Supplementary text for: Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets

Martin R. Smith (martin.smith@durham.ac.uk)

2019-01-10

Contents

This document has been generated from an R markdown file, which contains the source code used to generate figures and conduct analyses, and is provided in the Electronic Supplementary Material that accompanies the main article [1].

---

1 Tree distance metrics

A number of metrics area available to quantify the similarity between two undirected topologies (i.e. unrooted trees with no edge lengths).

1.1 SPR metric

The subtree pruning and regrafting (SPR) distance [2] counts the number of SPR rearrangements necessary to transform Tree A into Tree B.

1.2 Path difference metric

The length of a path from one tip to another in a tree is the number of edges within the tree that must be crossed to navigate from one tip to the other.

Given two trees, is possible to calculate the difference in path length between each pair of tips.

The path difference metric [3] is the square root of the sum of squares of each of these differences.

1.3 Partition metric

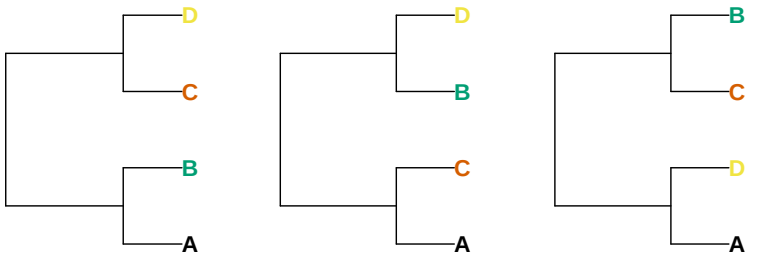
The Robinson-Foulds (RF or 'partition') metric [3,4] measures the symmetric difference between two trees by adding the number of bipartitions that are present in tree A (but not tree B) to the number of bipartitions present in tree B (but not tree A).

It is most useful when the trees to be compared are very similar; it has a low range of integer values, limiting its ability to distinguish between trees [3].

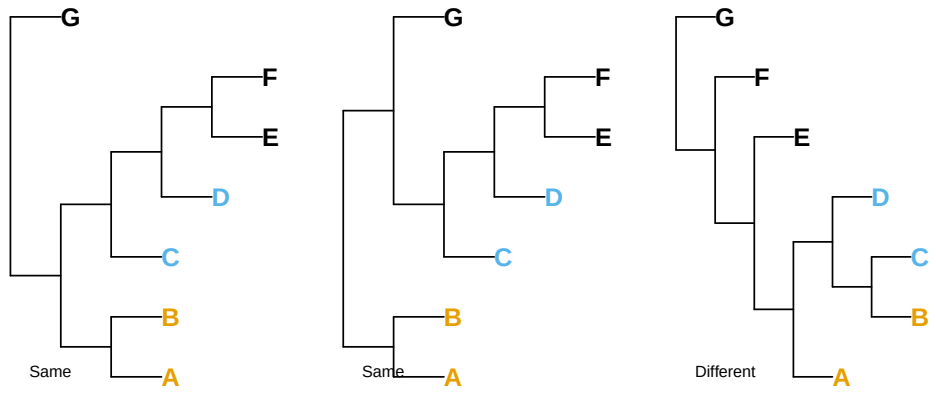
1.4 Quartet metric

Instead of partitions, symmetric differences can be measured by counting the number of four-taxon statements (quartets) that differ between two trees [5,6].

For any four tips A, B, C and D, a bipartition on a bifurcating tree will separate tip A and either B, C or D from the other two tips. That is to say, removing all other tips from the tree will leave one of these three trees:



Thus two of the random trees below share the quartet `(A, B), (C, D)`, whereas the third does not; these four tips are divided into `(A, D), (B, C)`.



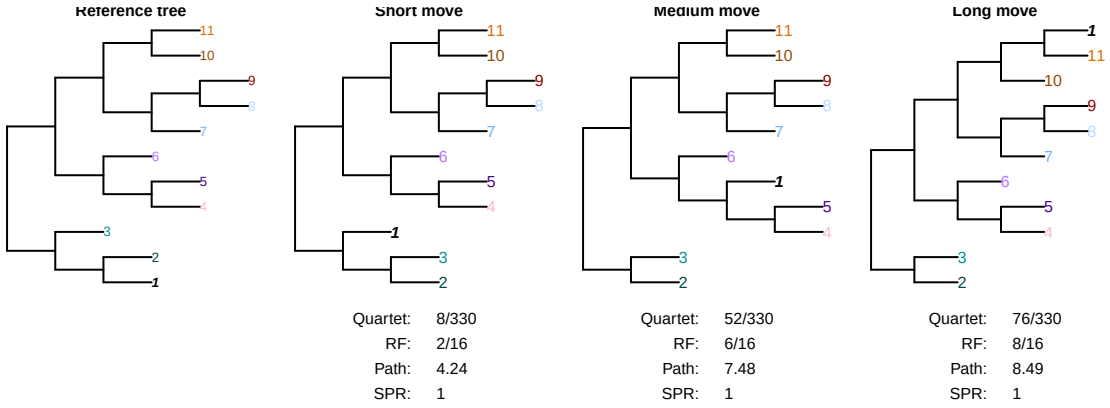
There are $\binom{n}{4}$ groups of four taxa in a tree with n tips; for each of these groups, one of the three trees above will be consistent with a given tree. As such, two identical trees will have a quartet distance of 0, and a random pair of trees will have an expected $\binom{n}{4}/3$ quartets in common. Because quartets are not independent of one another, no pair of trees with six or more tips can have all $\binom{n}{4}$ quartets in common [3].

2 Desired behaviour of tree distance metrics

The advantages of the quartet symmetric difference over other tree distance metrics [2] are best illustrated by examining a set of example trees.

2.1 Moving a single taxon

If trees differ only in the location of a single taxon (see taxon 1 in the trees below), then the distance between two trees should correspond to the distance that this taxon has been moved.



