Bioinformatics, 37(16), 2021, 2325–2331
doi: 10.1093/bioinformatics/btab096
Advance Access Publication Date: 9 February 2021
Original Paper

OXFORD

Phylogenetics

PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data

Jacob L. Steenwyk [1,*], Thomas J. Buida III [2], Abigail L. Labella [1], Yuanning Li [1], Xing-Xing Shen [3] and Antonis Rokas [1,*]

[1]Department of Biological Sciences, Vanderbilt University, VU Station B #35-1634, Nashville, TN 37235, USA, [2]9 City Place #312, Nashville, TN 37209, USA and [3]Ministry of Agriculture Key Lab of Molecular Biology of Crop Pathogens and Insects, Institute of Insect Sciences, Zhejiang University, Hangzhou 310058, China

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on December 6, 2020; revised on January 13, 2021; editorial decision on February 1, 2021; accepted on February 5, 2021

Abstract

Motivation: Diverse disciplines in biology process and analyze multiple sequence alignments (MSAs) and phylogenetic trees to evaluate their information content, infer evolutionary events and processes and predict gene function. However, automated processing of MSAs and trees remains a challenge due to the lack of a unified toolkit. To fill this gap, we introduce PhyKIT, a toolkit for the UNIX shell environment with 30 functions that process MSAs and trees, including but not limited to estimation of mutation rate, evaluation of sequence composition biases, calculation of the degree of violation of a molecular clock and collapsing bipartitions (internal branches) with low support.

Results: To demonstrate the utility of PhyKIT, we detail three use cases: (1) summarizing information content in MSAs and phylogenetic trees for diagnosing potential biases in sequence or tree data; (2) evaluating gene–gene covariation of evolutionary rates to identify functional relationships, including novel ones, among genes and (3) identify lack of resolution events or polytomies in phylogenetic trees, which are suggestive of rapid radiation events or lack of data. We anticipate PhyKIT will be useful for processing, examining and deriving biological meaning from increasingly large phylogenomic datasets.

Availability and implementation: PhyKIT is freely available on GitHub (https://github.com/JLSteenwyk/PhyKIT), PyPi (https://pypi.org/project/phykit/) and the Anaconda Cloud (https://anaconda.org/JLSteenwyk/phykit) under the MIT license with extensive documentation and user tutorials (https://jlsteenwyk.com/PhyKIT).

Contact: jacob.steenwyk@vanderbilt.edu or antonis.rokas@vanderbilt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Multiple sequence alignments (MSAs) and phylogenetic trees are widely used in numerous disciplines, including bioinformatics, evolutionary biology, molecular biology and structural biology. As a result, the development of user-friendly software that enables biologists to process and analyze MSAs and phylogenetic trees is an active area of research (Kapli *et al.*, 2020).

In recent years, numerous methods have proven useful for diagnosing potential biases and inferring biological events in genome-scale phylogenetic (or phylogenomic) datasets. For example, methods that evaluate sequence composition biases in MSAs (Phillips and Penny, 2003), signatures of clock-like evolution in phylogenetic trees (Liu *et al.*, 2017), phylogenetic treeness (Lanyon, 1988; Phillips and Penny, 2003), taxa whose long branches may cause variation in their placement on phylogenetic trees (Struck, 2014), and others have assisted in summarizing the information content in phylogenomic datasets and improved phylogenetic inference (Doyle *et al.*, 2015; Felsenstein, 1978; Liu *et al.*, 2017; Philippe *et al.*, 2011; Salichos and Rokas, 2013; Smith *et al.*, 2018; Walker *et al.*, 2019).

Other methodological innovations include identifying significant gene–gene covariation of evolutionary rates, which has been shown to accurately and sensitively identify genes that have shared functions, are coexpressed, and/or are part of the same multimeric complexes (Clark *et al.*, 2012; Sato *et al.*, 2005). Furthermore, gene–gene covariation serves as a powerful evolution-based genetic screen for predicting gene function (Brunette *et al.*, 2019). Lastly, a recently developed method has enabled the identification of unresolved internal branches or polytomies in species trees (One Thousand Plant Transcriptomes Initiative, 2019; Sayyari and Mirarab, 2018); such branches can stem from rapid radiation events or from lack of data (Rokas and Carroll, 2006).

Despite the wealth of information in MSAs and phylogenetic trees, there is a dearth of tools that enable researchers to conduct these analyses in a unified framework. For example, to utilize the functions mentioned in the previous paragraphs, a combination of web-server applications, 'hard-coded' scripts available through numerous repositories and supplementary material, standalone software and/or extensive programming in languages including R, Python or C is currently required (Brown *et al.*, 2017; Cock *et al.*, 2009; Hernández *et al.*, 2018; Huerta-Cepas *et al.*, 2016; Junier and Zdobnov, 2010; Kück and Longo, 2014; One Thousand Plant Transcriptomes Initiative, 2019; Revell, 2012; Struck, 2014; Talevich *et al.*, 2012; Wolfe and Clark, 2015). As a result, integrating these functions into bioinformatic pipelines can be challenging, reducing their accessibility to the scientific community.

