUSCLE5, or Clustal-Omega (Edgar, 2022; Katoh & Standley, 2013; Sievers & Higgins, 2018) and can be executed using any of the following commands:

```
# Alignment with MAFFT
$ mafft --auto input.fa > output.fa
# Alignment with MUSCLE5
$ muscle5 -align input.fa -output output.fa
# Alignment with Clustal-Omega
$ clustalo -i input.fa -o output.fa
```

Subsequently, multiple sequence alignments can be trimmed using ClipKIT or trimAl (Capella-Gutiérrez et al., 2009; Steenwyk et al., 2020). Although ClipKIT can be run with default parameters, recent benchmarking studies have demonstrated that all ClipKIT modes perform well (Steenwyk et al., 2020). Here, we demonstrate ClipKIT usage with default parameters. If using trimAl, the same benchmarking studies suggest using the "gappyout" parameter (Tan et al., 2015).

```
# Trimming with ClipKIT
$ clipkit input.fa -o output.fa
# Trimming with trimAl
$ trimal -in input.fa -out output.fa -gappyout
```

Users may want to create codon-based alignments for phylogenomic inference from DNA sequences. To do so, nucleotide sequences must be threaded on top of protein sequence alignment (Fig. 2). The PhyKIT function "thread_dna" (aliases: "pal2nal", "p2n"; *https://jlsteenwyk.com/PhyKIT/usage/index.html#protein-to-nucleotide-alignment*) can achieve this. The input files are the multiple sequence alignment protein and the unaligned corresponding nucleotide sequences. To ensure the correct matching between protein and nucleotide sequences, sequences must be named identically in both sequence files. The thread_dna function is executed using the following command:

```
# Thread nucleotide sequences onto a protein alignment
$ pk_thread_dna -p protein_alignment.faa -n nucleotide_sequences.fna
```

It is also common to first trim a multiple sequence alignment and then thread the resulting nucleotide sequence onto the trimmed protein alignment. PhyKIT can do so using the log file outputted from ClipKIT. To output the ClipKIT log file, execute the following command, and then use it as an argument to PhyKIT:

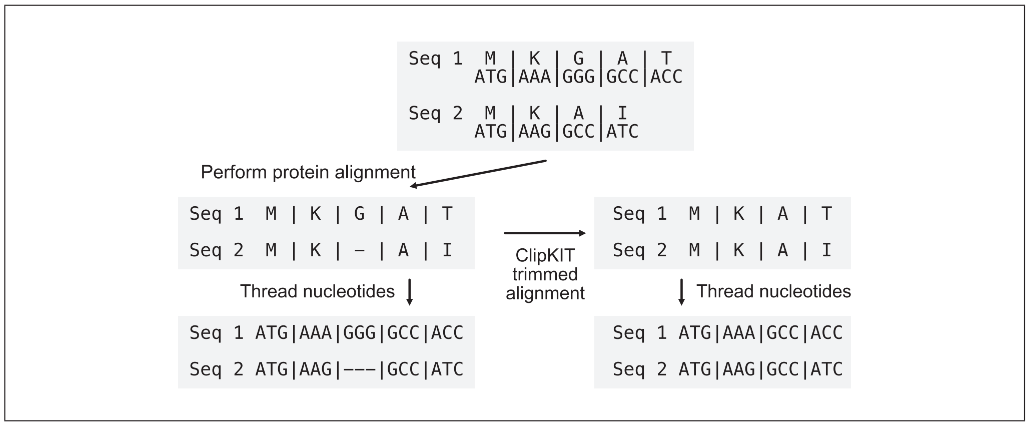


**Figure 2** Generating codon-based alignments. The “thread_dna function” facilitates the threading nucleotide sequences on top of protein alignments, generating codon-based multiple sequence alignments. The protein alignment can be untrimmed or ClipKIT trimmed alignment (Steenwyk et al., 2020). If trimmed, a ClipKIT log file is required as an additional input argument.

**Table 2** Recoding Schemes Available in PhyKIT

| Recoding scheme name | Nucleotides or amino acids | Description | Reference |
|---|---|---|---|
| RY-nucleotide | Nucleotides | Two characters, one for purines and another for pyrimidines | Phillips et al. (2004) |
| SandR-6 | Amino acids | Six characters; based on JTT substitution matrix | Susko & Roger (2007) |
| KGB-6 | Amino acids | Six characters; based on WAG substitution matrix | Kosiol et al. (2004) |
| Dayhoff-6 | Amino acids | Six characters; based on Dayhoff (or PAM250) matrix | Embley et al. (2003) |
| Dayhoff-9 | Amino acids | Nine characters; based on Dayhoff (or PAM250) matrix | Hernandez & Ryan (2021) |
| Dayhoff-12 | Amino acids | Twelve characters; based on Dayhoff (or PAM250) matrix | Hernandez & Ryan (2021) |
| Dayhoff-15 | Amino acids | Fifteen characters; based on Dayhoff (or PAM250) matrix | Hernandez & Ryan (2021) |
| Dayhoff-18 | Amino acids | Eighteen characters; based on Dayhoff (or PAM250) matrix | Hernandez & Ryan (2021) |
| <file path> | Either | Custom recoding scheme specified using a two-column file, in which the first column is the recoded character and the second is the character in the alignment | NA |

```
# Thread nucleotide sequences onto a trimmed protein alignment
$ clipkit output.fa -o output.fa --log
# Thread nucleotide sequences onto a trimmed protein alignment
$ pk_thread_dna -p protein_alignment.faa -n nucleotide_sequences.fna -l clipkit.log
```

Alignment recoding—the practice of recoding amino acids and nucleotides to reduced characters sets—can at times combat issues associated with long branch attraction and saturation by multiple substitutions (Foster et al., 2022; Giacomelli et al., 2022; Hernandez & Ryan, 2021). PhyKIT can recode alignments using eight different recoding schemes (Table 2) using the “alignment_recoding” function (aliases: “aln_recoding”,

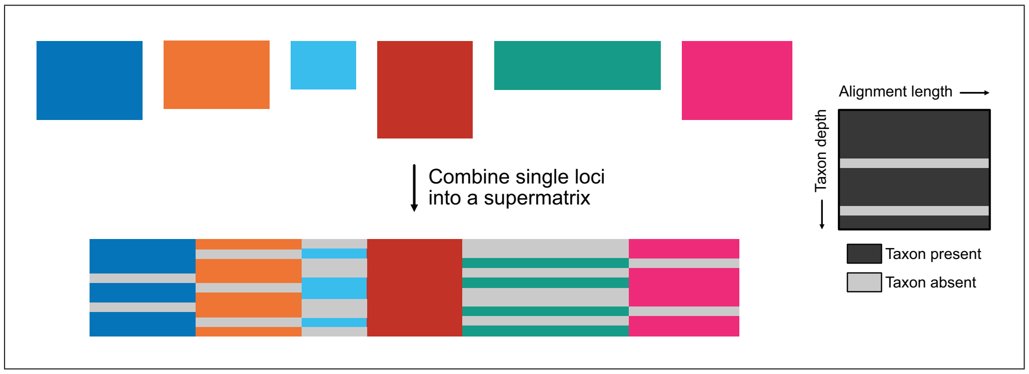


**Figure 3** Creating a phylogenomic supermatrix. The "create_concat" function generates a concatenated supermatrix from individual multiple sequence alignments of single genes in FASTA format for phylogenomic analyses. Also generated is a file summarizing taxon occupancy information for each gene, as well as partition files summarizing the gene boundaries in the concatenation matrix (with a key at right defining how the information is presented).

"recode"; *https://jlsteenwyk.com/PhyKIT/usage/index.html#alignment-recoding*). Alternatively, users can recode alignments using custom character schemes.

```
# Recode alignments
$ pk_aln_recoding input.fa -c <recoding scheme>
# <recoding scheme> can either be one of the eight available
# coding schemes or a file that has the custom coding scheme
```

Custom recoding schemes should be specified using a two-column file. The first column is the recoded character and the second is the current character in the alignment. For example, the recoding scheme for converting the four nucleotides into a two-character scheme would be as follows:

```
# Specify custom recoding scheme
$ cat custom_recoding_scheme.txt
 R       A
 R       G
 Y       T
 Y       C
```

### *Constructing a concatenated supermatrix*

The resulting trimmed multiple sequence alignments can be combined into a supermatrix using the PhyKIT function "create_concat" (Fig. 3; *https://jlsteenwyk.com/PhyKIT/usage/index.html#create-concatenation-matrix*):

```
# Create concatenation matrix
$ pk_create_concat -a alignment_list.txt -p output_prefix
```

create_concat requires two arguments. One argument, "`-a`", is a single-column file with the (absolute or relative) paths to the alignments that will be concatenated. Relative or absolute file paths should be included in the file. An example of how the `alignment_list.txt` should be formatted is as follows:

```
# First five lines of the alignment_list.txt file
$ head -n 5 alignment_list.txt
Alignment0.fa
Alignment1.fa
Alignment2.fa
Alignment3.fa
Alignment4.fa
```

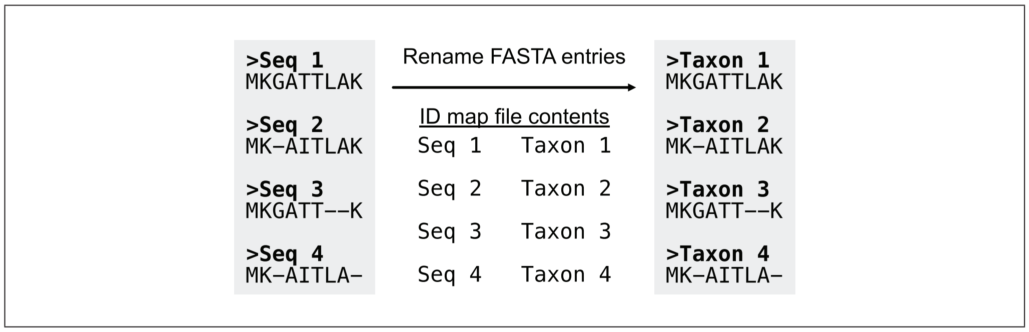


**Figure 4** Renaming entry headers in a FASTA file. Taking as input a FASTA file and an identifier map file, the PhyKIT function "rename_fasta" can rename entries in a FASTA file.

The other argument, "`-p`", will be used as a prefix for the output files. PhyKIT will generate three output files:

- `output_prefix.fa`: the concatenated supermatrix
- `output_prefix.partition`: a description of partition boundaries in RAxML-style format
- `output_prefix.occupancy`: a description of taxon occupancy per partition, including a detailed list of which taxa are present or missing

A common error users experience is when one organism is represented by multiple strings across the alignments specified in `alignment_list.txt.` This may be due to, for example, gene identifiers being included in the FASTA header. PhyKIT requires the same organism to be represented by the same string in the FASTA header. In this way, PhyKIT can determine which sequences belong to the same organism and should therefore be concatenated together. If required, the function "rename_fasta_entries" (alias: "rename_fasta") can rename FASTA entry names (Fig. 4; *https://jlsteenwyk.com/PhyKIT/usage/index.html#rename-fasta-entries*). To do so, a two-column file referred to as an "identifier map", or "ID-map", file must be provided. The ID-map file has two tab-delimited columns: the first column is the current FASTA entry name, and the second column is the new FASTA entry name in the resulting output alignment. For example, the file could be formatted as follows:

```
# First five lines of the idmap file for renaming FASTA entries
$ head -n 5 idmap.txt
speciesA|gene043    speciesA
speciesB|gene367    speciesB
speciesC|gene589    speciesC
speciesD|gene251    speciesD
speciesE|gene417    speciesE
```

Subsequently, PhyKIT can be used to rename the FASTA entries using the following command:

```
# Renaming FASTA entries
$ pk_rename_fasta input.fa -i idmap.txt [-o output.fa]
```

Once the concatenated gene matrix has been generated, the resulting file can be used as input to software that reconstructs evolutionary histories using maximum-likelihood or Bayesian frameworks. The partition file can be used if separate evolutionary models will be used for each locus partition.

