Fragmentary Gene Sequences Negatively Impact Gene Tree and Species Tree Reconstruction

Erfan Sayyari,[1] James B. Whitfield,[2] and Siavash Mirarab*,[1]

[1]Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA

[2]Department of Entomology, University of Illinois, Urbana, IL

*Corresponding author: E-mail: smirarab@ucsd.edu.

Associate editor: Rasmus Nielsen

Abstract

Species tree reconstruction from genome-wide data is increasingly being attempted, in most cases using a two-step approach of first estimating individual gene trees and then summarizing them to obtain a species tree. The accuracy of this approach, which promises to account for gene tree discordance, depends on the quality of the inferred gene trees. At the same time, phylogenomic and phylotranscriptomic analyses typically use involved bioinformatics pipelines for data preparation. Errors and shortcomings resulting from these preprocessing steps may impact the species tree analyses at the other end of the pipeline. In this article, we first show that the presence of fragmentary data for some species in a gene alignment, as often seen on real data, can result in substantial deterioration of gene trees, and as a result, the species tree. We then investigate a simple filtering strategy where individual fragmentary sequences are removed from individual genes but the rest of the gene is retained. Both in simulations and by reanalyzing a large insect phylotranscriptomic data set, we show the effectiveness of this simple filtering strategy.

Key words: **species tree estimation, gene trees, phylogenomics, fragmentary data, insect phylogeny.**

Introduction

Genome-scale reconstruction of species trees has become the standard practice in phylogenetics. A typical phylogenomic analysis starts by sequencing hundreds to thousands of loci using one of several sequencing strategies (e.g., transcriptomics, targeted amplicon sequencing, hybrid enrichment, etc.). Data from multiple loci may be then concatenated together to build a supermatrix, which is then analyzed using standard phylogenetic methods such as maximum likelihood (ML). The concatenation approach ignores potential discordance between gene trees and the species tree (Maddison and Wiens 1997; Degnan and Rosenberg 2009), and has been proven statistically inconsistent (Roch and Steel 2014) under the multispecies coalescent (MSC) model (Pamilo and Nei 1988; Rannala and Yang 2003). An alternative approach, gaining in popularity, is to first estimate a gene tree for each locus (independently from other loci) and to then combine the gene trees using a summary method (Edwards 2009). Several existing summary methods have been proven statistically consistent under the idealized conditions when gene trees are considered error-free; examples of summary methods used in practice include ASTRAL (Mirarab et al. 2014a; Mirarab and Warnow 2015), STAR (Liu et al. 2009), NJst/ASTRID (Liu and Yu 2011; Vachaspati and Warnow 2015), and MP-EST (Liu et al. 2010). Although alternative approaches such as coestimation (Heled and Drummond 2010; Liu 2008) and site-based MSC-based methods (Bryant et al. 2012; Chifman and Kubatko 2014) exist, these methods have been less frequently used, perhaps due to their computational requirements (Bayzid and Warnow 2012; Zimmermann et al. 2014).

Despite their growing application to real data (Jarvis et al. 2014; Wickett et al. 2014; Prum et al. 2015; Rouse et al. 2016), the accuracy of summary methods is directly impacted by the accuracy of the input gene trees (Patel et al. 2013; Roch and Warnow 2015; Mirarab et al. 2016; Springer and Gatesy 2016; Xu and Yang 2016). A well-studied source of gene tree estimation error (or uncertainty) is statistical noise due to lack of phylogenetic signal in short loci (Mirarab et al. 2014c; Xu and Yang 2016). This has motivated the development of methods for detecting and removing low signal genes (Salichos and Rokas 2013; Xu and Yang 2016) or binning of loci to larger units (Mirarab et al. 2014c; Bayzid et al. 2015). However, other factors, such as long branch attraction and missing data may also impact gene tree accuracy (Gatesy and Springer 2014; Springer and Gatesy 2016), and these have been less thoroughly studied (but see Liu et al. 2015; Edwards et al. 2016).

