Biases of tree-independent-character-subsampling methods

Mark P. Simmons a,⇑, John Gatesy b

a Department of Biology, Colorado State University, Fort Collins, CO 80523, USA
b Department of Biology, University of California, Riverside, CA 92521, USA

A R T I C L E   I N F O
Article history:
Received 9 December 2015
Revised 16 March 2016
Accepted 15 April 2016
Available online 19 April 2016

Keywords:
Character subsampling
Gnetales
Observed Variability
Phylogenetic bias
Substitution rates
Tree Independent Generation of Evolutionary Rates

A B S T R A C T
Observed Variability (OV) and Tree Independent Generation of Evolutionary Rates (TIGER) are quick and easy-to-apply tree-independent methods that have been proposed to provide unbiased estimates of each character’s rate of evolution and serve as the basis for excluding rapidly evolving characters. Both methods have been applied to multiple phylogenomic datasets, and in many cases the authors considered their trees inferred from the OV- and TIGER-delimited sub-matrices to be better estimates of the phylogeny than their trees based on all characters. In this study we use four sets of simulations and an empirical phylogenomic example to demonstrate that both methods share a systematic bias against characters with more symmetric distributions of character states, against characters with greater observed character-state space, and against large clades in the context of character conflict. As a result these methods can favor convergences and reversals over synapomorphy, exacerbate long-branch attraction, and produce mutually exclusive phylogenetic inferences that are dependent upon differential taxon sampling. We assert that neither OV nor TIGER should be relied upon to increase the ratio of phylogenetic to non-phylogenetic signal in a data matrix. We also assert that skepticism is warranted for empirical phylogenetic results that are based on OV- and/or TIGER-based character deletion wherein a small clade is supported after deletion of characters, yet is contradicted by a larger clade when the entire data matrix was analyzed.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

In phylogenomic datasets that consist of tens or hundreds of thousands of characters, taxon sampling and character quality are generally greater concerns than character number. Character quality is important because of the widespread recognition that biases in these datasets may lead to incorrect, yet highly supported, molecular-evolution and phylogenetic inferences (e.g., Wong et al., 2008; Philippe et al., 2011; Gatesy and Springer, 2014; Simmons and Gatesy, 2015). Hence automated methods that are thought to reduce these biases by excluding characters are widely applied. These methods include excluding ambiguously aligned regions (e.g., Castresana, 2000; Capella-Gutiérrez et al., 2009; but see Tan et al., 2015), excluding third-codon positions (e.g., Chiari et al., 2012; Wickett et al., 2014; but see Källersjö et al., 1999), and analyzing amino-acid rather than nucleotide characters (e.g., Katz and Grant, 2015; Zapata et al., 2015; but see Simmons and Freudenstein, 2002).

Rather than excluding all silent substitutions or all third-codon positions a priori, an attractive possibility is to estimate the rate for each individual character along a continuum (rather than assigning it to a pre-defined rate category) in a tree-independent manner such that rate estimates are not dependent upon, and hence not potentially biased by, any particular topology. Two alternative methods have been introduced by authors who have claimed to do just this. Goremykin et al. (2010) described Observed Variability (OV) and Cummins and McInerney (2011) presented Tree Independent Generation of Evolutionary Rates (TIGER).

1.1. Observed Variability

Goremykin et al. (2010:329) introduced OV as a method that “…can arguably improve the net result of phylogeny reconstruction, especially for deeper nodes, where the inference of phylogeny is especially obscured by multiple substitutions and the resulting long-branch attraction.” In addition to OV they introduced a gamma-rate (Yang, 1993) sorter. They reported that OV outperformed the gamma-rate sorter, and most empirical phylogenetics papers that have cited Goremykin et al. (2010) only applied OV. Hence we focus on their OV sorter here.
OV is calculated for each character individually without reference to any other characters, so it does not take into account character congruence. The OV-score for a given character is simply the number of pairwise character-state matches (scored as 0’s) relative to the number of character-state mismatches (scored as 1’s) among all terminals for which the character is scored (i.e., excluding terminals with gaps or missing data). Hence, all else equal, lower OV scores are given to characters with fewer observed character states as well as those with more asymmetric distributions of character states. For example, in a matrix with 100 terminals a character with two adenines and 98 thymines has an OV score of just 0.04 because the vast majority of pairwise comparisons are matches of thymines with other thymines. On the other hand a character with 50 cytosines and 50 guanines has an OV score of 0.5 because half of the pairwise comparisons are mismatches between cytosines and guanines.

OV has been broadly applied to delete putatively fast evolving characters from genomic-scale datasets, including those based on mitochondrial (e.g., Lavrov et al., 2013; Liu et al., 2014; Meiklejohn et al., 2014), plastid (e.g., Zhong et al., 2011; Goremykin et al., 2013; Sun et al., 2015), and nuclear (Xi et al., 2013, 2014) genomes. Typically characters are deleted in blocks of 250–1000 positions at a time, and a range of deletions are explored, from <5% up to 50% of the parsimony-informative characters (Xi et al., 2013, 2014) or even 50% of all characters (Sun et al., 2015). The most extreme case that we know of is Xi et al.’s (2014) deletion of all but 5000 of their 142,590 parsimony-informative characters in one of their analyses.

For many authors, the preferred phylogenetic hypothesis is that based on an OV-selected sub-matrix rather than that based on all of the data (e.g., Zhong et al., 2011; Goremykin et al., 2013; Xi et al., 2014). But other authors have expressed concern about OV-based results. Drew et al. (2014:379) noted that OV “...essentially collapses branches by reducing the number of characters,” and that this can even apply to “...virtually universally accepted clades...” Likewise, Meiklejohn et al. (2014:321) concluded that OV-based character deletion “...eliminated valuable evolutionary signal, resulting in reduced resolution of relationships at a range of taxonomic levels.”

More recently, Simmons and Gatesy (2015) demonstrated that Xi et al.’s (2014) application of OV to eliminate half of the parsimony-informative characters from their dataset of 310 nuclear genes resulted in worse performance (as measured by congruence between gene trees with well-established reference clades, the overall success of resolution (number of clades correctly resolved minus number of clades incorrectly resolved), and the averaged overall success of resolution, which incorporates resampling support (Simmons and Webb, 2006)) relative to the complete datasets. Furthermore, the OV-slow parsimony-informative characters had a much lower amount of possible synapomorphy (maximum number of steps possible – minimum number of steps possible in a parsimony context; Farris, 1989) than the OV-fast parsimony-informative characters. The concatenation-based analyses of OV-slow characters actually had lower ensemble retention indices (Farris, 1989) on their most parsimonious trees than did the concatenation-based analyses of OV-fast characters on their most parsimonious trees.

Simmons and Gatesy (2015) also demonstrated that Xi et al.’s (2014) application of OV to analyze the 5000 most conserved characters resulted in a disparate likelihood tree topology (gymnosperms paraphyletic, monocots polyphyletic, eudicots paraphyletic, Amborella and Nuphar nested within the angiosperms) with all clades consisting of more than five terminals (in an unrooted context) being effectively unsupported. The cause of the effectively unsupported large clades was that none of the 5000 parsimony-informative characters, all of which were binary, included the minority character state in more than two terminals. Finally, Simmons et al. (2016) demonstrated that when Xi et al.’s (2014) 310 gene trees were inferred using the OV-slow characters, they had substantially greater average pairwise topological incongruence than when they were inferred from all characters.

1.2. Tree Independent Generation of Evolutionary Rates

Cummins and McInerney (2011:833) introduced TIGER as a method that “...estimates the relative evolutionary rate of each homologous character,” with “...the similarity between characters as a proxy for evolutionary rate.” Hence, in contrast to OV, TIGER is based on character congruence and more congruent characters are expected to be slower evolving than less congruent characters. This is a natural application of the idea that character congruence is a measure of phylogenetic signal (Archie, 1989; Faith and Cranston, 1991). The partition-agreement scores produced by TIGER range from zero to one, with higher scores indicating higher congruence.

Partition-agreement scores are calculated such that characters with state distributions that are inclusive of other characters’ state distributions receive higher scores. Hence, for a character with two adenines and 98 thymines in a 100-terminal matrix, the thymines (i.e., the character state present in the majority of the terminals) are more likely to be inclusive of other characters’ state distributions than would another character with 50 cytosines and 50 guanines. So OV and TIGER would prefer the same character in this example.

Like OV, TIGER has been applied to selectively delete putatively fast-evolving characters from genomic datasets (e.g., Morgan et al., 2014; Xi et al., 2014; Katz and Grant, 2015). But unlike OV, TIGER has also been used to partition characters into rate categories for parametric analyses (e.g., Rota and Wahlberg, 2012; Heikkilä et al., 2014; Nakov et al., 2014). Typically character removal is performed by excluding one or more of the ten rate categories that TIGER delimited, with different numbers of characters assigned to each category (e.g., Greene et al., 2014; Morgan et al., 2014). The percentage of characters removed ranged from just 0.6% of all characters (Owen et al., 2015) to 76% of the variable characters (Feuda and Smith, 2015).

Most authors who applied TIGER to partition rate categories or remove characters favored their TIGER-based results. But others have expressed concern. Morgan et al. (2014:642) noted that removal of even one of their 20 TIGER-delimited rate categories “...resulted in an increase in phylogenetic conflict for the remaining 10 [of 13] mtGenes, suggesting that removal of site category 20 could be removing necessary phylogenetic signal.” Katz and Grant (2015:411) noted that TIGER-based character deletion “...yields consistent topology with lower support for most clades.” Sharma et al. (2015:3) “...observed major loss of phylogenetic signal upon removing sites ranked in one or more of the fastest evolving bins (of 10 equally sized bins), yielding a basal polytomy for two different matrices and the non-monophyly of scorpions.”

As with OV, Simmons and Gatesy (2015) demonstrated that Xi et al.’s (2014) application of TIGER to eliminate half of the parsimony-informative characters from their dataset of 310 nuclear genes resulted in worse performance for the same three measures cited above for OV relative to the complete datasets, albeit generally not quite as bad as OV. This pattern also applied to the amount-of-possible-synapomorphy measure. But there was no consistent pattern for the ensemble-retention index (Farris, 1989) for the concatenation-based analyses of TIGER-fast and TIGER-slow characters. Finally, as with OV, Simmons et al. (2016) demonstrated that when Xi et al.’s (2014) 310 gene trees were inferred using the TIGER-slow characters, they had substantially greater average pairwise topological incongruence than when they were inferred from all characters.
1.3. Biases of OV and TIGER

As described above, both OV and TIGER favor characters with asymmetrical distributions of character states among terminals over characters with more symmetrical distributions of character states. We believe that this preference is often unjustified, can be negatively determinate to phylogenetic inference, and hence constitutes a bias. A nearly symmetrical character-state distribution (i.e., about the same number of terminals with each of the observed character states) may be caused by numerous substitutions such that the observed state frequency for that character approaches the equilibrium state frequency among other characters (Xia et al., 2003). In this context the preferences of OV and TIGER are appropriate. But the same character-state distribution may be simply caused by that character having a single synapomorphy for a large clade (in an unrooted context). In this second context the preference of OV and TIGER is inappropriate, such that successive application of these methods to delete characters will purge synapomorphies from the matrix for large clades.