### *Constructing a dataset for two-step coalescence*

Popular coalescence-based software requires a single file populated with single-locus phylogenies as input. To generate this file, the best-fitting substitution model for the multiple sequence alignment as well as the associated single-locus phylogenetic tree must first be inferred. Two popular software tools for these steps are ModelTest-NG and

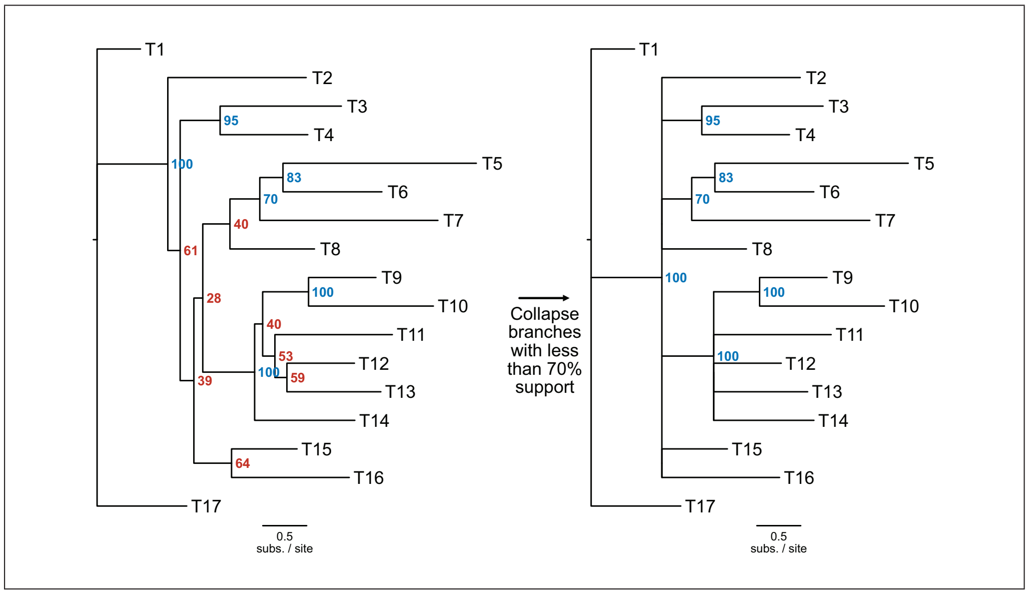


**Figure 5** Collapsing poorly supported branches in a phylogeny. Poorly supported branches in a phylogenetic tree can be collapsed using the "collapse_branches" function. In this example, branches with less than 70% support shown in the tree on the left (in red) were collapsed, resulting in the tree on the right.

ModelFinder within IQ-TREE (Darriba et al., 2020; Kalyaanamoorthy et al., 2017; Minh et al., 2020), which can be executed using either of the following commands:

```
# ModelTest-NG
$ modeltest-ng -i input.fa -d aa
# IQ-TREE
$ iqtree -s input.fa -m MF
```

IQ-TREE can also be executed to resemble jModelTest and ProtTest (Darriba et al., 2012, 2011) by changing `-m MF` to `-m TESTONLY`.

Subsequently, RAxML-NG or IQ-TREE can infer the single-locus phylogeny using the best-fitting substitution model (Kozlov et al., 2019; Minh et al., 2020), and bipartition support can be examined using 100 bootstrap replicates or 1,000 ultrafast bootstrap approximations, respectively, using the following commands:

```
# ModelTest-NG
$ raxml-ng --msa prot21.fa --model LG+G4 --prefix output_prefix --bs-trees 100
# IQ-TREE
$ iqtree -s example.phy -m LG+G4 -bb 1000
# Note: LG+G4 should be replaced by the best-fitting substitution model.
```

To account for uncertainty in single-locus phylogenies, bipartitions with low support can be collapsed using the PhyKIT function "collapse_branches" (Fig. 5); aliases: "collapse", "cb"; *https://jlsteenwyk.com/PhyKIT/usage/index.html#collapse-bipartitions*). To use this function, users must provide the path to the tree file and define a threshold for collapsing bipartitions, which is specified using the `-s/--support` argument.

```
# Collapse branches with bipartition support values less than 75
$ pk_collapse input.tre -s 75 [-o output.tre]
```

The resulting single-locus phylogenies can be combined into a single file using the "cat" function. However, software that takes the resulting file as input—such as ASTRAL, Asteroid, or TREE-QMC—typically requires the names of the same organism to be represented by the same string in each tree file. If needed, the PhyKIT function "rename_tree_tips" (alias: "rename_tips") can rename the leaves in a phylogenetic tree using

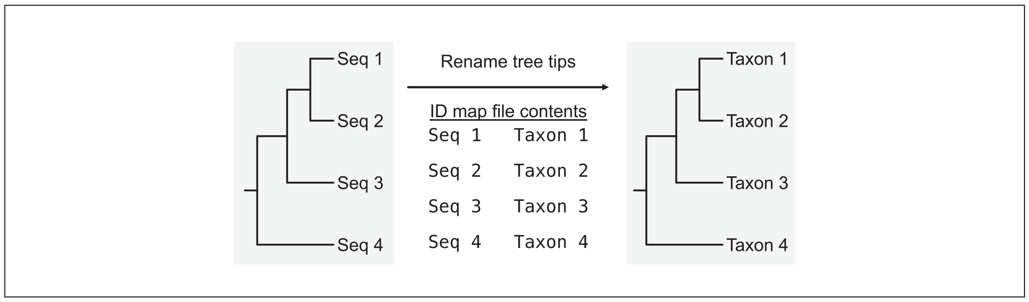


**Figure 6** Renaming taxa in a phylogenetic tree. Taking as input a Newick tree file and an identifier map file, the PhyKIT function "rename_tips" can rename tips in a phylogeny.

an ID-map file, which is the same two-column file format used in rename_fasta_entries (Fig. 6; *https://jlsteenwyk.com/PhyKIT/usage/index.html#rename-tree-tips*).

```
# First five lines of the idmap file for renaming tree tips
$ head -n 5 idmap.txt
speciesA|gene043    speciesA
speciesB|gene367    speciesB
speciesC|gene589    speciesC
speciesD|gene251    speciesD
speciesE|gene417    speciesE
```

To rename tree tips, use the following command:

```
# Renaming tips in a phylogenetic tree
$ pk_rename_tree input.tre -i idmap.txt [-o output.tre]
```

The resulting collection of single-locus phylogenetic trees can be used as input into software that summarizes them using methods that are consistent with the multispecies coalescence model.

*BASIC PROTOCOL 3*

## DETECTING ANOMALIES IN ORTHOLOGY RELATIONSHIPS

When constructing phylogenomic data matrices, it is important to consider that errors can be introduced during every step (Steenwyk et al., 2023a). Basic Protocols 3-5 help diagnose and sometimes ameliorate diverse sources of error. In this protocol, we will discuss how to detect two types of errors in orthology inference: hidden paralogy and spurious ortholog inference.

### *Hidden paralogy and clan check*

Phylogenomics typically relies on single-copy orthologs because these presumably have not undergone a history of duplication and loss (Li et al., 2017; Waterhouse et al., 2018). Hidden paralogy refer to orthologous groups of genes that contain orthologs and paralogs that have undergone asymmetric patterns of duplication and loss (Fernández et al., 2019; Martín-Durán et al., 2017; Steenwyk et al., 2023a). As a result, their evolutionary history can be distinct from the species tree.

Hidden paralogy can be detected on the basis of single-loci that do not recover the monophyly of "incontestable" clades, which are defined as lineages broadly accepted to be monophyletic and that are often free from phylogenetic controversies (Philippe et al., 2009; Rodríguez-Ezpeleta et al., 2007). For example, it is well accepted that fungi and animals form separate clades (Liu et al., 2023; Ocaña-Pallarès et al., 2022), and thus that individual loci that do not recapitulate the monophyly of each lineage may have been subject to complex patterns of duplication and loss. Hidden paralogs also can differ in their phylogenetic information content compared to loci that are not hidden paralogs (Mulhair et al., 2022). Alternatively, hidden paralogy may not be present, but the phylogenetic signal of a single gene may be insufficient to infer such ancient divergences.

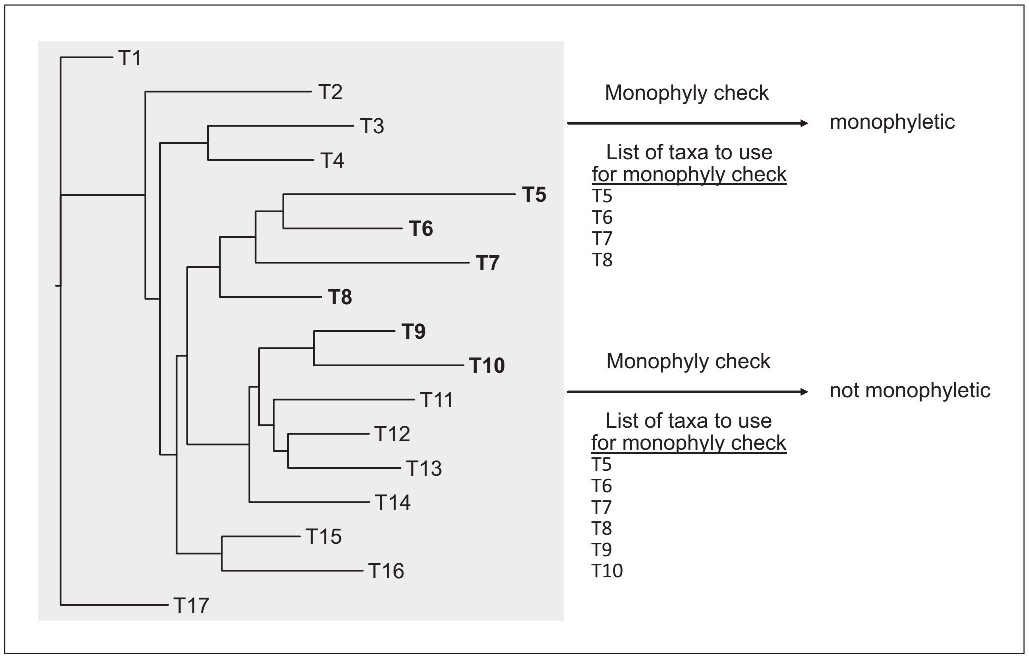


**Figure 7** Examining the exclusive monophyly of taxa. This can be a helpful method for detecting hidden paralogy. The "monophyly_check" function in PhyKIT makes it possible to examine the exclusive monophyly of a lineage. In the example shown, T5, T6, T7, and T8 form an exclusive monophyly; thus, PhyKIT will report "monophyletic" from a monophyly check of those taxa (top). However, the same set of taxa as well as T9 and T10 do not form an exclusively monophyletic clade; thus, PhyKIT will report "not_monophyletic" for those taxa (bottom). PhyKIT will also report additional information, including the taxa that are in the same lineage as those in the input file; in this case, that includes T11, T12, T13, and T14.

The PhyKIT function "monophyly_check" (alias: "is_monophyletic") can determine whether a set of species define a monophyletic group (Fig. 7; *https://jlsteenwyk.com/PhyKIT/usage/index.html#monophyly-check*). The function takes as input two files: a tree file and a single-column file with tip names to examine for monophyly. To facilitate high-throughput processing, tip names not present in the tree will be excluded when examining monophyly. This function can be executed as follows:

```
# Monophyly check
$ pk_monophyly_check input.tre list_of_taxa.txt
```

The output of this function will have six columns: (1) a string reporting if the taxa listed form a monophyletic group; (2) the average, (3) maximum, (4) minimum, and (5) the standard deviation of bipartition support values in the clade of interest; and (6) the names of taxa that are monophyletic with the taxa in `list_of_taxa.txt`.

Users may also be interested in examining the exclusive monophyly of multiple sets of taxa, which is an analysis that is often referred to as "clan check" (Mulhair et al., 2022; Siu-Ting et al., 2019). The PhyKIT function "hidden_paralogy_check" (alias: "clan_check") can conduct this analysis (Fig. 8; *https://jlsteenwyk.com/PhyKIT/usage/index.html#hidden-paralogy-check*). To do so, rather than a single column of taxa, each lineage must be provided as a row, and each tip name should be separated by a space (this is termed a "clade file"). For example, suppose it is anticipated that tips T5, T6, T7, and T8 are expected to be monophyletic, and so are T9, T10, T11, and T12. In this case, the clade file should be formatted as follows:

```
# The format of a clade file
$ cat clades.txt
T5 T6 T7 T8
T9 T10 T11 T12
```

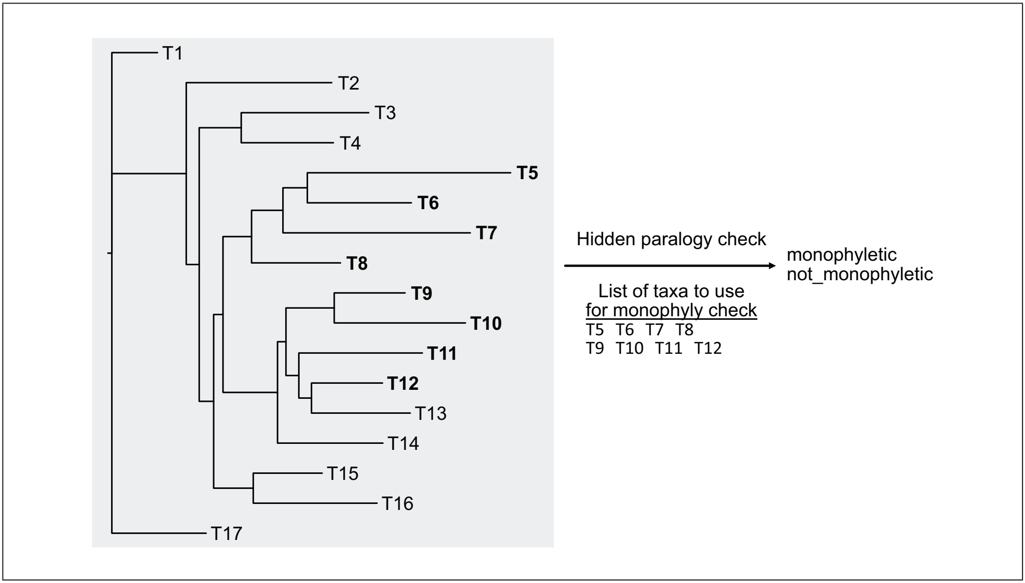


**Figure 8** Checking the exclusive monophyly of different expected clades. Unlike with the "monophyly_check" function (Fig. 8), here hidden paralogy can be examined for multiple clades with one command, using the "hidden_paralogy_check" function. In this example, the taxa T5, T6, T7, and T8 do form an exclusively monophyletic clade, which will be reported by PhyKIT as "monophyletic." However, the other taxa—T9, T10, T11, and T12—do not form an exclusively monophyletic clade and will therefore be reported as "not_monophyletic."