The effect of missing data on the accuracy of single-locus or supermatrix tree reconstruction has been thoroughly studied (Philippe et al. 2004; Wiens 2006; Lemmon et al. 2009; Wiens and Morrill 2011; Simmons 2012). In a summary method pipeline, missing data come in two forms, as previously noted by Hosner et al. (2016). A species may be fully missing from some of the loci; we refer to this scenario (type I in Hosner et al. [2016]) as *missing tips* and to the patterns of presence/absence resulting from it as *taxon occupancy*. Alternatively, a species may be present with only partial data for some of the loci, and we refer to this scenario as

Article

fragmentary data (type II in Hosner et al. [2016]). These two forms of missing data may have very different impacts on the species tree reconstruction. Missing tips may negatively impact the summary method when the species tree is being inferred from a set of taxonomically incomplete (partial) gene trees, whereas, fragmentary data may negatively impact the gene tree inference step (Simmons 2014; Hosner et al. 2016). Although some studies have examined the impact of missing tips on summary methods (Xi et al. 2016; Hovmöller et al. 2013; Huang and Knowles 2016), to our knowledge, only Hosner et al. (2016) have examined impacts of both types of missing data.

Current high-throughput genomic sequencing methods vary considerably in the size of the raw sequencing reads they generate, with there generally being a positive relationship between read length and the error rate in the sequence (Nagarajan and Pop 2013). Sequence assembly (generating larger contigs) also varies greatly in efficiency and accuracy, and most applications, due to computational difficulty, rely on heuristics (Lemmon and Lemmon 2013). High-quality sequence generation and assembly also can be compromised by challenges with organism size and availability, particularly for highly diverse taxa of small body size such as insects (Richards and Murali 2015). Finally, transcriptomic data sets (Misof et al. 2014; Wickett et al. 2014) can have length variation in the assembled genes because of alternative splicing and the coverage of individual genes may also be affected by expression levels. Thus, phylogenomic and phylotranscriptomic studies that strive to provide thorough taxon sampling of diverse lineages at a reasonable cost often contain fragmentary data, at least for a subset of the taxa. Since fragmentary data may negatively impact both the gene alignment (Nguyen et al. 2015) and gene tree estimation (Lemmon et al. 2009; Hosner et al. 2016), it is important to study effects of fragmentary data on the species tree reconstruction and ways to ameliorate the impacts.

One approach to deal with negative impacts of fragmentation, used in some phylotranscriptomic studies (Wickett et al. 2014), is to remove each species from those genes where it is fragmentary. Filtering fragmentary data creates missing tips and thus presents a trade-off between fragmentation and taxon occupancy. Because of this trade-off, it is not clear whether filtering fragmentary data is overall beneficial to the accuracy; if indeed beneficial, it is not clear what level of filtering is warranted. Note that a similar trade-off does not face a concatenation analysis because no gene tree is ever estimated in such analyses and removing fragmentary sequences while keeping the respective genes only creates more missing data (supplementary fig. S1, Supplementary Material online) without any obvious benefit (except perhaps in the alignment step).

In this paper, we study effects of fragmentary data on species and gene tree reconstruction using summary methods. In line with observations of Hosner et al. (2016), but using simulations in addition to real data, we demonstrate the negative impact of fragmentary data. Unlike Hosner et al. (2016) who deal with fragmentary data by removing genes that show low phylogenetic signal, we study the strategy of filtering specific species from individual genes. Given a filtering threshold (e.g., 20%), we remove from each gene alignment any species that has nongap characters in less than the given threshold (e.g., 20%). This form of filtering retains the gene and can arguably result in better utilization of the data because the nonfragmentary sequences are retained. We test our proposed filtering method on an insect data set and corroborate our findings in simulations.