To facilitate the integration of these methods into bioinformatic pipelines, we introduce PhyKIT, a UNIX shell toolkit with 30 functions (Supplementary Table S1) that have broad utility for analyzing and processing MSAs and phylogenetic trees. Exemplary functions implemented in PhyKIT include measuring topological similarity of phylogenetic trees, creating codon-based MSAs, concatenating sets of MSAs into phylogenomic datasets, editing and/or viewing alignments and phylogenetic trees and identifying putatively spurious homologs in MSAs. We highlight three uses of PhyKIT: (1) calculating diverse statistics that summarize the information content and potential biases (e.g. sequence- or phylogeny-based biases) in MSAs and phylogenetic trees; (2) creating a gene–gene covariation network and (3) inferring the presence of polytomies from phylogenomic data. The diverse functions implemented in PhyKIT will likely be of interest to bioinformaticians, molecular biologists, evolutionary biologists and others.

2 Materials and methods

PhyKIT is a command line tool for the UNIX shell environment written in the Python programming language (https://www.python.org/). PhyKIT requires few dependencies [Biopython (Cock *et al.*, 2009) and SciPy (Virtanen *et al.*, 2020)] making it user-friendly to install and integrate into existing bioinformatic pipelines. Online documentation of PhyKIT comes complete with tutorials that detail use cases for various functions. Lastly, PhyKIT is modularly designed to allow straightforward integration of additional functions in future versions.

PhyKIT has 30 different functions that help process and analyze MSAs and phylogenetic trees (Supplementary Table S1). The 30 functions can be grouped into broad categories that assist in conducting analyses of MSAs and phylogenies or in processing/editing them. For example, 'analysis' functions help examine information content biases, gene–gene covariation and polytomies in phylogenomic datasets; 'processing/editing' functions help prune tips from phylogenies, collapse poorly supported bipartitions in phylogenetic trees, concatenate sets of MSAs into a single data matrix or create codon-based alignments from protein alignments and their corresponding nucleotide sequences.

Detailed information about each one of PhyKIT's functions and tutorials for using the software can be found in the online documentation (https://jlsteenwyk.com/PhyKIT). Here, we focus on three specific groups of functions implemented in PhyKIT that enable researchers to summarize information content in phylogenomic datasets, create gene–gene evolutionary rate covariation networks and identifying polytomies in phylogenomic data.

2.1 Evaluating information content and biases in phylogenomic datasets

MSAs and phylogenetic trees are frequently examined to evaluate their information content and potential biases in characteristics such as sequence composition or branch lengths (Doyle *et al.*, 2015; Liu *et al.*, 2017; Phillips and Penny, 2003; Philippe *et al.*, 2011; Shenet *al.*, 2016a; Smith *et al.*, 2018; Struck, 2014). PhyKIT implements numerous functions for doing so. We demonstrate the application of 14 functions:

(1)*Alignment length*. The length of anMSA, which is associated with robust bipartition support and tree accuracy (Shenet *al.*, 2016a; Walker *et al.*, 2019).

(2)*Alignment length with no gaps*. The length of anMSA after excluding sites with gaps, which is associated with robust bipartition support and tree accuracy (Shenet *al.*, 2016a).

(3)*Degree of violation of a molecular clock (DVMC)*. A metric used to determine the clock-like evolution of a gene using the standard deviation of branch lengths for a single gene tree (Liu *et al.*, 2017). DVMC is calculated using the following formula:

$$\mathrm{DVMC} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(i_j - -i)^2},$$

where N represents the number of tips in a phylogenetic tree, $i_j$being the distance between the root of the tree and species j, and $-i$ represents the average root to tip distance. DVMC can be used to identify genes with clock-like evolution for divergence time estimation (Liu *et al.*, 2017).

(4)*Internal branch lengths*. Summary statistics of internal branch lengths in a phylogenetic tree are reported including mean, median, 25thpercentile, 75th percentile, minimum, maximum, standard deviation and variance values. Examination of internal branch lengths is useful in evaluating phylogenetic tree shape.

(5)*Long branch score*. A metric that examines the degree of taxon-specific long branch attraction (Struck, 2014; Weigert *et al.*, 2014). Long branch scores of individual taxa are calculated using the following formula:

$$LB_i = \left(\frac{-PD_i}{-PD_{all}} - 1\right) \times 100,$$

where $-PD_i$ represents the average pairwise patristic distance of taxon i to all other taxa, $-PD_{all}$ represents the average patristic distance across all taxa and LB_i represents the long branch score of taxon i. Long branch scores can be used to evaluate heterogeneity in tip-to-root distances and identify taxa that may be susceptible to long branch attraction.

(6)*Pairwise identity*. Pairwise identity is a crude approximation of the evolutionary rate of a gene and is calculated by determining the average number of sites in an MSA that are the same character state between all pairwise combinations of taxa. This can be used to group genes based on their evolutionary rates (e.g. faster-evolving genes versus slower-evolving ones) (Chen *et al.*, 2017).

(7)*Patristic distances*. Patristic distances refer to all distances between all pairwise combinations of tips in a phylogenetic tree (Fourment and Gibbs, 2006), which can be used to evaluate the rate of evolution in gene trees or taxon sampling density in species trees.

(8)*Parsimony-informative sites*. Parsimony-informative sites are those sites in an MSA that have a least two character states (excluding gaps) that occur at least twice (Kumar *et al.*, 2016); the number of parsimony-informative sites is associated with robust bipartition support and tree accuracy (Shenet *al.*, 2016a; Steenwyk *et al.*, 2020).

(9)*Variable sites*. Variable sites are those sites in an MSA that contain at least two different character states (excluding gaps) (Kumar *et al.*, 2016); the number of variable sites is associated with robust bipartition support and tree accuracy (Shenet *al.*, 2016a).

