PLOS ONE

RESEARCH ARTICLE

Can quartet analyses combining maximum likelihood estimation and Hennigian logic overcome long branch attraction in phylogenomic sequence data?

Patrick Kück[1,2]*, Mark Wilkinson[2], Christian Groß[2,3], Peter G. Foster[2], Johann W. Wägele[1]

1 Zoologisches Forschungsmuseum Alexander Koenig, Bonn, 53113, Germany, **2** The Natural History Museum, London, SW7 5BD, United Kingdom, **3** Delft University of Technology, Delft, 2628 CD, The Netherlands

* patrick_kueck@web.de

Citation: Kück P, Wilkinson M, Groß C, Foster PG, Wägele JW (2017) Can quartet analyses combining maximum likelihood estimation and Hennigian logic overcome long branch attraction in phylogenomic sequence data? PLoS ONE 12(8): e0183393. https://doi.org/10.1371/journal.pone.0183393

Editor: Arndt von Haeseler, Max F Perutz Laboratories GmbH, AUSTRIA

Received: April 10, 2017

Accepted: August 3, 2017

Published: August 25, 2017

Data Availability Statement: The software script as well as the corresponding manual and example files can be downloaded from https://github.com/PatrickKueck/Penguin. Otherwise, all relevant data are within the paper and its Supporting Information files.

Funding: This work was funded in part by PK's IEF Marie Curie Fellowship and by a grant (43915) from the John Templeton Foundation to MW.

Abstract

Systematic biases such as long branch attraction can mislead commonly relied upon model-based (i.e. maximum likelihood and Bayesian) phylogenetic methods when, as is usually the case with empirical data, there is model misspecification. We present *PhyQuart*, a new method for evaluating the three possible binary trees for any quartet of taxa. *PhyQuart* was developed through a process of reciprocal illumination between a priori considerations and the results of extensive simulations. It is based on identification of site-patterns that can be considered to support a particular quartet tree taking into account the Hennigian distinction between apomorphic and plesiomorphic similarity, and employing corrections to the raw observed frequencies of site-patterns that exploit expectations from maximum likelihood estimation. We demonstrate through extensive simulation experiments that, whereas maximum likeilihood estimation performs well in many cases, it can be outperformed by *PhyQuart* in cases where it fails due to extreme branch length asymmetries producing long-branch attraction artefacts where there is only very minor model misspecification.

Introduction

Reconstructing what happened is a central task of any historical science [1]. In biology, phylogenetic relationships are an important component of the history of life, some knowledge of which is a precondition of comparative methods [2]. The centrality of phylogeny in biology justifies the substantial continuing interest in reconstructing the Tree of Life, e.g. [3–5].

Modern techniques of nucleotide sequencing and the exponential growth of molecular databases increasingly provide data sets featuring hundreds of species and thousands of nucleotides in phylogenetic studies. The availability of whole genomes in the order of billions of nucleotides makes all-encompassing phylogenetic analyses possible for the first time [6]. The new age of phylogenomics gives reason to hope that congruence in phylogenetic analysis can finally be achieved through the reduction of stochastic sampling errors [7]. However, there is

Competing interests: The authors have declared that no competing interests exist.

considerable concern about increased accumulation of systematic errors due to reliance upon simple substitution models that may not adequately consider variation in substitution rate, compositional heterogeneity and the erosion of phylogenetic signal, e.g. [8–13] and which may be inconsistent.

Generally, systematic errors are increasingly important and apparent as more data are analysed because stochastic effects become less prominent, eventually yielding maximally supported, but incorrectly resolved phylogenetic relationships [14–16]. Numerous studies have shown that model misspecification can reduce the accuracy of phylogeny inference, e.g. [11–13, 17–31].

Systematic bias is particularly a molecular data problem due to the small number of possible character states [32] and the absence of complexity that might otherwise allow better distinction between homologies and homoplasies. Recent phylogenomic studies demonstrate how sensitive probabilistic tree reconstruction methods are to model assumptions and data composition. For example, the position of myriapods within the arthropod tree of life [33–37], the phylogeny within Chelicerata [38–40], the relationship within Lophotrochozoa [13, 41–43] or the relationship between Placozoa, Porifera, Cnidaria and Ctenopohora within the Metazoa [44–48] are remarkably sensitive to methods of analyses. Recent simulation studies show that even a slight model misspecification, such as that arising from approximating among site rate heterogeneity using discrete categories, can cause incorrect topologies in maximum likelihood (ML) analyses [12, 49].

A major source of systematic bias, and probably the most frequently cited reason for incorrect placements of taxa in phylogenetic reconstructions, is long branch attraction (LBA). First described by [50] as a problem of parsimony and compatibility methods, later studies revealed that even more robust, probabilistic tree reconstruction methods such as ML and Bayesian inference (BI) can fail to find the correct tree because of LBA, e.g. [8, 21, 23, 32, 41, 49, 51–63].

LBA is commonly understood as an incorrect phylogenetic reconstruction of two or more highly-divergent (long branch) lineages as sister (rooted) or adjacent (unrooted) groups due to the accumulation of convergent split signal (chance similarities) and the simultaneous loss of apomorphic characters shared with the actual close relatives, e.g. [32, 50, 61, 64]). [12] have shown that the probability of incorrect phylogenetic inferences increases with increasing heterogeneity of only inner edges and that unbalanced length differences between internal and terminal branches can have a negative effect on the tree reconstruction process when internal branch lengths are either too short or too long. Our usage of the term LBA is equivalent to the characterisations of [23] and [32]: "...conditions under which bias in finite data set analyses and/or statistical inconsistency arise due to the combination of short and long branches".

Different strategies exist for ameliorating LBA in phylogenetic analyses. Possibilities include the analysis of only slowly evolving sequences to reduce branch lengths [65] or the addition of slowly-evolving taxa to divide long internal branches [66]. However, slowly evolving sequences are sometimes not available, not least because of extinction [32], and exclusion of rapidly-evolving taxa reduces taxon sampling, which is often considered undesirable [67–74]. The exclusion of complete long-branched groups might successfully reduce LBA, but is not helpful if the relationship of those taxa is of importance to the study in question. Another frequently used strategy is the removal of sequence positions inferred to be fast evolving, e.g. [75–78], or entire classes of putatively fast evolving sites such as third codon positions in protein coding nucleotide data sets, which are potentially saturated by multiple substitutions [53, 79]. Conversion of nucleotides to more slowly evolving character states such as amino acid residues or purines and pyrimidines [60] is another strategy. One likely reason for misspecifications in modern probabilistic substitution models is the usual assumption of time reversibility. The direction of character evolution along a tree is not considered by these models and therefore

these analyses do not incorporate an important step of Hennigian phylogenetic inference, the distinction between new (apomorphic) and old (plesiomorphic) homologies [49].

The susceptibility of ML to systematic biases in cases where these is model misspecification motivates us to ask: is it possible to develop alternative techniques that are less effected than is ML by, for example, extreme branch length asymmetries? Here we introduce *PhyQuart*, a new, quartet-based algorithm which considers two alternative directions of character evolution along the internal branch of a quartet tree to discern between potentially apomorphic and plesiomorphic split-supporting site-patterns, and ML to estimate the expected number of convergent split-supporting site-patterns. This combination of Hennigian logic and ML estimation represents a completely new strategy for the evaluation of sequence data. A quartet tree comprising one internal and four external branches is the smallest phylogenetically informative unrooted tree. It is sometimes helpful to focus on quartets because of their computational simplicity: there are only three alternative topologies to be investigated and far fewer potential site-patterns (the basic empirical data from which inferences are to be made) than in alignments containing many taxa. Despite this helpful computational simplicity, it is widely believed that quartet analyses exacerbates LBA (because it is the opposite of adding taxa so as to break up long branches) and thus represents the most difficult taxon sampling context in which to overcome LBA [80]. Through extensive quartet simulations, including cases with strong branch length differences, we demonstrate the efficiency of our new approach in detecting phylogenetically informative and conflicting signals and compare its performance to ML alone when there is a (unrealistically) small degree of model misspecification. The *PhyQuart* algorithm is implemented in a command line driven software script.

1 The method: Concept and algorithm

1.1 Concept

The *PhyQuart* algorithm takes as input an alignment and outputs normalised split-support for alternative quartet trees based on a site-pattern classification and using observed and expected (based on ML inference) frequencies of split-supporting site-patterns, considering the Hennigian distinction between phylogenetically informative (apomorphic) and uninformative (plesiomorphic) character.

Here we define some basic concepts and provide a brief overview of our approach. A more complete and formal description of the *PhyQuart* algorithm is given in the next section.

We decided to use the established terms plesiomorphy and apomorphy to distinguish between old (pelsiomorphic) and new (apomorphic) shared homologous character states. The alternative would have been to invent some new term for the same meaning, which clearly is not a better option. Both terms describe a simple fact that is observed everywhere where evolution takes place: an old state is modified and transformed to a new state. Or, something that did not exist appears de novo. This is not different from saying that there is a sequence (a complex character) in which a nucleotide at a specific site position of an alignment is substituted by a new nucleotide, which would be the apomorphic detail. The discovery that it makes a difference whether in molecular evolutionary processes the polarity in time is considered or not has recently been published by Kück & Wägele [49]. Phylogenetically informative split-supporting site-patterns are only those site-patterns which contain apomorphies. Further, we define an informative split-supporting sitre-pattern as "putative synapomorphy" when a shared apomorphic character similarity between two taxa on one side of a split is assumed to be present in the most recent common ancestor (internal node of a tree).

The goal of the *PhyQuart* algorithm is to identify among all split-supporting site-patterns those that support polarized splits with characters that are probably putative synapomorphiies.

A split is a bipartition of a set of species or sequences [62, 64, 81]. A sequence position supports some split if no pair of taxa separated by split share the same character state [82]. For the quartet of taxa A-D there are three phylogenetically informative quartet splits AB|CD, AC|BD and AD|BC, corresponding to the single internal branches of the three possible unrooted binary quartet trees. Let W-Z correspond to different character states (e.g., nucleotides or amino acid residues). Sites with the character distribution {WXYZ} support all three quartet trees and thus do not differentially support any of them, whereas a sequence position differentially supports one quartet split/tree if two taxa have the same character state and the other two taxa have some other character state(s). Thus, sites with the character state distribution {XXYY} (symmetric) or {XXYZ} and {YZXX} (two possible asymmetric) split-supporting site-patterns are counted as differential split-support for the quartet tree AB|CD (Fig 1a and 1b).

We denote the polarity of (i.e., the direction of character transformation along) the internal branch of a given quartet tree AB|CD using parentheses, e.g. AB(CD) and (AB)CD which indicate the direction is towards CD and towards AB respectively. This distinction enables a classification of split-supporting site-patterns into potentially phylogenetic informative (apomorphic) and uninformative (plesiomorphic) site-patterns [83] contingent on the assumed polarity. Thus, the asymmetric split-supporting site-pattern {XXYZ} is interpreted as supporting the polarized quartet tree (AB)CD because the shared character state similarity of A and B appears apomorphic but as uninfomative for the polarized quartet tree AB(CD) because the similarity of A and B appears plesiomorphic (Fig 1c). Because the *PhyQuart* algorithm assumes no *a priori* knowledge of polarity (placement of the root) each possible polarity for each of the three quartet trees is evaluated separately for each quartet of taxa (for a total of six evaluations) to find the best supported quartet tree based exclusively on putative synapomorphic character states.