We studied an empirical transcriptomic data set of insects consisting of 1,478 protein-coding genes of 144 taxa, where 27% of the alignment is gaps (Misof et al. 2014). In 90% of the genes, there are 115–141 species present, and aligned protein coding sequences were between 134 and 890 amino acids in 90% of the genes. Insects represent a species-rich lineage of organisms with generally small body size and can be challenging for production of high-quality phylogenomic data due to the difficulty of obtaining sufficient tissue (and thus DNA) for tiny, rare taxa required for full taxon representation. The insect lineage contains within its history several questions of broad evolutionary and scientific interest beyond its high species diversity, such as the evolution of wings and flight, various forms of metamorphosis, and multiple origins of social behavior (Misof et al. 2014). Thus, for reconstructing the history of insects in sufficient detail to draw conclusions about biological questions of interest, a relatively full taxon sampling from all insect groups was desired. Transcriptomic data were used as an achievable way of providing such taxon representation with comparative genomic data for phylogeny estimation. Transcriptomic data typically include a large number of genes, but with the cost of having significant amounts of fragmentary data. Misof et al. (2014) used concatenated analysis rather than coalescent-based gene tree summary methods, presumably because of the highly variable quality of the individual gene trees. The large size of this data set and the high amount of fragmentary data make it especially well-suited for our analyses. In addition, the insect phylogeny has received considerable attention over the years, so that we have some prior expectation of relationships among some lineages, providing a perspective on the accuracy of the phylogenetic results.

In simulations, we study impacts of fragmentary data and the filtering strategy on the accuracy of gene trees and consequently the accuracy of the species trees. Our simulated data set (see Materials and Methods) simulates gene tree discordance due to ILS and we use estimated gene trees (with error); we also randomly inject fragmentation in gene alignments with patterns that emulate the biological insect data set. We infer gene trees from the original sequences, unfiltered alignments, and alignments after filtering fragments (thresholds: 10–80%) using both RAxML (Stamatakis 2014) and FastTree2 (Price et al. 2010). We infer species trees using ASTRAL-II (Mirarab and Warnow 2015) using 50, 200, or 1,000 gene trees.

Results

We start with a simulation study and then analyze the empirical insect data set.

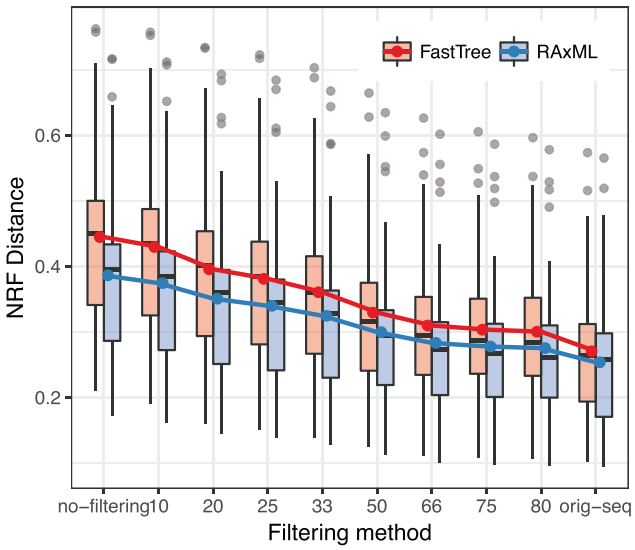

Fig. 1. NRF distances between true and estimated gene trees in the simulated data set. The x-axis shows the filtering thresholds; from left to right, more aggressive filtering is applied. Leftmost boxes are for no filtering and the rightmost boxes are for gene trees in absence of fragmentary data. For each threshold, box plots and average distances (lines) over 48 replicates are shown.

Simulation Results

Impact on Gene Trees

Comparing the original data sets that include no fragmentation (*orig-seq*) and those with injected fragmentation (*no-filtering*) shows that the presence of fragmentary data dramatically increases gene tree error (fig. 1), as measured by the average normalized Robinson-Foulds distance (NRF) between true gene trees and the estimated gene trees. However, progressively applying more aggressive filtering gradually decreases gene tree error (fig. 1) and the extent of improvements depends on the filtering threshold ($P \ll 10^{-5}$; ANOVA). With no filtering, the NRF distance is as high as 0.44 for FastTree and 0.39 for RAxML. Filtering fragmentary data gradually reduces the NRF distance. At the 75% threshold, the average NRF distance is reduced to 0.30 or 0.28, respectively, for FastTree and RAxML, which is only slightly lower than the average NRF error with no fragmentation (0.27 or 0.25, respectively) but these small differences are still statistically significant ($P \ll 10^{-5}$). Overall, RAxML gene trees are significantly more accurate than FastTree trees ($P \ll 10^{-5}$). Note that to compute the NRF distance, true and estimated trees are restricted to the same set of leaves, normalizing by the remaining branches; as supplementary figure S2, Supplementary Material online demonstrates, reductions in the NRF after filtering cannot be attributed to the shrinking leaf set due to filtering.