(10)*Relative composition variability*. Relative composition variability is the average variability in the sequence composition among taxa in an MSA. Relative composition variability is calculated using the following formula:

$$\text{Relative composition variability} = \sum\nolimits_{i=1}^{c}\sum\nolimits_{j=1}^{n}\frac{|c_{ij} - -c_i|}{s \times n},$$

where c is the number of different character states per sequence type, n is the number of taxa in an MSA, $c_{ij}$is the number of occurrences of the ith character state for the jth taxon, $-c_i$ is the average number of the ith c character state across n taxa and s refers to the total number of sites (characters) in an MSA. Relative composition variability can be used to evaluate potential sequence composition biases in MSAs, which in turn violate assumptions of site composition homogeneity in standard models of sequence evolution (Phillips and Penny, 2003).

(11)*Saturation*. Saturation refers to when an MSA contains many sites that have experienced multiple substitutions in individual taxa. Saturation is estimated from the slope of the regression line between patristic distances and pairwise identities. Saturated MSAs have reduced phylogenetic information and can result in issues of long branch attraction (Lake, 1991; Philippe *et al.*, 2011).

(12)*Total tree length*. Total tree length refers to the sum of internal and terminal branch lengths and is calculated using the following formula:

$$\text{total tree length} = \sum\nolimits_{i=1}^{a} l_i + \sum\nolimits_{j=1}^{b} l_j,$$

where $l_i$is the branch length of the ith branch of a internal branches and l_j is the branch length of the jth branch of b terminal branches. Total tree length measures the inferred total amount or rate of evolutionary change in a phylogenetic tree.

(13)*Treeness*. Treeness (also referred to as stemminess) is a measure of the inferred relative amount or rate of evolutionary change that has taken place on internal branches of a phylogenetic tree (Lanyon, 1988; Phillips and Penny, 2003) and is calculated using the following formula:

$$\text{treeness} = \sum\nolimits_{u=1}^{b} \frac{l_u}{l_t}$$

where l_u is the branch length of the uth branch of b internal branches and l_t refers to the total branch length of the phylogenetic tree. Treeness can be used to evaluate how much of the total tree length is observed among internal branches.

(14)*Treeness divided by relative composition variability*. This function combines two metrics to measure both composition bias and other biases that may negatively influence phylogenetic inference. High treeness divided by relative composition variability values have been shown to be less susceptible to sequence composition biases and are associated with robust bipartition support and tree accuracy (Phillips and Penny, 2003; Shen*et al.*, 2016a).

2.2 Calculating gene–gene evolutionary rate covariation or coevolution

Genes that share similar rates of evolution through speciation events (or coevolve) tend to have similar functions, expression levels, or are parts of the same multimeric complexes (Clark *et al.*, 2012; Sato *et al.*, 2005). Thus, identifying significant coevolution between genes (i.e. identifying genes that are significantly correlated in their evolutionary rates across speciation events) can be a powerful evolution-based screen to determine gene function (Brunette *et al.*, 2019).

To measure gene–gene evolutionary rate covariation, PhyKIT implements the mirror tree method (Pazos and Valencia, 2001; Sato *et al.*, 2005), which examines whether two trees have correlated branch lengths. Specifically, PhyKIT calculates the Pearson correlation coefficient between branch lengths in two phylogenetic trees that share the same tips and topology. To account for differences in taxon representation between the two trees, PhyKIT first automatically determines which taxa are shared and prunes one or both such that the same set of taxa is present in both trees. PhyKIT requires that the two input trees have the same topology, which is typically the species tree topology inferred from whole genome or proteome data. Thus, the user will typically first estimate a gene's branch lengths by constraining the topology to match that of the species tree. When running this function, users should be aware that many biological factors, such as horizontal transfer (Doolittle and Bapteste, 2007), incomplete lineage sorting (Degnan and Salter, 2005) and introgression/hybridization (Sang and Zhong, 2000), can lead to gene histories that deviate from the species tree. In these cases, constraining a gene's history to match that of a species may lead to errors in the covariation analysis.

Due to factors including time since speciation and mutation rate, correlations between uncorrected branch lengths result in a high frequency of false positive correlations (Chikina *et al.*, 2016; Clark *et al.*, 2012; Sato *et al.*, 2005). To ameliorate the influence of these factors, PhyKIT first transforms branch lengths into relative rates. To do so, branch lengths are corrected by dividing the branch length in the gene tree by the corresponding branch length in the species tree. Previous work revealed that one or a few outlier branch length values can be responsible for false positive correlations and should be removed prior to analysis (Clark *et al.*, 2012). Thus, PhyKIT removes outlier data points defined as having corrected branch lengths greater than five (i.e. removing gene tree branch lengths that are five or more times greater than their corresponding species tree branch lengths). Lastly, values are converted into relative rates using a Z-transformation. The resulting relative rates are used when calculating Pearson correlation coefficients.

2.3 Identifying polytomies in phylogenomic data

Rapid radiations or diversification events have occurred throughout the tree of life including among mammals, birds, plants and fungi (Jarvis *et al.*, 2014; Li *et al.*, 2020; Liu *et al.*, 2017; One Thousand Plant Transcriptomes Initiative, 2019). Polytomies correspond to internal branches whose length is 0 (or statistically indistinguishable from 0) and can be driven either by biological (e.g. rapid radiations) or analytical (e.g. low amount of data) factors. Thus, polytomies are useful for inferring rapid radiation or diversification events and exploring incongruence in phylogenies (Li *et al.*, 2020; One Thousand Plant Transcriptomes Initiative, 2019; Sayyari and Mirarab, 2018).