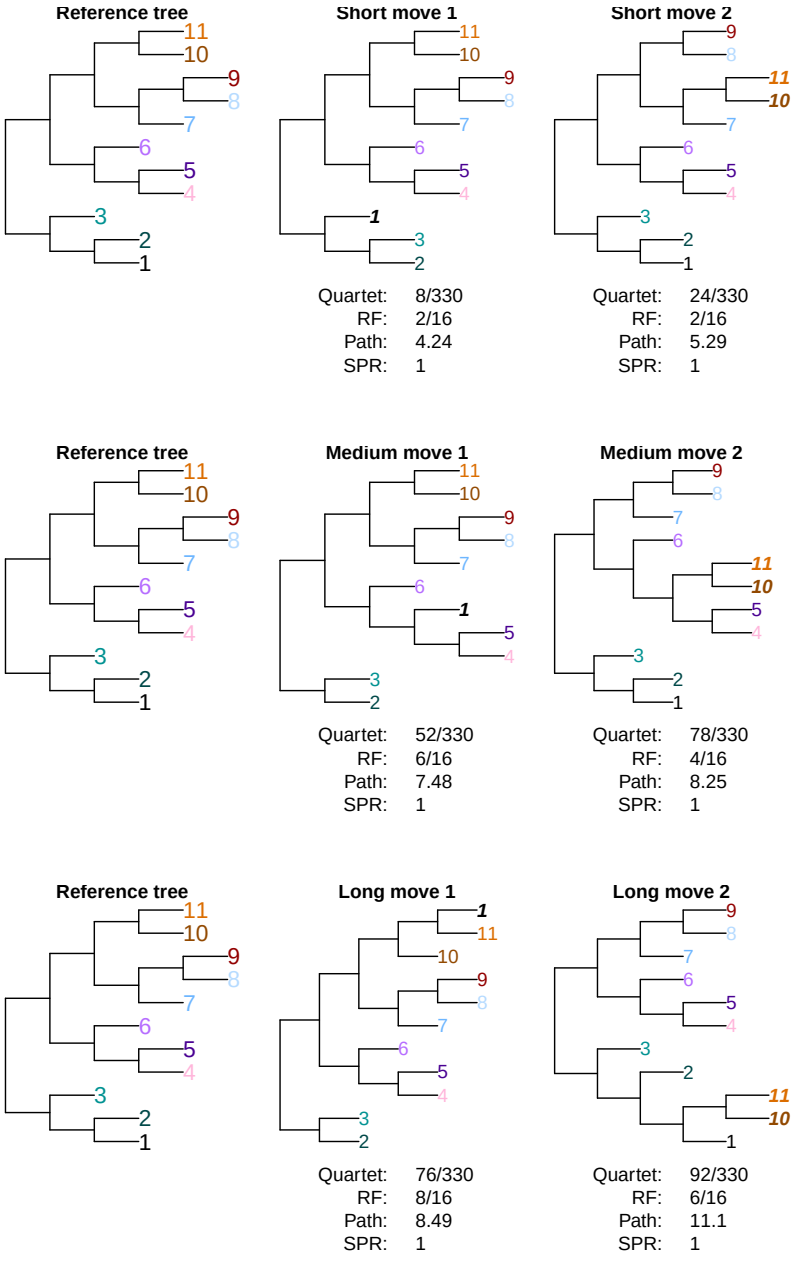
The subtree pruning and regrafting (SPR) distance does not distinguish between these trees, as they differ only in the placement of a single tip. The Robinson-Foulds, path difference and quartet metrics, in contrast, recognize trees in which this tip has been moved further as more distant from the starting tree.

2.2 Moving two taxa

Intuitively, moving a pair of tips on a tree should lead to higher tree distances than moving a single tip. In the case of a short move, the RF distance does not differ whether one or two tips are moved. For larger

moves, however, the RF distance is *less* when two tips are moved than when a single tip is moved. The path and quartet metrics perform as expected.

The trees below differ from a reference tree in the position of a single tip (tip 1), or a pair of tips (tips 10 and 11), which have been moved a short, medium or long distance from their original positions.