Impact on the Species Tree

Regardless of the number of genes, adding fragmentation to the data increases the NRF distance between true and estimated species trees (fig. 2*a*) significantly ($P = 0.00014$ for RAxML and $P \ll 10^{-5}$ for FastTree). For example, with 1,000 genes, the average species tree NRF distance increases from 0.030 with no fragmentation to 0.057 (90% increase) with fragments and no filtering for FastTree gene trees and from 0.023 to 0.037 (60% increase) for RAxML gene trees.

Filtering fragmentary data has nonmonotonic impacts on the species tree error (fig. 2*a*). As the filtering threshold increases, the average species tree estimation error initially tends to drop but eventually starts to increase again. The optimal threshold depends on the number of genes and in most conditions varies between 25% and 33% (supplementary table S1, Supplementary Material online). Limiting ourselves to all thresholds up to 33%, we observe that the accuracy of the species tree tends to gradually improve as a result of increased filtering when at least 200 genes are available; however, improvements are statistically significant only with 1,000 genes ($P = 0.00461$ and $P = 0.0417$, respectively, for FastTree and RAxML) and not for 200 ($P = 0.102$, and $P = 0.314$, for FastTree and RAxML). With 1,000 genes, filtering is never worse than no filtering, even with extremely aggressive filtering. At the 50% threshold, the NRF distance reduces from 0.037 to 0.028 for RAxML gene trees and from 0.057 to 0.036 for FastTree trees.

Gene Tree versus Species Tree Error

Reducing gene tree error by increased filtering is only beneficial to the species tree estimation when taxon occupancy is not dramatically sacrificed (fig. 2*b*). As we go from no filtering to filtering up to 33%, the species tree error and gene tree error both tend to decrease at first. Further increases in the filtering threshold continue to reduce the gene tree error, but those reductions don't always translate to improvements in species tree error, and in fact, can increase it. This is perhaps partly because improvements in gene tree error eventually become small with each increase in the filtering threshold. More importantly, the taxon occupancy continues to decrease with more filtering and lack of occupancy may offset the benefits of reduced gene tree error. The average taxon occupancy drops from 88% to 77% and then to 65% as we increase the filtering threshold from 33% to 50% and then to 66%.

Empirical Results

Reconstructed ASTRAL-II insect trees (supplementary figs. S4–S11, Supplementary Material online) change in topology and support based on our choice of gene tree estimation method (RAxML vs. FastTree) and the treatment of fragmentary data. Before presenting results in detail, we start by describing our approach in judging accuracy on the biological data set.

Evaluation and Expected Relations

In order to gain some indication of how accurate the species tree results are, we first surveyed some relationships that have been previously considered to be well established on the basis of evidence, and also identified those that have been previously found but not consistently supported (table 1). Among the former are the monophyly of Hexapoda among the Pancrustacea, monophoply of "true insects," monophyly of the Dicondylia (Zygentoma + Pterygota),