To identify polytomies, a modified approach to a previous strategy was implemented (Sayyari and Mirarab, 2018). More specifically, the support for three alternative topologies is calculated among all gene trees from a phylogenomic dataset. For example, in species tree *((A, B),C), D);*, if examining the presence of a polytomy at the ancestral bipartition of tips *A*, *B* and *C*, PhyKIT will determine the number of gene trees that support *((A, B),C);*, *((A, C),B);* and *((B, C),A);* using the rooted gene trees provided by the user. Equal support for the three topologies (i.e. the presence of a polytomy) among a set of gene trees is assessed using a Chi-squared test. Failing to reject the null hypothesis is indicative of a polytomy (Sayyari and Mirarab, 2018). Note that this approach is distinct from the approach of Sayyari and Mirarab to identify polytomies because PhyKIT uses a gene-based signal rather than a quartet-based signal. The difference between the two methods is that each gene contributes equally to the inference of a polytomy when a gene-based signal is used, whereas genes with greater taxon representation (which contain a greater number of quartets) will contribute a greater signal during polytomy identification when a quartet-based signal is used. From a technical perspective, both approaches are simple to implement and require only a single line of code in the commandline.

3 Results

We outline three example uses of PhyKIT: (1) summarizing information content and identifying potential biases in animal, plant, yeast and filamentous fungal phylogenomic datasets (Laumer *et al.*, 2019; One Thousand Plant Transcriptomes Initiative, 2019; Shen*et al.*, 2016b; Steenwyk *et al.*, 2019), (2) constructing a network of significant gene–gene covariation, which reveals genes of shared functions from empirical data spanning ~550 million years of evolution among fungi (Shen *et al.*, 2020) and (3) illustrating how to identify polytomies using simulated and empirical data (Steenwyk *et al.*, 2019).

3.1 Summarizing information content and biases in phylogenomic data

Examining information content in phylogenomic datasets can help diagnose potential biases that stem from low signal-to-noise ratios, multiple

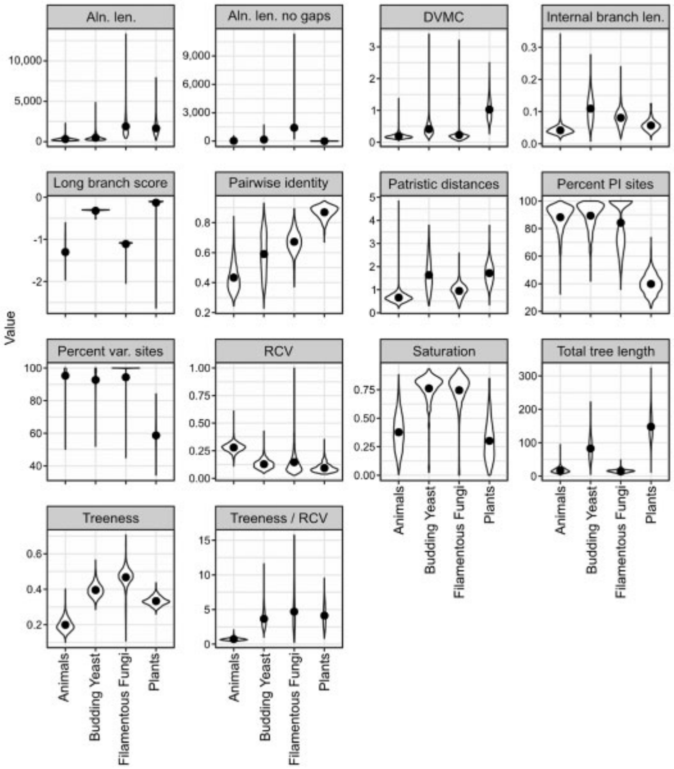

Fig. 1. Summary of information content in four empirical phylogenomic datasets.Fourteenexemplary metrics implemented in PhyKIT help summarize the information content and identify potential biases in phylogenomic datasets. Each graph displays a violin plot with a black point representing the mean. Error bars indicate one standard error above and below the mean; however, these are difficult to see in nearly all graphs because they were often near the mean. Abbreviations are as follows: Aln. len.: alignment length; Aln. len. no gaps: alignment length excluding sites with gaps; DVMC: degree of violation of a molecular clock; Internal branch len.: average internal branch length; patristic distances: average patristic distance in a gene tree; percent PI sites: percentage of parsimony-informative sites in an MSA; percent var. sites: percentage of variable sites in an MSA; RCV: relative composition variability

substitutions, nonclock-like evolution and other biological or analytical factors. To demonstrate the utility of PhyKIT to summarize the information content in phylogenomic datasets, we calculated 14 different metrics known to help diagnose potential biases in phylogenomic datasets or be associated with accurate and well supported phylogenetic inferences (Doyle *et al.*, 2015; Felsenstein, 1978; Liu *et al.*, 2017; Phillips and Penny, 2003; Philippe *et al.*, 2011; Shenet *al.*, 2016a; Smith *et al.*, 2018; Struck, 2014) using four empirical phylogenomic datasets from animals (201 tips; 2891 genes) (Laumer *et al.*, 2019), budding yeast (332 taxa; 2408 genes) (Shen *et al.*, 2018), filamentous fungi (93 taxa; 1668 genes) (Steenwyk *et al.*, 2019) and plants (1124 taxa; 403 genes) (One Thousand Plant Transcriptomes Initiative, 2019) (Figure 1 and Supplementary Table S1).