With an appropriately formatted clade file, users can next determine if a single-locus has signatures of hidden paralogy. (However, users should note that other causes, such as lack of signal, may also result in this type of signature.)

```
# Clan check
$ pk_hidden_paralogy_check input.tre -c clades.txt
```

The output will report whether the specified tips form an exclusively monophyletic group or not. The output will have the same number of rows as specified in the clade file. For example, in the exemplary clade file that examines the monophyly of two sets of taxa, there will be two rows of output. The first row corresponds to the result for T5, T6, T7, and T8; the second row corresponds to the result for T9, T10, T11, and T12. In the "toy" example (Fig. 8), PhyKIT will report that the first group is exclusively monophyletic and that the second group is not.

### *Spurious homolog/ortholog detection*

Erroneously inferred sequence homology and orthology can often manifest as long terminal branches (Shen et al., 2018). The PhyKIT function "spurious_sequence" (aliases: "spurious_seq", "ss") can be used to detect these (Fig. 9; *https://jlsteenwyk.com/PhyKIT/usage/index.html#spurious-homolog-identification*). Long terminal branches are defined as having X times the median length of all branches in a phylogeny (the default threshold value of X is 20, and this can be modified with the `-f/--factor` argument). This function is executed as follows:

```
# Identify putatively spurious homologs/orthologs
$ pk_spurious_seq input.tre [-f 20]
```

The output of this function will have four columns: (1) name of the tip that is a putatively spurious homolog/ortholog; (2) length of branch leading to putatively spurious sequence, (3) threshold used to identify putatively spurious sequences, and (4) median branch length in the phylogeny. The `-f` argument allows users to modify the threshold value X that defines a "long terminal branch." Putatively erroneous homologs/orthologs

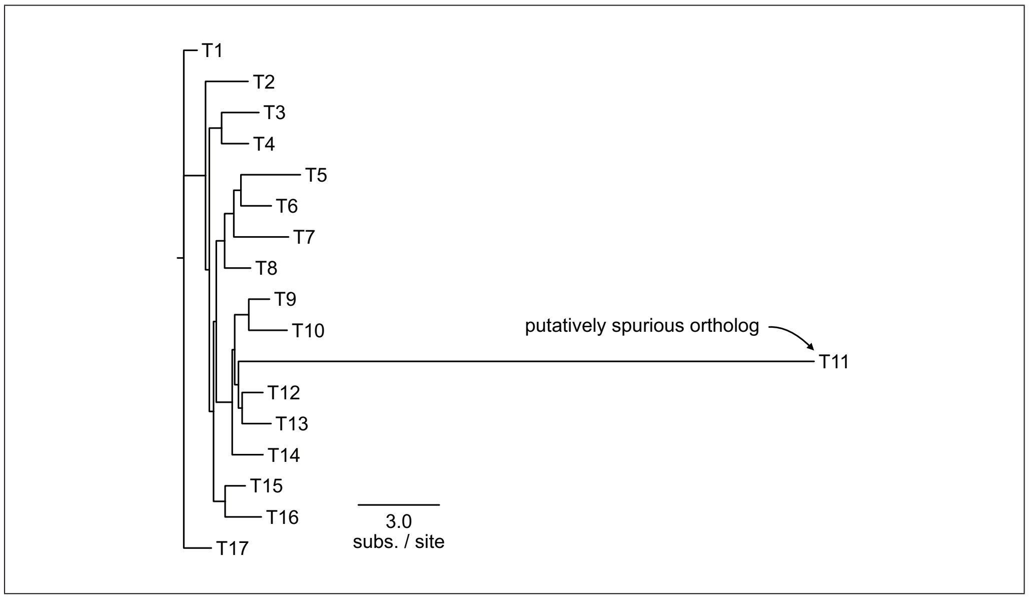


**Figure 9** Identifying putatively spurious orthologs. Taxa with outlier long branches can be a signature of spurious orthology/homology. The PhyKIT function "spurious_seq" can help identify putatively incorrect orthologs. In this example, PhyKIT will identify T11 as a putatively spurious ortholog.

can subsequently be removed from the larger set of sequences before continuing with downstream analyses. To remove sequences from a FASTA file, the BioKIT function "remove_fasta_entry" can be used (Steenwyk et al., 2022a).

## QUANTIFYING BIASES IN PHYLOGENOMIC DATA MATRICES AND RELATED MEASURES

***BASIC PROTOCOL 4***

Phylogenomic data matrices may contain biases at the level of the composition of taxa, genes, and sites. PhyKIT employs methods to diagnose potential sources of bias at each of these levels (Table 3). Taxa that may be sources of bias can potentially be removed from phylogenomic data matrices to, for example, examine the stability of a phylogenomic tree inferred using the full dataset (Aberer et al., 2013; Li et al., 2021). At the level of genes, users may want to also select a subset for specific downstream analyses. For example, the stability of a species tree can be examined using phylogenomic subsampling, the practice of subsetting full phylogenomic data matrices for the "best" genes according to a metric and reinferring the species tree (Bjornson et al., 2023; Edwards, 2016; Mongiardino Koch, 2021; Salichos & Rokas, 2013; Shen et al., 2016), or genes that evolved in a more clock-like manner can be identified for downstream molecular clock analysis (Liu et al., 2017; Smith et al., 2018).

### Measuring Bias at the Level of Taxa

#### *Relative composition variability, taxon (RCVT)*

Convergent evolution of amino acid or nucleotide usage can challenge phylogenomic inference due to similar sequence changes occurring in independent lineages (Fig. 10; Steenwyk et al., 2023a). For example, high-salt-adapted Methanonatronarchaeia and Haloarchaea, two distantly related lineages, both have compositional skews toward highly acidic amino acids, which can lead to the erroneous inference of phylogenetic affinity (Martijn et al., 2020). Accordingly, users may want to identify taxa with potential compositional biases. The PhyKIT function "relative_composition_variability_taxon" (aliases: "rel_comp_var_taxon", "rcvt") can quantify biases in individual taxa by calculating RCVT (*https://jlsteenwyk.com/PhyKIT/usage/index.html#relative-composition-variability-taxon*).

**Table 3** Summary of Whether High or Low Values are Desirable for Phylogenomic Subsampling

| Feature being subsampled | Metric for subsampling | PhyKIT function | Higher/lower values are better |
|---|---|---|---|
| Taxon | Relative composition variability, taxon (RCVT) | relative_composition_variability_taxon; rel_comp_var_taxon; rcvt | Lower |
| Taxon or gene | Long branch score | long_branch_score; lb_score; lbs | Lower |
| Gene | Alignment length | alignment_length; aln_len; al | Higher |
| | Alignment length, no gaps | alignment_length_no_gaps; aln_len_no_gaps; alng | Higher |
| | Pairwise identity | pairwise_identity; pairwise_id, pi | Context dependent |
| | Relative composition variability | relative_composition_variability; rel_comp_var; rcv | Lower |
| | Variable sites | variable_sites; vs | Higher |
| | Average (or median) bipartition support value | bipartition_support_stats; bss | Higher |
| | Evolutionary rate | evolutionary_rate, evo_rate | Context dependent |
| | Total tree length | total_tree_length; tree_len | Context dependent |
| | Treeness | treeness; tness | Higher |
| | Saturation | saturation; sat | Higher |
| | Treeness/relative composition variability | treeness_over_rcv; toverr; tor | Higher |
| | Degree of violation of the molecular clock | degree_of_violation_of_a_molecular_clock; dvmc | Lower |
| Sites | Compositional bias | compositional_bias_per_site; comp_bias_per_site; cbps | Lower |
| | Evolutionary rate | evolutionary_rate_per_site; evo_rate_per_site; erps | Lower |

RCVT is derived from a similar metric, relative composition variability (RCV), which was designed to quantify compositional biases in multiple sequence alignments (Phillips & Penny, 2003). RCVT has been adapted to quantify compositional biases in individual taxa rather than across a whole alignment. Specifically, a bias score is calculated for each taxon in an alignment and higher scores indicate higher biases. For example, consider the following "toy" alignment in which the first two sequences are skewed toward all guanines or only guanines and cytosines.

```
# Sequence T1 and T2 are compositionally biased
$ cat toy_alignment.fa
>T1
gggggggggggggggggggg----------------ggggggg
>T2
---------------gggggccccccccccccccccgggggggg-
>T3
-----atgcatgcatgcatgcatgcatgcatgc-----------
>T4
```

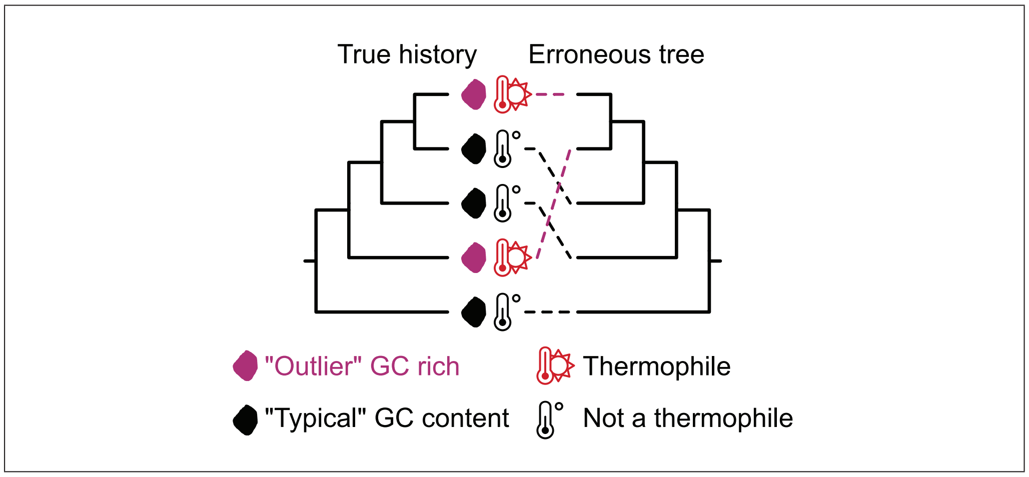


**Figure 10** Convergent sequence evolution and associated compositional biases can lead to erroneous phylogenomic inference. A phylogeny depicting the true history among exemplary microbes reveal how two taxa independently became GC rich, which is associated with a thermophilic lifestyle. GC-rich compositional biases can lead to erroneous phylogenetic inferences, such as the sister relationship between the GC-rich taxa here. The “rcvt” function in PhyKIT can help identify taxa with compositional sequence biases.

```
-----atgcatgcatgcatgcatgcatgcatgc-----------
>T5
-----atgcatgcatgcatgcatgcatgcatgc-----------
```

Use the “rcvt” function to quantify biases in each taxon.

```
# Quantify relative compositional biases
$ pk_rcvt toy_alignment.fa
```

```
T1    0.1436
T2    0.0782
T3    0.0509
T4    0.0509
T5    0.0509
```

Because T1 and T2 have the highest values, they have the greatest compositional bias in their associated sequences (Table 3). In contrast, T3, T4, and T5 have less bias and lower values. Note that these values should be interpreted relative to one another. Thus, T1 is more compositionally biased than T2, and T3, T4, and T5 have equally low relative compositional bias. To identify taxa that may introduce phylogenomic error, researchers can use outlier-detection analysis—such as identifying values outside of the interquartile ranges—across RCVT values to identify taxa with atypical sequences. This analysis can be done via scripting in R or Python.