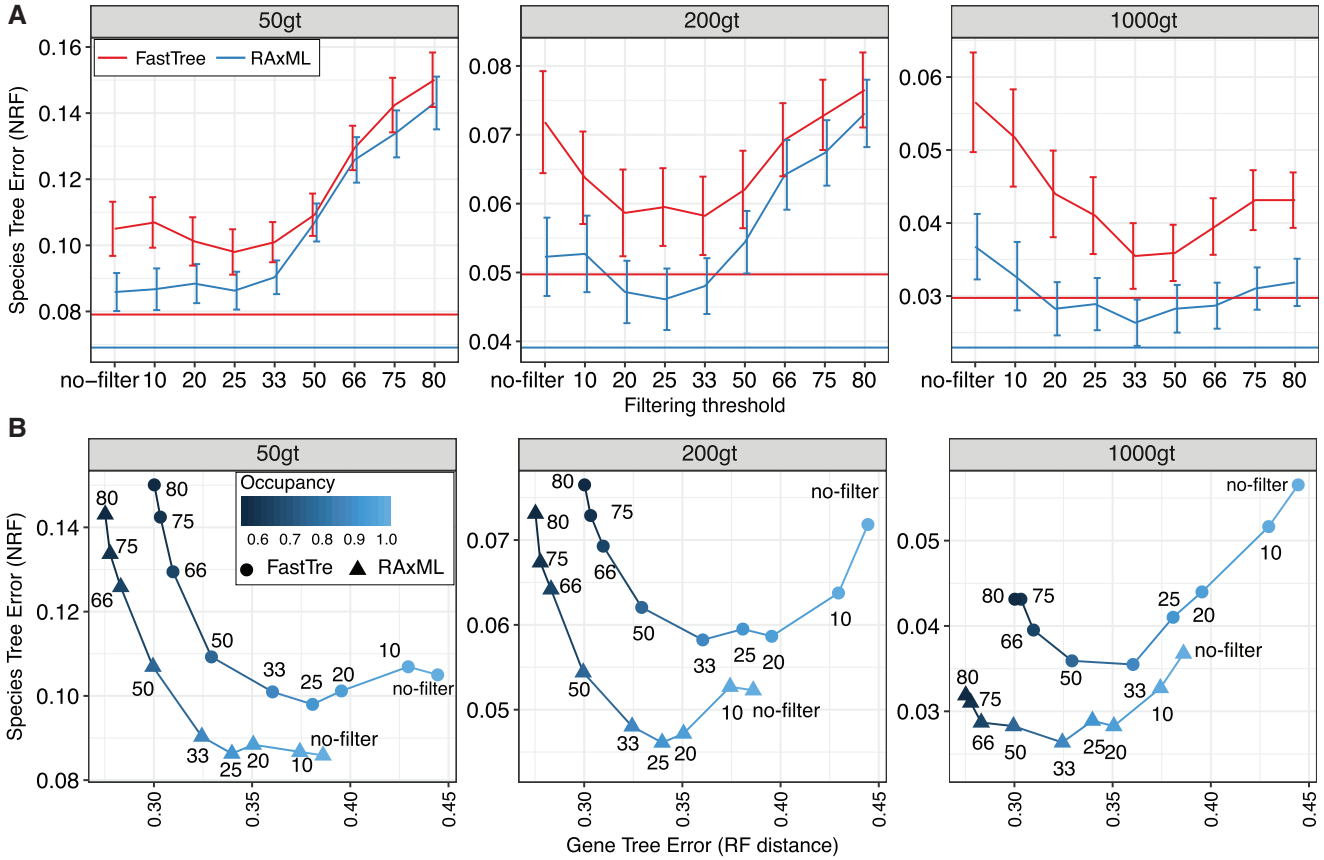

Fig. 2. Species tree error in simulation data sets. (*a*) NRF error of estimated ASTRAL species trees for different numbers of genes (boxes) and varying filtering thresholds (x-axis) with both RAxML and FastTree gene trees. The horizontal lines indicates the error rate of ASTRAL in the absence of fragmentary data. The error bars in these figures indicate the standard errors around the average. (*b*) Correlation between the gene tree and the species tree error. The y-axis shows the average species tree error (NRF distance) and x-axis shows the NRF distance between true and estimated gene trees. Shades represent the average occupancy of species, and fragmentary filtering thresholds are noted next to the dots. Results from RAxML and FastTree gene trees are distinguished by dot shape.

monophyly of Pterygota, monophyly of Neoptera, monophyly of Holometabola, monophyly of Antliophora (Diptera + Mecoptera + Siphonaptera), and monophyly of Amphiesmenoptera (Trichoptera + Lepidoptera). Among the latter are the monophyly of the Paleoptera (Ephemeroptera + Odonata), monophyly of the Polyneoptera, sister groups relationships between the Grylloblattodea and Mantophasmatodea, between Hemiptera and Thysanoptera, and between Coleoptera and Strepsiptera, and the position of Hymenoptera as sister to the remaining holometabolous orders. Misof et al. (2014) found two other relationships without previous strong support: the sister group relationship between Psocodea and Holometabola, and a clade containing Mantophasmatodea + Grylloblattodea plus Embiidina + Phasmida. We discuss our results in the context of these prior expectations and findings.