Examination of the distributions of the values of the 14 different metrics revealed inter- and intra-dataset heterogeneity (Fig. 1). For example, inter-dataset heterogeneity was observed among animal and plant datasets, which had the lowest and highest average pairwise identity across alignments, respectively; intra-dataset heterogeneity was observed in the uniform distribution of pairwise identities in the budding yeast datasets. Similarly, inter-dataset heterogeneity was observed in estimates of saturation where the budding yeast and filamentous fungal MSAs were less saturated by multiple substitutions than the plant and animal datasets; intra-data heterogeneity was also observed in all four datasets. Varying degrees of inter- and intra-dataset heterogeneity was observed for other information content statistics, which may be due biological (e.g. mutation rate) or analytical factors (e.g. taxon sampling, distinct alignment, trimming and tree inference strategies).

In summary, PhyKIT is useful for examining the information content of phylogenomic datasets. For example, the generation of different phylogenomic data submatrices by selecting subsets of genes or taxa with certain properties (e.g. retention of genes with the highest numbers of parsimony-informative sites or following removal of taxa with high long branch scores) can facilitate the exploration of the robustness of species tree inference or estimating time since divergence (Li *et al.*, 2020; Liu *et al.*, 2017; Salichos and Rokas, 2013; Shen *et al.*, 2018,2020; Steenwyk *et al.*, 2019; Walker *et al.*, 2019).

3.2 A network of gene–gene covariation reveals neighborhoods of genes with shared function

Genes with similar evolutionary histories often have shared functions, are coexpressed or are parts of the same multimeric complexes (Clark *et al.*, 2012; Sato *et al.*, 2005). Using PhyKIT, we examined gene–gene covariation using 815 genes spanning 1107 genomes and ~563 million years of evolution among fungi (Shen *et al.*, 2020). By examining 331 705 pairwise combinations of genes, we found 298 strong signatures of gene–gene covariation (defined as $r > 0.825$). The two genes with the strongest signatures of covariation were *SEC7* and *TAO3* ($r = 0.87$), suggesting that their protein products have similar or shared functions. Supporting this hypothesis, Sec7 p contributes to cell-surface growth in the model yeast *Saccharomyces cerevisiae* (Novick and Schekman, 1979) and genes with the Sec7 domain are transcriptionally coregulated with yeast-hyphal switches in the human pathogen *Candida albicans* (Song *et al.*, 2008). Similarly, Tao3p in both *S. cerevisiae* and *C. albicans* is part of a RAM signaling network, which controls hyphal morphogenesis, polarized growth and cell-cycle related processes including cell separation, cell proliferation and phase transitions (Bogomolnaya *et al.*, 2006; Song *et al.*, 2008).

Complex relationships of gene–gene covariation can be visualized as a network (Fig. 2). Examination of network neighborhoods identified groups of genes that have shared functions and are parts of the same multimeric complexes. For example, the proteins encoded by *NDC80* and *NUF2* are part of the same kinetochore-associated complex termed the NDC80 complex—which is required for efficient

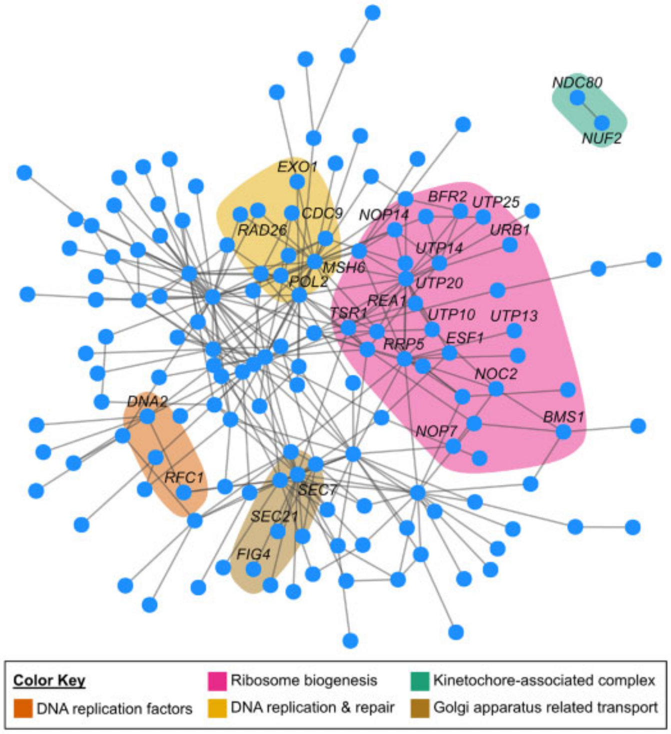

Fig. 2. Gene–gene covariation network inferred from ∼550 million years of evolution across 1107 fungi.A network of significant gene–gene coevolution identifies network neighborhoods representative of associated functional categories. For example, the NDC80 and NUF2 genes (toward the top right of the network) were identified to be significantly coevolving with one another ($r = 0.84$, $P < 0.01$, Pearson's correlation test); they both encode proteins that are part of the same multimeric kinetochore-associated complex (green). Similarly, genes that are DNA replication factors (orange), contribute to DNA replication and repair processes (yellow), participate in Golgi apparatus-related transport (brown) or ribosome biogenesis (pink) were found to be neighbors in the network. Network visualization was done with the igraph package, v1.2.4.2 (Hunter and Cohen, 2007), in R, v3.6.2 (https://www.r-project.org/)

mitosis (Sundin *et al.*, 2011)—and significantly covary with one another ($r = 0.84$). Similarly, multiple genes that encode proteins involved in DNA replication and repair (i.e. *POL2*, *MSH6*, *RAD26*, *CDC9* and *EXO1*) were part of the same network neighborhood, consistent with previous work suggesting an intimate interplay between DNA replication and multiple DNA repair pathways (Boiteux and Jinks-Robertson, 2013; Lujan *et al.*, 2012; Tsubouchi and Ogawa, 2000). Other network neighborhoods of genes with shared function such as ribosome biogenesis, Golgi apparatus-related transport and control of DNA replication were identified (Fig. 2).