Putative synapomorphy can be phylogenetically misleading if the similarity is not homologous but rather evolved convergently. For example, given a polarized quartet tree (AB)CD, similarity would be present in split-supporting site-patterns with characters shared between taxon C and D: {XXYY} and {XZYY}. However, observed split-supporting site-patterns for (AB)CD can be either synapomorphic (inherited from a common ancestor of C and D) or convergently evolved along the terminal branches of C and D. Suppose that given polarized quartet trees (AB)CD and AB(CD) of quartet tree *q1* are incorrect and thus the correct quartet tree is one of the two alternative topologies (*q2* or *q3*). *PhyQuart* uses ML to estimate how much apomorphic support for each polarization (number of split-supporting site-patterns {XXYY} and {XZYY} for (AB)CD and {XXYY} and {XXYZ} for AB(CD)) would have evolved with the branch lengths of the underlying data if *q2* were the correct tree, or if *q3* were the correct tree. These values are equivalent to the number of parallel substitutions on unrelated branches that are expected if *q1* is the correct tree. We take the mean of these two values as the estimate of expected convergently-evolved, misleading support for each polarized quartet tree of *q1* (Fig 1d). The ML inference of expected convergence for each polarized quartet tree is implemented using the P4 package [84] with individually optimised branch length and model parameters for each of the alternative quartet trees. The estimated number of misleading sites is then subtracted from the observed number of supporting positions to get an estimate of the synapomorphic support for each polarized quartet tree.

Multiple substitutions among longer branches may lead to underestimation of the support for the correct quartet tree. To reduce the effects of such underestimation we use correction factors (ω) based on the frequencies of the four singleton patterns (i.e. {XXXY}, {XXYX}, {XYXX}, {YXXX}) that are intended to make the corrected support values closer to what would be expected if external branches were of equal length. Correction factors are applied to both observed (ω_{obs}) and ML inferred expected (ω_{exp}) frequencies of site-patterns. In each

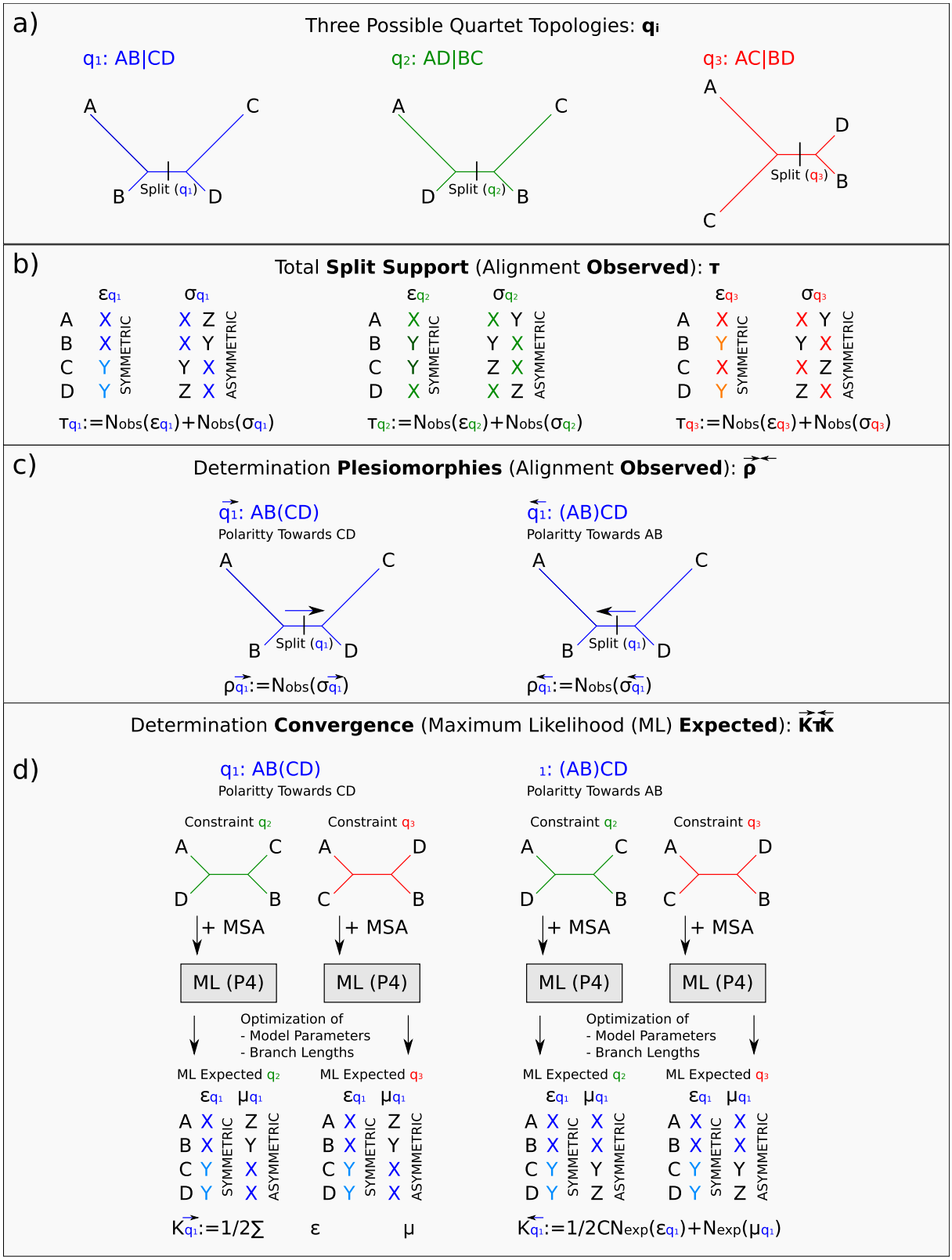

Fig 1. Flowchart of the *PhyQuart* algorithm. Simplified flowchart showing a) each of the three possible quartet relationships for a set of 4 sequences (q_1, q_2, q_3), b) the site-pattern classification of observed (N_{obs}) symmetric (ξ_{q_i}) and asymmetric (σ_{q_i}) support (τ_{q_i}), c) the determination of plesiomorphic (old) split-supporting site-patterns given two different polarities of character

transformation along the internal branch of each possible quartet tree, $\rho_{\vec{q}_1}$ and $\rho_{\overleftarrow{q}_1}$, and d) estimation of expected convergent split-supporting site-patterns ($\kappa_{\vec{q}_1}$, $\kappa_{\overleftarrow{q}_1}$) supporting quartet q_1 in ML split pattern estimations using branch length and model optimization on constraint topologies of the other two possible quartet relationships (q_2, q_3).

https://doi.org/10.1371/journal.pone.0183393.g001

case, the correction factor reduces the split-supporting site-patterns in proportion to the complement of four times the frequency of the least frequent singleton pattern (observed or expected) divided by the sum of the frequencies of all four singleton patterns. The frequencies of observed singleton site-patterns and the corresponding correction factor (ω_{obs}) are constant for all quartet trees and polarities, whereas ω_{exp} can differ for ML expected frequencies. Thus by itself ω_{obs} has no effect on relative support for different quartet trees, rather it is ω_{exp} that drives the effect of the correction. When there are strong branch length asymmetries ω_{exp} and thus the corrected estimate of convergent split-supporting site-patterns will be high. Thus, the correction is important in cases of unequal branch lengths (as evidenced by differences in singleton frequencies) such as can produce LBA.

Let P be a polar quartet tree. Let S_{obs} be the sum of the observed numbers of symmetric and asymmetric site-patterns supporting P and let M_{obs} be the smallest number (the minimum) taken over all four singletons, and T_{obs} be the total number of observed singleton site-patterns. Let $\omega_{obs} = 1 - (4M_{obs}/T_{obs})$. Similarly, using ML estimation for the two contrary quartet trees (that conflict with P), let S_{exp1} and S_{exp2} each be the sum of the expected number of symmetric and asymmetric site-patterns supporting P and let M_{exp1} and M_{exp2} be the smallest numbers and let T_{exp1} and T_{exp2} be the total numbers respectively of expected singleton site-patterns. Let $S_{exp} = (S_{exp1} + S_{exp2})/2$ and let $\omega_{exp} = 1 - (4M_{exp1}/T_{exp1} + 4M_{exp2}/T_{exp2})/2$. Then the *PhyQuart* score for any P is $(S_{obs} - (S_{obs} * \omega_{obs})) - (S_{exp} - (S_{exp} * \omega_{exp}))$ and the score for each quartet is the highest of the scores for its polarized quartets normalised so that the scores for all three alternative quartets sum to one.

1.2 Algorithm

1.2.1 Observed split support of each quartet tree. First, the algorithm counts the total number of observed split-supporting site-patterns (τ) in a given set of four aligned sequences of length L for each of the three possible quartet topologies $x \in Q := \{q_1, q_2, q_3\}$. All site-patterns ($s$) with symmetric ($\xi$) and asymmetric ($\sigma$) split-support for a given quartet relationship are taken into account (Fig 1a and 1b).

$$Q := \{q_1, q_2, q_3\} \tag{1}$$

$$\tau_x := \sum_{i=1}^{L} 1_{\{s_i \in (\xi_x \vee \sigma_x)\}, x \in Q} \tag{2}$$

1.2.2 Determination of plesiomorphic split signal. To identify potentially plesiomorphic split-supporting site-patterns of a given quartet tree ($x \in Q$), two different polarities are specified: $Q_{polar} := \{\vec{x}, \overleftarrow{x}\}$. Each polarity defines one of the two possible directions of character transformation along the internal branch of a given quartet tree. Quartet-supporting positions based on symplesimorphic split-supporting site-patterns are counted separately for each polarized quartet tree $z \in Q_{polar}$.

The right pointing direction of a quartet tree ($z = \vec{x}$) defines asymmetric z split-supporting site-patterns (σ_z) as apomorphic (ϱ_z) whenever identical character states are only shared between taxa A and B in quartet tree z:= AB(CD).

For example, site-pattern $s \coloneqq \{XXYZ\}$ contains asymmetric split-supporting site-patterns ($s = \sigma_z$) based on a plesiomorphic character state ($\sigma_z = \varrho_z$) if polarity $z\coloneqq$ AB(CD). Otherwise, with the left pointing direction of a quartet tree ($z = \overleftarrow{x}$, $z\coloneqq$ (AB)CD, site-pattern $s = \sigma_z$, but $\sigma_z \neq \varrho_z$, and the site-pattern is interpreted as apomorphic.