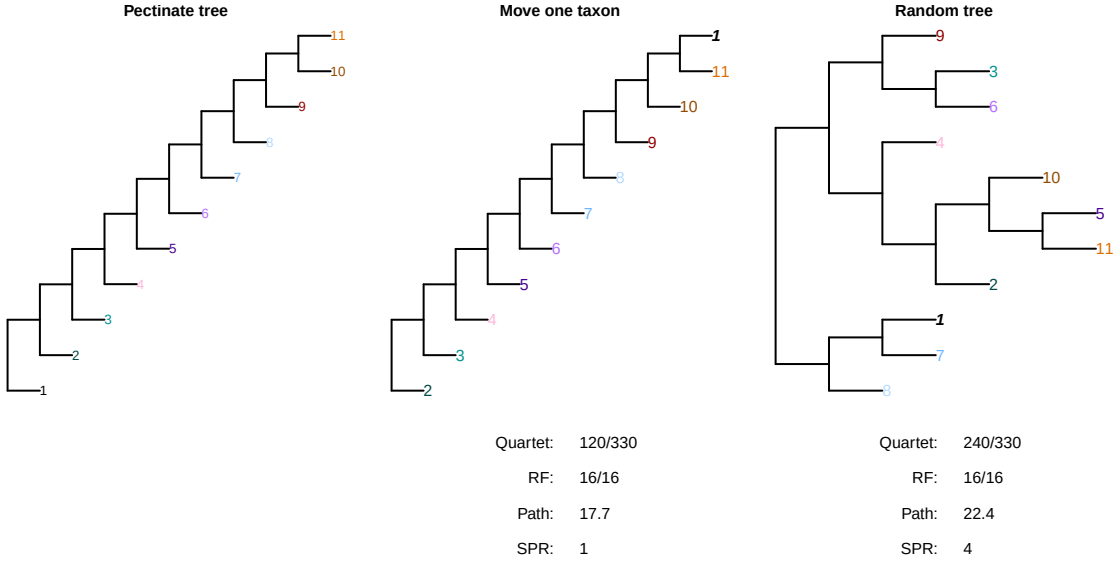


2.3 Maximum distance

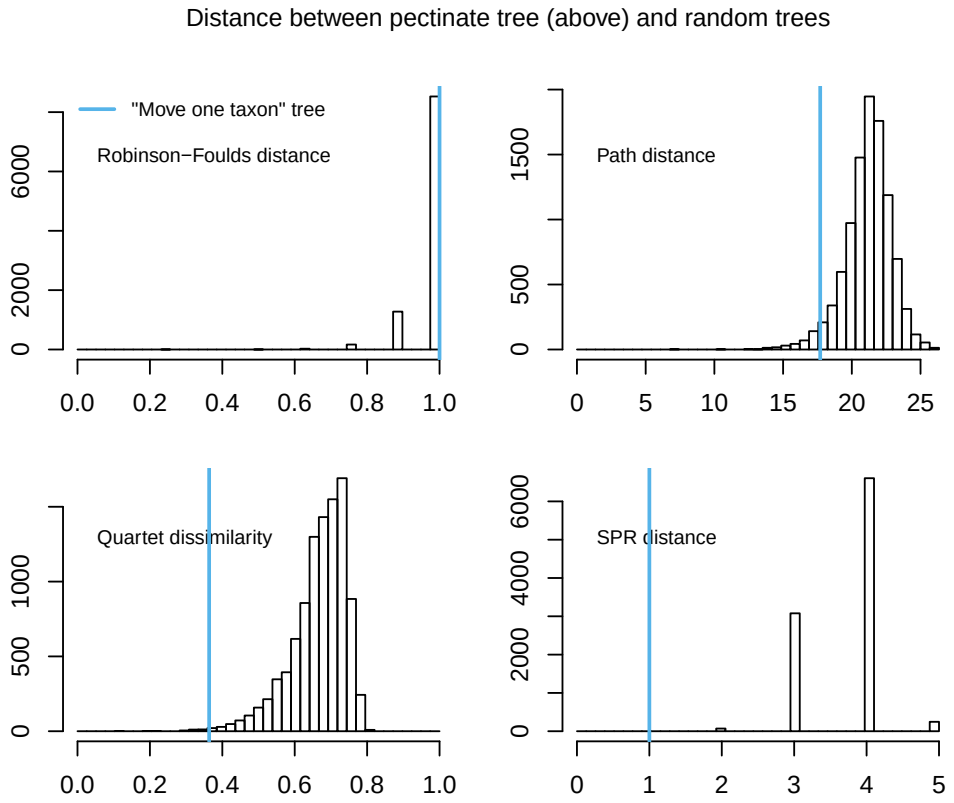
A distance metric should distinguish slightly-perturbed trees from random trees and those that are more different from the starting tree than expected by chance.

The Robinson-Foulds metric can reach its maximum value when a single taxon is relocated from the most

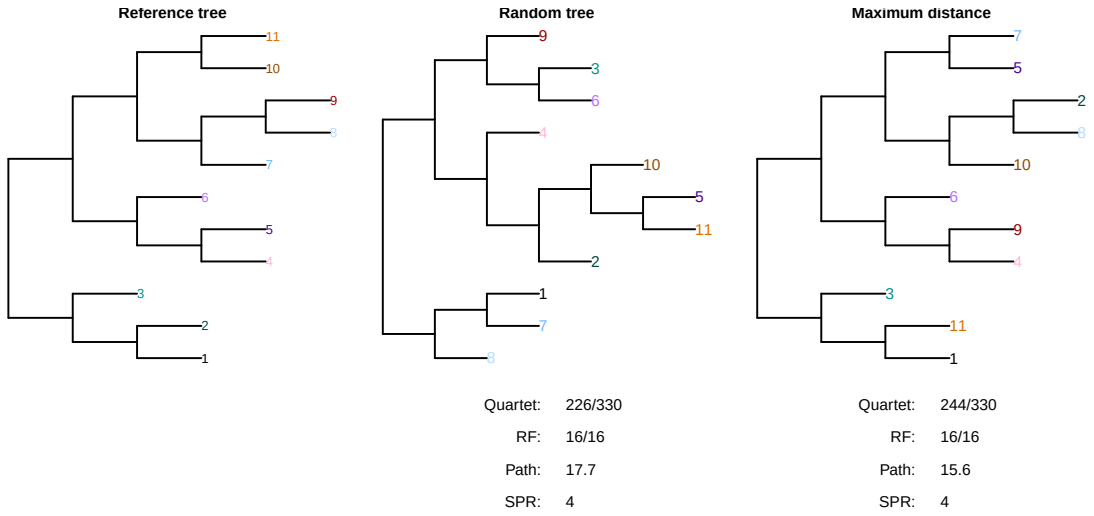
basal to the most derived point of a pectinate tree, representing a maximal value despite retaining relationship information about all other taxa.



A notable proportion of random trees receive a lower RF distance from the original tree, even though they do not show any structural similarity. This is not the case with the quartet symmetric difference.



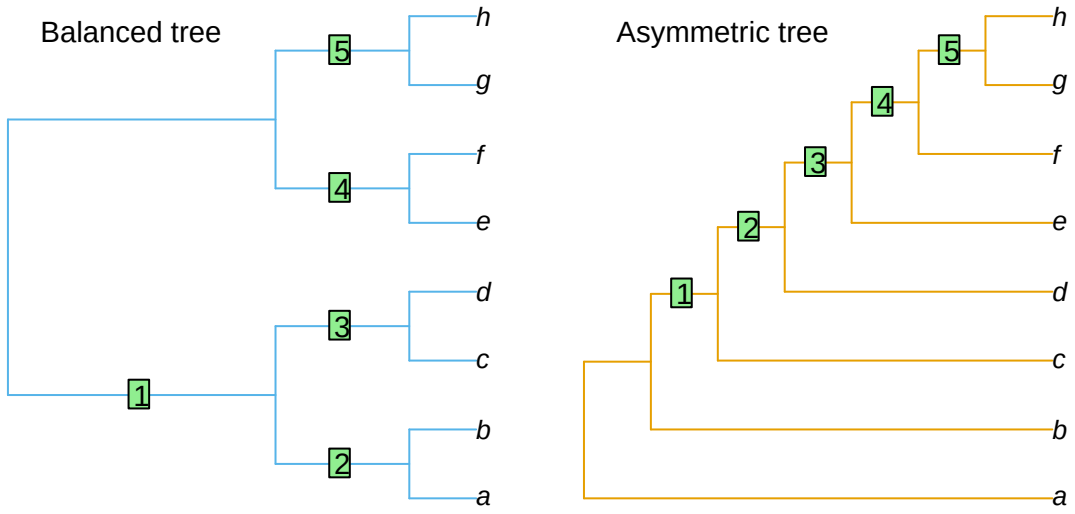
An advantage of the quartet symmetric distance is that the normalized metric of a random tree is $\frac{2}{3}$ [3,6]. As such, trees that are more different than expected by chance can be readily recognized, as their distance metric will be greater than $\frac{220}{330}$. The 'maximum distance' tree depicted below was identified using the R package `TreeSearch` [7], using the quartet difference from the reference tree as an optimality criterion.



2.4 Unit equivalence

A further shortcoming of the RF metric is that not all partitions represent an equivalent amount of information. A partition distance of 1 could mean that two trees differ in an uninformative partition, or a more informative partition. All quartets, in contrast, are equally informative.