### *Long branch score*

Long terminal branches can lead to an artifact called long branch attraction whereby distantly related taxa can be inferred as more closely related (Bergsten, 2005; Susko & Roger, 2021). As a result, users may be interested in identifying taxa that can be a source of long branch attraction bias. The long branch score offers a way to calculate long branch scores for individual taxa and is the mean distance between an individual taxon compared to all other taxa divided by the average distance across every pairwise combination of taxa (Struck, 2014); thus, higher values indicate higher risk for long branch attraction compared to lower values (Table 3). Long branch scores may be calculated in an entire phylogenomic data matrix. and the resulting distribution of long branch scores per taxon can be used to identify taxa that may contribute to long branch attraction artifacts.

To provide intuition for the long branch score, consider the following toy phylogeny in which taxon C has a long terminal branch relative to all other branches *(((A:1,B:1):0.5,C:3):0.25,D:1);*. To quickly visualize the long branch, print the phylogeny in American Standard Code for Information Interchange (ASCII) format, using the PhyKIT function "print_tree" (aliases: "print", "pt"; *https://jlsteenwyk.com/PhyKIT/usage/index.html#print-tree*).

```
# Visualize the toy tree
$ pk_print_tree toy.tre
_________________________ A
______________|
   _| ________|__________ B
_| _|
_| _|____________________________________ C
_|
_|____________ D
```

To calculate long branch scores with PhyKIT, use the "long_branch_score" function (aliases: "lb_score", "lbs"; *https://jlsteenwyk.com/PhyKIT/usage/index.html#long-branch-score*).

```
# Quantify long branch scores per taxon
$ pk_lb_score toy.tre -v
 A   -10.8434
 B   -10.8434
 C   27.7108
 D   -6.0241
```

Here, we use an additional argument, `-v/--verbose`, which also reveals that taxon C has the highest long branch score, indicating it is on the longest branch; thus, lower values are more desirable. Similarly to RCVT analysis, outlier-detection analysis may be useful to identify taxa that may introduce long branch attraction artifacts. If the `-v/--verbose` argument is not used, the summary statistics among the long branch scores per taxon will be reported.

```
# Quantify summary statistics of long branch score
$ pk_lb_score toy.tre
mean: -0.0
median: -8.4337
25th percentile: -10.8434
75th percentile: 2.4096
minimum: -10.8434
maximum: 27.7108
standard deviation: 18.6131
variance: 346.446
```

Together, the RCVT and long branch score can help researchers identify taxa that may be sources of phylogenomic error.

### Phylogenomic Subsampling Using the Information Content of Alignments

Phylogenomic subsampling can help identify branches in a phylogenomic tree that are unstable (Bjornson et al., 2023; Edwards, 2016; Steenwyk et al., 2023a). Rather than randomly subsampling full data matrices, most studies featuring this approach use metrics associated with phylogenetic signal. Here, we demonstrate how to calculate diverse metrics that collectively summarize the phylogenetic information content in multiple sequence alignments. For sake of brevity, we only briefly introduce each metric, and point

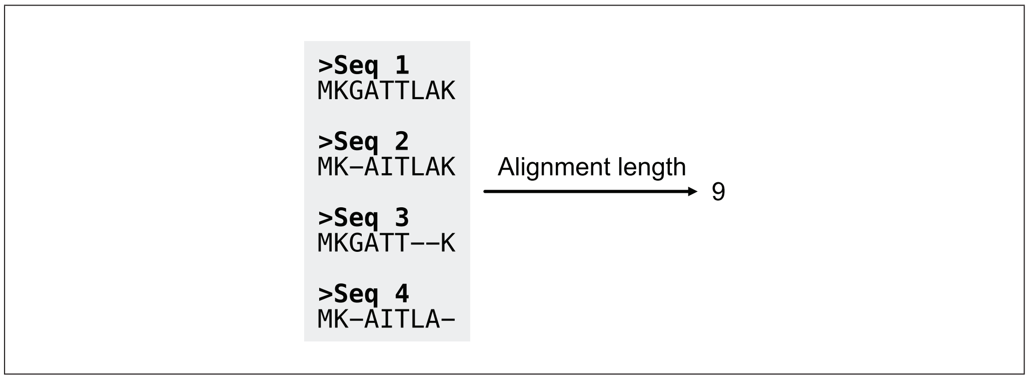


**Figure 11** Calculating the length of alignments. The PhyKIT function "aln_len" can calculate the length of alignments.

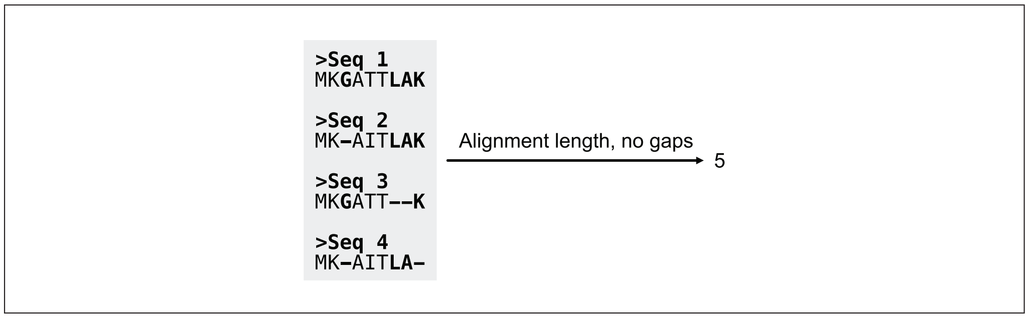


**Figure 12** Calculating the length of alignments excluding sites with gaps. The PhyKIT function "aln_len" calculates the length of alignments; sites in the alignment that contain a gap character are depicted in boldface.

users to the relevant documentation for more information. Functions are alphabetically organized to help users quickly find each function. Brief description and explanations of whether high or low values are reflective of greater or lower phylogenetic signal are provided in Table 3.

### *Alignment length*

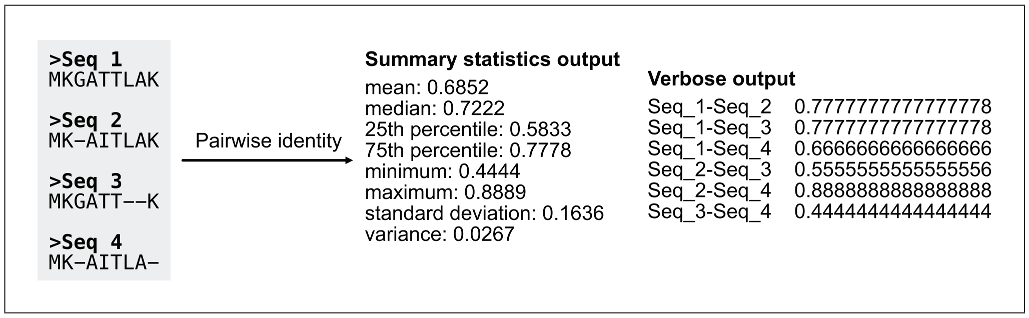
Longer alignments are more informative and have been associated with stronger phylogenetic signal (Shen et al., 2016). To calculate the length of an alignment, use the "alignment_length" (aliases: "aln_len" and "al") function in PhyKIT (Fig. 11; *https://jlsteenwyk.com/PhyKIT/usage/index.html#alignment-length*).

```
# Calculate alignment length
$ pk_aln_len input.fa
```

### *Alignment length, no gaps*

Gappy sites in alignments may offer little phylogenetic signal and even contribute to noise. The length of an alignment excluding sites with gaps can be calculated with the PhyKIT function "alignment_length_no_gaps" (Fig. 12; aliases: "aln_len_no_gaps", "alng"; *https://jlsteenwyk.com/PhyKIT/usage/index.html#alignment-length-no-gaps*).

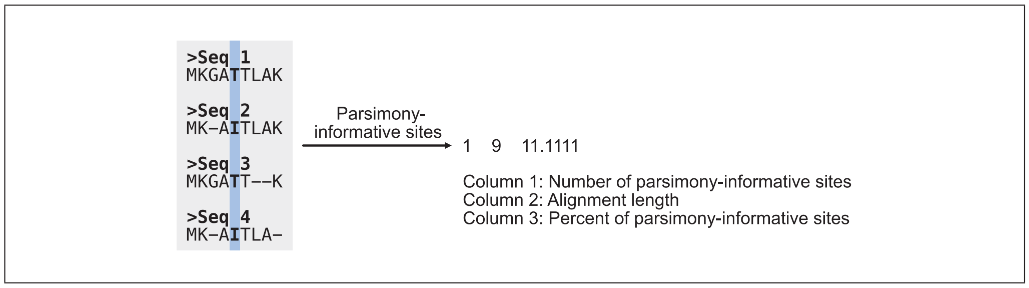
```
# Calculate alignment length excluding sites with gaps
$ pk_alng input.fa
# PhyKIT reports three-tab delimited values.
# col1: number of sites without gaps
# col2: total number of sites
# col3: percentage of sites without gaps
```


```
>Seq 1
MKGATTLAK

>Seq 2
MK-AITLAK

>Seq 3
MKGATT--K

>Seq 4
MK-AITLA-
```

```
Summary statistics output
mean: 0.6852
median: 0.7222
25th percentile: 0.5833
75th percentile: 0.7778
minimum: 0.4444
maximum: 0.8889
standard deviation: 0.1636
variance: 0.0267
```

```
Verbose output
Seq_1-Seq_2   0.7777777777777778
Seq_1-Seq_3   0.7777777777777778
Seq_1-Seq_4   0.6666666666666666
Seq_2-Seq_3   0.5555555555555556
Seq_2-Seq_4   0.8888888888888888
Seq_3-Seq_4   0.4444444444444444
```

**Figure 13** Calculating pairwise sequence identities. Pairwise sequence identities can be calculated using the “pairwise_id” function. The output is summary statistics of sequence similarity for each pairwise comparison. Alternatively, users can use the “-verbose” option to output the sequence similarity between each pairwise comparison of sequences. Users can also exclude sites with gaps during pairwise identity calculations using the “-e/–exclude_gaps” function.


```
>Seq 1
MKGATTLAK

>Seq 2
MK-AITLAK

>Seq 3
MKGATT--K

>Seq 4
MK-AITLA-
```

```
1   9   11.1111

Column 1: Number of parsimony-informative sites
Column 2: Alignment length
Column 3: Percent of parsimony-informative sites
```

**Figure 14** Determining the number and percentage of parsimony-informative sites in an alignment. The PhyKIT function “pis” calculates the number and percentage of parsimony-informative sites in an alignment. Here, the alignment of length nine has one site that is parsimony informative (denoted in blue).

### *Pairwise identity*

Pairwise identity is the average fraction of identical columns for each pairwise combination of sequences in an alignment. As a result, pairwise identity provides an estimate of evolutionary rate (Chen et al., 2017): faster-evolving sequences will have lower pairwise identity values, whereas slower-evolving sequences will have high pairwise identities. The PhyKIT function “pairwise_identity” (aliases: “pairwise_id”, “pi”) can calculate pairwise identities (Fig. 13; *https://jlsteenwyk.com/PhyKIT/usage/index.html#pairwise-identity*).

```
# Calculate pairwise identity
$ pk_pairwise_id input.fa [-e/--exclude_gaps -v/--verbose]
# Exclude sites with gaps while calculating pairwise identities.
# Obtain identity values for each pair with the verbose option.
```

### *Parsimony-informative sites*

The presence of more sites in an alignment providing information about parsimony (i.e., sites that have at least two non-gap characters that appear at least twice) is associated with increasing phylogenetic signal (Shen et al., 2016; Steenwyk et al., 2020). The PhyKIT function “parsimony_informative_sites” (alias: “pis”) can calculate the number and percentage of parsimony-informative sites in an alignment (Fig. 14; *https://jlsteenwyk.com/PhyKIT/usage/index.html#parsimony-informative-sites*).

```
# Determine the number of parsimony informative sites
$ pk_pis input.fa
# PhyKIT reports three-tab delimited values.
# col1: number of parsimony informative sites
```

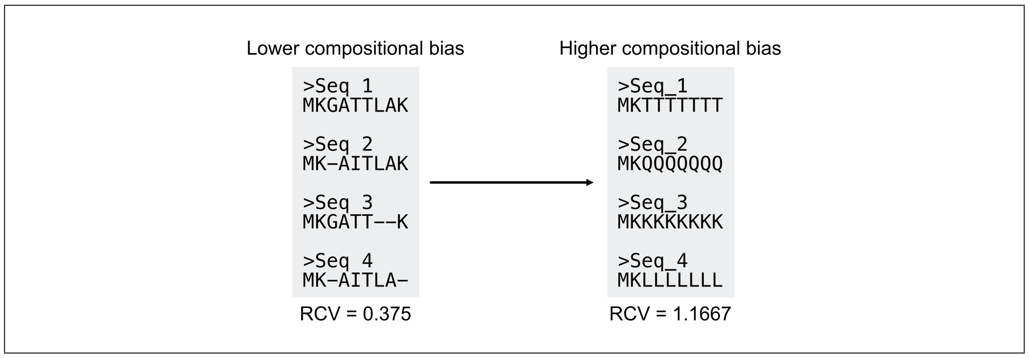


**Figure 15** Calculating the relative compositional bias of alignments. The function "rcv" calculates relative compositional biases in alignments. Low RCV values indicate lower compositional biases, whereas higher values reflect greater biases.