The total number of observed ϱ_z sites of a split-supporting site-pattern for a given polarized quartet relationship z (given a sequence length L) is defined as ρ_z (Fig 1c).

$$Q_{polar}\coloneqq\{\vec{q}_1, \overleftarrow{q}_1, \vec{q}_2, \overleftarrow{q}_2, \vec{q}_3, \overleftarrow{q}_3\} \tag{3}$$

$$\rho_z\coloneqq\sum_{i=1}^{L} 1_{\{s_i\in\varrho_z\},z\in Q_{polar}} \tag{4}$$

1.2.3 Determination of convergent split signal. Contrary to the identification of plesiomorphic split-supporting site-patterns observed in a given alignment of sequence length L, the total amount of potentially convergent split-supporting site-patterns (κ_z) for a given polarity z ($z \in Q_{polar}\coloneqq\{\vec{q}_1, \overleftarrow{q}_1, \vec{q}_2, \overleftarrow{q}_2, \vec{q}_3, \overleftarrow{q}_3\}$) of a quartet tree x ($x \in Q \coloneqq \{q_1, q_2, q_3\}$) is determined by ML estimation of symmetric ($\xi_{exp_{\pi(z)}}$) and asymmetric (σ_{exp_z}) split-supporting site-pattern frequencies, which support tree x based on constraint topologies of the other two possible quartet relationships y ($y \in Q \setminus \{\pi(z)\}$). Thereby, π is defined as the projection of Q_{polar} onto Q (π: $Q_{polar} \rightarrow Q$), saying that $\pi(\vec{q}_i) = q_i = \pi(\overleftarrow{q}_i)$ for i = 1, 2, 3. Note that polarity is not relevant for ML inferences, but ML estimated site frequencies depend on branch lengths.

Site-pattern frequencies of each constrained topology y are calculated by ML using branch length and model parameter optimization on the basis of the original quartet alignment and a defined substitution model. Estimated frequencies of each possible site-pattern are multiplied by the original alignment length L to get the expected number of sites for a pattern in a given alignment.

For each polarity of a given quartet z ($z \in Q_{polar}$), the ML expected number of the potentially convergent split-supporting site-patterns (chance similarities) (κ_z) is defined by the mean number of ML estimated split symmetric ($\xi_{exp_{\pi(z)}}$) and asymmetric site-patterns (σ_{exp_z}), supporting polarized tree z.

Given tree z, the expected number of chance similarities (e.g. for (CD)) is estimated with the number of characters (μ_z) shared by C and D in the two other quartet-topologies y ($y \in Q \setminus \{\pi(z)\}$), where they are not adjacent and thus cannot be sister-taxa. We use for each split-group the average (κ_z) of these two values (Fig 1d).

For example, the asymmetric split-supporting site-pattern $\sigma_z \coloneqq \{YZXX)\}$ shares identical character states only between taxa C and D in $z \coloneqq \{AB(CD)\}$, therefore: $\sigma_z = \mu_z$. Otherwise, if $\sigma_z \coloneqq \{XXYZ)\}$, then: $\sigma_z \neq \mu_z$.

$$\kappa_z\coloneqq\frac{1}{2}\sum_{y\in Q\setminus\{\pi(z)\}}(\xi^y_{exp_{\pi(z)}} + \mu^y_z), z \in Q_{polar} \tag{5}$$

1.2.4 Further noise reduction using correction factor ω. Singleton site-pattern frequencies can be used as an approximation for terminal branch lengths. Fast evolving sequences will have more of these than slower ones. Four different singleton site-patterns are possible, {YXXX}, {XYXX}, {XXXY}, and {XXYX}, each of them contributing to one of the four terminal branch lengths.

To further reduce the impact of noise upon the identified number of split-supporting site-patterns for a given polarity z ($z \in Q_{polar} := \{\vec{q}_1, \overleftarrow{q}_1, \vec{q}_2, \overleftarrow{q}_2, \vec{q}_3, \overleftarrow{q}_3\}$) of a quartet tree x ($x \in Q := \{q_1, q_2, q_3\}$), the algorithm reduces for each polarity (z) of a given tree (x) the total number (τ_x) of counted symmetric (ξ_x) and asymmetric (σ_x) split-supporting site-patterns as well as the number of plesiomorphic (ρ_z) and convergent split-supporting site-patterns (κ_z).

The correction factor (ω) is defined as one minus the ratio of four times the smallest number of the singleton site-patterns (ϕ) to the total number of singleton site-patterns (N). The total number of tree x supporting split signal (τ_x) as well as the the number of plesiomorphic split-supporting site-patterns (ρ_z) for a given polarity z of tree x ($z \in Q_{polar}$) are reduced in relation to single substituted site-pattern frequencies of the original quartet alignment (ω_{obs}).

$$\omega_{obs} := 1 - \left(\frac{4\phi}{N}\right) \tag{6}$$

The correction factor (ω_{exp_z}) for convergent split-supporting site-patterns (κ_z) of a given polar quartet tree (z) is specified by the mean of the two single correction factors (ω_z), which are derived (in the same manner as described in Eq 7) from the ML-estimated singleton site-pattern frequencies of the other two quartet topologies y ($y \in Q \setminus \{\pi(z)\}$).

$$\omega_{exp_z} := \frac{1}{2} \sum_{y \in Q \setminus \{\pi(z)\}} \omega_z^y, z \in Q_{polar} \tag{7}$$

1.2.5 Determination of potential apomorphic split signal (θ). Only the actual number of potentially synapomorphic, split-supporting site-patterns is counted as phylogenetic signal. To identify the number of potentially synapomorphic split-supporting site-patterns for each possible polarized quartet tree (z), the total number of observed split-supporting site-patterns (τ_z) as well as the number of potentially plesiomorphic (ρ_z) and convergent (κ_z) split-supporting site-patterns are adjusted by the correction factor ω_{obs} and ω_{exp_z}. Afterwards, the remaining (synapomorphic) split signal is calculated for each polar quartet tree by subtracting the corrected phylogenetic uninformative plesiomorphic and convergent split-supporting site-patterns from the corrected number of observed split-supporting site-patterns.

$$\theta_z := (\tau_{\Pi(z)} * \omega_{obs}) - (\rho_z * \omega_{obs}) - (\kappa_z * \omega_{exp_{\pi(z)}}), z \in Q_{polar} \tag{8}$$

1.2.6 Final quartet weighting (λ) for polarized topologies based on best polar quartet tree support values. After the assignment of the actual, potentially synapomorphic split supporting site-patterns all three quartet topologies x ($x \in Q := \{q_1, q_2, q_3\}$) are scored ($\delta_x$) related to their higher number of split-supporting site-patterns given both possible polarities ($\theta_a, \theta_b \in Q_{polar} := \{\vec{x}, \overleftarrow{x}\}, x \in Q$). For example, if the obtained score of polarity $\theta_{(AB)CD} > \theta_{AB(CD)}$, then $\delta_{AB|CD} := \theta_{(A,\,B)CD}$.

$$\delta_x := \begin{cases} \theta_a & \text{if } \theta_a \geq \theta_b \\ \theta_b & \text{if } \theta_a < \theta_b \end{cases}, (\theta_a, \theta_b \in Q_{polar} := \{\vec{x}, \overleftarrow{x}\}, x \in Q) \tag{9}$$

Finally, each quartet tree x ($x \in Q := \{q_1, q_2, q_3\}$) is weighted ($\lambda_x$) equal the difference between the actual number of split-supporting site-patterns (δ_x) and the lowest number of split-supporting site-patterns given all three quartets (δ_{lowest_x}), normalised by the sum of single

quartet weights. For example, if $\delta_{lowest_x} = \delta_{AC|BD}$, then $\lambda_{AB|CD} := \delta_{AB|CD} - \delta_{AC|BD}$.

$$\lambda_x := \frac{\delta_x \setminus \delta_{lowest_x}}{\sum_{i \in Q} \lambda_i}, x \in Q \quad (10)$$

1.3 Software implementation

The algorithm introduced in this study is implemented in a new software tool called PENGUIN, a command line driven PERL script that runs on Windows PCs, Mac OS and Linux operating systems and can be easily implemented into automatic process pipelines. A PERL interpreter must be present in order to execute the software. PENGUIN is freely available (i.e., open-source) and released under the terms of the GNU General Public License (GPL) 3.0. The software script as well as the corresponding manual and example files can be downloaded from https://github.com/PatrickKueck/Penguin.

PENGUIN reads files of multiple sequence alignments in FASTA and PHYLIP format. If the alignment consists of more than four sequences, a clan input file comprising four predefined clans (sensu [85]) of one or more taxa must be provided in plain TEXT format. If specified, PENGUIN analyses all possible quartet combinations of one taxon from each predefined clan. PENGUIN does not allow multiple records of the same taxon name within given input file(s) and mismatches between taxa included in a predefined clan file and a multiple sequence alignment are just left unanalysed. The script can handle both nucleotides and amino-acid sequences. Sequence sites with indels (gap or '-'), ambiguity or missing characters are always excluded from the analysis. Under default, PENGUIN excludes all forbidden site positions separately for each quartet of sequences drawn from a given multiple sequence alignment. This has the advantage that sequence positions do not have to be deleted from the full alignment and can be used in cases of other quartets that do not have such ambiguities in these positions. Alternatively, site exclusion can be performed on the complete sequence alignment in advance of the quartet establishment. However, the performance of our algorithm has only been tested on nucleotide data without simulated indels, ambiguities, or missing data.

PENGUIN writes information on split support for each possible quartet relationship between four taxa or clans in plain TEXT files. Obtained discrepancies in topological split support of the three possible quartet topologies of a set of four clans are also presented as split network and triangle graphs. A further vector network shows the distribution of best, second best, and third best resolved quartet trees.

Detailed information about single analysis output is provided by the PENGUIN script manual.