Consider a balanced and an unbalanced eight-taxon tree:



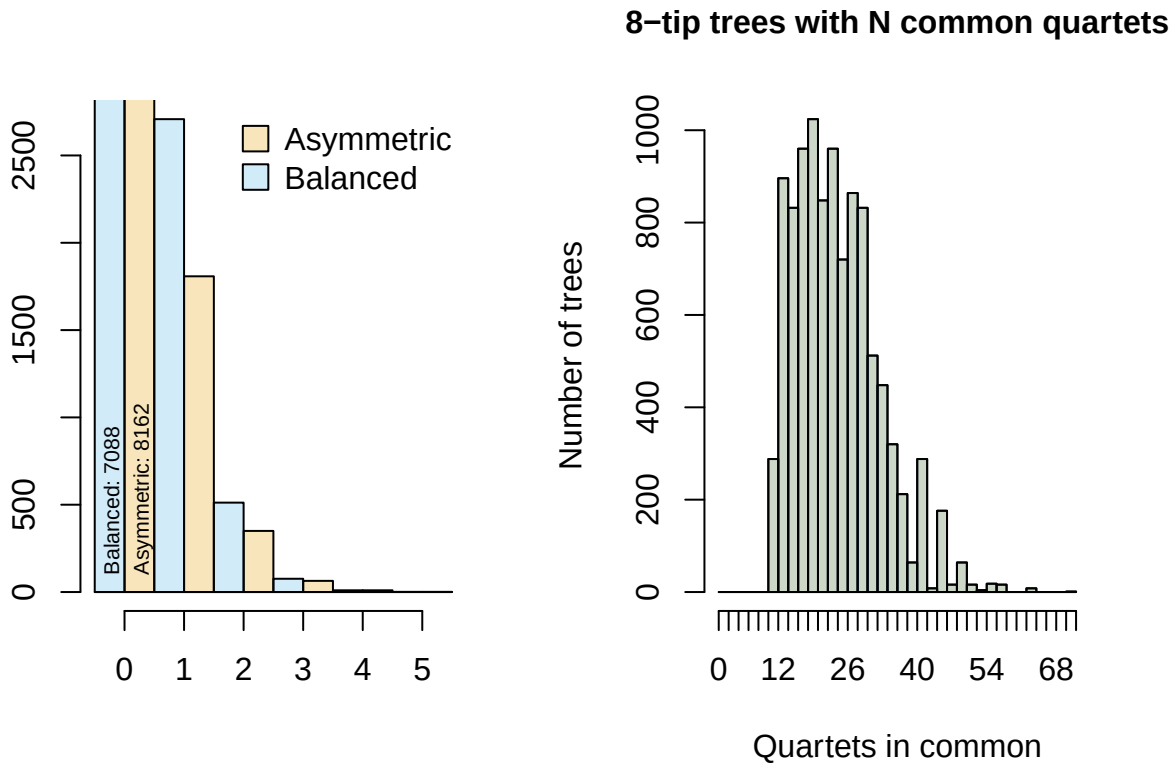
Each tree divides the eight taxa into five bipartition splits.
The information content (Shannon entropy) of a split can be calculated based on what proportion of eight-tip trees contain the split in question. This is a function of the evenness of the split:

	Matching trees	p(Match in random tree)	Information content / bits
Partition size: 2:6	945	0.0909	3.459432
Partition size: 3:5	315	0.0303	5.044394
Partition size: 4:4	225	0.0216	5.529821

In the first tree, split 1 is even, dividing four taxa from four others (`4:4`); splits 2–5 are maximally uneven (`2:6`). The total information content of these five splits is 19.37, whereas that of the five splits in the second

tree, of sizes `2:6`, `3:5`, `4:4`, `3:5` and `2:6`, is 22.54. Put another way, a random tree will on average share more partitions with the balanced tree (whose partitions are predominantly uneven and thus likely to be matched) than the asymmetric tree (which contains more even partitions that are less likely to occur in a random tree).

Of the 10 395 eight-tip trees, many more bear at least one partition in common with a balanced tree than with an asymmetric tree, whereas the distribution of quartets is identical:



2.5 Unresolved trees

Whereas the path distance and SPR metrics are only defined on bifurcating trees, symmetric difference approaches can be applied to trees that contain polytomies – i.e. not every node is resolved as bifurcating.

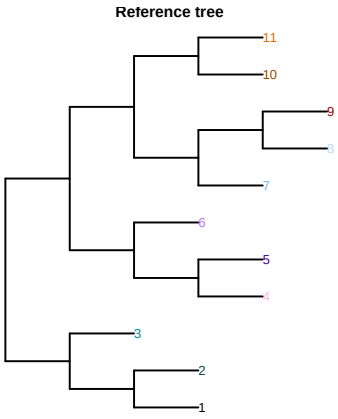
2.6 Conclusion

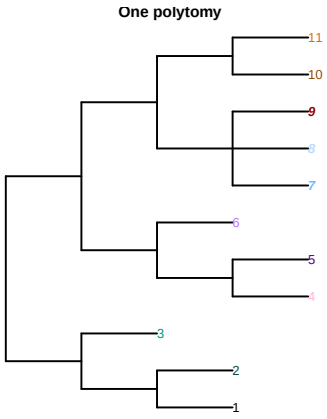
Quartet dissimilarity is the only available metric of tree distance that fulfils all of the following desiderata:

- Allocates trees higher distances if a clade moves greater distances
- Allocates trees higher distances if a the clade that is moved is larger
- Distinguishes contradicted from unresolved information in trees that are not fully bifurcating (resolved)
- Identifies pairs of trees that are more random than expected by chance
- Does not reach its maximum value after relatively trivial rearrangements

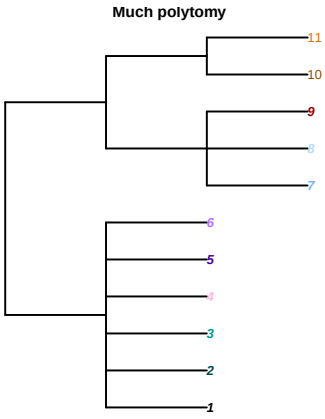
3 Calculating resolution and accuracy

One way to modify a tree topology is to reduce its resolution by collapsing nodes, without changing any of the relationships presented within the tree. The trees below have been derived from a reference tree by collapsing one and many nodes:





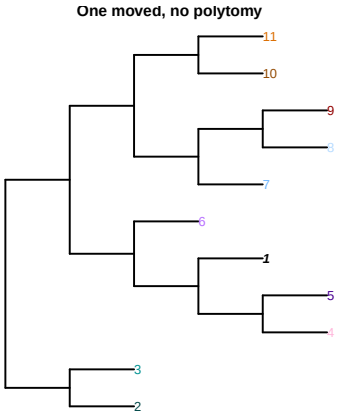
Quartets contradicted	0/330
Quartets unresolved	8/330
Partitions contradicted	NA/8
Partitions unresolved	NA/8



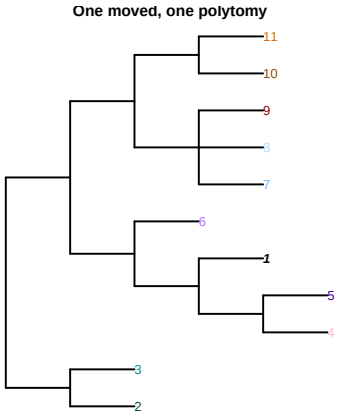
Quartets contradicted	0/330
Quartets unresolved	123/330
Partitions contradicted	NA/8
Partitions unresolved	NA/8

These trees do not contain any quartets or partitions that are not present in the reference tree, though they do contain a number of unresolved quartets and partitions.

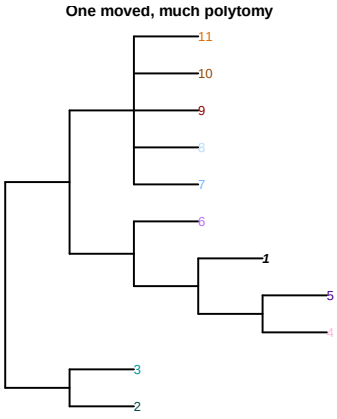
We can alternatively choose to change the topology, and then collapse some nodes. The following trees represent the same loss of resolution as the previous trees, but applied to a tree in which one tip (tip 1) has been moved relative to the reference tree:



Quartets contradicted	52/330
Quartets unresolved	0/330
Partitions contradicted	NA/8
Partitions unresolved	NA/8



Quartets contradicted	52/330
Quartets unresolved	8/330
Partitions contradicted	NA/8
Partitions unresolved	NA/8

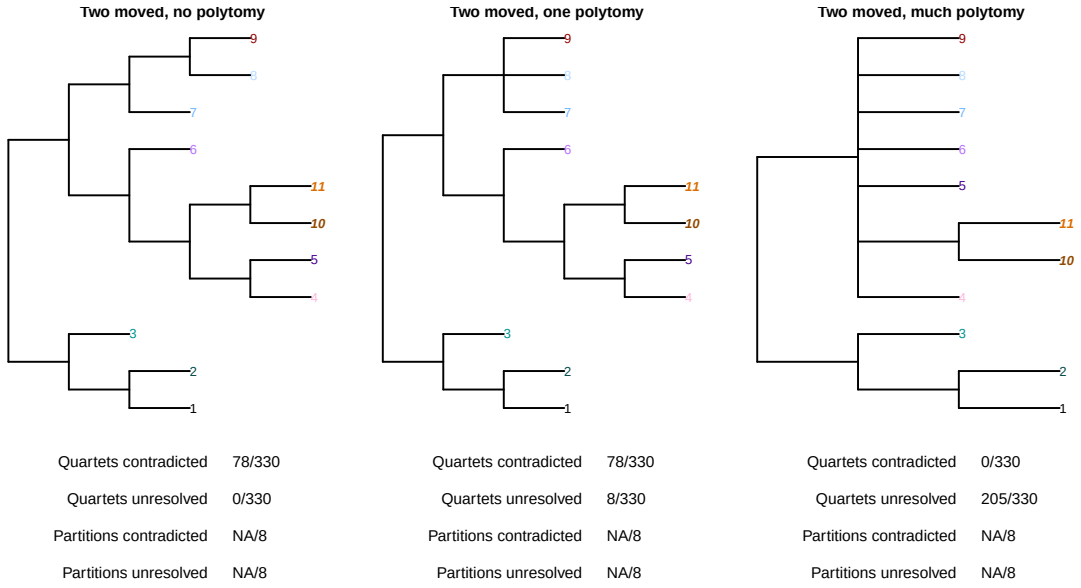


Quartets contradicted	52/330
Quartets unresolved	65/330
Partitions contradicted	NA/8
Partitions unresolved	NA/8

This causes results in trees that contradict a number of partitions or quartets that occurred in the original reference tree.

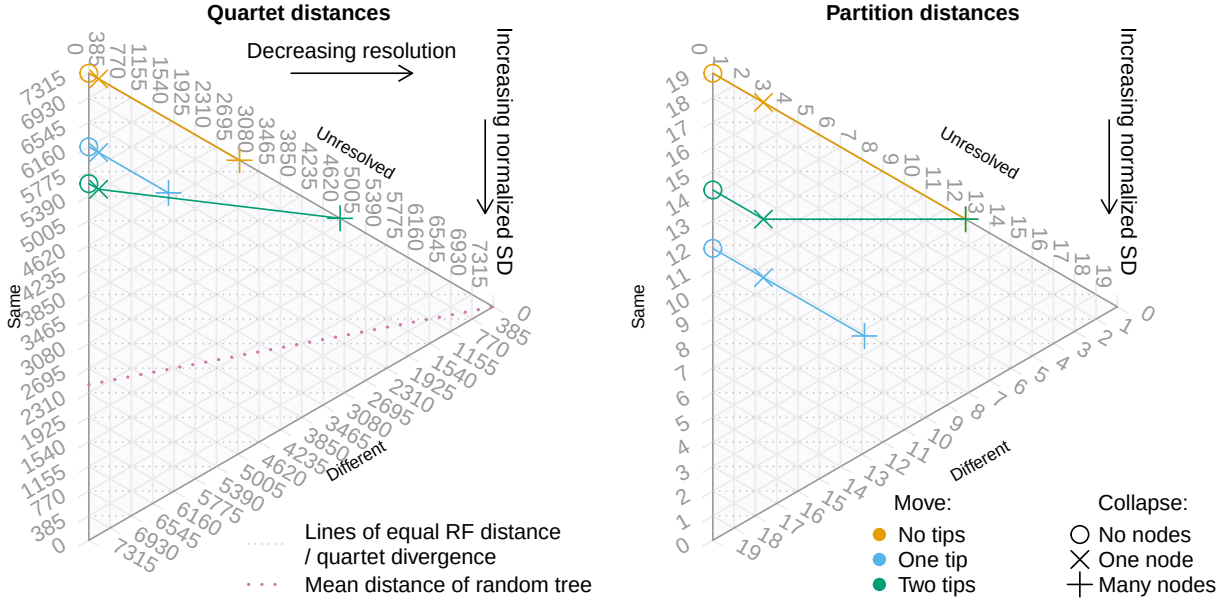
We could introduce a larger change to the tree topology by moving a 'cherry', i.e. two adjacent taxa (10 &