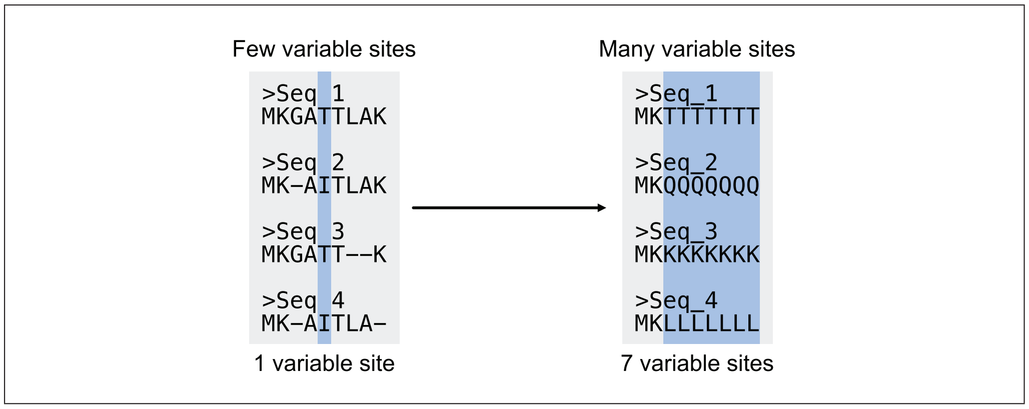


**Figure 16** Calculating the number and percentage of variable sites. The function "variable_sites" determines the number of variable sites in an alignment. It also returns information on the alignment length and the percentage of variable sites.

```
# col2: total number of sites
# col3: percentage of parsimony informative sites
```

### *Relative composition variability*

As previously noted, compositional biases can introduce phylogenomic errors (Martijn et al., 2020). Relative composition variability (RCV) measures the compositional bias of an alignment (Fig. 15). Statistically, RCV describes the average variability in sequence composition among taxa (Phillips & Penny, 2003). Lower RCV values indicate that the alignment has lower compositional bias. The PhyKIT function "relative_composition_variability" (aliases: "rel_comp_var", "rcv") can calculate RCV for an alignment (*https://jlsteenwyk.com/PhyKIT/usage/index.html#relative-composition-variability*).

```
# Determine the number of parsimony informative sites
$ pk_rcv input.fa
```

### *Variable sites*

The number of variable sites is associated with phylogenetic signal (Shen et al., 2016). The PhyKIT function "variable_sites" (alias: "vs") can calculate the number and percentage of variable sites in an alignment (Fig. 16; *https://jlsteenwyk.com/PhyKIT/usage/index.html#variable-sites*).

```
# Determine the number of variable sites
$ pk_vs input.fa
# PhyKIT reports three-tab delimited values.
```

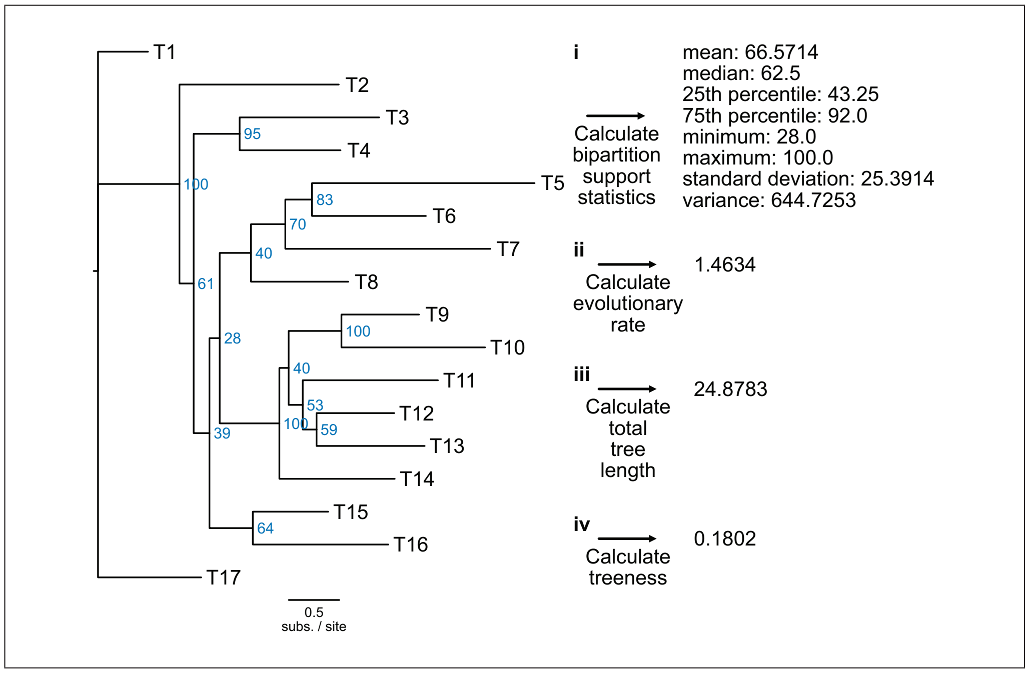


**Figure 17** Calculating tree-based measures of information content. (i) The function “bipartition_support_stats” calculates summary statistics for the support values in a phylogenetic tree. Here, bipartition support values are depicted in blue. This argument also has a verbose option that allows users to acquire the underlying distribution of values that were used to calculate summary statistics. (ii) Branch length information can be used to estimate the rate of gene evolutionary across a phylogeny. The function “evo_rate” can calculate the evolutionary rate of a gene. (iii) The function “tree_len” can calculate the sum of branch lengths in a phylogeny. (iv) Treeness is a measure of signal-to-noise whereby higher values indicate a greater signal-to-noise ratio. The function “treeness” can calculate treeness.

```
# col1: number of variable sites
# col2: total number of sites
# col3: percentage of variable sites
```

## Phylogenomic Subsampling Using the Information Content in Phylogenetic Trees

Phylogenomic subsampling may also be conducted based on the features of phylogenetic trees. Here, we demonstrate how to use PhyKIT to calculate diverse metrics that can guide subsampling of full data matrices. As in the previous section, each metric is briefly discussed, and we refer users to the documentation for more information. Functions are alphabetically organized to help users refer back to each function.

### *Bipartition support statistics*

Single-locus phylogenetic trees that have high overall bootstrap values can help robustly infer ancient divergences (Salichos & Rokas, 2013). The underlying concept is that high support values are indicative of greater certainty in tree topology and, thus, genes with stronger phylogenetic signal. The PhyKIT function “bipartition_support_stats” (alias: “bss”) calculates summary statistics of support values in a phylogenetic tree (Fig. 17; *https://jlsteenwyk.com/PhyKIT/usage/index.html#bipartition-support-statistics*).

```
# Calculate summary statistics of bipartition support
$ pk_bss input.tre
```

### *Evolutionary rate*

Like pairwise identity, a measure of evolutionary rate using alignment information, the PhyKIT function “evolutionary_rate” (alias: “evo_rate”) can calculate evolutionary

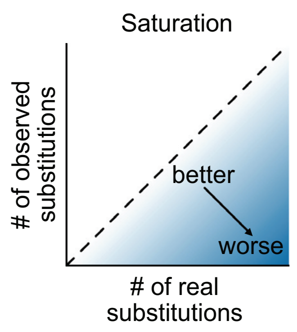


**Figure 18** Cartoon depiction of saturation. Saturation occurs when the number of observed substitutions underestimates the number of real substitutions. As a result, a perfect 1:1 ratio of observed substitutions and real substitutions would be indicative of data that lack saturation.

rate using tree-based information (Fig. 17; *https://jlsteenwyk.com/PhyKIT/usage/index.html#evolutionary-rate*). Specifically, evolutionary rate is the total tree length divided by the number of tips in the tree (Telford et al., 2014).

```
# Estimate evolutionary rate
$ pk_evo_rate input.tre
```

### *Total tree length*

Alternatively, total tree length, the sum of all branch lengths, can be used as a proxy for evolutionary rate. The PhyKIT function "total_tree_length" (alias: "tree_len") can be used to calculate the total tree length of a phylogeny (Fig. 17; *https://jlsteenwyk.com/PhyKIT/usage/index.html#total-tree-length*).

```
# Calculate total tree length
$ pk_tree_len input.tre
```

### *Treeness*

The signal-to-noise ratio in a phylogenetic tree can be calculated using the metric treeness (Phillips & Penny, 2003). Treeness (which is also referred to as stemminess) is the portion of tree distance among internal branches. Higher treeness values are more desirable because they reflect a higher signal-to-noise ratio. The PhyKIT function "treeness" (alias: "tness") can calculate treeness in a phylogeny (Fig. 17; *https://jlsteenwyk.com/PhyKIT/usage/index.html#treeness*).

```
# Calculate treeness
$ pk_tness input.tre
```

## Combining the Information Content in Alignments and Trees for Phylogenomic Subsampling

Other measures of phylogenomic subsampling combine information content in multiple sequence alignments and phylogenetic trees.

### *Saturation*

With multiple sequence alignments that have undergone numerous substitutions, such that the distances between them are underestimated, saturation is at play (Philippe et al., 2011; Fig. 18). The PhyKIT function "saturation" (alias: "sat") can quantify the level of saturation by multiple substitutions (*https://jlsteenwyk.com/PhyKIT/usage/index.html#saturation*). Data with no saturation will have a value of 1, whereas completely saturated data will have a value of 0.

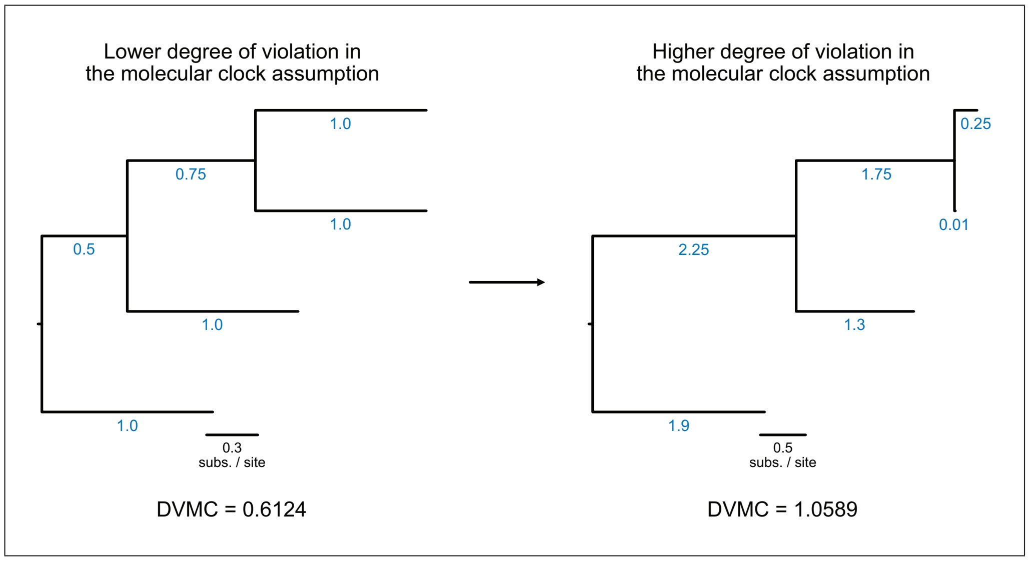

Lower degree of violation in the molecular clock assumption
1.0
0.75
1.0
0.5
1.0
1.0
0.3 subs. / site
DVMC = 0.6124
Higher degree of violation in the molecular clock assumption
0.25
1.75
0.01
2.25
1.3
1.9
0.5 subs. / site
DVMC = 1.0589

**Figure 19** Calculating the degree of violation of the molecular clock assumption. The degree of violation in the molecular clock is a helpful metric to identify genes that violates the assumption that there is a “molecular clock” representing a relatively constant rate of change across time and taxa. Phylogenies with low degrees of violation of this assumption will have broadly similar branch lengths (indicated in blue type) (left), whereas phylogenies that substantially violate the assumption will have highly variable branch lengths (right).

```
# Calculate saturation
$ pk_tness -a input.fa -t input.tre
```

### *Treeness/RCV*

Combining the two metrics treeness and RCV—specifically, treeness divided by RCV—makes it possible to identify loci that harbor a high signal-to-noise ratio and are not very susceptible to compositional bias (Phillips & Penny, 2003). Thus, higher treeness/RCV values are observed among loci with high signal-to-noise ratios and lower compositional biases. The PhyKIT function “treeness_over_rcv” (aliases: “toverr”, “tor”) can calculate treeness/RCV.

```
# Calculate treeness/RCV
$ pk_tor -a input.fa -t input.tre
# PhyKIT reports three-tab delimited values.
# col1: treeness/RCV
# col2: treeness
# col3: RCV
```

## Subsampling for Time Tree Analysis

The accuracy of divergence time estimates can be improved by using genes that evolve in a molecular-clock-like manner (with a constant rate of sequence change; Smith et al., 2018). The PhyKIT function “degree_of_violation_of_a_molecular_clock” (alias: “dvmc”) can quantify how much a gene deviates from a clock-like pattern of evolution (Fig. 19; *https://jlsteenwyk.com/PhyKIT/usage/index.html#degree-of-violation-of-the-molecular-clock*).

```
# Calculate degree of violation of a molecular clock (or DVMC)
$ pk_dvmc input.tre
```

Lower values are indicative of a lower degree of violation of the molecular clock assumption; thus, lower values are more desirable for downstream divergence time analysis.