2 Performance

The *PhyQuart* algorithm was tested and its performance compared with ML using 172,800 simulations. Varying amounts of nucleotide sequence data was simulated using INDELible v.1.03 [86] on quartet trees with different combinations of fixed (at 0.1), alternative internal (BL1 = {0.01, 0.02}), and more (BL2) or less highly varied (BL3) terminal branch lengths (Fig 2) under the GTR model of sequence evolution. Among site rate variation (ASRV) was modelled using a continuous Γ-rate distribution with different shape parameters and a fixed proportion of invariant sites (0.3). Simulations did not include indels. Table 1 summarises the parameters employed in the analyses. ML trees were inferred from simulated data with PhyML_3.0_linux64 [87, 88], using a mixed-distribution model (GTR+Γ+I) with the model parameters (α, I) used in the simulation but with the simulated continuous gamma

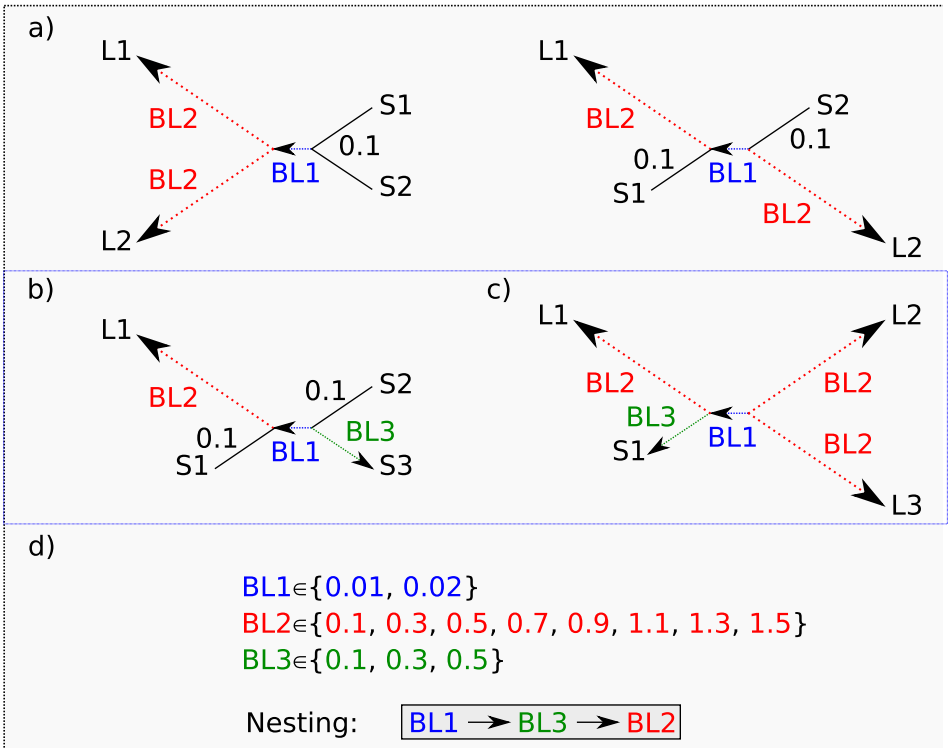

Fig 2. Quartet simulation setups. Simulation setups for quartet analyses testing effects of given a very short internal branch (BL1) with a) stepwise elongation of two adjacent (Farris-topology, left) and non-adjacent (Felsenstein-topology, right) terminal long branches (BL2), b) elongation of one terminal branch (BL2) using different lengths for one of the three short branches (BL3), and c) stepwise elongation of three terminal branches, with different lengths of the remaining short terminal branch (BL3).

https://doi.org/10.1371/journal.pone.0183393.g002

distribution approximated by a discrete gamma with four relative substitution rate categories and the relative rates and base compositions estimated from the data. This difference in gamma (continuous or discrete) and any small differences in the simulated and estimated relative rates and base compositions are the only model misspecifications involved in the ML inference. Thus we expect ML to perform well in most cases. The ML estimation of split site-

Table 1. Defined model parameters for data simulation (INDELible) and ML analyses used in *PhyQuart* and PhyML.

Simulation Setup	Seq. Length	I	Γ (α Shape Parameter)					
1-elongated branch	(BL2) = 250 kbp	0.3	0.1	0.3	0.5	0.7	1.0	2.0
2-elongated branches	(BL2) = 250 kbp	0.3	0.1	0.3	0.5	0.7	1.0	2.0
2-elongated branches	(BL2) < 250 kbp	0.3			0.5		1.0	2.0
3-elongated branches	(BL2) = 250 kbp	0.3	0.1	0.3	0.5	0.7	1.0	2.0
GTR Substitution Rates								
INDELible:	C↔T: 0.3; T↔A: 0.8; T↔G: 0.6; C↔A: 0.5; G↔C: 0.4; G↔A: 1.0							
PhyQuart:	Estimated							
PhyML:	Estimated							
Nucleotide Frequencies								
INDELible:	T: 0.35; C: 0.15; A: 0.35; G: 0.15							
PhyQuart:	Estimated							
PhyML:	Estimated							

https://doi.org/10.1371/journal.pone.0183393.t001

pattern frequencies in the *PhyQuart* algorithm used the same model but with all parameters estimated from the underlying data. All analyses were performed and evaluated with a Perl pipeline. We generated 100 multiple sequence alignments for each combination of internal and terminal branch lengths and recorded the frequencies of correct and incorrect tree reconstructions from these replicates.

2.1 Elongation of two terminal branches

Our first quartet simulations test the classical quartet LBA problem [50] with two alternative topologies in which two terminal long branches were either adjacent (termed "Farris" topologies), or non-adjacent (termed "Felsenstein" topologies) (Fig 2a). We examined the stepwise elongation of the long terminal branches (BL2 = {0.1 → 1.5} in steps of 0.2) with the other two terminal short branches kept constant (length = 0.1) and two alternative internal branch lengths (BL1 = {0.01, 0.02}) analysing a wide range of sequence lengths (0.5, 1, 2, 5, 10, 20, 50, 100, 250 thousands of base pairs (kbp)). For data sets of 250 kbp, we simulated sequences with six different rate heterogeneity parameters (α = {0.1, 0.3, 0.5, 0.7, 1.0, 2.0}) whereas for shorter sequences we analysed three different heterogeneties (α = {0.5, 1.0, 2.0}).

With the longest simulated sequences (250 kbp), ML mostly performs very well in reconstructing Farris topologies, but as the ratio of long to short branches increases reconstruction success for Felsenstein topologies decreases precipitously (Fig 3). In contrast, *PhyQuart* successfully reconstructs Felsenstein topologies in the majority of replicates, independent of simulated model parameter and branch length conditions and, except for strongly heterogeneous data sets (α = 0.1), *PhyQuart* outperforms ML especially with the shortest internal branches (BL1 = {0.01}) (Fig 3a). While not as successful as ML in reconstructing simulated Farris topologies, *PhyQuart* successfully reconstructs these in a majority of cases when $\alpha > 0.1$, while both reconstruction methods often failed in cases of high branch length heterogeneity for data sets simulated with α = 0.1 (Fig 3b). Except for very strong heterogeneous data simulations (α = 0.1), ML outperformed *PhyQuart* in identifying correct Farris topologies if terminal branches exceeded a length 70 times higher as the internal branch (BL2 ≥ {0.7}). Contrary to ML, the *PhyQuart* algorithm consistently recovered correct Farris and Felsenstein topologies in the majority of the (250 kbp long) replicates, even in simulations with very low internal branch signal of the correct tree (BL1 = {0.01}) if $\alpha > 0.1$ (Fig 3). Reconstruction successes for all Felsenstein and Farris topology simulations based on sequence lengths of 250 kbp are given in the supplementary file S1 Fig. Comparison of the *PhyQuart* reconstruction results of this setup with and without implementation of the correction factor ω are given in the supplementary information S2 Fig.

Reconstruction success decreases with sequence length when branch lengths are heterogeneous. ML and *PhyQuart* correctly recovered Felsenstein and Farris topologies in the majority of data replicates given a wide range of internal and terminal branch conditions if sequence length exceeds 50 kbp (Fig 4). Considering the reconstruction success for different rate heterogeneities and for Farris as well as Felsenstein topologies, ML slightly outperforms *PhyQuart* in cases of strong branch length differences if sequence are shorter than 50 kbp for longer internal branch lengths (BL1 = {0.02}) whereas *PhyQuart* outperfoms ML with the shorter internal branch (BL1 = {0.01}). *PhyQuart* often outperforms ML if sequences are longer than 50 kbp (Fig 4). Detailed summaries of our analyses with sequence lengths shorter than 250 kbp are given in the supplementary information S3 Fig.

2.2 Elongation of one terminal branch

Our second quartet simulation experiments (Fig 2b) investigate reconstruction success when there is one long and three short terminal branches. These experiments also used two

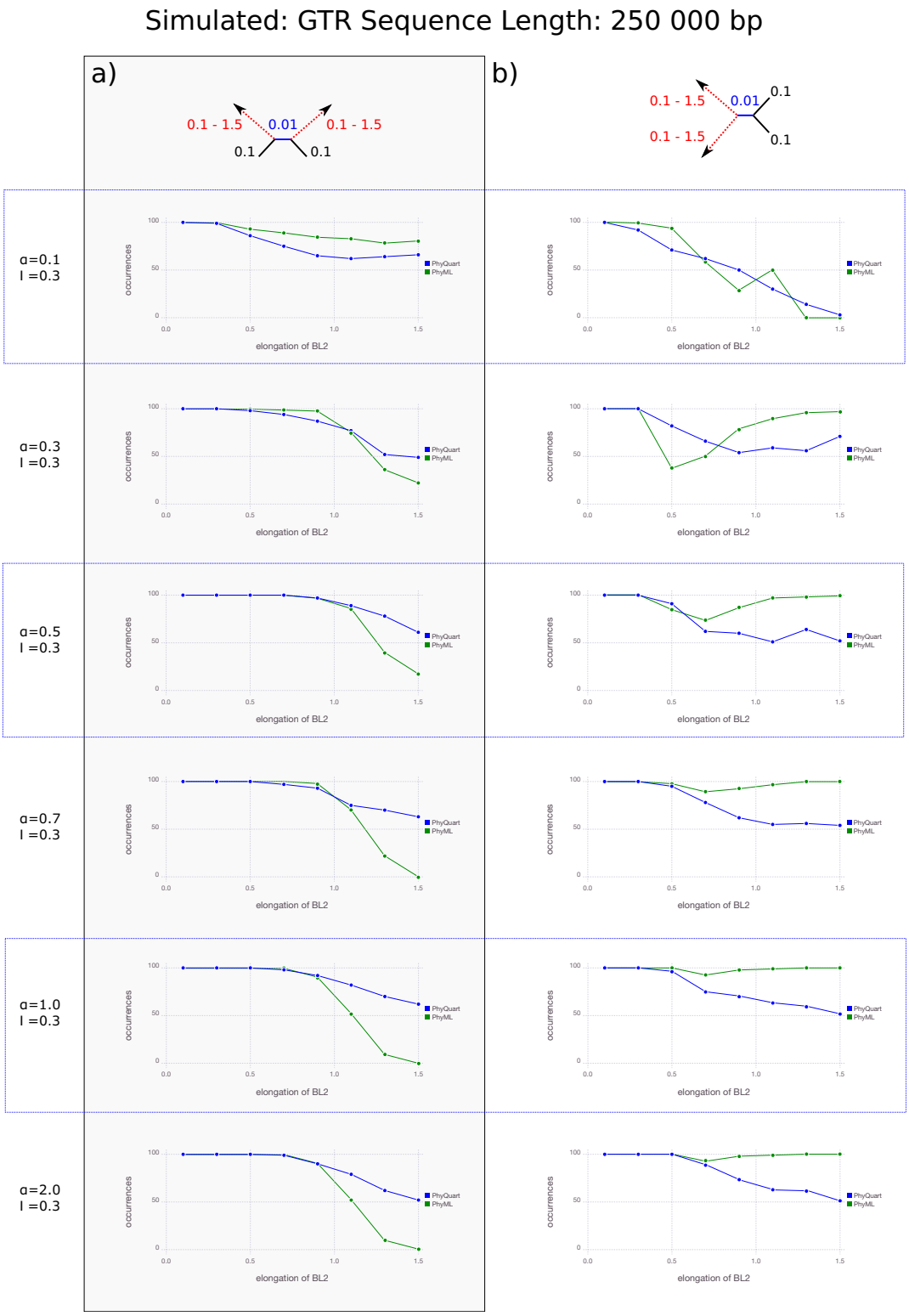

Fig 3. Quartet reconstruction success given stepwise elongation of two terminal branches if sequence lengths equal 250 kbp. Plots visualizing the reconstruction success of *PhyQuart* (blue) and ML (green) given stepwise elongation of two terminal branches (BL2, x-axis) and a fixed, very short internal branch length (BL1 = {0.01}) for 100 (250 kbp long) data replicates (y-axis). The plots present the summarized reconstruction success for (a) Felsenstein-, and (b) Farris-topologies of given $\alpha = \{0.1, 0.3, 0.5, 0.7, 1.0, 2.0\}$ and an invariable site proportion (I) of 0.3. A detailed overview of all simulation results of this setup is given as supplementary information S1 Fig.