Taken together, these results indicate PhyKIT is a useful tool for evaluating gene–gene covariation and predicting genes' functions (Brunette *et al.*, 2019; Clark *et al.*, 2012; Sato *et al.*, 2005). Thus, we anticipate PhyKIT will be helpful for evaluating gene–gene covariation and conducting evolution-based screens for gene functions across the tree of life.

3.3 Identifying polytomies in phylogenomic datasets

Rapid radiations or diversification events have occurred throughout the tree of life (Jarvis *et al.*, 2014; Li *et al.*, 2020; Liu *et al.*, 2017; One Thousand Plant Transcriptomes Initiative, 2019). One approach to identifying rapid radiations is by testing for the existence of polytomies in species trees (Li *et al.*, 2020; One Thousand Plant Transcriptomes Initiative, 2019; Sayyari and Mirarab, 2018). Polytomies can also arise when the amount of data at hand is insufficient for resolution (Walsh *et al.*, 1999). To demonstrate the utility of PhyKIT to identify polytomies, we examined the ability of our approach to identify a simulated polytomy (Fig. 3A). PhyKIT was able to conservatively identify the simulated polytomy demonstrating the efficacy of our approach.

We next examined if there is evidence of polytomies in the evolutionary history of filamentous fungi from the genera *Aspergillus* and *Penicillium*. We examined three branches. The first two branches—one dating back ∼110 million years ago (Fig. 3B) and another dating back ∼25 million years ago (Fig. 3C)—were not polytomies. In contrast, examination of a ∼60 million-year-old branch involving *Lanata-divaricata, Citrina*and *Exilicaulis* (Fig. 3D), which are major lineages (or sections) in the genus *Penicillium*, was consistent with a polytomy. Given the large number of gene trees used in our analysis ($n = 1668$), these results are consistent with a rapid radiation or diversification event in the history of *Penicillium* species.

In summary, these results suggest that PhyKIT is useful in identifying polytomies in simulated and empirical datasets. More broadly, these results support the notion that polytomies can be used to identify rapid radiation events. Beyond polytomy identification, PhyKIT can be used for exploring incongruence in phylogenies by calculating gene-support frequencies. Calculations of gene-based support among different topologies can be used in diverse applications, including identifying putative introgression/hybridization events and conducting phylogenetically-based genome-wide association (PhyloGWAS) studies (Pease *et al.*, 2016; Steenwyk *et al.*, 2019).

4 Conclusion

We developed PhyKIT, a comprehensive toolkit for processing and analyzing MSAs and trees in phylogenomic datasets. Executing functions implemented in PhyKIT would otherwise require extensive programming, multiple software and/or web-based applications (Supplementary Table S1); thus, PhyKIT offers users a way to streamline approaches and pipelines by relying on only one software. PhyKIT is freely available on GitHub (https://github.com/JLSteenwyk/PhyKIT), PyPi (https://pypi.org/project/phykit/) and the