## Measuring Bias at the Level of Sites

### *Compositional bias per site*

Compositional biases are known to negatively impact phylogenetic inferences (Phillips & Penny, 2003; Steenwyk et al., 2023a). The PhyKIT function "compositional_bias_per_site" (aliases: "comp_bias_per_site", "cbps") can quantify compositional biases in an alignment (*https://jlsteenwyk.com/PhyKIT/usage/index.html#compositional-bias-per-site*). Specifically, site-wise chi-squared ($\chi^2$) tests are conducted to detect compositional biases. Higher $\chi^2$ statistics indicate greater compositional biases. PhyKIT returns multi-test corrected *p*-values (derived from the Benjamini-Hochberg false discovery rate procedure) as well as uncorrected *p*-values.

```
# Calculate site-wise compositional biases
$ pk_comp_bias_per_site input.fa
# PhyKIT reports four-tab delimited values.
# col 1: index in alignment
# col 2: chi-squared statistic
# col 3: multi-test corrected p-value
# col 4: uncorrected p-value
```

### *Evolutionary rate per site*

When saturation is suspected to negatively influence phylogenetic reconstruction, fast-evolving sites are often removed (Eme et al., 2023; Steenwyk et al., 2023a). The PhyKIT function "evolutionary_rate_per_site" (aliases: "evo_rate_per_site", "erps") quantifies site-wise diversity as a proxy for site-wise evolutionary rate (*https://jlsteenwyk.com/PhyKIT/usage/index.html#evolutionary-rate-per-site*). Here, the greater the diversity, the greater the presumed evolutionary rate. This is conceptually similar to the use of pairwise identity in an alignment as a measure of evolutionary rate (Chen et al., 2014). Specifically, evolutionary rate per site is 1 minus the sum of the squared frequency of different characters at a given site. Values range from 0 (slow evolving; no diversity at the given site) to 1 (fast evolving; all characters appear only once).

```
# Calculate site-wise evolutionary rate
$ pk_evo_rate_per_site input.fa
# PhyKIT reports two-tab delimited values.
# col 1: index in alignment
# col 2: estimated evolutionary rate value
```

### Removing specific sites from an alignment

The resulting output from "compositional_bias_per_site" and "evolutionary_rate_per_site" can be used to guide site-specific trimming in a multiple sequence alignment. Site-specific trimming can be conducted using ClipKIT (Steenwyk et al., 2020). For the latter, ClipKIT implements a "cst" mode of trimming, which is an acronym for "custom-site trimming" (*https://jlsteenwyk.com/ClipKIT/advanced/index.html#custom-site-trimming-cst-mode*).

```
# Conduct site-specific trimming
$ clipkit input.fa -m cst -a auxiliary_file.txt
# -m specifies the cst mode
# auxiliary_file.txt specifies which sites to keep/remove
```

The auxiliary file is a two-column tab-delimited text file in which the first column is the site (starting at 1) and the second column specifies whether the site should be kept or trimmed using the strings "keep" or "trim".

```
# Conduct site-specific trimming
$ cat auxiliary_file.txt
```

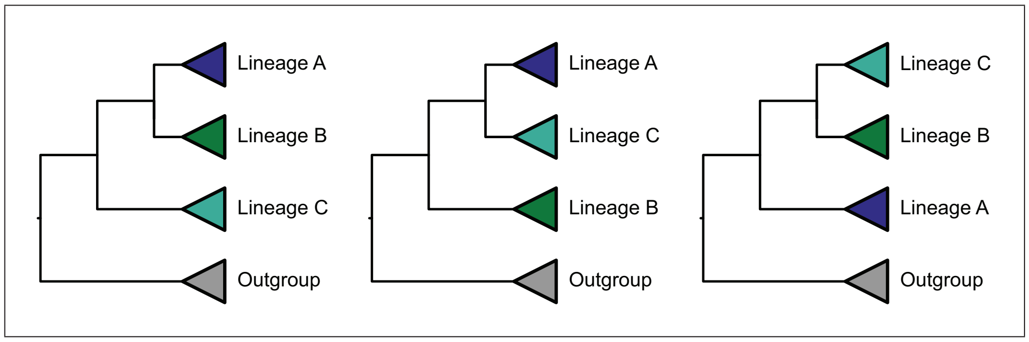


**Figure 20** Three possible topologies for a rooted quartet. When conducting a polytomy test, PhyKIT examines whether the three topologies occur in equal frequency among a set of single-locus phylogenies, which would be the signature of a polytomy; if not, there will be no such signature. Polytomies may represent a lack of resolution at a particular internode or, given sufficient data, be a signature of a radiation event.

```
1 keep
2 trim
3 keep
4 keep
5 keep
6 keep
```

Alternatively, users can specify sites that are only kept or trimmed using the `auxiliary_file.txt`. For example, the following would be equivalent to the auxiliary file described above:

```
# Conduct site-specific trimming
$ cat auxiliary_file.txt
2 trim
```

Similarly, users can also only specify sites to keep in the `auxiliary_file.txt`. The following would be equivalent to the two previous examples:

```
# Conduct site-specific trimming
$ cat auxiliary_file.txt
1 keep
3 keep
4 keep
5 keep
6 keep
```

In summary, this protocol demonstrates how to subsample phylogenomic data matrices at the level of taxa, genes, and sites. These analyses are aimed to facilitate identifying and ameliorating phylogenomic errors.

*BASIC PROTOCOL 5*

## IDENTIFYING POLYTOMIES

Polytomies can stem from radiation events or lack of resolution between three alternative topologies in a rooted quartet (Fig. 20; Sayyari & Mirarab, 2018). The signature of a polytomy (or multifurcation) is when each of the three topologies in a rooted quartet has equal or near-equal support, which can be tested for using a $\chi^2$ test (Steenwyk et al., 2021). This method has been successfully used to detect polytomies in various fungal and plant lineages (Li et al., 2021; One Thousand Plant Transcriptomes Initiative, 2019; Steenwyk et al., 2021).

To conduct a polytomy test, use the PhyKIT function "polytomy_test" (aliases: "polyt", "ptt"; *https://jlsteenwyk.com/PhyKIT/usage/index.html#polytomy-testing*). The polytomy_test function takes as input a file with the three groups of taxa to test the relationships.

```
# format of the groups file is:
# label group0 group1 group2
$ cat groups.txt
name_of_test T1;T2 T3 T4;T5
```

A single-column file with the names of the desired tree files to use for polytomy testing must also be specified.

```
# label group0 group1 group2
$ cat trees.txt
Input0.tre
Input1.tre
Input2.tre
Input3.tre
…
```

Using the two files, conduct a polytomy test.

```
# Conduct a polytomy test
$ pk_ptt -t trees.txt -g groups.txt
```

To test the process of conducting a polytomy test using real data, see the online documentation (*https://jlsteenwyk.com/PhyKIT/tutorials/index.html#identifying-signatures-of-rapid-radiations*).

## ASSESSING GENE-GENE COEVOLUTION AS A GENETIC SCREEN

***BASIC PROTOCOL 6***

Genes that coevolve tend to have shared function, be coexpressed, or be constituents of the same multimeric complex (Clark et al., 2012; Steenwyk et al., 2022c). Moreover, gene coevolution can be used to prioritize genes with a predicted functionality via guilt-by-association. For example, genes that are coevolving with other genes that function in DNA repair processes can be rapidly screened to identify additional DNA repair genes (Brunette et al., 2019).

Gene coevolution can be detected by the mirror principle whereby two gene trees have similar branch lengths across speciation events are likely to have coevolved (Steenwyk et al., 2022c). In other words, the two phylogenetic trees will have accelerated and decelerated in evolutionary rates in a coordinated manner (Fig. 21). To avoid false positives, gene tree branch lengths need to be corrected by the corresponding branch length in a

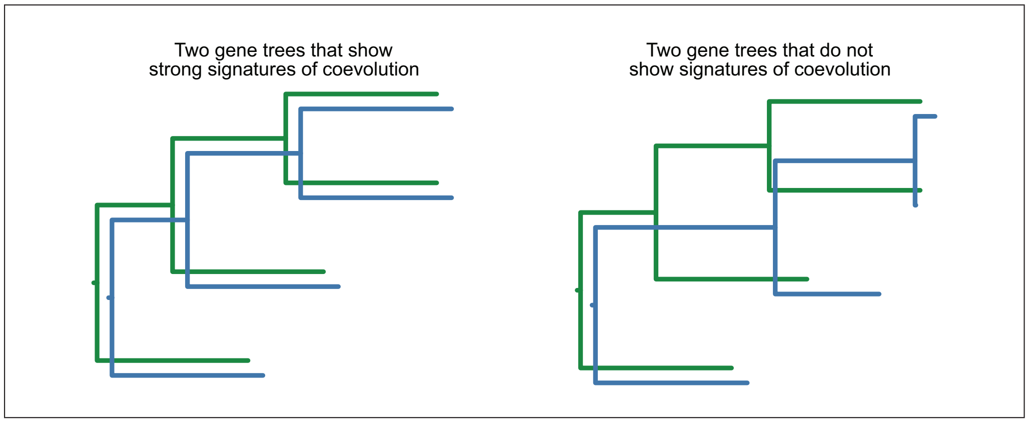

Two gene trees that show strong signatures of coevolution
Two gene trees that do not show signatures of coevolution

**Figure 21** Exemplary phylogenies that do and do not display signatures of coevolution. Single gene phylogenies are depicted in blue or green. Phylogenies that have similar branch lengths across speciation events harbor a signature of coevolution (left). Phylogenies that do not have similar branch lengths are not coevolving (right).

species tree (Clark et al., 2012; Steenwyk et al., 2022c). Moreover, this corrects to variation in branch lengths associated with differences in mutation and divergence time.

To quantify gene coevolution, the "cover" function in PhyKIT is used (*https://jlsteenwyk.com/PhyKIT/usage/index.html#covarying-evolutionary-rates*). The cover function takes as input two single-locus phylogenies and a reference species tree.

```
# Executing the cover function
$ pk_cover tree1.tre tree2.tre -r reference_tree.tre
# PhyKIT outputs two values
# col 1: coevolutionary coefficient (the strength of coevolution)
# col 2: p-value
```

PhyKIT requires that the three phylogenies have the same topology: that is, the two single-locus phylogenies should be constrained to match the reference tree. To perform constrained tree search using IQ-TREE (Minh et al., 2020), the following command can be used:

```
# constrained tree search
$ iqtree2 -s tree1.fa -te reference_tree.constrained_topology.tre -pre output_prefix -m TEST
  -keep-ident
```

The resulting phylogeny should be rooted following the reference tree using the "root_tree" (aliases: "root", "rt") function or other software. PhyKIT automatically accounts for variation in taxon representation between two single-locus phylogenies. Thereafter, gene coevolution can be quantified. A tutorial of gene-gene coevolution using real data is available in the online documentation (*https://jlsteenwyk.com/PhyKIT/tutorials/index.html#evaluating-gene-gene-covariation*).

## COMMENTARY

### Background Information

PhyKIT was initially developed to overcome the absence of transversal phyloinformatic tools for some key steps in data processing and analysis of phylogenomic data. Specifically, PhyKIT was developed to quantify biases in phylogenomic data matrices and calculate gene-gene coevolution. However, while developing PhyKIT, we recognized the opportunity to provide broader support to the phyloinformatic community. We were excited to provide a broadly applicable tool and were further excited (and surprised!) to see it be adopted by the community.

#### *Related tools*

Since initially releasing PhyKIT, we have developed several other tools that may be of interest to readers. Together, these tools are part of a broader ecosystem of bioinformatic tools that facilitate phylogenomic data analysis, processing, and more.