https://doi.org/10.1371/journal.pone.0183393.g003

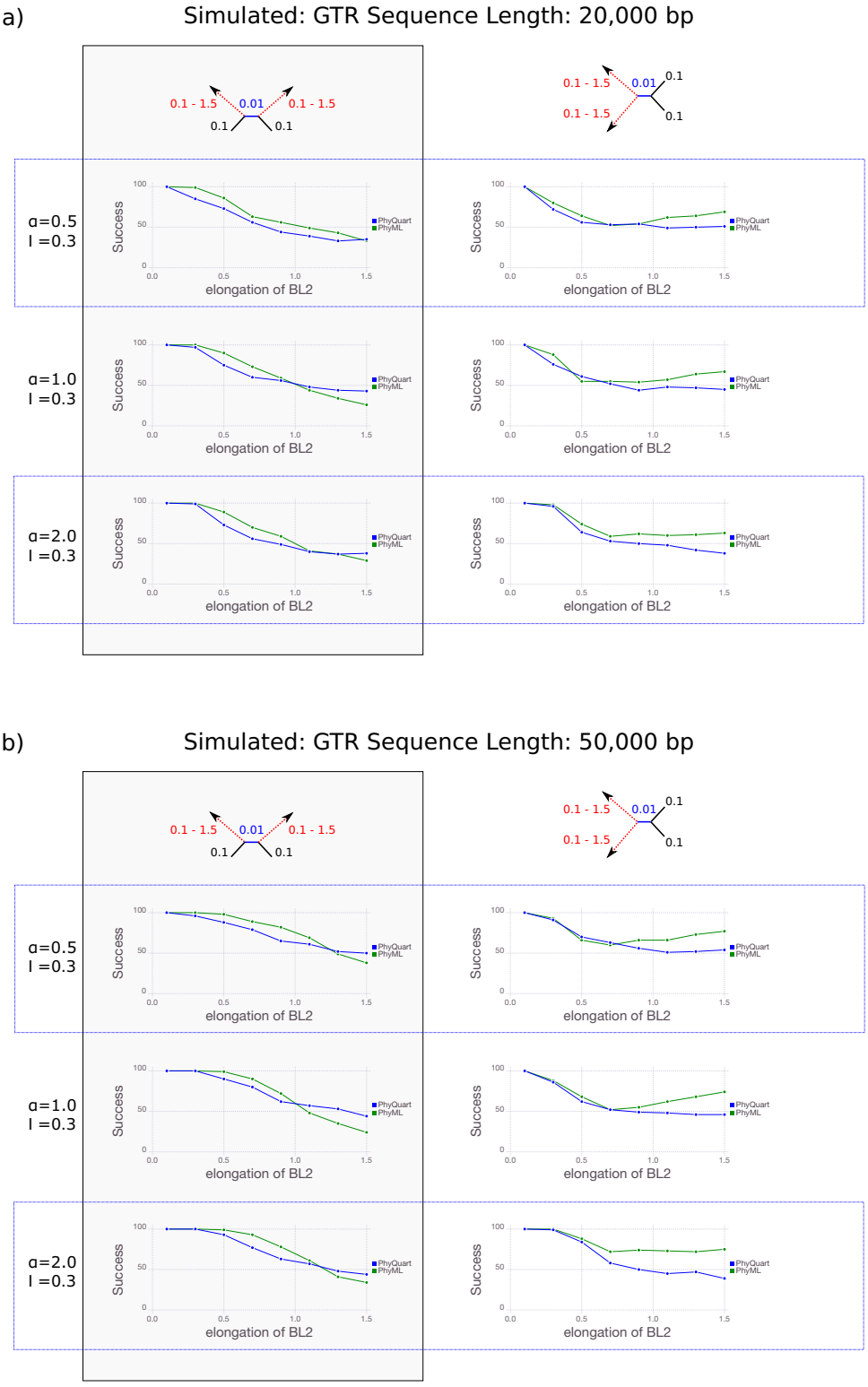

Fig 4. Quartet reconstruction success given stepwise elongation of two terminal branches if sequence lengths < 250 kbp. Reconstruction success of *PhyQuart* (blue) and ML (green) for different rate heterogeneities and for Farris as well as Felsenstein topologies under different lengths of two elongated terminal branches (BL2, x-axis), given a fixed internal branch length (BL1 = {0.01}), and 100 data replicates (y-axis). Reconstruction success for data sets of sequences <250 kbp are summarized for $\alpha = 0.5$, 1.0, and 2.0: a) 20 kbp, b) 50 kbp. A detailed overview of all simulation results of this setup is given as supplementary information S3 Fig.

https://doi.org/10.1371/journal.pone.0183393.g004

alternative internal branch lengths of BL1 = {0.01, 0.02}, stepwise elongation of the single long terminal branch (BL2 = {0.1 → 1.5} in steps of 0.2) with two of the remaining terminal branches kept constantly short (= 0.1) and the third branch also stepwise elongated (BL3 = {0.1, 0.3, 0.5}). Sequence lengths were 250 kbp with six alternative rate heterogeneities (α = {0.1, 0.3, 0.5, 0.7, 1.0, 2.0}). Both *PhyQuart* and ML performed well in all analyses independent of simulation parameters with sometimes slightly better performance of ML and vice versa. Fig 5a shows the similar reconstruction success of both methods given three equal short terminal branches. Detailed result plots of all "single long branch" simulation analyses are given in the supplementary information S4 Fig.

2.3 Elongation of three terminal branches

Our third quartet simulation experiments (Fig 2c) involve stepwise elongation of three terminal branches (BL2 = {0.1 → 1.5} in steps of 0.2) with a stepwise increase of the fourth terminal branch (BL3 = {0.1, 0.3, 0.5}), two alternative internal branch lengths (BL1 = {0.01, 0.02}), six rate heterogeneities (α = {0.1, 0.3, 0.5, 0.7, 1.0, 2.0}) and sequences of 250 kbp. As with other experiments, success decreases with increasing length differences between internal and terminal branches for both methods (Fig 5b). With the exception of data simulated with high among site rate variation (α = {0.1}), ML typically slightly outperformed *PhyQuart*. Detailed result plots of all performed analyses for cases with three long branches are given in the supplementary information. Reconstruction success of both methods was not strongly uninfluenced by the length of the shorter fourth terminal branch (BL3), but the shorter this branch (BL3) and the longer the internal branch length (BL1), the better the performance of both methods given three strongly elongated terminal branches. Detailed result plots of all "three long branch" simulation analyses are given in the supplementary information S5 Fig.

3 Discussion

Not without good reasons, modern molecular phylogenetics is dominated by the probabilistic, model-based ML and Bayesian methods. However, although these approaches have much to recommend them, they can fail to recover the correct tree and may instead recover the wrong tree with misleading high support when available models do not adequately represent underlying evolutionary dynamics. The robustness of ML to variation in evolutionary processes and the extent to which model misspecification results in systematic biases and statistical inconsistency are far from fully understood. However, we know that when evolutionary signal is eroded to the extent that is not, or is barely, distinguishable from confounding noise in the data, then phylogenetic methods are more susceptible to yielding biased estimates [79]. Therefore, we should be alert to potential errors when internal branches are short (and thus may have limited signal) and deep (and thus may have much signal erosion). Phylogenomic scale studies often address such cases, and through the application of large amounts of sequence data also run a greater risk of being substantially mislead by any systematic bias in the inadequately modelled data. Therefore, a major problem of phylogenomics is to determine if seemingly well-supported relationships are the result of systematic bias [16]. *PhyQuart* is motivated by this problem. Our results demonstrate that conventional ML inference can fail when there is strong branch length heterogeneity even when there is only seemingly very minor ML model misspecification. They also provide proof of concept for the idea that (at least for our simulated data and for long alignments) it is possible to design methods that can outperform conventional ML inference in those cases where ML does not perform well. These are the cases where accurate phylogenetic inference is most difficult and additional tools are most needed. *PhyQuart* is based on consideration of the evidential significance of observed site-patterns and