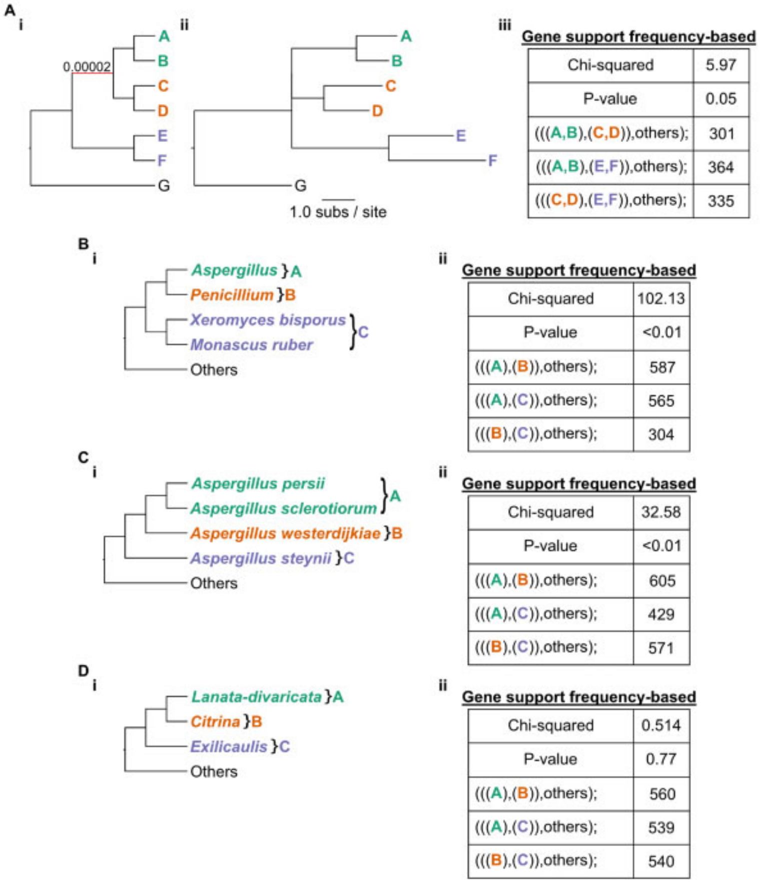

Fig. 3. Identifying polytomies from phylogenomic data. (**A**$_{\mathrm{i}}$) A cladogram of a simulated species phylogeny with tip names A–G. The red branch has a very short branch length of 2×10^{-5} substitutions per site. (**A**$_{\mathrm{ii}}$) Phylogram of the same phylogeny shows that all other branches are much longer ($\geq$1.0 substitutions per site). (**A**$_{\mathrm{iii}}$) After reconstructing the evolutionary history from 1000 alignments simulated from the phylogeny in A$_{\mathrm{ii}}$, the hypothesis of a polytomy was tested using gene-support frequencies for three alternative rooted topologies defined by the clades of green, orange and purple taxa. Failure to reject the null hypothesis of equal support among genes for each topology is indicative of a polytomy ($\chi^2 = 5.97$, P-value $= 0.05$, Chi-squared test). (**B–D**) The same approach was then used to examine if there is evidence for a polytomy at three different branches in a phylogeny of filamentous fungi. (D) Support for a polytomy ($\chi^2 = 0.514$, P-value $= 0.77$, Chi-squared test) was observed for the relationships between three different sections of *Penicillium* fungi. These results demonstrate the utility of gene-support frequencies for evaluating polytomies and examining incongruence in phylogenomic datasets

Anaconda Cloud (https://anaconda.org/JLSteenwyk/phykit) under the MIT license with extensive documentation and user tutorials (https://jlsteenwyk.com/PhyKIT). PhyKIT is a fast and flexible toolkit for the UNIX shell environment, which allows it to be easily integrated into bioinformatic pipelines. We anticipate PhyKIT will be of interest to biologists from diverse disciplines and with varying degrees of experience in analyzing MSAs and phylogenies. In particular, PhyKIT will likely be helpful in addressing one of the greatest challenges in biology, building, understanding and deriving meaning from the tree of life.

Acknowledgement

We thank the Rokas lab for helpful discussion and feedback.

Funding

J.L.S. and A.R. were funded by the Howard Hughes Medical Institute through the James H. Gilliam Fellowships for Advanced Study program. X.X.S. was supported by National Natural Science Foundation of China [32071665] and the Fundamental Research Funds for the Central Universities [2020QNA6019].

Conflict of Interest: none declared.

References

Bennett,D.J. *et al.* (2017) treeman: an R package for efficient and intuitive manipulation of phylogenetic trees. *BMC Res. Notes*, **10**, 30.

Bodenhofer,U. *et al.* (2015) msa: an R package for multiple sequence alignment. *Bioinformatics*, **31**, 3997–3999.

Bogomolnaya,L.M. *et al.* (2006) Roles of the RAM signaling network in cell cycle progression in *Saccharomyces cerevisiae*. *Curr. Genet.*, **49**, 384–392.

Boiteux,S. and Jinks-Robertson,S. (2013) DNA repair mechanisms and the bypass of DNA damage in *Saccharomyces cerevisiae*. *Genetics*, **193**, 1025–1064.

Borowiec,M.L. (2016) AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, **4**, e1660.

Brown,J.W. *et al.* (2017) Phyx: phylogenetic tools for unix. *Bioinformatics*, **33**, 1886–1888.

Brunette,G.J. *et al.* (2019) Evolution-based screening enables genome-wide prioritization and discovery of DNA repair genes. *Proc. Natl. Acad. Sci. USA*, **116**, 19593–19599.

Campanella,J.J. *et al.* (2003) MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinform.*, **4**, 29.

Chen,M.-Y. *et al.* (2017) Phylogenomic resolution of the phylogeny of laurasiatherian mammals: exploring phylogenetic signals within coding and noncoding sequences. *Genome Biol. Evol.*, **9**, 1998–2012.

Chikina,M. *et al.* (2016) Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Mol. Biol. Evol.*, **33**, 2182–2192.

Clark,N.L. *et al.* (2012) Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res.*, **22**, 714–720.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Degnan,J.H. and Salter,L.A. (2005) Gene tree distributions under the coalescent process. *Evolution*, **59**, 24–37.

Doolittle,W.F. and Bapteste,E. (2007) Pattern pluralism and the tree of life hypothesis. *Proc. Natl. Acad. Sci. USA*, **104**, 2043–2049.

Doyle,V.P. *et al.* (2015) Can we identify genes with increased phylogenetic reliability?*Syst. Biol.*, **64**, 824–837.

Felsenstein,J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.*, **27**, 401–410.

Fourment,M. and Gibbs,M.J. (2006) PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evol. Biol.*, **6**, 1.

Hernández,Y. *et al.* (2018) BpWrapper: bioPerl-based sequence and tree utilities for rapid prototyping of bioinformatics pipelines. *BMC Bioinform.*, **19**, 76.

Huerta-Cepas,J. *et al.* (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.

Hunter,J.E. and Cohen,S.H. (2007) Package: igraph. *Educ. Psychol. Meas.*

Jarvis,E.D. *et al.* (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, **346**, 1320–1331.