Goremykin et al. (2010:324) explicitly acknowledged this effect, but only reported it to be problematic in the context of molecular dating:

"Finally, with regard to the OV-based sorting, it should be noted that sites which support internal splits amongst large groups of species will tend to have higher OV scores than sites that support splits of few versus many species. This will have the effect of reducing internal branch lengths disproportionately compared to external branch lengths."

Characters with highly asymmetrical character-state distributions (e.g., 98/2 for a 100-terminal matrix) may, on average, have fewer steps than those with more symmetrical state distributions (e.g., 50/50), but that does not mean that they have greater phylogenetic signal. Farris (1989) clearly recognized this when he proposed the amount-of-possible-synapomorphy (APS) measure and introduced the retention index as a better measure of homoplasy than the consistency index (Kluge and Farris, 1969). The APS for a character with a state distribution of 98/2 is just 1 (i.e., maximum number of steps = 2 – minimum number of steps = 1), whereas the APS for a character with a state distribution of 50/50 is 49 (i.e., 50 – 1).

We expect that characters with more asymmetrical character-state distributions have, on average, fewer steps than characters with more symmetrical state distributions on any particular tree topology (irrespective of whether it is the true tree or a random tree). Therefore it is reasonable to expect that the former characters will have a higher ensemble consistency index. After all, the consistency index quantifies homoplasy in the context of the minimum number of steps, and if the maximum number of steps is just above the minimum number of steps then the worst possible observed consistency index is $\geq 0$. For example, the worst possible CI for a 98/2 character is 0.5 (i.e., 1/2), whereas the worst possible CI for a 50/50 character is 0.02 (i.e., 1/50). This is the same reason that the ensemble consistency index is generally higher on trees inferred from matrices with fewer terminals (Sanderson and Donoghue, 1989). Hence it should be no surprise that the ensemble consistency index would increase as TIGER is used to successively delete more characters, as documented by Cummins and McInerney (2011).

Farris (1989) addressed this limitation of the consistency index by introducing the retention index, which measures the amount of retained synapomorphy. If there are two steps for a 98/2 character then the amount of retained synapomorphy is zero. As Farris (1989:418) stated, “Similarity in this character is then irrelevant to the groupings of the tree; all apparent apomorphic similarities in the character are dismissed as nonhomologies.” On the other hand, if there are two steps for a 50/50 character then 98% of the amount of possible synapomorphy is retained. So, OV’s and TIGER’s shared bias in favor of characters with asymmetrical state distributions may be an effective way to decrease the average number of steps per character, but this has no necessary relationship to retained synapomorphy, which is the more important criterion.

A second context wherein OV’s and TIGER’s character preference is inappropriate is, all else equal, favoring binary characters over multistate characters (or, more generally, favoring characters with a lower observed character-state space over characters with a greater observed state space). All else equal, the greater the observed state space, the more pairwise mismatches identified by OV. Likewise, the greater the observed state space, the less likely that character’s state distributions are to be inclusive of other characters’ state distributions for TIGER.

Even if characters with more observed character states have evolved faster than characters with fewer observed states on average, this does not mean that they have more homoplasy or retain less phylogenetic signal (Naylor et al., 1995; Ortí and Meyer, 1996; Källersjö et al., 1999; Simmons et al., 2004a, 2004b, 2006). The simple reason is that, all else equal, constrained state space increases the probability of convergences and reversals. For example, if there are two observed steps in a character with a possible state space of two then that character must be homoplasious. On the other hand, if there are two observed steps in a character with a possible state space of four then, if every possible substitution is equally likely, there is only a 0.33 chance that the character is homoplasious.

As described above, purging characters that provide synapomorphies for large clades is problematic because those clades lose support and eventually collapse. But a third context wherein OV’s and TIGER’s character preference is inappropriate is conflict between characters that provide synapomorphies for larger clades (in an unrooted context) with characters that provide synapomorphies for contradictory smaller clades. All else equal, both methods will favor the latter group of characters, and hence are biased in favor of more symmetrical tree topologies.

Simmons and Gacesy (2015) identified this bias as a primary contributor to Xi et al.’s (2014) inference that their nucleotide data from 310 genes support the clade of (Amborella, Nuphar) as sister to the other flowering plants that they sampled rather than Amborella alone as the sister. The former phylogenetic inference entails a 2/44 split (i.e., 44 other terminals were sampled) whereas the latter phylogenetic inference entails a 5/41 split (i.e., four non-flowering plants were sampled, so these four taxa and Amborella represent one side of the split).

1.4. Foci of this study

In this study we test a series of hypotheses on how OV’s and TIGER’s bias in favor of characters with asymmetrical distributions of character states over characters with more symmetrical distributions of character states among terminals can be negatively determinate to phylogenetic inference. We do so using four sets of simulations as well as an empirical phylogenomic example. Taken together, these simulations address all three contexts described above in Section 1.3 wherein OV’s and TIGER’s character preference is inappropriate.

In addition to comparing the performance of phylogenetic inference from all simulated characters and OV- and TIGER-selected sub-matrices of putatively fast- and slow characters, we also compare the performance of the OV-slow sub-matrices with the TIGER-slow sub-matrices. Our hypothesis for all simulations as well as the empirical example is that the TIGER-slow sub-matrices would be
significantly less biased than the OV-slow sub-matrices. The basis for this hypothesis is that TIGER, unlike OV, is based on a measure of character congruence, even though the measure is biased. Note that this is not an endorsement of TIGER.

2. Materials and methods

2.1. Choice of simulation conditions

All four sets of simulations were performed on a completely asymmetrical (a.k.a., pectinate) topology, in all but one case all branches in the simulation tree are of equal length, in all but one case the Jukes and Cantor (1969) model was used to simulate all of the nucleotide substitutions, and in all but one case no rate heterogeneity among characters was simulated in a given data matrix. These simulation conditions are not meant to emulate empirical data. Rather we selected them to minimize complexity while still enabling us to test our \textit{a priori} hypotheses regarding the biases entailed in OV- and TIGER-selected subsamples of characters. Likewise, our decision to analyze relatively small subsets (500–2000 characters) of the 10,000 characters simulated for each data matrix does not emulate how most authors apply OV or TIGER to their empirical data. Rather we did this to focus on the putatively most conserved parsimony-informative characters identified by these methods, wherein we expected the greatest bias when they are analyzed in isolation.

Our decision to use completely asymmetrical simulation topologies rather than symmetrical topologies was originally based on the expectation that the former typically represent more difficult cases for accurate phylogenetic inference (e.g., Simmons et al., 2007), and hence they represent a more severe test of the phylogenetic signal contained within the OV- and TIGER-based sub-matrices. But another relevant factor to consider is that we expect OV- and TIGER-based sub-matrices of slow characters to perform particularly poorly when applied to asymmetrical lineages because of their bias in favor of characters with more asymmetrical distributions of character states. For example, if a method, on average, favors characters with a character-state distribution of 2/14 over characters with character-state distributions of 3/13–8/8 and there are only two terminal clades (as in an unrooted completely asymmetrical 16-terminal tree), then many of those favored characters probably include convergences or reversals. On the other hand, for a completely symmetrical 16-terminal tree there are eight two-terminal clades. Hence, all else equal (including branch lengths), the ratio of synapomorphy to convergence and reversal in characters with a character-state distribution of 2/14 is probably greater in the symmetrical tree.

We believe that our decision to use equal branch lengths for three of the four simulations is conservative in that there should be no necessary tendency for convergent autapomorphies to dominate synapomorphies on internal branches. Had we used longer terminal than internal branches, we expect that phylogenies inferred from the OV- and TIGER-slow matrices would be even more divergent from the simulated topologies because of these methods’ bias in favor of characters with more asymmetrical distributions of character states.

Although we did not simulate rate heterogeneity among characters, this does not mean that all characters had the same number of steps in a given simulation because the simulations applied a stochastic model. Most of our hypotheses regarding the biases of OV and TIGER are independent of rate heterogeneity among characters. In these cases direct incorporation of rate heterogeneity in our simulation model would have been a confounding effect. Our empirical example does incorporate rate heterogeneity though, given variation in selective constraints on different nucleotide sites.

2.2. Simulation one

The first set of simulations was performed to test our hypotheses that both OV and TIGER are biased against characters that could provide synapomorphies for large clades as well as those characters with more observed character states. Furthermore, we tested our hypothesis that OV and TIGER do so without necessarily identifying characters with the highest levels of homoplasy (as quantified by the ensemble retention index; RI) or identifying the characters with greatest (potential) phylogenetic signal (as quantified by the APS as well as the number of correctly resolved clades in the strict consensus of the inferred trees).

All simulations were performed using the Evolver program with the MCbase.dat file from the PAML ver. 4.1 package (Yang, 2007). The 10,000 characters were simulated using the Jukes–Cantor model with no rate heterogeneity among sites. The 32-terminal matrices include no gaps or missing data. All branches on the completely asymmetrical simulation topology were of equal length. Three alternative sets of branch lengths were used in different sets of simulations: 0.05, 0.1, and 0.2. These branch lengths were selected to ensure that many characters are homoplasious, as would be expected for empirical studies wherein OV or TIGER might be applied. Given that there are 32 terminal branches and 29 internal branches, each of the 10,000 characters, on average, is expected to have 3.05 substitutions for the 0.005-branch-length simulations, 6.1 substitutions for the 0.1-branch-length simulations, and 12.2 substitutions for the 0.2-branch-length simulations. One hundred replicate data matrices were simulated for each set of branch lengths.

PAUP* ver. 4.0b10 (Swofford, 2001) was used to exclude parsimony-uninformative characters from the simulated matrices. This step was performed to ensure that the slowest-characters and fastest-characters sub-matrices are directly comparable in the sense that they both contain the same number of parsimony-informative characters. The parsimony-informative characters were then independently analyzed by TIGER ver. 1.02 as well as the sorter.pl Perl script from the OV version 1.1 package, both of which sorted the characters from slow to fast. PAUP* was then used to create sub-matrices of the 500 and 1000 slowest characters as well as the 500 and 1000 fastest characters.