11):



3.1 Visualizing these data

The number of quartets or partitions that are unresolved, different, or identical to the reference tree can be visualized using ternary diagrams:



In these plots, the vertical direction corresponds to the normalized symmetric distance. Collapsing nodes decreases the resolution (movement in the horizontal direction), but can increase accuracy; the balance between resolution lost and accuracy gained determined whether the collapsing of nodes increases or decreases net divergence.

4 Using ternary diagrams to inform tree reconstruction techniques

This means of visualization provides a helpful way to understand how effective different methods of phylogenetic reconstruction are on particular trees.

Here I have taken a representative dataset simulated from a 22-tip reference tree [8], and analysed the dataset in TNT v1.5 [9] under equal weights parsimony and implied weights (with concavity constants of 1, 2, 3, 5 and 10), and in MrBayes v3.2.2 [10] using the Markov K model [11].

For each parsimony analysis, I recorded a strict consensus of all optimal trees, then proceeded to collapse groups with a Bootstrap GC support under -95, -90, -85...

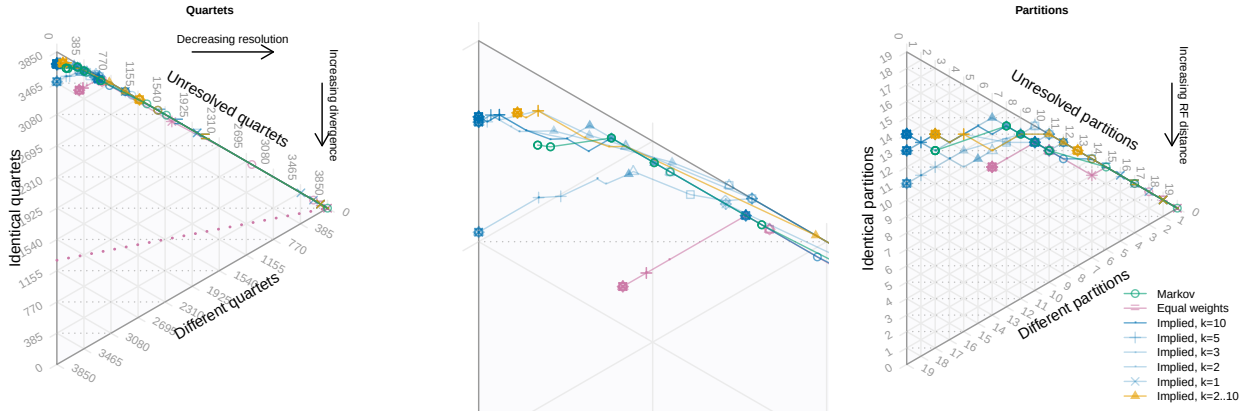
For each Bayesian analysis, I collapsed all groups whose posterior probability was under x, at 20 uniformly spaced values of x from 0.5 to 1.0.

These trees represent a progressive loss of resolution (precision) from the optimal tree, allowing an exploration of the relationship between resolution and accuracy. In each plot, resolution decreases from left to right.

Congreve and Lamsdell [8] argue that equal weighting is the optimal superior method because it resolves the fewest incorrect bipartitions – that is, its most-resolved tree is the closest to the top-right edge of the ternary diagram. By this measure, all methods are improved by collapsing nodes until none remain.

On the view advocated here, the optimal tree is the one that has the lowest normalized distance from the generative tree, which corresponds to the greatest position in the vertical direction. (The normalizing constant is the maximum possible number of partitions or quartets that could be resolved, not the number that are actually resolved in a pair of trees.)

By this measure, collapsing the least-supported nodes leads to an increase in tree quality, as predicted by Goloboff [12]: nodes with low support are likely to be incorrect. Collapsing better-supported nodes, however, reduces tree quality: nodes with high support are likely to be correct.



With this particular dataset, the optimal tree is not perfectly resolved in any method. In fact, of all 100 datasets, the best available tree was only perfectly resolved in a 18% of cases. The best tree available under a specific method was perfectly bifurcating in 0–22% of cases:

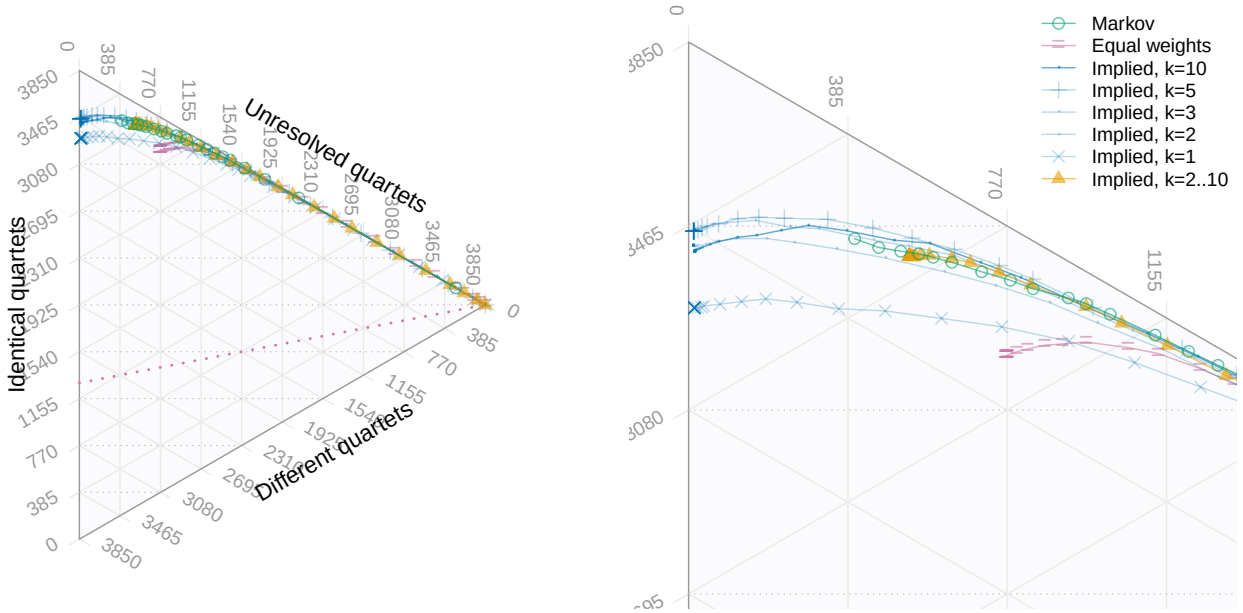
Equal weights	IW, k=1	IW, k=2	IW, k=3	IW, k=5	IW, k=10	IW consensus
2	16	22	18	16	9	2

With the 100, 350 and 1000 character datasets of O'Reilly *et al.* [13], the best available tree was almost never perfectly resolved.