Junier,T. and Zdobnov,E.M. (2010) The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*, **26**, 1669–1670.

Kapli,P. *et al.* (2020) Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.***21**, 428–444.

Kück,P. and Longo,G.C. (2014) FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front. Zool.*, **11**, 81.

Kumar,S. *et al.* (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets.*Mol. Biol. Evol.*, **33**,1870–1874.

Lake,J.A. (1991) The order of sequence alignment can bias the selection of tree topology. *Mol. Biol. Evol.*, **8**, 378–385.

Lanyon,S.M. (1988) The stochastic mode of molecular evolution: what consequences for systematic investigations?*Auk*, **105**, 565–573.

Laumer,C.E. *et al.* (2019) Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. R. Soc. B Biol. Sci.*, **286**, 20190831.

Li,Y. *et al.* (2020) A genome-scale phylogeny of fungi; insights into early evolution, radiations, and the relationship between taxonomy and phylogeny. *bioRxiv*, *2020.08.23.262857*.

Liu,L. *et al.* (2017) Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proc. Natl. Acad. Sci. USA*, **114**, E7282–E7290.

Louca,S. and Doebeli,M. (2018) Efficient comparative phylogenetics on large trees. *Bioinformatics*, **34**, 1053–1055.

Lujan,S.A. *et al.* (2012) Mismatch repair balances leading and lagging strand DNA replication fidelity. *PLoS Genet.*, **8**, e1003016.

Novick,P. and Schekman,R. (1979) Secretion and cell-surface growth are blocked in a temperature-sensitive mutant of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA*, **76**, 1858–1862.

One Thousand Plant Transcriptomes Initiative (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, **574**, 679–685.

Paradis,E. *et al.* (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng. Des. Sel.*, **14**, 609–614.

Pease,J.B. *et al.* (2016) Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLOS Biol.*, **14**, e1002379.

Philippe,H. *et al.* (2011) Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.*, **9**, e1000602.

Phillips,M.J. and Penny,D. (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.*, **28**, 171–185.

Revell,L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, **3**, 217–223.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Rokas,A. and Carroll,S.B. (2006) Bushes in the tree of life. *PLoS Biol.*, **4**, e352.

Rozas,J. *et al.* (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.*, **34**, 3299–3302.

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

Sang,T. and Zhong,Y. (2000) Testing hybridization hypotheses based on incongruent gene trees. *Syst. Biol.*, **49**, 422–434.

Sato,T. *et al.* (2005) The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, **21**, 3482–3489.

Sayyari,E. and Mirarab,S. (2018) Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes*, **9**,132.

Shen,X.-X. *et al.* (2016a) A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. *Genome Biol. Evol.*, **8**, 2565–2580.

Shen,X.-X. *et al.* (2020) Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum.*Ascomycota Sci. Adv.*, **6**, eabd0079.

Shen,X.-X. *et al.* (2016b) Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. *G3*, **6**, 3927–3939.

Shen,X.-X. *et al.* (2018) Tempo and mode of genome evolution in the budding yeast subphylum. *Cell*, **175**, 1533–1545.e20.

Smith,S.A. *et al.* (2018) So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLoS One*, **13**, e0197433.

Song,Y. *et al.* (2008) Role of the RAM network in cell polarity and hyphal morphogenesis in *Candida albicans*. *Mol. Biol. Cell*, **19**, 5456–5477.

Steenwyk,J.L. *et al.* (2019) A robust phylogenomic time tree for biotechnologically and medically important fungi in the genera *Aspergillus* and *Penicillium*. *mBio*, **10**,e00925-19.

Steenwyk,J.L. *et al.* (2020) ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.*, **18**, e3001007.

Struck,T.H. (2014) TreSpEx--detection of misleading signal in phylogenetic reconstructions based on tree information. *Evol. Bioinform.*, **10**, EBO.S14239.

Sundin,L.J.R. *et al.* (2011) The NDC80 complex proteins Nuf2 and Hec1 make distinct contributions to kinetochore-microtubule attachment in mitosis. *Mol. Biol. Cell*, **22**, 759–768.

Suyama,M. *et al.* (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**,W609–W612.

Talevich,E. *et al.* (2012) Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinform.*, **13**, 209.

Tsubouchi,H. and Ogawa,H. (2000) Exo1 roles for repair of DNA double-strand breaks and meiotic crossing over in *Saccharomyces cerevisiae*. *Mol. Biol. Cell*, **11**, 2221–2233.

Virtanen,P. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**,261–272.

Walker,J.F. *et al.* (2019) Characterizing gene tree conflict in plastome-inferred phylogenies. *PeerJ*, **7**, e7747.

Walsh,H.E. *et al.* (1999) Polytomies and the power of phylogenetic inference. *Evolution*, **53**, 932.

Wang,L.-G. *et al.* (2020) Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.*, **37**, 599–603.

Weigert,A. *et al.* (2014) Illuminating the base of the annelid tree using transcriptomics. *Mol. Biol. Evol.*, **31**, 1391–1401.

Wolfe,N.W. and Clark,N.L. (2015) ERC analysis: web-based inference of gene function via evolutionary rate covariation. *Bioinformatics*, **31**, 3835–3837.

Xia,X. (2013) DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. *Mol. Biol. Evol.*, **30**, 1720–1728.

Yu,G. *et al.* (2017) ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*, **8**, 28–36.

Zhang,C. *et al.* (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.*, **19**, 153.