Each sub-matrix of parsimony-informative characters was analyzed with Fitch (1971) parsimony as implemented in TNT ver. 1.1 (June 2015; Goloboff et al., 2008) as well as maximum likelihood (Felsenstein, 1973) as implemented in PhyML ver. 2012.04.12 (Guindon et al., 2010). Tree searches were performed in TNT using 1000 random-addition-sequence (RAS) tree-bisection-reconnection (TBR) searches, each of which held up to 1000 equally parsimonious trees. Unsupported branches were collapsed using TBR-collapsing (Goloboff and Farris, 2001), followed by calculation of the strict consensus. Tree searches were performed in PhyML using 500 RAS subtree-pruning-regrafting (SPR) searches and the same Jukes–Cantor model that was used to simulate the characters.

We quantified the results using the five measures described below. Given that we used 100 simulation replicates and the degree of statistical power was not a limitation, we simply treated non-overlapping plus and minus 95% confidence intervals as indicating statistically significant differences, and did so for all five sets of results that we quantified.

First, we calculated the Robinson and Foulds (1981) distance (hereafter “RF distance”) of the strict consensus (for parsimony) and the single fully resolved reported tree (for likelihood) to the simulation topology by using the RF Distances Filter script (Simmons et al., 2016), which runs DendroPy (Sukumaran and Holder, 2010) as a daughter process. We scaled the observed RF distance to the maximum possible RF distance by using the scaling
function (RF distance) \((2 \times ((\text{number of shared terminals}) - 3))\), as described by Rosenberg and Kumar (2001). Our hypothesis was that there would be no significant difference in RF distance between trees inferred from the fast and slow matrices.

A potential concern regarding our use of strict consensus trees for parsimony in the context of RF distances is that collapsed branches are, in effect, treated as half correct even if all or most of the fully resolved trees contradict a given branch in the reference tree (Rannala et al., 1998). This treatment of polytomies may be regarded as an advantage to parsimony when comparing the parsimony with the likelihood results in the simulations, given that the likelihood implementation that we applied (PhyML) only ever produces fully resolved trees—even from ambiguous data (Simmons and Norton, 2013, 2014). But we do not expect this treatment of polytomies to be an advantage to the OV- or TIGER-fast characters given that, on average, far fewer internal branches were collapsed for all four sets of simulations (simulation one: 0% for fast, 21% for slow; simulation two: 0.04% for fast, 29% for slow; simulation three: 1.8% for fast, 19% for slow; simulation four: 0% for fast, 11% for slow).

Second, we recorded the number of cherries (i.e., two-terminal cladets; McKenzie and Steel, 2000) on a most parsimonious tree by using the “cherry” command in APE ver. 2.3-1 (Paradis et al., 2004). This command is only applicable to fully bifurcating trees so, rather than using the strict consensus (when applicable), the first most parsimonious tree saved was used instead. OV and TIGER determine each character’s rate in an unrooted context and consequently the topological bias of these methods should also be measured in an unrooted context. Therefore, although our trees were arbitrarily rooted using the first terminal, we did not expect Colless’ (1982)index of tree asymmetry. Colless’ (1982)index does not perform well in an unrooted context (Kupczok, 2011) and there is no equivalent index for unrooted trees (A. Kupczok, pers. comm., 2015). Although limited in power, cherry number is positively correlated with tree symmetry (Holton et al., 2014). Because our inferred parsimony and likelihood trees were arbitrarily rooted, the expected number of cherries on the pectinate simulation tree is one.

Our hypothesis was that the trees inferred from the fast matrices would have significantly fewer cherries than would trees inferred from the slow matrices. The basis for this hypothesis is our expectation that OV and TIGER would preferentially purge characters that include synapomorphies along internal branches (particularly those toward the center of the fully asymmetrical tree) while retaining characters that include convergent apomorphies between terminal branches in the slow matrices. Hence convergences among terminal branches can overcome synapomorphies for internal branches, and the inferred trees are more symmetrical than the simulated fully asymmetrical topology.

Third, we recorded the amount of possible synapomorphy (APS) as a measure of the asymmetry of character-state distributions. For example, the APS for a binary character with a character-state distribution of 30/2 would be one, whereas the APS for a binary character with a character-state distribution of 16/16 would be 15. The values that we present below are averages over all characters in each data matrix. Our hypothesis, for the reasons described above in Section 1.3, was that the fast matrices would have significantly higher APS than the slow matrices.

Fourth, we recorded the RI to quantify homoplasy in the most parsimonious tree(s). Our hypothesis, for the reasons described above in Section 1.3, was that there would be no significant difference in RI between the fast and slow matrices.

Fifth, we recorded the average observed character-state space (i.e., the number of observed character states per character; Naylor et al., 1995; Simmons et al., 2004a, 2004b). Our hypothesis, for the reasons described above in Section 1.3, was that the fast matrices would, on average, have higher observed character-state space than the slow matrices.

2.3. Simulation two

The second set of simulations was performed to test our hypothesis that, in cases of character conflict, OV and TIGER are biased in favor of those characters that support the smaller clade and hence the smaller-clade topology. Eighteen-terminal trees were simulated with all branches of equal length (0.1). Given that there are 18 terminal branches and 15 internal branches, the expected average number of substitutions per character is 3.3. The Jukes–Cantor model without rate heterogeneity was used to simulate 5000 characters onto each of two tree topologies. The two tree topologies differed for just a single terminal (“WILD” for wildcard; Nixon and Wheeler, 1991; Fig. 1), which was alternately placed on adjacent branches such that the two tree topologies differ by a single nearest-neighbor-interchange branch swap.

Each pair of trees included one completely asymmetrical topology and one topology with an additional two-terminal clade (Fig. 1). In the 1 vs. 2 simulations the WILD terminal was alternately sister (albeit in an unrooted context) to one or two terminals. In the 1 vs. 4 simulations the WILD terminal was alternately sister to one or four terminals. In the 1 vs. 8 simulations the WILD terminal was alternately sister to one or eight terminals. 2matrix.pl ver. 1.0 (Salinas and Little, 2014) was used to concatenate the 5000 characters simulated onto each of the trees, for a grand total of 10,000 characters in each of the 100 replicate matrices. As with the first set of simulations, PAUP* was used to exclude parsimony-uninformative characters, and the remaining parsimony-informative characters were then independently sorted by OV and TIGER. PAUP* was then used to create sub-matrices of the 500, 1000, and 2000 slowest characters as well as the 500, 1000, and 2000 fastest characters.

We quantified the results by recording where in the inferred parsimony and likelihood trees the WILD terminal was resolved: either as sister to its single sister terminal in the tree that was used to simulate the first block 5000 characters, or as part of the clade of two, four, or eight terminals (for the 1 vs. 2, 1 vs. 4, and 1 vs. 8 simulations, respectively) that it was sister to in the tree that was used to simulate the second block of 5000 characters (Fig. 1). The former inferred topologies were assigned a score of zero and the latter inferred topologies were assigned a score of one. All other topologies that did not conform to either of these alternatives (e.g., because of a polytomy in the parsimony strict consensus) were scored as 0.5 to indicate neutrality.

Our first hypothesis was that the optimal trees inferred from the slow matrices would resolve the WILD terminal as sister to the single terminal that it was sister to in the tree that was used to simulate the first block 5000 characters, or as part of the clade of two, four, or eight terminals (for the 1 vs. 2, 1 vs. 4, and 1 vs. 8 simulations, respectively) that it was sister to in the tree that was used to simulate the second block of 5000 characters (Fig. 1). The former inferred topologies were assigned a score of zero and the latter inferred topologies were assigned a score of one. All other topologies that did not conform to either of these alternatives (e.g., because of a polytomy in the parsimony strict consensus) were scored as 0.5 to indicate neutrality.

Our second hypothesis was that the bias of OV and TIGER would be more pronounced as the discrepancy in number of terminals that WILD was simulated as sister to increased in the second block of 5000 characters, such that 1 vs. 2 < 1 vs. 4 < 1 vs. 8. That is, the greater the discrepancy in state distributions among characters that provide synapomorphies for the alternative clades, the greater
the chance that characters that provide synapomorphies for the larger clade would be excluded from the slow sub-matrices and be assigned to the fast sub-matrices.

2.4. Simulation three

The third set of simulations was performed to test our hypothesis that OV and TIGER can exacerbate borderline cases of long-branch attraction (Felsenstein, 1978) among distantly related terminals or clades by purging synapomorphies from the intervening internal branches that separate the two lineages with long branches. Ten-terminal fully asymmetrical trees were simulated with all seven internal branches and eight of the ten terminal branches of equal length (alternately 0.05, 0.1, or 0.2). By contrast, branch lengths of 1.0 were set for the terminal branches leading to taxa two and ten, which are separated by all seven internal branches (Fig. 2). The expected average number of substitutions per character is 2.75 when the 0.05 branch lengths were applied to 15 of the 17 branches, 3.5 when the 0.1 branch lengths were applied to 15 of the 17 branches, and 5 when the 0.2 branch lengths were applied to 15 of the 17 branches. All 10,000 characters were simulated using the Jukes–Cantor model without rate heterogeneity in each of the 100 replicate matrices.

As with the first and second sets of simulations, PAUP* was used to exclude parsimony-uninformative characters, and the remaining parsimony-informative characters were then independently sorted by OV and TIGER. PAUP* was then used to create sub-matrices of the 500 and 1000 slowest characters as well as the 500 and 1000 fastest characters.

We quantified the results by recording the scaled RF distance between the inferred parsimony and likelihood trees and the simulated tree. Given that the long-branch terminals were at opposite ends of the fully asymmetrical simulated tree, their resolution as sister groups, which contradicts all seven internal branches, would result in a scaled RF distance of 1.0.

Our hypothesis regarding OV and TIGER exacerbating long-branch attraction in the third set of simulations was based on our expectation that OV and TIGER would preferentially purge characters that provide synapomorphies for five of the seven internal branches while retaining characters that have autapomorphies, including convergent autapomorphies, on the two long terminal branches leading to taxa two and ten. That is, synapomorphies along five of the seven internal branches would result in less asymmetric character-state distributions (3/7, 4/6, and 5/5) than would convergences between taxa two and ten (8/2). Furthermore, we hypothesized that long-branch attraction in the slow matrices would apply not only to parsimony, but to likelihood as well—despite our use of the exact same model for simulating the data as well as likelihood inference. Our basis for this second hypothesis is that biased OV and TIGER character sub-sampling creates slow matrices that are not representative of the simulated data with respect to distributions of apomorphies.