*ClipKIT, an alignment trimming software (Steenwyk et al.*, 2020*)*:

• Documentation: *https://jlsteenwyk.com/ClipKIT/*

• Source code: *https://github.com/JLSteenwyk/ClipKIT*

*BioKIT, a broadly applicable toolkit for broad genomic analysis (*Steenwyk et al., 2022a*)*:

• Documentation: *https://jlsteenwyk.com/BioKIT/*

• Source code: *https://github.com/JLSteenwyk/BioKIT*

*OrthoSNAP, an algorithm to identify single-copy orthologs nested within larger multi-copy gene families (*Steenwyk et al., 2022b*)*:

• Documentation: *https://jlsteenwyk.com/orthosnap/*

• Source code: *https://github.com/JLSteenwyk/orthosnap*

*orthofisher, software for sequence similarity search using Hidden Markov Models (*Steenwyk & Rokas, 2021b*)*:

• Documentation: *https://jlsteenwyk.com/orthofisher/*

• Source code: *https://github.com/JLSteenwyk/orthofisher*

*treehouse, a graphical user interface tool for pruning large phylogenies (Steenwyk & Rokas*, 2019*)*:

• Documentation and source code: *https://github.com/JLSteenwyk/treehouse* *ggpubfigs, a ggplot2 extension (https://ggplot2.tidyverse.org/) for making colorblind-friendly and publication-quality scientific figures (*Steenwyk & Rokas, 2021a*)*:

• Documentation and source code: *https://github.com/JLSteenwyk/ggpubfigs*

These tools, alongside other software, have been incorporated into a Snakemake workflow, OrthoFlow (*https://github.com/rbturnbull/orthoflow*), to enable phylogenomic analysis using one command (Turnbull et al., 2023).

***Future directions***

Further development of PhyKIT will be guided by the best practices and advances in the field. Currently, alternative data types (e.g., synteny or structure information) are becoming more common in phylogenomics (Parey et al., 2023; Schultz et al., 2023; Steenwyk & King, 2023). The utility of PhyKIT for alternative data types will be explored. We anticipate that certain metrics will be easily transferred to other data types (e.g., bipartition support statistics or treeness) because they handle phylogenetic trees, not the underlying data used to infer them.

***Glossary***

**Bootstrap replicates:** In the context of phylogenetics, each replicate is a resampling (with replacement) of sites from the full alignment to generate an alignment of equal size; these replicates are then used to reinfer a phylogeny and evaluate support for the phylogeny inferred using the full alignment.

**Concatenation:** The phylogenomic method of combining sequences from multiple loci into a single sequence for each species and using the resulting supermatrix for species tree inference.

**Hidden paralogy:** Asymmetric loss of paralogs in some lineages, leading to mistaken identification of paralogs as orthologs.

**Long branch attraction:** A phylogenetic artifact in which rapidly evolving taxa/lineages are erroneously inferred to be closely related.

**Multispecies coalescence:** The phylogenomic method of using single-locus phylogenies, which may differ from each other, to infer a species tree.

**Orthologs or orthologous genes:** Genes in different species that originated from a common ancestor through speciation.

**Orthology inference:** Identifying genes among organisms that evolved from a common ancestral gene.

**Paralogs or paralogous genes:** Genes that are related by duplication.

**Phylogenomic subsampling:** The process of selecting a subset of a complete phylogenomic data matrix to reconstruct phylogenetic trees, often aiming to reduce noise and improve signal or evaluating the stability of the inferred phylogeny.

**Radiation events:** Rapid speciation events that result in a succession of short internal branches in a phylogeny.

**Single-copy orthologs:** Genes present as a single copy in the genome across a set of taxa and originate from speciation events.

**Spurious ortholog inference:** Incorrect identification of genes as orthologous, often due to errors in sequence analysis or interpretation.

## Troubleshooting

Users can submit GitHub issues for support (*https://github.com/JLSteenwyk/PhyKIT/issues*) or contact the lead developer via email (*https://jlsteenwyk.com/contact.html*).

## Author Contributions

**Jacob L. Steenwyk:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; validation; visualization; writing—original draft; writing—review and editing. **Gemma I Martínez-Redondo:** Validation; writing—review and editing. **Rosa Fernández:** Funding acquisition; methodology; writing—review and editing. **Thomas Buida:** Conceptualization; methodology; software; validation; writing—review and editing. **Emile Gluck-Thaler:** Writing—review and editing. **Xing-Xing Shen:** Writing—review and editing. **Toni Gabaldón:** Funding acquisition; writing—review and editing. **Antonis Rokas:** Funding acquisition; methodology; writing—review and editing.

## Acknowledgements

PhyKIT has benefitted from the kindness and generosity of the research community. A special thanks to Drs. Michael F. Seidl, Jason Stajich, Toby Baril, and Heroen Verbruggen for helpful comments and suggestions.

JLS is a Howard Hughes Medical Institute Awardee of the Life Sciences Research Foundation. Research in AR's lab is supported by grants from the US

National Science Foundation (DEB-2110404), the US National Institutes of Health/National Institute of Allergy and Infectious Diseases (R01 AI153356), and the Burroughs Wellcome Fund. GIMR acknowledges the support of the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya and ESF Investing in your future (grant 2021 FI_B 00476). X-XS was supported by the National Key R&D Program of China (2022YFD1401600) and the National Science Foundation for Distinguished Young Scholars of Zhejiang Province (LR23C140001). RF acknowledges support from the following sources of funding: Ramón y Cajal fellowship (grant agreement no. RYC2017-22492 funded by MCIN/AEI /10.13039/501100011033 and ESF 'Investing in your future'), the European Research Council (this project has received funding from the European Research Council [ERC] under the European's Union's Horizon 2020 research and innovation programme, grant agreement no. 948281) and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (AGAUR 2021-SGR00420).

## Conflicts of Interest

JLS is an advisor for ForensisGroup Inc. AR is a scientific consultant for LifeMine Therapeutics, Inc.

## Data Availability Statement

No data is provided as part of this study. PhyKIT tutorials (*https://jlsteenwyk.com/PhyKIT/tutorials/index.html*) come with available test data for learning to use PhyKIT.

## Literature Cited

Aberer, A. J., Krompass, D., & Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice. *Systematic Biology*, *62*, 162–166. https://doi.org/10.1093/sysbio/sys078

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., & Smith, K. (2011). Cython: The best of both worlds. *Computing in Science & Engineering*, *13*, 31–39.

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*, 163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x

Bjornson, S., Upham, N., Verbruggen, H., & Steenwyk, J. (2023). Phylogenomic inference, divergence-time calibration, and methods for characterizing reticulate evolution. *Biology and Life Sciences*, Retrieved from: https://www.preprints.org/manuscript/202309.0905/v1

Brunette, G. J., Jamalruddin, M. A., Baldock, R. A., Clark, N. L., & Bernstein, K. A. (2019). Evolution-based screening enables genome-wide prioritization and discovery of DNA repair genes. *Proceedings of the National Academy of Sciences*, *116*, 19593–19599. https://doi.org/10.1073/pnas.1906559116

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Caurcel, C., Laetsch, D. R., Challis, R., Kumar, S., Gharbi, K., & Blaxter, M. (2021). MolluscDB: A genome and transcriptome database for molluscs. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*, 20200157. https://doi.org/10.1098/rstb.2020.0157

Chen, M.-Y., Liang, D., & Zhang, P. (2017). Phylogenomic resolution of the phylogeny of laurasiatherian mammals: Exploring phylogenetic signals within coding and noncoding sequences. *Genome Biology and Evolution*, *9*, 1998–2012. https://doi.org/10.1093/gbe/evx147

Chen, W., Lee, M. K., Jefcoate, C., Kim, S. C., Chen, F., & Yu, J. H. (2014). Fungal cytochrome P450 monooxygenases: Their distribution, structure, functions, family expansion, and evolutionary origin. *Genome Biology and Evolution*, *6*, 1620–1634. https://doi.org/10.1093/gbe/evu132

Clark, N. L., Alani, E., & Aquadro, C. F. (2012). Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Research*, *22*, 714–720. https://doi.org/10.1101/gr.132647.111

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Coombe, L., Warren, R. L., Wong, J., Nikolic, V., & Birol, I. (2023). ntLink: A toolkit for *de novo* genome assembly scaffolding and mapping using long reads. *Current Protocols*, *3*, e733. https://doi.org/10.1002/cpz1.733

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution*, *37*, 291–294. https://doi.org/10.1093/molbev/msz189

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: More models, new heuristics and parallel computing. *Nature Methods*, *9*, 772–772. https://doi.org/10.1038/nmeth.2109

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2011). ProtTest 3: Fast selection of best-fit models of protein evolution.

*Bioinformatics*, *27*, 1164–1165. https://doi.org/10.1093/bioinformatics/btr088

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, *6*, 361–375. https://doi.org/10.1038/nrg1603

Edgar, R. C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications*, *13*, 6968. https://doi.org/10.1038/s41467-022-34630-w

Edwards, S. V. (2016). Phylogenomic subsampling: A brief review. *Zoologica Scripta*, *45*, 63–74. https://doi.org/10.1111/zsc.12210

Eisen, J. A. (1998). Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, *8*, 163–167. https://doi.org/10.1101/gr.8.3.163

Embley, M., Der Giezen, M. V., Horner, D. S., Dyal, P. L., & Foster, P. (2003). Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*, 191–203. https://doi.org/10.1098/rstb.2002.1190

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., de Anda, V., Schön, M. E., Seitz, K. W., Dombrowski, N., Lewis, W. H., Homa, F., Saw, J. H., Lombard, J., Nunoura, T., Li, W. J., Hua, Z. S., Chen, L. X., Banfield, J. F., John, E. S., Reysenbach, A. L., … Ettema, T. J. G. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, *618*, 992–999. https://doi.org/10.1038/s41586-023-06186-2

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*, 238. https://doi.org/10.1186/s13059-019-1832-y

Fernández, R., Gabaldón, T., & Dessimoz, C. (2019). Orthology: definitions, inference, and impact on species phylogeny inference. Retrieved from: https://arxiv.org/abs/1903.04530

Fernández, R., Tonzo, V., Simón Guerrero, C., Lozano-Fernandez, J., Martínez-Redondo, G. I., Balart-García, P., Aristide, L., Eleftheriadi, K., & Vargas-Chávez, C. (2022). MATEdb, a data repository of high-quality metazoan transcriptome assemblies to accelerate phylogenomic studies. *Peer Community Journal*, *2*, e58. https://doi.org/10.24072/pcjournal.177

Foster, P. G., Schrempf, D., Szöllősi, G. J., Williams, T. A., Cox, C. J., & Embley, T. M. (2022). Recoding amino acids to a reduced alphabet may increase or decrease phylogenetic accuracy. *Systematic Biology*, *72*(3), 723–737.

Gatesy, J., Meredith, R. W., Janecka, J. E., Simmons, M. P., Murphy, W. J., & Springer, M. S. (2017). Resolution of a concatenation/coalescence kerfuffle: Partitioned coalescence support and a robust family-level tree for Mammalia. *Cladistics*, *33*, 295–332. https://doi.org/10.1111/cla.12170

Giacomelli, M., Rossi, M. E., Lozano-Fernandez, J., Feuda, R., & Pisani, D. (2022). Resolving tricky nodes in the tree of life through amino acid recoding. *iScience*, *25*, 105594. https://doi.org/10.1016/j.isci.2022.105594

Green, R. E., Braun, E. L., Armstrong, J., Earl, D., Nguyen, N., Hickey, G., Vandewege, M. W., St John, J. A., Capella-Gutierrez, S., Castoe, T. A., Kern, C., Fujita, M. K., Opazo, J. C., Jurka, J., Kojima, K. K., Caballero, J., Hubley, R. M., Smit, A. F., Platt, R. N., … Ray, D. A. (2014). Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, *346*, 1254449. https://doi.org/10.1126/science.1254449

Han, Y., & Molloy, E. K. (2023). Improving quartet graph construction for scalable and accurate species tree estimation from gene trees. *Genome Research*, *33*, 1042–1052.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hernandez, A. M., & Ryan, J. F. (2021). Six-state amino acid recoding is not an effective strategy to offset compositional heterogeneity and saturation in phylogenetic analyses. *Systematic Biology*, *70*, 1200–1212. https://doi.org/10.1093/sysbio/syab027

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., … Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, *346*, 1320–1331. https://doi.org/10.1126/science.1253451

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*, 587–589. https://doi.org/10.1038/nmeth.4285

Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, *21*, 428–444. https://doi.org/10.1038/s41576-020-0233-0

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780. https://doi.org/10.1093/molbev/mst010

Kosiol, C., Goldman, N., & Buttimore, N. H. (2004). A new criterion and method for amino acid classification. *Journal of Theoretical Biology*, *228*, 97–106. https://doi.org/10.1016/j.jtbi.2003.12.010

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*, 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Li, Y., Steenwyk, J. L., Chang, Y., Wang, Y., James, T. Y., Stajich, J. E., Spatafora, J. W., Groenewald, M., Dunn, C. W., Hittinger, C. T., Shen, X. X., & Rokas, A. (2021). A genome-scale phylogeny of the kingdom Fungi. *Current Biology*, *31*, 1653–1665.e5. https://doi.org/10.1016/j.cub.2021.01.074