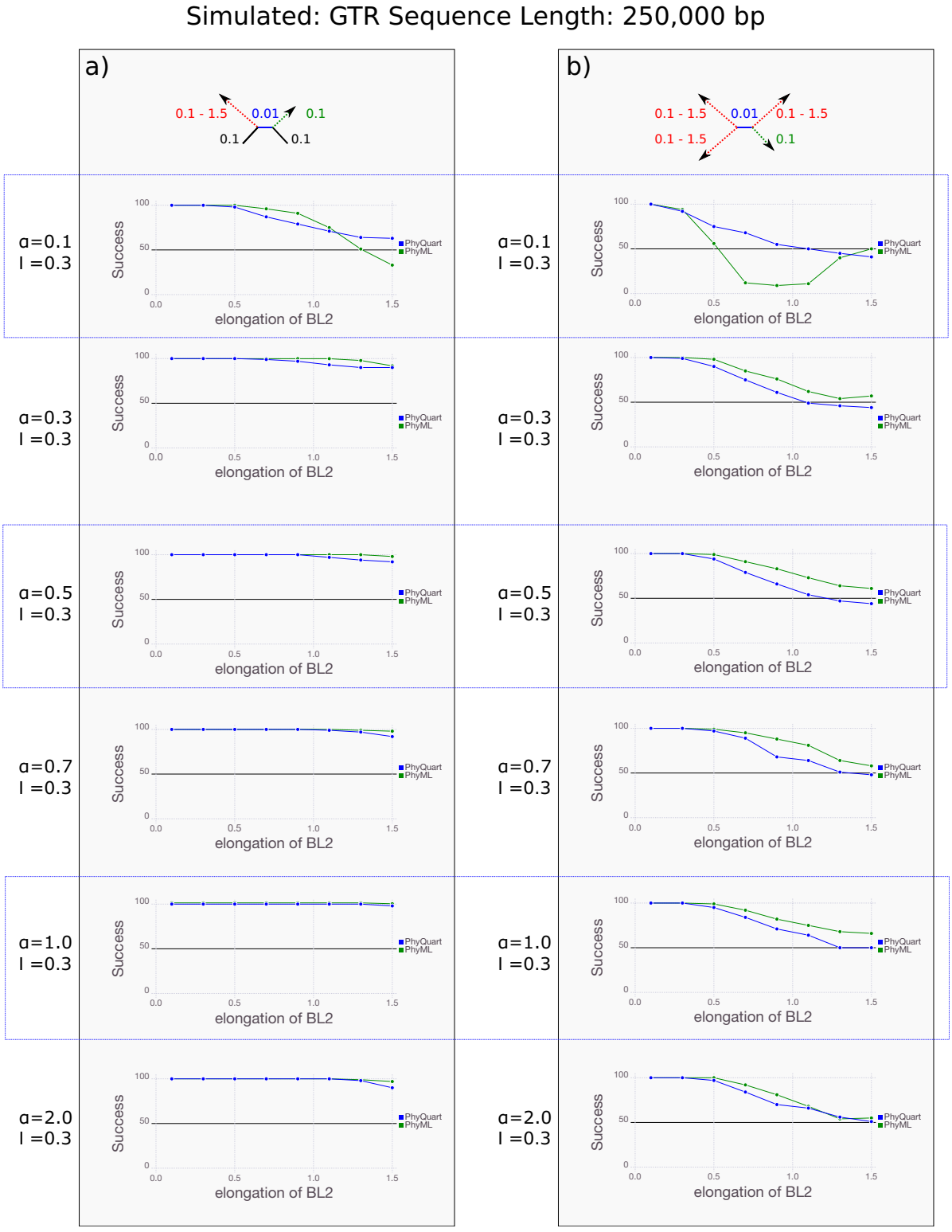

Fig 5. Quartet reconstruction success given stepwise elongation of one or three terminal branches if sequence lengths equal 250 kbp. Reconstruction success of *PhyQuart* (blue) and ML (green) for different rate heterogeneities under different lengths of a) a single long terminal branch (BL2, x-axis) and b) three long terminal branches (BL2, x-axis), given 100 data replicates (y-axis) of 250 kbp length and a fixed alternative internal branch length of BL1 = {0.01}, summarized for α = {0.1, 0.3, 0.5, 0.7, 1.0, 2.0}. A detailed overview of all simulation results of both setups is given as supplementary information S4 and S5 Figs.

https://doi.org/10.1371/journal.pone.0183393.g005

combines ML estimation (to help correct for convergence) with *Hennigian* logics which are disregarded in conventional ML analyses, together with a simple approach to reducing apparent support in proportion to branch length asymmetries.

Our quartet simulations, show that *PhyQuart* and ML are either very successful or, if branch length heterogeneity is very high, are moderately successful (i.e., in 50% of simulations) in identifying correct topologies if either one or three terminal branches are long. In the classic LBA problem, with two long and two short terminal branches in a quartet, *PhyQuart* is quite successful in inferring correct topologies from very heterogenous sequence data if the alignment is large (more than 50 kbp) and can outperform ML, overcoming both long branch attraction and repulsion, independent of the chosen simulation assumptions. In the simulations, rate heterogeneity is rather less important for reconstruction success using *PhyQuart* than using ML. Except with very heterogenous sequence data ($\alpha = 0.1$), *PhyQuart* was successful in the majority of simulated cases even when internal branches were kept very short. The simulations show that the reconstruction success of ML decreases with increasing branch length differences even when there is only very minor model misspecification, whereas the performance of *PhyQuart* is only slightly affected by more extreme branch length conditions. It might be expected then that estimated ML models will often be much more inadequate with real, strongly heterogeneous data whereas the *PhyQuart* site-pattern analysis would be less affected by strongly heterogeneous rates of substitution and branch length inequalities. Certainly that possibility is worth investigating. The overall reconstruction success of *PhyQuart* is worst when if the substitution rate heterogeneity of underlying data is extremely high ($\alpha = 0.1$) and two adjacent-group sequences have very long branches compared to the internal branch. However—as shown by our simulation studies—the observed phylogenetic reconstruction success of ML is even worse for such data. Of course, despite conducting almost 173,000 simulations we have only considered a limited range of possible simulations on just four taxa and we have not taken into account the possibility of other sources of error that may result in or exacerbate model misspecification in real data, such as substantial alignment errors (e.g. [10, 89–94]), non-randomly distributed missing data (e.g. [95–98]), and compositional heterogeneity (e.g. [48, 84, 99–104]).

It must be stressed that the restriction of our comparison of *PhyQuart* with ML to quartet analysis is a substantial one given that quartet analysis is considered to exacerbate LBA. Thus we cannot generalise from our results to say that *PhyQuart* will ever outperform conventional ML applied, as it usually is, to larger phylogenetic trees, but this merits investigation if *PhyQuart* is to be of any practical use and further simulation studies investigating this are under way. Despite its potential drawbacks, the benefit of the the computational simplicity of quartet analyses is two-fold, allowing consideration of the evidential significance and calculation of expected frequencies of a small number of site patterns in the development of the *PhyQuart* score, and the ability to obtain and compare these scores for all three quartet tees and thereby get an indication of the strength of the signal detected by *PhyQuart*. Thus, *PhyQuart* support for possible quartet trees can be used directly as a quality measure for how good a data set fits to alternative quartet relationships before ML tree inference or for existing/published tree topologies. Based on our simulations, we suggest that in cases where, for any quartet of taxa, there are two long and two short terminal branches (and thus the potential for classical LBA) and ML and *PhyQuart* both provide good support the same relationship we can be more confident that the ML inference is not the result of LBA. Conversely, where *PhyQuart* and ML provide good support for conflicting relationships or in cases in which *PhyQuart* shows strong contradictory split support for at least one of the other two alternative quartet trees, then we should be more concerned that ML might be being misled by LBA. This does not directly imply that *PhyQuart* supports the correct quartet topology, but it should be seen as an

indication that the initially ML-reconstructed topology should be handled with caution. Furthermore, it can be stated that the higher the conflict of *PhyQuart* support for a given quartet tree, the more suspicious is the phylogenetic value of the data.

However, *PhyQuart* is likely to be useful only with large alignments such as in phylogenomic supermatrices and some next generation data types such as RADseq, and is not recommended for shorter sequences such as single gene analyses where stochastic errors in the split split-supporting site-pattern estimation are expected to dominate when trying to infer short internal branches. However, there is probably substantial room for improving the *PhyQuart* approach. For example, to estimate the amount of potentially convergent split signal for a given quartet tree *PhyQuart* uses a simple mean score derived from the two alternative quartet trees. Given that at most only one of these alternative quartet trees could be correct, this scoring function can be expected to differ from the actual number of convergences. The correction factor ω, which is used to reduce the counted observed and ML estimated number of positions with relevant site-patterns to approach a more balanced branch-length ratio, depends on the smallest number of observed singleton site-patterns and the total number of these. In our simulation, this reduces the impact of systematic bias in *PhyQuart*, especially in reconstructing quartet topologies with moderate and strong branch length differences (a detailed comparison of reconstruction success with and without correction factor ω for simulations with both Felsenstein and Farris topologies using alignments of 250 kbp sequence length are given in the supplementary information S1 Fig). However, due to varying substitution rates along branches and differences in multiple substitutions, the number of observed singleton site-patterns is unlikely to be linearly correlated with the number of split-supporting site-patterns and this might be expected to leads to underestimation of ω. Additionally, *PhyQuart* currently ignores potentially useful information in ambiguity states (e.g. [105–107]), or indel events (e.g. [108–110]). Another desiderable extension is for *PhyQuart* to be able to handle data partitions. Clearly, *PhyQuart* is not perfect, but it points the way to new split-supporting site-pattern based methods that allow users to investigate conflicting signals in macromolecular sequence data.

Whereas our simulations have focussed upon proof of concept using only quartets, the PENGUIN software allows the analysis of all quartets of taxa in larger trees, or form predefined quartets of multitaxon clans, and provides a new tool for evaluating contradicting signals that can be used to assess the robustness of relationships within a more complex tree. Generally, it can be stated that the higher the observed contradictory split signal, the more questionable is the reliability of the corresponding branch in a tree and the more suspicious are any high support values for that branch. The PENGUIN software allows users to produce a graphical output summarising signal strengths found for each sequence quartets. This may also be useful for identifying individual rogue taxa that are difficult to place due to ambiguous or weak phylogenetic signal [111]. This characteristic of rogue taxa should become visible when multiple quartets selected from predefined multi-taxon clans are analysed. We also see potential for *PhyQuart* to be used in combination with quartet-based supertree methods (e.g. [112, 113]), of which there are many, and for development of networks summarising conflicting signal. Because, unlike ML, the method makes use of the distinction between plesiomorphy and apomorphy it may provide information on the probable location of the root in trees or networks independent of any consideration of outgroups. The availability of split support information for all three possible quartet relationships and two alternative directions of character evaluation along the innermost branch can be seen as an advantage of the *PhyQuart* approach over conventional ML quartet analyses. The information can be further used in supertree techniques to improve the selection of highly informative and thus appropriate quartets (e.g. quartet topologies without much signal conflict). New ideas on how to use the *PhyQuart* information to build supertrees (e.g. through translation into pairwise support distance matrices

based on quartet analyses of multiple-taxon clans) have already been successfully tested in recent test studies and will be published soon.

Supporting information

S1 Fig. Complete results of 4-taxon simulations of 250 kbp long sequences given two elongated branches. Complete results of 4-taxon simulations based on stepwise BL2 elongations of two adjacent or non-adjacent terminal branches given 250 kbp long nucleotide alignment data. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.
(PDF)

S2 Fig. Complete results of 4-taxon simulations of 250 kbp long sequences given two elongated branches with and without using correction factor. Complete *PhyQuart* results of 4-taxon simulations with and without using correction factor ω based on stepwise BL2 elongations of two adjacent or non-adjacent terminal branches given 250 kbp long nucleotide alignment data. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.
(PDF)

S3 Fig. Complete results of 4-taxon simulations of sequences shorter $<$ 250 kbp given two elongated branches. Complete results of 4-taxon simulations based on stepwise BL2 elongations of two adjacent or non-adjacent terminal branches given nucleotide alignment data $<$ 250 kbp. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.
(PDF)

S4 Fig. Complete results of 4-taxon simulations of 250 kbp long sequences given a single elongated branch. Complete results of 4-taxon simulations based on stepwise BL2 elongations of one terminal branch given 250 kbp long nucleotide alignment data. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.
(PDF)

S5 Fig. Complete results of 4-taxon simulations of 250 kbp long sequences given three elongated branches. Complete results of 4-taxon simulations based on stepwise BL2 elongations of three terminal branches given 250 kbp long nucleotide alignment data. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.
(PDF)

Acknowledgments

We would like to thank Arndt von Haesseler, Bui Quang Min, and one anonymous reviewer for their constructive revision comments. Further, we would like to thank Seraina Klopfstein for helpful advices about the manual description, and Marco Thill for inspiring discussions about script programming.

Author Contributions

Conceptualization: Patrick Kück.