2.5. Simulation four

The fourth set of simulations was performed to test our hypothesis that OV and TIGER are biased in favor of characters that have a low observed character-state space—even when they evolved more rapidly than characters that have a higher observed character-state space. Two sets of 5000 characters each were simulated onto the same completely asymmetrical 32-terminal tree topology. The first set of 5000 characters each were simulated using the Jukes–Cantor model (i.e., four-state characters) with no rate variation among characters. All branches were of equal length (0.2), so the expected average number of substitutions for each
character is 12.2. This high rate of evolution was selected to ensure that the observed character-state space for these characters would approach the actual character-state space. The second set of 5000 characters was simulated using the Felsenstein (1981) model with nucleotide frequencies of 0.5 adenine and 0.5 thymine (i.e., two-state characters). All branches were of an extreme and equal length of 20, so the expected number of substitutions for each character on the tree is 3220, which we expected to effectively randomize the two-state characters and consequently eliminate phylogenetic signal.

2matrix.pl was used to concatenate the 5000 multistate characters together with the 5000 binary characters for each of the 100 replicate matrices. PAUP* was used to exclude parsimony-uninformative characters, and the remaining parsimony-informative characters were then independently sorted by OV and TIGER. PAUP* was then used to create sub-matrices of the 500, 1000, and 2000 slowest characters as well as the 500, 1000, and 2000 fastest characters.

In contrast to the three other sets of simulations, the Felsenstein (1981) model, with nucleotide frequencies estimated from the data, was used for likelihood inference. The model incorporated gamma rate heterogeneity with a likelihood estimate of the alpha parameter value and two rate classes. This model was chosen to approximate the models used to simulate the two blocks of 5000 characters without applying character partitions.

We quantified the results using the five measures described below. The basis for our hypotheses regarding all five measures is that we expected the slow matrices to consist primarily of the randomized binary characters whereas we expected the fast matrices to consist primarily of the phylogenetically informative multistate characters.

First, we recorded the scaled RF distances between the parsimony and likelihood inferred trees and the simulated tree. Our hypothesis was that the trees inferred from the fast matrices would be significantly more congruent with the simulated tree than the trees inferred from the slow matrices. Despite the incorporation of ~5000 effectively random characters, we hypothesized that trees inferred from all ~10,000 characters would be very similar to the simulated topology because the other half of the characters provide well-structured phylogenetic signal with numerous synapomorphies for each internal branch (Wenzel and Siddall, 1999).

Second, we recorded the APS. In the context of simulation four, unlike simulation one, we expected the APS to be high for the slow matrices because both character states in the effectively randomized binary characters should, on average, be present in about equal frequency. But we expected the APS to be higher still for the fast matrices because of their greater observed character-state space.

Third, we recorded the RI. Our hypothesis was that the slow matrices would have a significantly lower RI on their inferred trees than the fast matrices would on their inferred trees. Fourth, we recorded the average observed number of character states, and our hypothesis was that the slow matrices would have a significantly lower observed state space than the fast matrices. Fifth, we recorded the average score that TIGER assigned to all 5000 four-state characters as well as all 5000 two-state characters. Our hypothesis was that TIGER would, on average, assign the 5000 effectively randomized binary characters significantly higher partition-agreement scores than it would the 5000 phylogenetically informative multistate characters.

2.6. Empirical example

We used Zhong et al.’s (2011) data matrix of 33,636 nucleotide characters from exons that were sampled from 16 seed-plant plastid genomes as our empirical example. The matrix was downloaded from http://gbe.oxfordjournals.org/content/3/1340/suppl/DC1 and, following Zhong et al. (2011), we excluded all gapped positions (using PAUP*), which resulted in a matrix of 33,290 characters, though Zhong et al. (2011) reported the matrix as containing 33,289 characters. Of the 33,290 characters, 10,587 are parsimony-informative. Zhong et al. (2011) alternatively analyzed nucleotide characters from all three codon positions, first and second positions only, and third positions only. Our re-analyses are restricted to the all-three-codon-positions set of 33,290 characters.

Zhong et al. (2011) sorted their 33,289 characters using OV and analyzed the complete matrix as well as sub-matrices that successively excluded 250 characters at each cut. They presented their phylogenetic results regarding three alternative theories for the sister group of the Gnetales, after excluding 0–4750 OV-fast characters in their Fig. 2A. The three competing theories that they presented in this figure are as follows: Gnceup = (((Gnetales)(Cupressophyta)), Gnepine = (((Gnetales)(Pinaceae)), and Antho-

phyte = (((Gnetales)(angiosperms)). The Gncup and Gnepine theo-

ries are shown in Fig. 3A. A fourth competing theory that they presented in their Fig. 1 is Gnetifer = (((Gnetales)(Cupressophyta)

(Pinaceae)).

We extended Zhong et al.’s (2011) character-subsampling analyses by applying both OV and TIGER, independently of each other, with blocks of 500 fast characters sequentially excluded, and up to 10,000 characters excluded. OV and TIGER were run on the complete matrix of 33,290 characters rather than on the parsimony-informative characters only. In addition to comparing the OV and TIGER results with each other, we also compared their phylogenetic-inference results for six well supported reference clades to those generated from 10 sets of matrices wherein parsimony-informative characters were randomly excluded. For this we separated the 10,587 parsimony-informative characters into a sub-matrix and then used the random-integer-set generator from https://www.random.org/integer-sets/ to generate a given number (500–10,000 in intervals of 500) of non-repeating integers between 1 and 10,587. The selected parsimony-informative characters were then excluded using PAUP* and 2matrix.pl was used to concatenate the remaining parsimony-informative characters with all 22,703 parsimony-uninformative characters. The six reference clades used are: (angiosperms), (Cycas, Ginkgo), (Cupresso-

phyta, Gnetales, Pinaceae), (Cupressophyta), (Pinaceae), and (Gnetales). All six of these reference clades have been highly supported by both plastome- and nuclear-gene-based phylogenomic datasets (e.g., Xi et al., 2013; Ruhfel et al., 2014; Wickett et al., 2014).

Following Zhong et al. (2011), we analyzed the matrices using PhyML with the GTR + 4-rate gamma substitution model without specifying any partitions among the included characters. Tree searches were performed in PhyML using 500 RAS SPR searches for the most likely tree, as well as 1000 bootstrap (Felsenstein, 1985) pseudoreplicates that each consisted of a single SPR search. 2.7. Taxon duplication for hypothesis testing

The three primary competing theories regarding the sister group of the Gnetales that have been supported by phylogenomics analyses are Gnceup, Gnepine, and Gnetifer. Ruhfel et al. (2014) reported 100% bootstrap support for the Gnceup theory in their plastome-based analyses and found that this result was robust to two different character-sampling strategies (all three codon positions, first and second positions only) as well as three different character-coding strategies (DNA, AA, purine-pyrimidine coding). Alternatively, Wickett et al. (2014) reported conflicting results from their analyses of up to 852 nuclear genes wherein the Gnepine theory was generally strongly supported in their
Fig. 3. Empirical example from Zhong et al. (2011) wherein different theories for the sister-group of the Gnetales are tested based on plastome data. (A) Most likely phylogenetic tree inferred from all 33,290 nucleotide characters. (B) The same tree from A (compressed horizontally) showing taxon replication of all terminals except the Gnetales and Cupressophyta to test our hypothesis that OV and TIGER would favor the Gnecup theory (a 7-taxon clade) over the Gnepine theory (a 12-taxon clade). (C) The same tree from A showing taxon replication of all terminals except the Gnetales and Pinaceae to test our hypothesis that OV and TIGER would favor the Gnepine theory (a 6-taxon clade) over the Gnecup theory (a 15-taxon clade). Note that taxon replicates were only used for OV and TIGER character sorting. All tree searches were performed on matrices of the original 16 taxa.
supermatrix analyses, yet the Gnetifer theory was generally strongly supported in their coalescent analyses.

We do not know which, if any, of these three theories is correct and we do not have a strong preference for any of them. Our re-analyses of Zhong et al.'s (2011) plastome dataset is not meant to improve upon Ruhfel et al.'s (2014) more thorough (with respect to both taxon and character sampling) plastome-based analyses. Rather, our analyses were performed to provide an empirical example to test the following two hypotheses. First, we hypothesized that OV- and TIGER-selected slow characters do not consistently provide stronger phylogenetic signal than the equivalent number of randomly selected parsimony-informative characters.

Second, we hypothesized that OV- and TIGER-generated slow branches would alternatively support the Gneup or the Gneup process theory based on the character conflict in these data (as demonstrated by Zhong et al. (2011); Fig. 2A) together with changes in the number of Cupressophyta and Pinaceae terminals sampled. That is, if the ratio of the Cupressophyta to Pinaceae terminals sampled is high, then progressive deletion of OV- and TIGER-identified fast characters would favor the Gneup theory because characters that provide synapomorphies for the larger Gneup clade would have been preferentially eliminated. Vice versa, if the ratio of Pinaceae to Cupressophyta terminals sampled is high, then progressive deletion of OV- and TIGER-identified fast characters would favor the Gnep theory because characters that provide synapomorphies for the larger Gnep clade would have been preferentially eliminated.

To test our second hypothesis, we alternatively triplicated all 16 original terminals except the three Gnetales and four Cupressophyta terminals (i.e., 34 total terminals were included in the manipulated matrix) or triplicated all 16 original terminals except the three Gnetales and three Pinaceae terminals (i.e., 36 total terminals were included in the manipulated matrix). In the former case a binary character for which the only variation is a synapomorphy for (Gnetales, Cupressophyta) would have a character-state distribution of 7:27, whereas a synapomorphy for (Gnetales, Pinaceae) would have a character-state distribution of 12:22 (Fig. 3B). In the latter case a binary character for which the only variation is a synapomorphy for (Gnetales, Cupressophyta) would have a character-state distribution of 15:21, whereas a synapomorphy for (Gnetales, Pinaceae) would have a character-state distribution of 12:22 (Fig. 3C). So our hypothesis would be corroborated if successive deletion of fast characters identified by OV and TIGER results in the clade (Gnetales, Cupressophyta) being inferred in the former case, but the clade (Gnetales, Pinaceae) being inferred in the latter case.

The terminal-duplication approach described above was only used for the OV and TIGER character-ordering steps. The terminal replicates were removed prior to tree searches, such that all tree searches were performed on 16-terminal matrices. Hence our terminal-duplication approach should only have an affect on how OV and TIGER ordered the 33,290 characters.

In order to manipulate the size of bipartitions (clades), we chose to replicate rather than removing terminals because of two confounding factors that could affect phylogenetic inference after removing terminals. First, subsampling the Cupressophyta, Pinaceae, and/or Gnetales clades could result in long-branch attraction where there was none (or less) previously. This long-branch attraction can occur when terminals that formerly broke up long branches are excluded (Graybeal, 1998). Second, arbitrary selection of which terminals to remove and which terminals to keep may be determinate to the results based on differential distributions of missing/inapplicable data among terminals as well as variation in evolutionary rates on different branches.