Li, Z., de la Torre, A. R., Sterck, L., Cánovas, F. M., Avila, C., Merino, I., Cabezas, J. A., Cervera, M. T., Ingvarsson, P. K., & van de Peer, Y. (2017). Single-copy genes as molecular markers for phylogenomic studies in seed plants. *Genome Biology and Evolution*, *9*, 1130–1147. https://doi.org/10.1093/gbe/evx070

Liu, H., Steenwyk, J. L., Zhou, X., Schultz, D. T., Kocot, K. M., Shen, X.-X., Rokas, A., & Li, Y. (2023). A genome-scale Opisthokonta tree of life: Toward phylogenomic resolution of ancient divergences. *Evolutionary Biology*, Retrieved from: http://biorxiv.org/lookup/doi/10.1101/2023.09.20.556338

Liu, L., Zhang, J., Rheindt, F. E., Lei, F., Qu, Y., Wang, Y., Zhang, Y., Sullivan, C., Nie, W., Wang, J., Yang, F., Chen, J., Edwards, S. V., Meng, J., & Wu, S. (2017). Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proceedings of the National Academy of Sciences*, *114*, E7282–E7290. Retrieved from: https://pnas.org/doi/full/10.1073/pnas.1616744114

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, *1*, e323. https://doi.org/10.1002/cpz1.323

Martijn, J., Schön, M. E., Lind, A. E., Vosseberg, J., Williams, T. A., Spang, A., & Ettema, T. J. G. (2020). Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nature Communications*, *11*, 5490. https://doi.org/10.1038/s41467-020-19200-2

Martín-Durán, J. M., Ryan, J. F., Vellutini, B. C., Pang, K., & Hejnol, A. (2017). Increased taxon sampling reveals thousands of hidden orthologs in flatworms. *Genome Research*, *27*, 1263–1272. https://doi.org/10.1101/gr.216226.116

Martínez-Redondo, G. I., Vargas-Chávez, C., Eleftheriadi, K., Benítez-Álvarez, L., Vázquez-Valls, M., & Fernández, R. (2024). MATEdb2, a collection of high-quality metazoan proteomes across the Animal Tree of Life to speed up phylogenomic studies. Retrieved from: http://biorxiv.org/lookup/doi/10.1101/2024.02.21.581367

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, *30*, i541–i548. https://doi.org/10.1093/bioinformatics/btu462

Mongiardino Koch, N. (2021). Phylogenomic subsampling and the search for phylogenetically reliable loci. *Molecular Biology and Evolution*, *38*, 4025–4038. https://doi.org/10.1093/molbev/msab151

Mulhair, P. O., McCarthy, C. G. P., Siu-Ting, K., Creevey, C. J., & O'Connell, M. J. (2022). Filtering artifactual signal increases support for Xenacoelomorpha and Ambulacraria sister relationship in the animal tree of life. *Current Biology*, *32*, 5180–5188.e3. https://doi.org/10.1016/j.cub.2022.10.036

Ocaña-Pallarès, E., Williams, T. A., López-Escardó, D., Arroyo, A. S., Pathmanathan, J. S., Bapteste, E., Tikhonenkov, D. V., Keeling, P. J., Szöllősi, G. J., & Ruiz-Trillo, I. (2022). Divergent genomic trajectories predate the origin of animals and fungi. *Nature*, *609*, 747–753. https://doi.org/10.1038/s41586-022-05110-4

One Thousand Plant Transcriptomes Initiative. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, *574*, 679–685. https://doi.org/10.1038/s41586-019-1693-2

Parey, E., Louis, A., Montfort, J., Bouchez, O., Roques, C., Iampietro, C., Lluch, J., Castinel, A., Donnadieu, C., Desvignes, T., Floi Bucao, C., Jouanno, E., Wen, M., Mejri, S., Dirks, R., Jansen, H., Henkel, C., Chen, W. J., Zahm, M., … Guiguen, Y. (2023). Genome structures resolve the early diversification of teleost fishes. *Science (New York, N.Y.)*, *379*, 572–575. https://doi.org/10.1126/science.abq4257

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*, *9*, e1000602. https://doi.org/10.1371/journal.pbio.1000602

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., da Silva, C., Wincker, P., le Guyader, H., Leys, S., Jackson, D. J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G., & Manuel, M. (2009). Phylogenomics revives traditional views on deep animal relationships. *Current Biology*, *19*, 706–712. https://doi.org/10.1016/j.cub.2009.02.052

Philippe, H., Vienne, D. M. D., Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, Retrieved from: https://doi.org/10.5852/ejt.2017.283

Phillips, M. J., Delsuc, F., & Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, *21*, 1455–1458. https://doi.org/10.1093/molbev/msh137

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

Raghavan, V., Kraft, L., Mesny, F., & Rigerte, L. (2022). A simple guide to *de novo* transcriptome assembly and annotation. *Briefings in Bioinformatics*, *23*, bbab563. https://doi.org/10.1093/bib/bbab563

Rodríguez-Ezpeleta, N., Brinkmann, H., Burger, G., Roger, A. J., Gray, M. W., Philippe, H., & Lang, B. F. (2007). Toward resolving the eukaryotic tree: The phylogenetic positions of Jakobids and Cercozoans. *Current Biology*, *17*, 1420–1425. https://doi.org/10.1016/j.cub.2007.07.036

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

Sayyari, E., & Mirarab, S. (2018). Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes*, *9*, 132. https://doi.org/10.3390/genes9030132

Schultz, D. T., Haddock, S. H. D., Bredeson, J. V., Green, R. E., Simakov, O., & Rokhsar, D. S. (2023). Ancient gene linkages support ctenophores as sister to other animals. *Nature*, *618*, 110–117. Retrieved from: https://doi.org/10.1038/s41586-023-05936-6

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., Haase, M. A. B., Wisecaver, J. H., Wang, M., Doering, D. T., Boudouris, J. T., Schneider, R. M., Langdon, Q. K., Ohkuma, M., Endoh, R., Takashima, M., Manabe, R.-I., Čadež, N., Libkind, D., & Rokas, A. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. *Cell*, *175*, 1533–1545.e20. https://doi.org/10.1016/j.cell.2018.10.023

Shen, X.-X., Salichos, L., & Rokas, A. (2016). A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. *Genome Biology and Evolution*, *8*, 2565–2580. https://doi.org/10.1093/gbe/evw179

Sierra-Patev, S., Min, B., Naranjo-Ortiz, M., Looney, B., Konkel, Z., Slot, J. C., Sakamoto, Y., Steenwyk, J. L., Rokas, A., Carro, J., Camarero, S., Ferreira, P., Molpeceres, G., Ruiz-Dueñas, F. J., Serrano, A., Henrissat, B., Drula, E., Hughes, K. W., Mata, J. L., & Hibbett, D. (2023). A global phylogenomic analysis of the shiitake genus *Lentinula*. *Proceedings of the National Academy of Sciences*, *120*, e2214076120. https://doi.org/10.1073/pnas.2214076120

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*, 135–145. https://doi.org/10.1002/pro.3290

Siu-Ting, K., Torres-Sánchez, M., San Mauro, D., Wilcockson, D., Wilkinson, M., Pisani, D., O'Connell, M. J., & Creevey, C. J. (2019). Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. *Molecular Biology and Evolution*, *36*, 1344–1356. https://doi.org/10.1093/molbev/msz067

Smith, S. A., Brown, J. W., & Walker, J. F. (2018). So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. *PLOS ONE*, *13*, e0197433. https://doi.org/10.1371/journal.pone.0197433

Steenwyk, J., & King, N. (2023). From genes to genomes: Opportunities and challenges for synteny-based phylogenies. *Preprints*, Retrieved from: http://doi.org/10.20944/preprints202309.0495.v1

Steenwyk, J. L., Balamurugan, C., Raja, H. A., Gonçalves, C., Li, N., Martin, F., Berman, J., Oberlies, N. H., Gibbons, J. G., Goldman, G. H., Geiser, D. M., Houbraken, J., Hibbett, D. S., & Rokas, A. (2024). Phylogenomics reveals extensive misidentification of fungal strains from the genus *Aspergillus*. *Microbiology Spectrum*, *12*, e03980–23. https://doi.org/10.1128/spectrum.03980-23

Steenwyk, J. L., Buida, T. J., Gonçalves, C., Goltz, D. C., Morales, G., Mead, M. E., LaBella, A. L., Chavez, C. M., Schmitz, J. E., Hadjifrangiskou, M., Li, Y., & Rokas, A. (2022a). BioKIT: A versatile toolkit for processing and analyzing diverse types of sequence data. *Genetics*, *221*, iyac079. https://doi.org/10.1093/genetics/iyac079

Steenwyk, J. L., Buida, T. J., Labella, A. L., Li, Y., Shen, X.-X., & Rokas, A. (2021). PhyKIT: A broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics*, *37*, 2325–2331. https://doi.org/10.1093/bioinformatics/btab096

Steenwyk, J. L., Buida, T. J., Li, Y., Shen, X.-X., & Rokas, A. (2020). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLOS Biology*, *18*, e3001007. https://doi.org/10.1371/journal.pbio.3001007

Steenwyk, J. L., Goltz, D. C., Buida, T. J., Li, Y., Shen, X.-X., & Rokas, A. (2022b). OrthoSNAP: A tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees. *PLOS Biology*, *20*, e3001827. https://doi.org/10.1371/journal.pbio.3001827

Steenwyk, J. L., Li, Y., Zhou, X., Shen, X.-X., & Rokas, A. (2023a). Incongruence in the phylogenomics era. *Nature Reviews Genetics*, *24*, 834–850. Retrieved from: https://doi.org/10.1038/s41576-023-00620-x

Steenwyk, J. L., Phillips, M. A., Yang, F., Date, S. S., Graham, T. R., Berman, J., Hittinger, C. T., & Rokas, A. (2022c). An orthologous gene coevolution network provides insight into eukaryotic cellular and genomic structure and function. *Science Advances*, *8*, eabn0105. https://doi.org/10.1126/sciadv.abn0105

Steenwyk, J. L., & Rokas, A. (2021a). ggpubfigs: Colorblind-friendly color palettes and ggplot2 graphic system extensions for publication-quality scientific figures. *Microbiology Resource Announcements*, *10*, e00871–21. https://doi.org/10.1128/MRA.00871-21

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

Steenwyk, J. L., & Rokas, A. (2019). Treehouse: A user-friendly application to obtain subtrees from large phylogenies. *BMC Research Notes*, *12*, 541. https://doi.org/10.1186/s13104-019-4577-5

Steenwyk, J. L., Rokas, A., & Goldman, G. H. (2023b). Know the enemy and know yourself: Addressing cryptic fungal pathogens of humans and beyond. *PLOS Pathogens*, *19*, e1011704. https://doi.org/10.1371/journal.ppat.1011704

Struck, T. H. (2014). TreSpEx—detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolutionary Bioinformatics*, *10*, EBO.S14239. https://doi.org/10.4137/EBO.S14239

Susko, E., & Roger, A. J. (2021). Long branch attraction biases in phylogenetics. *Systematic Biology*, *70*, 838–843. https://doi.org/10.1093/sysbio/syab001

Susko, E., & Roger, A. J. (2007). On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution*, *24*, 2139–2150. https://doi.org/10.1093/molbev/msm144

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., & Dessimoz, C. (2015). Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology*, *64*, 778–791. https://doi.org/10.1093/sysbio/syv033

Telford, M. J., Lowe, C. J., Cameron, C. B., Ortega-Martinez, O., Aronowicz, J., Oliveri, P., & Copley, R. R. (2014). Phylogenomic analysis of echinoderm class relationships supports Asterozoa. *Proceedings of the Royal Society B: Biological Sciences*, *281*, 20140479. https://doi.org/10.1098/rspb.2014.0479

Thornton, J. W., & DeSalle, R. (2000). Gene family evolution and homology: Genomics meets phylogenetics. *Annual Review of Genomics and Human Genetics*, *1*, 41–73. https://doi.org/10.1146/annurev.genom.1.1.41

Turnbull, R., Steenwyk, J. L., Mutch, S. J., Scholten, P., Salazar, V. W., Birch, J. L., & Verbruggen, H. (2023). OrthoFlow: Phylogenomic analysis and diagnostics with one command. In review. Retrieved from: https://www.researchsquare.com/article/rs-3699210/v1

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*, 543–548. https://doi.org/10.1093/molbev/msx319

Zhang, C., Scornavacca, C., Molloy, E. K., & Mirarab, S. (2020). ASTRAL-Pro: Quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution*, *37*, 3292–3307. https://doi.org/10.1093/molbev/msaa139

Zhao, D., Liu, J., & Yu, T. (2023). Protocol for transcriptome assembly by the TransBorrow algorithm. *Biology Methods and Protocols*, *8*, bpad028. https://doi.org/10.1093/biomethods/bpad028