Data curation: Patrick Kück.

Formal analysis: Patrick Kück.

Funding acquisition: Patrick Kück, Mark Wilkinson, Johann W. Wägele.

Investigation: Patrick Kück.

Methodology: Patrick Kück, Christian Groß, Johann W. Wägele.

Project administration: Patrick Kück.

Software: Patrick Kück, Peter G. Foster.

Supervision: Patrick Kück, Johann W. Wägele.

Validation: Patrick Kück.

Visualization: Patrick Kück.

Writing – original draft: Patrick Kück, Johann W. Wägele.

Writing – review & editing: Patrick Kück, Mark Wilkinson, Peter G. Foster, Johann W. Wägele.

References

1. Haldane JBS. Possible worlds and other essays. Chatto and Windus, London; 1927.
2. Harvey PH, Brown LAJ, Smith MJ, Nee S. New uses for new phylogenies. Oxford University Press, New York; 1996.
3. Soltis PS, Soltis DE. Molecular systematics: assembling and using the tree of life. Taxon. 2001; 50(3): 663–677. https://doi.org/10.2307/1223700
4. Cracraft J, Donoghue MJ. Assembling the Tree of Life. Oxford University Press, New York; 2004.
5. Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proceedings of the National Academy of Sciences. 2015; 112:12764–12769. https://doi.org/10.1073/pnas.1423041112
6. Fuellen G, Wägele JW, Giegerich R. Minimum conflict: a divide-and-conquer approach to phylogenetic estimation. Bioinformatics. 2001; 17(12):1168–1178. https://doi.org/10.1093/bioinformatics/17.12.1168 PMID: 11751225
7. Gee H. Ending incongruence. Nature. 2003; 425:782. https://doi.org/10.1038/425782a PMID: 14574398
8. Brinkmann H, Van Der Giezen M, Zhou Y, Poncelin De Raucourt G, Phillipe H. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol. 2005; 54(5): 743–757. https://doi.org/10.1080/10635150500234609 PMID: 16243762
9. Delsuc F, Brinkmann H, Phillipe H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 2005; 6(5):361–375. https://doi.org/10.1038/nrg1603 PMID: 15861208
10. Jeffroy O, Brinkmann H, Delsuc F, Phillipe H. Phylogenomics: the beginning of incongruence? Trends Gene. 2006; 22:225–231. https://doi.org/10.1016/j.tig.2006.02.003
11. Nesnidal MP, Heimkampf M, Bruchhaus I, Hausdorf B. Compositional heterogeneity and phylogenomic inference of Metazoan relationships. Mol Biol Evol. 2010; 27(9):2095–2104. https://doi.org/10.1093/molbev/msq097 PMID: 20382658
12. Kück P, Mayer C, Wägele JW, Misof B. Long branch effects distort Maximum Likelihood phylogenies in simulations despite selection of the correct model. PLoS ONE. 2012; 7(5):e36593. https://doi.org/10.1371/journal.pone.0036593 PMID: 22662120
13. Nesnidal MP, Heimkampf M, Meyer A, Witek A, Bruchhaus I, Ibersberger I, et al. New phylogenomic data support the monophyly of Lophophorata and an Ectoproct-Phoronid clade and indicate that Polyzoa and Kryptrochozoa are caused by systematic bias. BMC Evol Biol. 2013; 13:253. https://doi.org/10.1186/1471-2148-13-253 PMID: 24238092
14. Philippe H, Delsuc F, Brinkmann H, Lartillot N. Phylogenomics. Annu Rev Ecol Evol S. 2005; 36: 541–562. https://doi.org/10.1146/annurev.ecolsys.35.112202.130205
15. Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Phillipe H. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol. 2007; 56:389–399. https://doi.org/10.1080/10635150701397643 PMID: 17520503
16. Brinkmann H, Phillipe H. Animal phylogeny and large-scale sequencing: progress and pitfalls. J Syst Evol. 2008; 46(3):274–286.

17. Huelsenbeck JP, Hillis DM. Success of phylogenetic methods in the four-taxon case. Syst Zool. 1993; 42:247–264.
18. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol. 1994; 11(3):459–468. PMID: 8015439
19. Yang Z, Goldman N, Friday A. Comparison of models for nucleotide substitution used in Maximum-Likelihood phylogenetic estimation. Mol Biol Evol. 1994; 11(2):316–324. PMID: 8170371
20. Sullivan J, Holsinger KE, Simon C. Among-site rate variation and the phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol Biol Evol. 1995; 12:988–1001. PMID: 8524051
21. Lockhart PJ, Larkum AW, Steel MA, Waddell PJ, Penny D. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc Natl Acad Sci U S A. 1996; 93: 1930–1934. https://doi.org/10.1073/pnas.93.5.1930 PMID: 8700861
22. Huelsenbeck JP. Is the Felsenstein zone a fly trap? Syst Biol. 1997; 46(1):69–74. https://doi.org/10.1093/sysbio/46.1.69 PMID: 11975354
23. Sanderson MJ, Wojciechowski MF, Hu JM, Sher-Khan T, Brady SG. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol Biol Evol. 2000; 17(5): 782–797. https://doi.org/10.1093/oxfordjournals.molbev.a026357 PMID: 10779539
24. Omilian AR, Taylor DJ. Rate acceleration and long-branch attraction in a conserved gene of cryptic Daphniid (Crustacea) species. Mol Biol Evol. 2001; 18(12):2201–2212. https://doi.org/10.1093/oxfordjournals.molbev.a003767 PMID: 11719570
25. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, et al. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc Natl Acad Sci. 2002; 290:972–977.
26. Dacks JB, Marinets A, Doolittle W, Cavalier-Smith T, Logsdon JM Jr. Analyses of RNA polymerase II genes from free living protists: Phylogeny, long branch attraction, and the eukaryotic big bang. Mol Biol Evol. 2002; 19:830–840. https://doi.org/10.1093/oxfordjournals.molbev.a004140 PMID: 12032239
27. Lemmon AR, Moriarty EC. The importance of proper model assumptions in Bayesian phylogenetics. Syst Biol. 2004; 53:265–277. https://doi.org/10.1080/10635150490423520 PMID: 15205052
28. Gaucher EA, Miyamoto MM. A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. Mol Phylogenet Evol. 2005; 37:928–931. https://doi.org/10.1016/j.ympev.2005.03.027 PMID: 16291095
29. Nishihara H, Okada N, Hasegawa M. Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol. 2007; 8(9):R199. https://doi.org/10.1186/gb-2007-8-9-r199 PMID: 17883877
30. Fischer M, Steel M. Sequence length bounds for resolving a deep phylogenetic divergence. J Theor Biol. 2009; 256:247–252. https://doi.org/10.1016/j.jtbi.2008.09.031 PMID: 18955066
31. Hallström BM, Janke A. Mammalian Evolution May not Be Strictly Bifurcating. Mol Biol Evol. 2010; 27(12):2804–2816. https://doi.org/10.1093/molbev/msq166 PMID: 20591845
32. Bergsten J. A review of long-branch attraction. Cladistics. 2005; 21:163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x
33. Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W. Mitochondrial protein phylogeny joins myriapods with chelicerates. Nature. 2001; 413:154–157. https://doi.org/10.1038/35093090 PMID: 11557978
34. Pisani D. The timetree of life. Oxford University Press, New York; 2009.
35. Meusemann K, von Reumont BM, Simon S, Roeding F, Kück P, Strauss S, et al. A phylogenomic approach to resolve the arthropod tree of life. Mol Biol Evol. 2010; 27(11):2451–2464. https://doi.org/10.1093/molbev/msq130 PMID: 20534705
36. Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, et al. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc R Soc B. 2010; 278:298–306. https://doi.org/10.1098/rspb.2010.0590 PMID: 20702459
37. Fernandez R, Edgecombe GD, Giribet G. Exploring phylogenomic relationships within Myriapoda: should high matrix occupancy be the goal? bioRxiv. 2015; p. 1–52.
38. Podsiadlowski L, Braband A. The complete mitochondrial genome of the sea spider *Nymphon gracile* (Arthropoda: Pycnogonida). BMC Genomics. 2006; 7:284. https://doi.org/10.1186/1471-2164-7-284 PMID: 17087824
39. Arabi J, Cruaud C, Couloux A, Hassanin A. Studying sources of incongruence in arthropod molecular phylogenies: Sea spiders (Pycnogonida) as a case study. Mol Phylogenet Evol. 2010; 333(5): 438–453.
40. Sharma PP, Kaluziak ST, R PPA, González GH, Wheeler WC, Giribet G. Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal). Mol Biol Evol. 2014; p. 1–86.

41. Struck TH, Wey-Fabrizius AR, Golombek A, Hering I, Weigert A, Bleidorn C, et al. Platyzoan Paraphyly Based on Phylogenomic Data Supports a Noncoelomate Ancestry of Spiralia. Mol Biol Evol. 2014; 31:1833–1849. https://doi.org/10.1093/molbev/msu143 PMID: 24748651

42. Laumer CE, Bekkouche N, Kerbl A, Goetz F, Neves RC, Sorensen MV, et al. Spiralian phylogeny informs the evolution of microscopic lineages. Curr Biol. 2015; 25(15):2000–2006. https://doi.org/10.1016/j.cub.2015.06.068 PMID: 26212884

43. Kocot KM, Struck TH, Merkel J, Waits DS, Todt C, Brannock PM, et al. Phylogenomics of Lophotrochozoa with consideration of systematic error. Syst Biol. 2016; https://doi.org/10.1093/sysbio/syw079

44. Dellaporta SL. Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum. PNAS. 2006; 103(23):8751–8756. https://doi.org/10.1073/pnas.0602076103 PMID: 16731622

45. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 2008; 452:745–749. https://doi.org/10.1038/nature06614 PMID: 18322464

46. Pick KS, Phillipe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, et al. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol. 2010; 27(9):1983–1987. https://doi.org/10.1093/molbev/msq089 PMID: 20378579

47. Nosenko T, Schreiber F, Adamska M, Adamski M, Eitel M, Hammel J, et al. Deep metazoan phylogeny: When different genes tell different stories. Mol Phylogenet Evol. 2013; 67(1):223–233. https://doi.org/10.1016/j.ympev.2013.01.010 PMID: 23353073

48. Whelan NV, Kocot KM, Moroz LL, Halanych KM. Error, signal, and the placement of Ctenophora sister to all other animals. PNAS. 2015; 112(18):5773–5778. https://doi.org/10.1073/pnas.1503453112 PMID: 25902535

49. Kück P, Wägele JW. Plesiomorphic character states cause systematic errors in molecular phylogenetic analyses: a simulation study. Cladistics. 2015; 0:1–18.

50. Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Syst Biol. 1978; 27(4):401–410. https://doi.org/10.1093/sysbio/27.4.401

51. Gaut S, Lewis PO. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol. 1995; 12(1):152–162. https://doi.org/10.1093/oxfordjournals.molbev.a040183 PMID: 7877489

52. Yang Z. Among-site rate variation and its impact on phylogenetic analyses. Tree. 1996; 11(9): 367–372. PMID: 21237881

53. Sullivan J, Swofford DL. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J Mammal Evol. 1997; 4:77–86. https://doi.org/10.1023/A:1027314112438

54. Pol D, Siddal ME. Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics. 2001; 17:266–281. https://doi.org/10.1006/clad.2001.0172

55. Sullivan J, Swofford DL. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst Biol. 2001; 50(5):723–729. https://doi.org/10.1080/106351501753328848 PMID: 12116942

56. Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Roger JS. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst Biol. 2001; 50(4):525–539. https://doi.org/10.1080/10635150117959 PMID: 12116651

57. Poe S. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst Biol. 2003; 52:423–428. https://doi.org/10.1080/10635150390197046 PMID: 12775529

58. Rosenberg MS, Kumar S. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol Biol Evol. 2003; 20(4):610–621. https://doi.org/10.1093/molbev/msg067 PMID: 12679548

59. Anderson FE, Swofford DL. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol Phylogenet Evol. 2004; 33:440–451. https://doi.org/10.1016/j.ympev.2004.06.015 PMID: 15336677

60. Phillips MJ, Delsuc F, Penny D. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol. 2004; 21(7):1455–1458. https://doi.org/10.1093/molbev/msh137 PMID: 15084674

61. Lartillot N, Brinkmann H, Phillipe H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 2007; 7(Suppl 1):S4. https://doi.org/10.1186/1471-2148-7-S1-S4 PMID: 17288577

62. Wägele JW, Mayer C. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol Biol. 2007; 7(1):147. https://doi.org/10.1186/1471-2148-7-147 PMID: 17725833

63. Susko E. Bayesian Long Branch Attraction Bias and Corrections. Syst Biol. 2015; 64(2):243–255. https://doi.org/10.1093/sysbio/syu099 PMID: 25432892

64. Felsenstein J. Inferring phylogenies. Sinauer Associates, Sunderland; 2004.

65. Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, et al. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature. 1997; 387:489–493. https://doi.org/10.1038/387489a0 PMID: 9168109

66. Hendy M, Penny D. A framework for the quantitive study of evolutionary trees. Syst Zool. 1989; 38: 297–309. https://doi.org/10.2307/2992396

67. Hillis DM. Inferring complex phylogenies. Nature. 1996; 383:130–131. https://doi.org/10.1038/383130a0 PMID: 8774876

68. Graybeal A. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol. 1998; 47:9–17. https://doi.org/10.1080/106351598260996 PMID: 12064243

69. Hillis DM. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol. 1998; 47:3–8. https://doi.org/10.1080/106351598260987 PMID: 12064238

70. Poe S. The effect of taxonomic sampling on accuracy of phylogenetic estimation: test case of a known phylogeny. Mol Biol Evol. 1998; 15:1086–1090. https://doi.org/10.1093/oxfordjournals.molbev.a026008

71. Rannala B, Huelsenbeck JP, Yang ZH, Nielsen R. Taxon sampling and the accuracy of large phylogenies. Syst Biol. 1998; 47:702–710. https://doi.org/10.1080/106351598260680 PMID: 12066312

72. Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol. 2002; 51:664–671. https://doi.org/10.1080/10635150290102357 PMID: 12228008

73. Zwickl DJ, Hillis DM. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 2002; 51:588–598. https://doi.org/10.1080/10635150290102339 PMID: 12228001

74. Huelsenbeck JP, Lander KM. Frequent inconsistency of parsimony under a simple model of cladogenesis. Syst Biol. 2003; 52:641–648. https://doi.org/10.1080/10635150390235467 PMID: 14530131

75. Hirt RP, Logsdon JM, Healy B, Dorey MW, Doolittle WF, Embley TM. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proceedings of the National Academy of Sciences. 1999; 96:580–585. https://doi.org/10.1073/pnas.96.2.580

76. Philippe H. Early–branching or fast–evolving eukaryotes? An answer based on slowly evolving positions. Proceedings of the Royal Society of London B: Biological Sciences. 2000; 267:1213–1221. https://doi.org/10.1098/rspb.2000.1130

77. Pisani D. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Systematic Biology. 2004; 53:978–989. https://doi.org/10.1080/10635150490888877 PMID: 15764565

78. Telford MJ, Lowe CJ, Cameron CB, Ortega-Martinez O, Aronowicz J, Oliveri P, et al. Phylogenomic analysis of echinoderm class relationships supports Asterozoa. Proc R Soc B. 2014; 281:20140479. https://doi.org/10.1098/rspb.2014.0479 PMID: 24850925

79. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. Phylogenetic inference. Sinauer Associates, Sunderland; 1996.

80. Wilkinson M, Cotton JA. Supertree methods for building the tree of life: divide-and-conquer approaches to large phylogenetic problems. No. 72 in The Systematic Association Special Volume Series. CRC Press, New York; 2006.

81. Dress AWM, Huson D, Moulton V. Analyzing and visulaizing sequence and distance data using SplitsTree. Discrete Appl Math. 1997; 71:95–109. https://doi.org/10.1016/S0166-218X(96)00059-5

82. Wilkinson M. Split support and split conflict randomization tests in phylogenetic inference. Syst Biol. 1998; 47(4):673–685. https://doi.org/10.1080/106351598260662 PMID: 12066310

83. Hennig W. Phylogenetic systematics. Annu Rev Entomol. 1965; 10:97–116. https://doi.org/10.1146/annurev.en.10.010165.000525

84. Foster PG. Modeling compositional heterogeneity. Syst Biol. 2004; 53:485–495. https://doi.org/10.1080/10635150490445779 PMID: 15503675

85. Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM. Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends in Ecology and Evolution. 2007; 22. https://doi.org/10.1016/j.tree.2007.01.002 PMID: 17239486

86. Fletcher W, Yang Z. INDELible: A flexible simulator of biological sequence evolution. Mol Biol Evol. 2009; 26(8):1879–1888. https://doi.org/10.1093/molbev/msp098 PMID: 19423664

87. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003; 52(5):696–704. https://doi.org/10.1080/10635150390235520 PMID: 14530136

88. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. PhyML 3.0: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59(3):307–321.

89. Susko E, Spencer M, Roger AJ. Biases in phylogenetic estimation can be caused by random sequence alignments. J Mol Evol. 2005; 61:351–359. https://doi.org/10.1007/s00239-004-0352-9 PMID: 16044245

90. Ogden TH, Rosenberg M. Multiple sequence alignment accuracy and phylogenetic inferrence. Syst Biol. 2006; 55(2):314–328. https://doi.org/10.1080/10635150500541730 PMID: 16611602

91. Wong KMA, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science. 2008; 319(5862):473–476. https://doi.org/10.1126/science.1151532 PMID: 18218900

92. Misof B, Misof K. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst Biol. 2009; 58(1):21–34. https://doi.org/10.1093/sysbio/syp006 PMID: 20525566

93. Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol. 2012; 29(1):1–5. https://doi.org/10.1093/molbev/msr177 PMID: 21772063

94. Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucl Acids Res. 2015; 43(W1): W7–W14. https://doi.org/10.1093/nar/gkv318 PMID: 25883146

95. Simmons MP, Freudenstein JV. Spurious 99% bootstrap and jackknife support for unsupported clades. Mol Phylogenet Evol. 2011; 61:177–191. https://doi.org/10.1016/j.ympev.2011.06.003 PMID: 21703355

96. Simmons MP. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics. 2012; 28:208–222. https://doi.org/10.1111/j.1096-0031.2011.00375.x

97. Simmons MP. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. Mol Phylogenet Evol. 2012; 62:472–484. https://doi.org/10.1016/j.ympev.2011.10.017 PMID: 22067131

98. Roure B, Baurain D, Phillipe H. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol Biol Evol. 2013; 30(1):197–214. https://doi.org/10.1093/molbev/mss208 PMID: 22930702

99. Embley T, Thomas R, Williams R. Reduced thermophilic bias in the 16S rDNA sequence from Thermus ruber provides further support for a relationship between Thermus and Deinococcus. Syst Appl Microbiol. 1993; 16(1):25–29. https://doi.org/10.1016/S0723-2020(11)80247-X

100. Lake JA. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proceedings of the National Academy of Sciences. 1994; 91(4):1455–1459. https://doi.org/10.1073/pnas.91.4.1455

101. Lockhart PJ, Steel MA, Hendy MD, Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol. 1994; 11(4):605–612. PMID: 19391266

102. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci. 2008; 105:20365–20361. https://doi.org/10.1073/pnas.0810647105

103. Foster PG, Cox CJ, Embley TM. The primary divisions of life: a phylogenomic approach employing composition?heterogeneous methods. Philos Tr R Soc B Biol Sci. 2009; 364:2197–2207. https://doi.org/10.1098/rstb.2009.0034

104. Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM. A congruent phylogenomic signal places eukaryotes within the Archaea. Proc R Soc B Biol Sci. 2012; 279:4870–4879. https://doi.org/10.1098/rspb.2012.1795

105. Huelsenbeck JP. When are fossils better than extant taxa in phylogenetic analysis? Syst Zool. 1991; 40:458–469. https://doi.org/10.2307/2992240

106. Dunn KA, McEachran JD, Honeycutt RL. Molecular phylogenetics of myliobatiform fishes (Chondrichthyes: Myliobatiformes), with comments on the effects of missing data on parsimony and likelihood. Mol Phylogenet Evol. 2003; 27:259–270. https://doi.org/10.1016/S1055-7903(02)00442-6 PMID: 12695090

107. Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. The effect of ambiguous data on phylogenetic estimates obtained by Maximum Likelihood and bayesian inference. Syst Biol. 2009; 58:130–145. https://doi.org/10.1093/sysbio/syp017 PMID: 20525573

108. Simmons MP, Ochoterena H. Gaps as characters in sequence-based phylogenetic analyses. Syst Biol. 2000; 49(2):369–381. https://doi.org/10.1080/10635159950173889 PMID: 12118412

109. Hartmann S, Vision TJ. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol Biol. 2008; 8:95:S13. https://doi.org/10.1186/1471-2148-8-95 PMID: 18366758

110. Ashkenazy H, Cohen O, Pupko T, Huchon D. Indel reliability in indel-based phylogenetic inference. Genome Biol Evol. 2014; 6(12):3199–3209. https://doi.org/10.1093/gbe/evu252 PMID: 25409663

111. Wilkinson M. Majority-rule reduced consensus trees and their use in bootstrapping. Mol Biol Evol. 1996; 13(3):437–444. https://doi.org/10.1093/oxfordjournals.molbev.a025604 PMID: 8742632

112. Bininda-Emonds ORP, Donoghue MJ. Phylogenomic Supertrees: Combining information to reveal the Tree of Life. Springer Science+Buisness Media, B.V.; 2004.

113. Snir S, Rao S. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. Mol Phylogenet Evol. 2012; 62:1–8. https://doi.org/10.1016/j.ympev.2011.06.021 PMID: 21762785