Our goal for replicating terminals is to demonstrate the bias of OV and TIGER in favoring characters that provide synapomorphies for smaller clades over characters that provide synapomorphies for larger clades. But we acknowledge that replicating terminals, rather than adding novel terminals to the lineage(s) in question, can result in unintended conservatively and liberally biased estimates of the rate of character evolution when analyzing submatrices of OV- and TIGER-identified slow characters. In both cases the bias would occur if variation in the replicated terminals for a given character is not representative of variation that would be observed for that character if novel terminals were added instead.

We present three alternative scenarios showing how replicating, rather than adding novel terminals, can affect OV-based rate estimates for a character in Fig. 4. In Fig. 4A we present the character states for a given character among the four originally sampled terminals within a given lineage that is to be duplicated. If the character-state distribution of the four originally sampled terminals is representative of four novel terminals that could have been added (Fig. 4B), then we do not expect any bias to apply to OV's rate assignment for this character. Alternatively, if the character-state distribution of the four originally sampled terminals results in more sum-of-pairs character-state differences than would have been obtained if four novel terminals had been added instead (Fig. 4C), then OV would overestimate the character's rate. The vice versa result can also occur (Fig. 4D). Note that all three of these scenarios are simplifications because character-state distributions in other lineages are not considered. We expect the ratio between conservative and liberal biases to be dependent upon the branch lengths leading to the novel terminals that could have been added as well as their phylogenetic distribution relative to the originally sampled terminals.

In contrast to OV, TIGER quantifies congruence among characters. By definition, the replicated terminals are entirely congruent for all of their characters. Hence we expect that replicating existing terminals, rather than adding novel terminals that have unsampled variation, will result in a global conservative bias wherein TIGER will assign overly high partition-agreement scores to all characters. We confirmed this hypothesis in the context of simulation one (see Section 2.2) by deleting terminals 26, 28, 30, and 32, replicating terminals 25, 27, 29, and 30, and then comparing the TIGER partition-agreement scores of the modified matrices with the

![Image](image-url)
original matrices. We did this for the first five replicates of the 0.1-branch length simulations. The average partition-agreement score across all parsimony-informative characters from the original matrix was 0.24 for the original matrices and 0.26 for the modified matrices, which corroborated our expectation.

All data matrices (simulated, empirical, all characters, and fast and slow sub-matrices) and optimal inferred trees from this study are posted as supplemental online data at: http://dx.doi.org/10.6084/m9.figshare.1617869.

3. Results

3.1. Simulation one

Results for simulation one averaged across all 100 replicate data matrices are presented in Fig. 5. The simulated topology was accurately reconstructed for all 100 replicate matrices when all 10,000 characters were analyzed using either parsimony or likelihood, irrespective of the simulated branch lengths (Fig. 5A and B). Likewise, parsimony-based analyses of the OV- and TIGER-fast sub-matrices always reconstructed the simulated topology (Fig. 5B). By contrast, the average scaled RF distance between the phylogenies inferred from the OV- and TIGER-slow matrices were significantly higher (averaging 0.22–0.58) for all cases except when the simulated branch lengths were 0.2 and 1000 characters were included in the OV-slow and TIGER-slow matrices. These results falsified our hypothesis that there would be no significant difference in RF distance between trees inferred from the fast and slow matrices.

Similar results were obtained from the likelihood trees, except the scaled RF distances for the OV- and TIGER-slow matrices were consistently higher than parsimony for all cases except when the simulated branch lengths were 0.2 and 1000 characters were included in the OV-slow and TIGER-slow matrices (Fig. 5B). Furthermore, in contrast to parsimony, phylogenetic trees inferred from the OV- and TIGER-fast matrices had significantly higher average scaled RF distances relative to the simulated topology when the simulated branch lengths were 0.2 and 500 (TIGER only) or 1000 (both OV and TIGER) characters were sampled in each sub-matrix.

Fig. 5. Results for simulation one averaged across all 100 replicate data matrices. Error bars are + and − 95% confidence intervals. Green bars are results for all 10,000 simulated characters (except for subfigures D and F, which are restricted to parsimony-informative characters), blue bars are for OV- and TIGER-selected fast characters, and peach bars are for OV- and TIGER-selected slow characters. Colored circles intersecting the X-axis indicate that all 100 replicate data matrices scored zero. (A) Scaled Robinson-Foulds distance between the parsimony-strict-consensus trees and the simulated tree. (B) Scaled Robinson-Foulds distance between the most likely trees and the simulated tree. (C) Number of cherries (i.e., two-terminal clades) on the first most parsimonious tree saved. (D) Average amount of possible synapomorphy per character within each data matrix. (E) Ensemble retention index of all characters on the most parsimonious trees. (F) Average observed character-state space per character within each data matrix. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
In comparing scaled RF distances between optimal trees (both parsimony- and likelihood-based) inferred from the OV-slow sub-matrices with the TIGER-slow sub-matrices, the TIGER-slow matrices performed significantly better for all pairwise comparisons (i.e., same simulated branch lengths and numbers of characters included) except when the simulated branch lengths were 0.2 and 1000 characters were included (Fig. 5A and B).

The average number of cherries (two-terminal clades) present in the inferred parsimony-based trees is presented in Fig. 5C. These results are very similar to those presented in Fig. 5A. Because the parsimony-based trees inferred from the slow sub-matrices always matched the simulated topology, the number of cherries for these rooted fully asymmetrical trees is one. By contrast, the number of cherries in the phylogenies inferred from the OV- and TIGER-slow matrices were significantly higher (averaging 4.6–12.3) for all cases except when the simulated branch lengths were 0.2 and 1000 characters were included in the OV- and TIGER-slow matrices. These results corroborated our hypothesis that trees inferred from the fast matrices would have significantly fewer cherries than would trees inferred from the slow matrices. As with the results presented in Fig. 5A, the number of cherries in the phylogenies inferred from the TIGER-slow matrices was significantly lower than in the phylogenies inferred from the OV-slow matrices for all pairwise comparisons except when the simulated branch lengths were 0.2 and 1000 characters were included.

The average APS (amount of possible synapomorphy) per parsimony-informative character in the fast and slow sub-matrices is presented in Fig. 5D. In every case the average APS is significantly greater in the matrices of all 10,000 characters (averaging 9.0–14.2) than for the characters included in the OV- and TIGER-slow matrices (averaging 1.3–8.4). Furthermore, within each of the three sets of branch lengths simulated, the average APS is significantly greater in the OV- and TIGER-fast matrices (averaging 11.9–19.1) than in the matrices of all 10,000 characters. These results corroborated our hypothesis that the fast matrices would have significantly higher APS than the slow matrices. The TIGER-slow matrices had significantly higher average APS than did the OV-slow matrices for all pairwise comparisons.

The average RI (ensemble retention index) on the inferred parsimony-based phylogenies is presented in Fig. 5E. For all pairwise comparisons between OV-fast and OV-slow inferred trees, the RI was significantly higher on the OV-fast trees (averaging 0.58–0.88 for OV-fast and 0.35–0.47 for OV-slow). Likewise the RI was significantly higher for the directly comparable set of all 10,000 simulated characters (averaging 0.55–0.87) relative to the OV-slow characters. By contrast, pairwise RI comparisons between TIGER-fast and TIGER-slow inferred trees always resulted in significant differences, but without TIGER-fast characters consistently having less homoplasy than TIGER-slow characters or vice versa. These results falsified our hypothesis that there would be no significant difference in RI between the fast and slow matrices. The RI for TIGER-slow matrices (averaging 0.48–0.84) was significantly higher than that for OV-slow matrices for all pairwise comparisons.

The average observed character-state space among parsimony-informative characters in the various matrices is presented in Fig. 5F. The observed state space was significantly higher in all pairwise comparisons for the OV- and TIGER-fast sub-matrices (averaging 3.7–4.0) than for the corresponding OV- and TIGER-slow sub-matrices (averaging 2.2–3.7). These results corroborated our hypothesis that the fast matrices would have significantly higher observed character-state space than the slow matrices. The directly comparable set of all parsimony-informative simulated characters (averaging 3.1–3.9) always had significantly lower average observed state space than the OV- and TIGER-fast sub-matrices yet always significantly higher than the OV- and TIGER-slow sub-matrices. Although the differences were statistically significant, only minor differences in average observed state space were found for all pairwise comparisons between OV-slow and TIGER-slow sub-matrices, and neither had consistently greater observed state space.

![Fig. 6. Results for simulation two averaged across all 100 replicate data matrices. Error bars are * and −95% confidence intervals. Green bars are results for all 10,000 simulated characters, blue bars are for OV- and TIGER-selected fast characters, and peach bars are for OV- and TIGER-selected slow characters. Colored ovals intersecting the X-axis indicate that all 100 replicate data matrices scored zero. The Y-axis is defined such that 0 = WILD terminal resolved in small clade with one other terminal and 1 = WILD terminal resolved in large clade with two, four, or eight other terminals, respectively. (A) Results for parsimony-strict-consensus trees. (B) Results for most likely trees. * indicates that neither of the expected clades were resolved in any of the 100 replicates, and the resolution (or lack thereof in the case of polytomies) was neutrally scored as 0.5. ^ indicates that neither of the expected clades were resolved in 91 and 93 replicates, respectively, and the resolution was neutrally scored as 0.5. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)](image)
3.2. Simulation two

Results for simulation two, wherein half of the characters were simulated on a tree topology in which the WILD terminal was sister to a single terminal and the other half of the characters were simulated onto a tree topology in which the WILD terminal was sister to two, four, or eight terminals, are presented in Fig. 6. In every pairwise comparison between the phylogeny inferred from the OV- or TIGER-fast sub-matrix with the phylogeny inferred from the corresponding OV- or TIGER-slow sub-matrix, the former resolved the WILD terminal with the larger clade significantly more often than did the latter. This result, which was obtained for both the parsimony (Fig. 6A) and likelihood (Fig. 6B) trees, corroborated our first hypothesis regarding simulation two. Not only were the differences statistically significant, they were generally at opposite extremes (at or approaching one and zero, respectively), which indicates consistency among most or even all of the 100 replicates. By contrast, no such consistency was observed among the 100 replicates when all 10,000 simulated characters were analyzed, with averages of 0.40–0.56.

Our second hypothesis for simulation two was that the bias of OV and TIGER would be more pronounced as the discrepancy in number of terminals that WILD was simulated as sister to increased in the second block of 5000 characters, such that 1 vs. 2 < 1 vs. 4 < 1 vs. 8. There is some evidence corroborating this hypothesis from both the fast and slow matrices when comparing the 1 vs. 2 with the 1 vs. 4 results in Fig. 6A and B. But the evidence is not strong, and we hypothesize that the primary reason for this unexpected result is that our measure for quantifying results approached extreme values even in the 1 vs. 2 comparison (i.e., fast-sub-matrix results approaching 1 whereas slow-sub-matrix results at or approaching zero).