	Best	Markov	Equal weights	IW, k=2	k=3	k=5	k=10	k=20	k=200	IW consensus
100	0.07	0.0	0.00	6.12	6.63	8.53	6.53	6.12	6.12	0
350	0.07	0.0	0.40	14.10	9.60	6.90	8.50	7.70	7.10	0
1000	0.11	0.2	1.61	12.92	12.31	11.20	12.61	10.19	11.60	0

4.1 Quartet metric

We can also examine the situation if we average across all 100 Congreve & Lamsdell datasets:



As the worst-supported nodes are progressively collapsed, the accuracy of implied weights trees begins to increase, decreasing the total distance of trees from the generative tree (thus meaning that the provide a better representation of the generative tree).

One question we can ask is how much we should reduce the resolution. After a certain point, the increase in accuracy gained by collapsing the least supported nodes no longer offsets the information lost by sacrificing resolution.

Bayesian and equal weights trees already produce incompletely resolved trees, and a further reduction of resolution does not improve their quality.

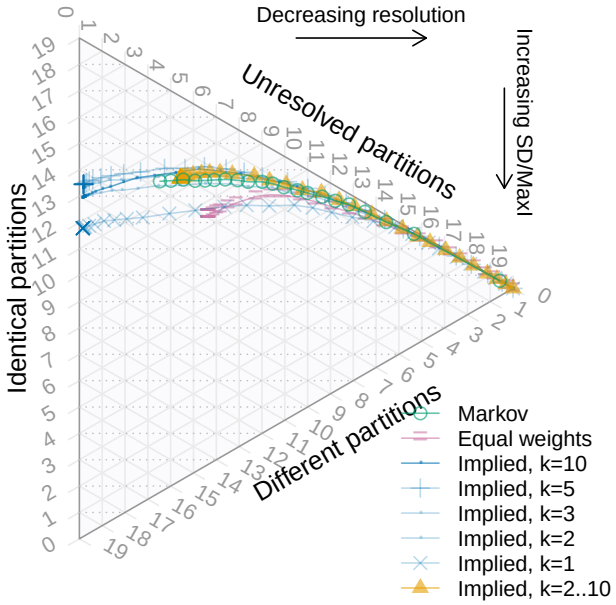
Under implied weights, averaged over these datasets, the optimum trade-off between accuracy and resolution comes when collapsing nodes with a Bootstrap GC support value below:

Equal weights	IW, k=1	IW, k=2	IW, k=3	IW, k=5	IW, k=10	IW consensus
-5	-20	-15	-15	-15	-10	-25

The only analyses to produce significantly different ($p = 0.01$) results from implied weights (at $k = 3$) or Bayesian are equal weights and implied weights with $k = 1$. Both these approaches are significantly worse.

4.2 Partition metric

We can run the same analysis counting partitions in place of quartets.



Under the partition metric, the most informative trees were found after collapsing nodes with a Bootstrap GC support value of:

Equal weights	IW, k=1	IW, k=2	IW, k=3	IW, k=5	IW, k=10	IW consensus
5	10	5	10	10	5	-5

The partition metric advocates a greater loss of resolution than the quartet metric. Otherwise, it too finds no statistically significant difference between the effectiveness of the methods, except again that equal weights, and implied weights with $k = 1$, are significantly worse.

5 Why small concavity constants are unsuitable

Of the implied weights concavity constants analysed above, $k = 1$ is strikingly (and significantly) worse than other values. It is worth recalling the mathematical underpinning for implied weights [14]:

Character penalty $= \frac{e}{e+k}$, where:

- e is the number of additional steps;
- k is the concavity constant.

The penalty can be normalized such that the first extra step in a character incurs a unit cost:

Normalized penalty $= (1 + k)\frac{e}{e+k}$.

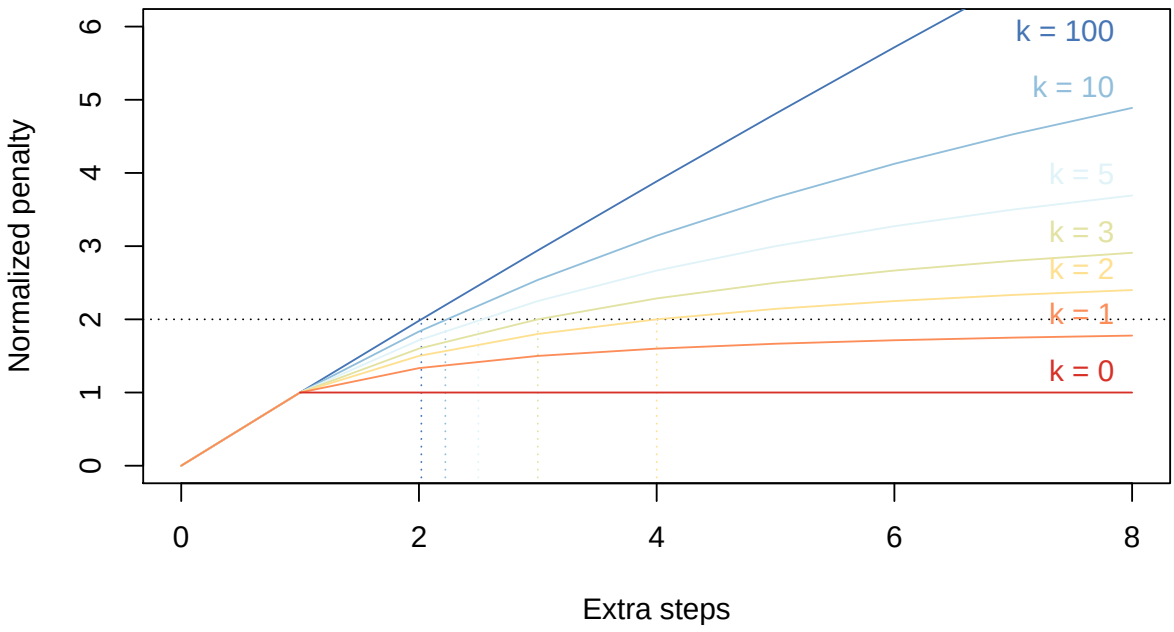
As $k \to \infty$, the penalty $\frac{e}{e+k} \to \frac{e}{k}$, and the normalized penalty $(1+k)\frac{e}{e+k} \to e$. This is to say, each subsequent step contributes the same amount to a tree's penalty; as $k \to \infty$, implied weights converges to equal weights.