We did not identify consistent statistically significant differences regarding how the WILD terminal was resolved in the inferred phylogenies between OV-slow and TIGER-slow matrices (Fig. 6). But in several cases the resolution of the WILD terminal was significantly less biased in phylogenies inferred from the TIGER-slow matrices than in phylogenies inferred from the OV-slow matrices.

Unexpected resolutions of the WILD terminal, which are inconsistent with its placement on both simulation topologies, were always (parsimony strict consensus; Fig. 6A) or generally (likelihood single fully resolved tree; Fig. 6B) obtained by the TIGER-slow 500- and 1000-character sub-matrices for the 1 vs. 8 set of simulations. Based on manual examination of a subset of the 100 replicate matrices, the primary cause for this result appears to be that all (500-character sub-matrices) or most (1000-character sub-matrices) of the character-state variation is among terminals 11–17, with no or relatively little character-state variation among terminals 1–10 and WILD. As a result the WILD terminal was generally in a large polytomy in the parsimony strict consensus whereas it was to a large extent arbitrarily resolved in the single fully resolved most likely tree identified by PhyML.

3.3. Simulation three

Results for simulation three, wherein there is potential for long-branch attraction between two terminals on opposite ends of the fully asymmetrical tree, are presented in Fig. 7. The scaled RF distance between the trees inferred from all 10,000 simulated characters and the simulated tree topology is zero for all cases except when parsimony was applied to the trees with 0.05 internal branch lengths. In that case the average scaled RF distance was nearly one because of long-branch attraction.

Our first hypothesis regarding simulation-three results was that OV and TIGER slow sub-matrices would exacerbate long-branch attraction. But average RF distances for phylogenies inferred from the sub-matrices did not show an overall consistent pattern when comparing OV- and TIGER-fast with OV- and TIGER-slow, parsimony with likelihood, or OV-slow vs. TIGER-slow (Fig. 7). Yet most such pairwise comparisons do represent statistically significant differences. The following three generalizations are not true in every applicable pairwise comparison.

![Fig. 7](image-url)

Fig. 7. Results for simulation three averaged across all 100 replicate data matrices. Error bars are ±95% confidence intervals. Green bars are results for all 10,000 simulated characters, blue bars are for OV- and TIGER-selected fast characters, and peach bars are for OV- and TIGER-selected slow characters. Colored ovals intersecting the X-axis indicate that all 100 replicate data matrices scored zero. (A) Scaled Robinson-Foulds distance between the parsimony-strict-consensus trees and the simulated tree. (B) Scaled Robinson-Foulds distance between the most likely trees and the simulated tree. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
First, phylogenies inferred from the OV- and TIGER-fast sub-matrices generally had lower average RF distances relative to the simulated topology than the phylogenies inferred from the corresponding OV- and TIGER-slow sub-matrices when the simulated internal branch lengths were 0.05 or 0.1. By contrast, the TIGER-slow sub-matrices consistently outperformed the TIGER-fast sub-matrices when the simulated internal branch lengths were 0.2. Likewise for the OV-slow sub-matrices, but only in the context of likelihood trees (Fig. 7B).

Second, phylogenies inferred from the OV- and TIGER-fast matrices generally had lower average RF distances with the simulated topology when the trees were inferred using parsimony (Fig. 7A) rather than likelihood (Fig. 7B).

Third, corroborating our second hypothesis regarding simulation-three results, long-branch attraction in the slow matrices applied not only to parsimony (Fig. 7A), but to likelihood as well in most cases (Fig. 7B). The exceptions are the 1000-character TIGER-slow matrices generated from the 0.1- and 0.2-internal-branch-lengths simulations.

Fourth, phylogenies inferred from TIGER-slow matrices were generally at least as congruent with the simulated topology as phylogenies inferred from OV-slow matrices. Yet most cases wherein TIGER-slow matrices significantly outperformed OV-slow matrices are from the 0.2-internal-branch-length simulations.

3.4. Simulation four

Results from simulation four, wherein 5000 multistate characters with phylogenetic signal were merged with 5000 effectively randomized binary characters, are presented in Fig. 8. The phylogenies inferred from all 10,000 characters were always identical to the simulated topology (Fig. 8A and B). This result corroborated our hypothesis that trees inferred from all ~10,000 characters would be very similar to the simulated topology.

Likewise, phylogenies inferred from the OV- and TIGER-fast sub-matrices were always identical to the simulation topology when they were analyzed using parsimony (Fig. 8A). Phylogenies inferred from the OV- and TIGER-slow sub-matrices were significantly less congruent, with average scaled RF distances of 0.68–0.88, irrespective of whether parsimony or likelihood were used to analyze the data. These results corroborated our hypothesis that trees inferred from the fast matrices would be significantly more congruent with the simulated tree than the trees inferred from the slow matrices. We did not identify any consistent differences

Fig. 8. Results for simulation four averaged across all 100 replicate data matrices. Error bars are ± 95% confidence intervals. Solid green bars are results for all 10,000 simulated characters (except for subfigures C and E, which are restricted to parsimony-informative characters), blue bars are for OV- and TIGER-selected fast characters, and peach bars are for OV- and TIGER-selected slow characters. Colored ovals intersecting the X-axis indicate that all 100 replicate data matrices scored zero. (A) Scaled Robinson-Foulds distance between the parsimony-strict-consensus trees and the simulated tree. (B) Scaled Robinson-Foulds distance between the most likely trees and the simulated tree. (C) Average amount of possible synapomorphy per character within each data matrix. (D) Ensemble retention index of all characters on the most parsimonious trees. (E) Average observed character-state space per character within each data matrix. (F) Average TIGER partition-agreement scores for the 5000 binary characters that were simulated on the tree with branch lengths of 20.0 as well as for the 5000 multistate characters that were simulated on the tree with branch lengths of 0.2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
in performance between phylogenies inferred from the OV-slow and TIGER-slow sub-matrices that were statistically significant. Furthermore the average differences were generally minor in scale. The average APS per parsimony-informative character in the fast and slow sub-matrices is presented in Fig. 8C. In all cases the average APS was substantially and significantly higher for the OV- and TIGER-fast sub-matrices (16.1–18.7) than for the OV- and TIGER-slow sub-matrices (7.5–10.8). These results corroborated our hypothesis that APS would be higher for the fast matrices. The average APS was always significantly higher for the TIGER-slow sub-matrices than for the OV-slow sub-matrices, but the differences were minor in scale.

The average RI on the inferred parsimony-based phylogenies is presented in Fig. 8D. In all cases the average RI was substantially and significantly higher on the most parsimonious trees for the OV- and TIGER-fast sub-matrices (0.47–0.58) than on the most parsimonious trees for the OV- and TIGER-slow sub-matrices (0.21–0.41). These results corroborated our hypothesis that the slow matrices would have a significantly lower RI on their inferred trees than the fast matrices would on their inferred trees. The average RI was always significantly higher for the TIGER-slow sub-matrices (0.34) than for the OV-slow sub-matrices (0.21–0.23).

The average observed character-state space among parsimony-informative characters in the various matrices is presented in Fig. 8E. In all cases the observed state space was substantially and significantly higher for the OV- and TIGER-fast matrices (4.0) than for the OV- and TIGER-slow sub-matrices (2.2–2.9). These results corroborated our hypothesis that the slow matrices would have a significantly lower observed state space than the fast matrices. The average observed state space was always significantly higher for the TIGER-slow sub-matrices (2.6–2.9) than for the OV-slow sub-matrices (2.2–2.5).

The average TIGER partition-agreement scores for all 5000 binary and all 5000 multistate characters are presented in Fig. 8F. Although the partition-agreement scores had a far greater spread for the 5000 multistate characters, the average partition-agreement score was higher for the 5000 binary characters. This result corroborated our hypothesis that TIGER would, on average, assign the 5000 effectively randomized binary characters significantly higher partition-agreement scores than it would the 5000 phylogenetically informative multistate characters.

3.5. Empirical example: reference clades

Results from our comparison of OV-, TIGER-, and randomly-selected (averaged over ten replicates) subsamples of characters (in blocks of 500 characters deleted at a time, with up to 10,000 characters deleted) with respect to their resolution and support for six reference clades are presented in Fig. 9. None of these analyses involved taxon replication.

Trees inferred from the matrices generated after randomly selected parsimony-informative characters were deleted clearly outperformed trees inferred from the matrices generated by excluding OV- and TIGER-fast characters for four of the six reference clades when high numbers of characters were deleted (≥9000 for the angiosperms clade, ≥4000 for the (Cupressophyta, Gnetales, Pinaceae) clade, ≥6500 for the Cupressophyta clade, and ≥9000 for the Pinaceae clade; Fig. 9). Random and TIGER-based character deletion also outperformed OV-based deletion when ≥9000 characters were deleted for the Gnetales reference clade (Fig. 9B). OV- and TIGER-based deletion only outperformed random-based deletion for the two-terminal reference clade (Cycas, Ginkgo), and both did so when ≥1500 characters were deleted (Fig. 9A). Taken together, the results for five of the six reference clades supported our hypothesis that OV- and TIGER-selected slow characters do not consistently provide stronger phylogenetic signal than the equivalent number of randomly selected characters. Our hypothesis was only refuted for the (Cycas, Ginkgo) clade.

![Fig. 9. Bootstrap support for the six reference clades from Zhong et al. (2011) with different numbers of fast characters deleted by using OV and TIGER (fast characters deleted), as well as randomly deleted parsimony-informative characters. Bootstrap percentages for the random deletions are based on averages across 10 replicates. Negative bootstrap percentages are for cases wherein the reference clade was contradicted. For example, ~100% bootstrap indicates that the reference clade was contradicted by at least one clad with 100% bootstrap support. (A) Bootstrap support for the (angiosperms), (Ginkgo, Cycas), and (Cupressophyta, Gnetales, Pinaceae) reference clades. Bootstrap support was 100% for the (angiosperms) clade for all 20 cuts using random deletion. (B) Bootstrap support for the (Cupressophyta), (Pinaceae) and (Gnetales) reference clades. Bootstrap support was 100% for the (Pinaceae) clade for all 20 cuts using random deletion as well as the (Gnetales) clade for all 20 cuts using both TIGER and random deletion. Bootstrap support for the (Cupressophyta) clade was ≥99.8% for all 20 cuts using random deletion.](image-url)
With respect to comparing the performance of OV- with TIGER-based deletion for the six reference clades, OV outperformed TIGER for the angiosperm clade (>8000 characters), the (Cycas, Ginkgo) clade (<3000 characters), and the Cupressophyta clade (>6500 characters). On the other hand, TIGER-based deletion outperformed OV-based deletion for the (Cycas, Ginkgo) clade (>8000 characters) and the Gnetales clade (>9000 characters). There was no clear winner for the (Cupressophyta, Gnetales, Pinaceae) or Pinaceae clades.

Unexpectedly, both OV and TIGER reported hundreds of parsimony-informative characters as being more slowly evolving than some variable but parsimony-uninformative characters. This result was obtained when >6000 characters were deleted by OV and within the first 500 characters deleted by TIGER (Table S1). A total of 598 variable but parsimony-uninformative characters were among the 10,000 characters deleted by OV, and a total of 780 such characters were among those deleted by TIGER (Table S1).

All 598 variable but parsimony-uninformative characters deleted by OV have three or four observed character states. By contrast, the majority of the 780 variable but parsimony-uninformative characters deleted by TIGER contain just two states, and hence each has a single autapomorphy. These include 77 autapomorphies for Nicotiana, 209 autapomorphies for Morus, 46 autapomorphies for Pinus, and 103 autapomorphies for Agathis. These unexpected results from TIGER appear to be related to how divergent a given terminal is from other terminals. For example, autapomorphies for Gnetum, which is on a long terminal branch within a highly divergent clade (Gnetales), were assigned partition-agreement scores of 0.98 by TIGER, whereas autapomorphies for Nicotiana were assigned partition-agreement scores of just 0.54.

3.6. Empirical example: taxon duplication

Resolution and bootstrap support for the clade consisting of (Gnetales) and its sister group are presented in Fig. 10, with OV-slow results in Fig. 10A and TIGER-slow results in Fig. 10B. The OV-slow results were identical irrespective of whether OV was applied to the original matrix of 16 terminals or to a matrix in which all 16 terminals were triplicated (i.e., 48 terminals). Bootstrap supports were identical because the same seed was used in all RAxML bootstrap analyses. Our OV-slow results (Fig. 10A) are consistent with those reported in Zhong et al.’s (2011) Fig. 6A. The Gnceup theory is supported when up to 2000 characters were deleted, after which the Gncepine theory is supported when up to

![Fig. 10. Bootstrap support for the clade consisting of the (Gnetales) and its sister group with different numbers of characters deleted by using OV (A) and TIGER (B).](image-url)
4000 characters were deleted. The results after 4500 characters were deleted are inconsistent and weakly supported.

Triplication of all 16 original terminals except the three Gnetales and four Cupressophyta terminals (i.e., “no Gneucup × 3” in Fig. 10) produced contrary OV-slow results when 2500–6000 characters were deleted, as the Gneucup theory was moderately to strongly supported (77–100% bootstrap) when up to 6000 characters were deleted. In this case the Gnepine theory was not supported in any of the OV-based deletions.

Triplication of all 16 original terminals except the three Gnetales and three Pinaceae terminals (i.e., “no Gnepine × 3” in Fig. 10) produced OV-based results that were more consistent with those produced when the method was applied to the 16 original terminals. Here the Gneucup theory was supported when up to 1500 characters were deleted, after which the Anthophyte theory was supported when 2000 characters were deleted, and the Gnepine theory was supported when 2500–3500 characters were deleted (Fig. 10A). Thereafter Cryptomeria (Cupressophyta) alone was supported as sister to the Gnetales, with 38–94% bootstrap support, when up to 7500 characters were deleted.

Taken together, these OV results generally supported our hypothesis that OV-generated slow matrices would alternatively support the Gneucup or the Gnepine theory in a predictable manner based on the number of Cupressophyta and Pinaceae terminals sampled. The best support for our hypothesis is the contrary results between “baseline × 3” (“baseline”) and “no Gneucup × 3” when 2500–5000 characters were deleted (Fig. 10A). The qualifier for our hypothesis is that we had predicted that “no Gnepine × 3” would support the Gnepine theory despite more severe character deletions than would “baseline,” which was not the case. Instead Cryptomeria alone was supported as the Gnetales’ sister group when 4000–7500 characters were deleted. But note that in the “no Gnepine × 3” matrices the triplicated Cryptomeria terminal represents the same number of terminals (three) as the Pinaceae clade. Hence support for the (Cryptomeria, Gnetales) clade is consistent with our broader hypothesis that OV-generated slow matrices would support alternative topologies depending on the number of terminals sampled in different lineages.

In contrast to the OV-based results, the sister group of the Gnetales differed for 17 of the 20 subsamplings when TIGER was applied to the original 16-terminal matrix and the triplicated 48-terminal matrix (Fig. 10B). In both cases the Gneucup theory was supported when up to 1500 characters were deleted. Thereafter, when TIGER was applied to the original 16-terminal matrix the Gnetifer theory was supported (80–100% bootstrap) when 2500–5000 characters were deleted, the Anthophyte theory was supported (100% bootstrap) when 5500–6500 characters were deleted, and then Cryptomeria alone was supported (99–100% bootstrap) as sister to the Gnetales when 7000–9500 characters were deleted. When TIGER was applied to the triplicated 48-terminal matrix the resolution and support (0–100% bootstrap) of the sister group to the Gnetales was still less consistent and never supported the Gneucup, Gnepine, or Gnetifer theories.

Triplication of all 16 original terminals except the three Gnetales and four Cupressophyta terminals (i.e., “no Gneucup × 3” in Fig. 10) resulted in the TIGER-slow matrices supporting (49–100% bootstrap) the Gneucup theory when up to 4500 characters were deleted, after which the Gnepine theory was supported (61, 76% bootstrap) when 5000–5500 characters were deleted (Fig. 10B). Resolution of the Gnetales sister group when 6000–10,000 characters were deleted was inconsistent.

Triplication of all 16 original terminals except the three Gnetales and three Pinaceae terminals (i.e., “no Gnepine × 3” in Fig. 10) resulted in TIGER-slow matrices supporting (76–95% bootstrap) the Gneucup theory when up to 4500 characters were deleted, after which the Gnetifer theory was supported (80–100%) when up to 6000–7500 characters were deleted. After which the Gnepine theory was not supported when up to 7500 characters were deleted. In this case the Gneucup theory was strongly supported (77–100% bootstrap) when up to 6000 characters were deleted, after which the Anthophyte theory was supported (99–100% bootstrap) as sister to the Gnetales when 7000–9500 characters were deleted. When TIGER was applied to the triplicated 48-terminal matrix the resolution and support (0–100% bootstrap) of the sister group to the Gnetales was still less consistent and never supported the Gneucup, Gnepine, or Gnetifer theories.

Taken together, these TIGER results only partially supported our hypothesis that TIGER-generated slow matrices would alternatively support the Gneucup or the Gnepine theory in a predictable manner based on the number of Cupressophyta and Pinaceae terminals sampled. First, a confounding factor is the unexpectedly different TIGER results for the “baseline” and “baseline × 3” analyses, respectively. Although in all cases a
single triplicated terminal was resolved as sister to Gnetales in the
"no Gnepine × 3" analyses, the Gnepine theory was never
supported by them. Instead the Gnepine theory was only
supported in the "no Gnecup × 3" analyses when 5000 or 5500
characters were deleted, which is the opposite of our expectation
(albeit only applicable to two character subsamples with 61% and
76% bootstrap support, respectively).

In contrast to the inconsistent resolution of the Gnetales’ sister
group based on OV- and TIGER deletion of slow characters, the
most likely trees based on matrices wherein parsimony-
informative characters had been randomly deleted supported the
Gnecup theory in eight or more of the 10 replicates, except when
9500 or 10,000 characters were deleted (Fig. 11).

The number of terminals in the Gnetales’ sister group, based on
the original 16-terminal matrix, is presented in Fig. S1A. The average
number of terminals is consistently ≥31 when the most likely
trees were inferred after random parsimony-informative charac-
ters were deleted. By contrast, the number of terminals in the sister
group generally decreased from four to one or two when OV or
TIGER were used to successively delete characters. The exception
to this trend was when OV was applied to the baseline (or entirely
triplicated) matrix of 16 terminals, wherein the general pattern
observed was a decrease from four to three terminals.

Bootstrap support for the clade consisting of the Gnetales and
their sister group after deletion of 500–10,000 characters is
presented in Fig. S1B. The bootstrap support for this clade generally
showed a predictable slow decrease from an average of 98% to 60%
as more randomly selected parsimony-informative characters
were deleted. The regularity of this decrease is partially accounted
for by each point being an average across 10 random sets. The other
trend lines, for OV- and TIGER-based results, are more
sporadic. Notably, bootstrap support for the focal clade declined
precipitously to <29% after ≥3000 characters were deleted from the
baseline (or entirely triplicated) matrix of 16 terminals by OV,
and Gnetales were not even resolved as a clade when 9000–
10,000 characters were deleted.

4. Discussion

4.1. Simulations

Our results from simulation one demonstrate the following five
points. First, phylogenetic inference from OV- and TIGER-fast
matrices can dramatically outperform phylogenetic inference from
slow matrices that include the same number of characters, in both
parsimony (Fig. 5A) and likelihood (Fig. 5B) contexts. Second, both
OV- and TIGER-slow matrices can favor more symmetrical topolo-
gies than the corresponding fast matrices (Fig. 5C). Third, OV- and
TIGER-fast matrices can contain dramatically more possible
synapomorphy per character than the corresponding slow matrices
(Fig. 5D). Fourth, OV-fast matrices, but not necessarily TIGER-fast
matrices, can have substantially more retained synapomorphy
(as measured by the RI) on their inferred trees than the corre-
sponding OV-slow matrices (Fig. 5E). Fifth, the observed
class-character state space can be substantially higher in both OV-
and TIGER-fast matrices than the corresponding slow matrices (Fig. 5F).

Our results from simulation two demonstrate that in cases of
character conflict both OV- and TIGER-slow matrices can be biased
in favor of characters that provide synapomorphies for smaller
clades, whereas both OV- and TIGER-fast matrices are biased in
favor of characters that provide synapomorphies for larger clades
(Fig. 6). These biases apply in both parsimony (Fig. 6A) and likeli-
hood (Fig. 6B) contexts.

Our results from simulation three demonstrate that using OV or
TIGER to delete putatively fast evolving characters can actually
exacerbate long-branch attraction between distantly related termi-
nals relative to analyzing all characters or even just the putatively
fast evolving characters (Fig. 7). This finding applied to both parsim-
ony (Fig. 7A) and likelihood (Fig. 7B) contexts. Goremykin et al.
(2010:329) asserted that OV “… can arguably improve the net
result of phylogeny reconstruction, especially for deeper nodes,
where the inference of phylogeny is especially obscured by multi-
ple substitutions and the resulting long-branch attraction.” We
expect that their assertion is correct under some conditions, but
as demonstrated by our simulation-three results, application of
OV can have the opposite effect as well.