At lower values of k, the penalty for extra steps decreases such that each subsequent additional step is penalized less than the previous one. As $k \to 0$, the penalty for the first step converges to one, and the penalty for subsequent steps converges to zero. At this extreme, the optimal tree is the one that maximises the number of characters that are convex. (A convex character is one that can be plotted onto the tree with no additional steps; its derived states each exhibit a unique origin.) All characters that are not convex are uninformative, as the number of steps beyond the first is irrelevant to their total contribution to tree score. This situation corresponds to clique analysis [15], a method that is no longer advocated for use in phylogenetic reconstruction.

The value $k = 1$ marks a significant point in the transition from parsimony analysis to clique analysis, because the highest cost that can be associated with a single character is less than twice the cost of a single extra step.

The cost associated with the first extra step is $\frac{1}{1+1} = \frac{1}{2}$. As the number of extra steps increases($e \to \infty$), the penalty increases towards its maximum value of $\frac{\infty}{\infty+1} \to 1$, i.e. just under twice the cost of the first step.

As such, given a pair of characters, a reconstruction that assigns infinitely many changes of one character, but no additional steps to the other, will be preferred to a reconstruction in which both characters undertake a single additional step.



Two characters with one extra step receive a total normalized penalty of two (dotted line). At progressively smaller values of k, a single character must exhibit increasingly more steps before it receives the same penalty. Once $k \leq 1$, no amount of steps in a single character will elicit a penalty equal to that which would be encountered if a second character undertook a single extra step.

Because non-convex characters are not entirely uninformative, this situation is not strictly equivalent to clique analysis. For example, trees that reconstruct fewer steps in a single non-convex character will still be preferred to those that reconstruct more steps in the same character; and a tree that imposes two extra steps on three characters receives the same penalty ($3 \times \frac{2}{2+1} = 2$) as one that imposes one extra step on four ($4 \times \frac{1}{1+1} = 2$).

Nevertheless, a value of $k = 1$ places significantly more emphasis on maximising the number of convex characters than on minimizing the total number of steps in any given (non-convex) character, behaviour that is more characteristic of clique analysis than parsimony analysis. This supports the long-standing recommendation [12] that low values of k should be avoided.

As there does not yet exist an objective method for selecting a single value of k for parsimony analysis, it has been proposed that nodes are recovered by a range of concavity constants are likely to be correct [16]. Studies that take a consensus of all trees found to be optimal under a range of concavity values (e.g. [17]) should consider discounting topologies that are only recovered under low values of k (*cf.* [18,19]).

References

1. Smith MR. in production Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. *Biology Letters; preprint at BioRxiv* (doi:10.1101/227942)

2. Penny D, Hendy MD. 1985 The use of tree comparison metrics. *Systematic Zoology* **34**, 75–82. (doi:10.2307/2413347)

3. Steel MA, Penny D. 1993 Distributions of tree comparison metrics—some new results. *Systematic Biology* **42**, 126–141. (doi:10.1093/sysbio/42.2.126)

4. Robinson DF, Foulds LR. 1981 Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)

5. Estabrook GF, McMorris FR, Meacham CA. 1985 Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology* **34**, 193–200. (doi:10.2307/2413326)

6. Day WH. 1986 Analysis of quartet dissimilarity measures between undirected phylogenetic trees. *Systematic Biology* **35**, 325–333. (doi:10.1093/sysbio/35.3.325)

7. Smith MR. 2018 TreeSearch: Phylogenetic Tree Search Using Custom Optimality Criteria. (doi:10.5281/zenodo.1042590)

8. Congreve CR, Lamsdell JC. 2016 Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. *Palaeontology* **59**, 447–465. (doi:10.1111/pala.12236)

9. Goloboff PA, Catalano SA. 2016 TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics* **32**, 221–238. (doi:10.1111/cla.12160)

10. Huelsenbeck JP, Ronquist F. 2001 MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755. (doi:10.1093/bioinformatics/17.8.754)

11. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50**, 913–925. (doi:10.1080/106351501753462876)

12. Goloboff PA. 1995 Parsimony and weighting: a reply to Turner and Zandee. *Cladistics* **11**, 91–104. (doi:10.1111/j.1096-0031.1995.tb00006.x)

13. O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. 2016 Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biology Letters* **12**, 20160081. (doi:10.1098/rsbl.2016.0081)

14. Goloboff PA. 1993 Estimating character weights during tree search. *Cladistics* **9**, 83–91. (doi:10.1111/j.1096-0031.1993.tb00209.x)

15. Wilkinson M. 1994 Three-taxon statements: when is a parsimony analysis also a clique analysis? *Cladistics* **10**, 221–223. (doi:10.1111/j.1096-0031.1994.tb00174.x)

16. Goloboff PA, Carpenter JM, Arias JS, Esquivel DRM. 2008 Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics* **24**, 758–773. (doi:10.1111/j.1096-0031.2008.00209.x)

17. Mirande JM. 2009 Weighted parsimony phylogeny of the family Characidae (Teleostei: Characiformes). *Cladistics* **25**, 574–613. (doi:10.1111/j.1096-0031.2009.00262.x)

18. Smith MR, Caron J-B. 2015 *Hallucigenia*'s head and the pharyngeal armature of early ecdysozoans. *Nature* **523**, 75–78. (doi:10.1038/nature14573)

19. Zhang X-G, Smith MR, Yang J, Hou J-B. 2016 Onychophoran-like musculature in a phosphatized Cambrian lobopodian. *Biology Letters* **12**, 20160492. (doi:10.1098/rsbl.2016.0492)