Our results from simulation four demonstrate that both OV-
and TIGER-slow matrices can be biased against characters
with greater observed character-state space (Fig. 5E and F). This
bias can be negatively determinate to phylogenetic inference in both
parsimony (Fig. 8A) and likelihood (Fig. 8B) contexts. The excluded
characters with greater observed state space can contain
more possible synapomorphy (Fig. 8C) and have more retained
synapomorphy (Fig. 8D) than the included characters with lower
state space.

Taken together, our simulation results demonstrate all three of
the contexts described above in Section 1.3 wherein OV’s and
TIGER’s character preference is inappropriate. First, results from
simulations one, three, and four demonstrate that their bias
against characters that may provide synapomorphies for large
clades (in an unrooted context) can be negatively determinate to
phylogenetic inference by favoring convergences and reversals
over synapomorphies (Figs. 5, 7 and 8). Second, results from simu-
lation four demonstrate that the shared bias of OV and TIGER
against characters with more observed character states can con-
found the number of observed states with rate of evolution, and
consequently be negatively determinate to phylogenetic inference
by preferentially deleting the characters with more retained
synapomorphy (Fig. 8). Third, results from simulation two dem-
strate that OV’s and TIGER’s bias against characters that provide
synapomorphies for large clades (in an unrooted context) can lead
to biased results in the context of character conflict for alternative
resolutions of a given terminal (Fig. 6).

In many simulations parsimony significantly outperformed
likelihood when applied to the OV- and TIGER-delimited
sub-matrices, despite likelihood using the same or a similar model
to infer the trees as was used to simulate the data (Figs. 5A, B, 7A, B,
and 8A, B). But an important qualification to these results is that
the 500-, 1000-, and 2000-character sub-matrices consisted only
of parsimony-informative characters. This restriction should have
no effect on parsimony, but it may have had a negatively determin-
ic effect on likelihood by biasing the distribution of autapom-
phies (a result of excluding variable but parsimony-uninformative
characters) and causing it to overestimate branch lengths (a result
of excluding invariant characters).

4.2. Empirical example

Results from our empirical example demonstrate that randomly-selected subsamples of parsimony-informative charac-
ters can contain greater phylogenetic signal than the OV- and
TIGER-selected slow characters (Fig. 9). Specifically, the
randomly-selected subsamples can have greater phylogenetic sig-
nal for larger clades (in an unrooted context), whereas OV- and
TIGER-slow characters can have greater phylogenetic signal for
smaller clades (also in an unrooted context). Furthermore, our
results demonstrate that phylogenetic inference from randomly-
selected subsamples of parsimony-informative characters can lead
to more stable inferred topologies than successively deleting char-
acters using OV or TIGER (Fig. 10 “baseline” vs. Fig. 11).

Results from our selective-terminal-replication experiments
demonstrate that this approach may not be problematic when
applied to OV-based character deletion (Fig. 10A “baseline” vs. “baseline × 3”), but that it does appear to be problematic when applied to TIGER-based character deletion because of the discrepant results between “baseline” and “baseline × 3” in Fig. 10B. Our selective-terminal-replication experiments demonstrated that phylogenetic inferences from both OV- and TIGER-slow submatrices can be sensitive to differential taxon sampling (based on our proxy of selective-terminal replication), such that well supported but mutually exclusive phylogenetic inferences can be obtained after deletion of the same number of characters (Fig. 10 “baseline × 3” vs. “no Gnecup × 3” vs. “no Gnepine × 3”). Specifically, in the context of character conflict, OV and TIGER are both biased in favor of clades that consist of few terminals and against conflicting clades that consist of many terminals. This empirically based result corroborates our simulation-two results (Fig. 6).

4.3. OV vs. TIGER

We hypothesized that TIGER, which incorporates an admittedly biased measure of character congruence, would produce less biased results than OV, which does not consider character congruence. This hypothesis was often, but certainly not always, supported based on comparison of the OV- and TIGER-slow results. For simulation one, TIGER-slow submatrices generally outperformed OV-slow submatrices with respect to the following four measures: higher congruence with the simulated topology (Fig. 5A and B), fewer artificial cherries (Fig. 5C), higher amount of possible synapomorphy (Fig. 5D), and higher retained synapomorphy (Fig. 5E). But no consistent differences were observed for character-state space (Fig. 5F). Likewise, we did not identify consistent differences regarding how the WILD terminal was resolved in the inferred phylogenies for simulation two (Fig. 6).

For simulation three, TIGER-slow submatrices generally performed at least as well as OV-slow submatrices with respect to congruence with the simulated topology (Fig. 7). For simulation four, TIGER-slow submatrices generally outperformed OV-slow submatrices with respect to the following three measures: higher amount of possible synapomorphy (Fig. 8C), higher retained synapomorphy (Fig. 8D), and higher observed character-state space (Fig. 8E). But no consistent differences were observed for congruence with the simulated topology (Fig. 8A and B).

Comparison of OV- and TIGER-slow results is problematic with respect to our selective-terminal-deletion experiments on the empirical example because of the divergent TIGER “baseline” and TIGER “baseline × 3” results (Fig. 10B). But comparison of the OV “baseline” with the TIGER “baseline” results indicates that the resolution of the sister group to Gnetales was more consistent when characters were successively deleted by TIGER than when they were successively deleted by OV, and that this resolution was generally more highly supported (Fig. 10A and B). Having stated that, it is dubious whether the high support from the TIGER submatrices is an advantage in this context given that the Anthophyta theory was supported by 100% bootstrap when 5500–6500 characters were deleted by TIGER, whereas Cryptomeria was supported by 99% or 100% bootstrap as the sole sister group of the Gnetales when 7000–9500 characters were deleted by TIGER.

OV and TIGER share the same bias in favor of characters with asymmetrical distributions of character states among terminals over characters with more symmetrical distributions of character states. But as demonstrated in both our simulation and empirical results, deletion of characters using the two methods can create mutually exclusive, well supported phylogenetic inferences—even when the same number of characters is deleted. We recommend against relying upon either method to improve phylogenetic inference relative to the complete data matrix. But if one insists upon using one of them, then we recommend that TIGER be applied instead of OV. Certainly both methods may be applied to test if they generate consistent phylogenetic inferences. Yet given their shared bias we consider testing for congruent OV- and TIGER-based results to be a dubious validation method.

5. Conclusions

Tree-independent measures of estimating the rate for each individual character along a continuum are intuitively appealing because they are not dependent upon any starting tree, and they are quick and easy to apply to phylogenomic data matrices after Goremykin et al. (2010) and Cummins and McInerney (2011) released executables that implement their procedures, which only require simple calculations.

Adding to the appeal is Goremykin et al.’s (2010:322) assertion that OV scores “... are free from any systematic bias in the estimation of the substitution rates for all sites in alignment that a wrong input topology might cause.” Likewise, Cummins and McInerney (2011:840) asserted that TIGER “... does not need any initial tree, therefore it is not biased toward any single tree and, instead, it picks out genuine signals in the data.” These assertions that OV and TIGER are not biased toward any specific tree topology may be true, but, as we have demonstrated here, they are certainly not free from any systematic bias, nor can they be trusted to reliably pick out genuine phylogenetic signal in many situations.

The systematic biases of OV and TIGER despite their tree-independent nature are analogous to two other cases that also show the same bias in favor of small clades. First, setting uniform priors on all possible tree topologies in Bayesian analyses (Yang and Rannala, 1997) is not biased toward any specific topology, but it does result in high, unjustified priors for small clades (Pickett and Randle, 2005; Randle and Pickett, 2010). Second, the frequency-within-replicates method of calculating bootstrap support within each replicate, wherein every optimal tree is given the same weight (Davis et al., 1998), may seem like a fair approach that does not overemphasize results from a given replicate (Felsenstein, 2004). But the reality is that, in cases of ambiguity, implementation of this approach, or saving a single fully resolved tree, can dramatically inflate bootstrap values for small clades (Goloboff and Farris, 2001; Goloboff and Pol, 2005; Simmons and Freudenstein, 2011).

Although OV- and TIGER-fast matrices often outperformed the corresponding OV- and TIGER-slow submatrices in our simulations (Figs. 5–8), we do not claim that characters identified as rapidly evolving by OV or TIGER necessarily contain more phylogenetic signal, on average, than those identified as slowly evolving. Nor do we advocate that authors should perform phylogenetic analyses only using their putatively fast evolving characters and then expect those fast-characters-only analyses to outperform those based on all characters or the slow characters only. We do not defend the OV- or TIGER-fast submatrices because the biased character subsampling in the OV- and TIGER-slow submatrices creates the exact opposite bias in the OV- and TIGER-fast submatrices. We simply assert that neither OV nor TIGER should be trusted to reliably increase the ratio of phylogenetic to nonphylogenetic signal in a data matrix.

We also assert that skepticism is warranted for empirical phylogenetic analyses that used OV- and/or TIGER-based character deletion to justify preference for a particular topology wherein a small clade is supported after deletion of characters, yet was contradicted by a larger clade when the entire data matrix was analyzed. Three examples, in addition to the Zhong et al. (2011) study that we re-examined here, are as follows. First, Goremykin et al.’s (2009:30) inference that Ceratophyllum is more closely related to the magnoliids clade (a 5/29 split) than it is to the eudicot clade (an 18/16 split) after OV-based deletion of 2500 characters, contra
Soltis et al. (2011). Second, Zhong et al.’s (2014) inference that Zygmenatales may be more closely related to Coleochaetales (a 5/25 split) than they are to the land-plant clade (a 7/23 split) after OV-based deletion of 9000 characters, contra Zhong et al. (2013). Ruhfel et al. (2014), and Wickett et al. (2014). Third, Goremykin et al.’s (2015) inference that the water lilies are more closely related to Amborella (a 4/32 split) than they are to the rest of the flowering plants (a 17/19 split) after OV-based deletion of 2250 characters, contra Ruhfel et al. (2014), Wickett et al. (2014), and Simmons and Gatyis (2015). We do not know that any of these phylogenetic inferences are wrong but do suggest that further investigation using methods independent from OV- and TIGER-based character deletion is warranted.

Acknowledgments

We thank two anonymous reviewers for suggestions with which to improve the manuscript; Ed Braun for helpful suggestions; Anne Kupczok for helpful discussions; and Zhenxiang Xi for sending the OV scripts. J.G. was supported by NSF grant DEB-1457735.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2016.03.018.

References