In addition to judging the quality of the species tree based on prior evidence, we also study the impact of the filtering on taxon occupancy, gene tree bootstrap support, and evolutionary diameter of the gene trees, measured by the tip-to-tip distance. A reduced taxon occupancy is clearly undesirable (even if inevitable). Reduced gene tree bootstrap support can be interpreted as a sign of increased uncertainty about gene trees and perhaps increased error. An increase in the evolutionary distance can be indicative of artificially long branches that can be inferred as a result of fragmentary data (Lemmon et al. 2009).

Occupancy

Filtering fragmentary data affects the taxon occupancy of different orders and species unevenly (fig. 3*a* and supplementary fig. S3, Supplementary Material online). Here, we measure occupancy of a clade by the percent of genes that have at least one of the species from the specific clade. Almost all clades have at least 50% occupancy, regardless of the threshold selected. However, the occupancy of clades for the filtering thresholds of 20–33% are similar, with a considerable drop at 50%, and a dramatic drop at 66% filtering or higher. At 50% filtering, the occupancy is above 70% for all orders.

Gene Trees

Filtering more fragmentary data improves gene tree bootstrap support for both RAxML and FastTree (fig. 3*b*). For example, with no filtering, the number of branches with 100% support in RAxML gene trees is only 5% but gradually

Table 1. Significant Clades in the Insect Phylogeny with References to Evidence Supporting Them.

Code	Clade Composition	Evidence	Selected References
A	Mecoptera+Siphonaptera	Fairly strong	Wiegmann et al. (2009)
B	Diptera + (Mecoptera+Siphonaptera)	Strong	Kristensen (1999); Wiegmann et al. (2009); Beutel et al. (2011)
C	Trichoptera+Lepidoptera	Strong	Kristensen et al. (2007); Kristensen (1975)
B/C	Clades B+C	Strong	Wiegmann et al. (2009); Beutel et al. (2009, 2014)
D	Neuropterida + (Coleoptera+Strepsiptera)	Fairly strong	Wiegmann et al. (2009); Beutel et al. (2011) Niehuis et al. (2012); Boussau et al. (2014)
D/B/C	Holometabola minus Hymenoptera	Fairly strong	Savard et al. (2006); Misof et al. (2007) McKenna and Farrell (2010); Heraty et al. (2011)
E	Holometabola (Endopterygota)	Strong	Wiegmann et al. (2009); Beutel et al. (2011) Niehuis et al. (2012); Trautwein et al. (2012)
F	Hemiptera+Thysanoptera	Fairly strong	Beutel et al. (2014)
G	Acercaria+Hymenoptera	Fairly strong	Beutel et al. (2014)
H	Mantophasmatodea+Grylloblattodea	Fairly strong	Terry and Whiting (2005); Cameron et al. (2006) Wipfler et al. (2011)
I (missing)	Clade H + (Embiidina+Phasmida)	Weak	Misof et al. (2014) but not in the final species tree of this study
J	Polyneoptera (Orthopteroidea)	Fairly strong	Yoshizawa (2011); Ishiwata et al. (2011) Letsch and Simon (2013)
K	Neoptera	Strong	Kristensen (1975); Klass (2009)
L	Pterygota (winged insects)	Strong	Kristensen (1991); Grimaldi and Engel (2005); Hasenfuss (2002) Klass (2009); Wheat and Wahlberg (2013)
M	Zygentoma+Pteygota	Fairly strong	Bitsch and Bitsch (2004); Engel and Grimaldi (2004) Klass (2009)
N	Insecta	Strong	Bitsch and Bitsch (2004); Engel and Grimaldi (2004) (Hennig et al. 1969; Klass 2009)
P	Psocodea+Holometabola	Weak	Only supported by Misof et al. (2014) and weakly by this study
Ingroup	Hexapoda	Fairly strong	Kjer et al. (2006); Klass (2009); Grimaldi (2010); Meusemann et al. (2010); Regier et al. (2010)

NOTE.—Letter codes refer to red-labeled nodes in figure 4, except node I, which is missing in our final species tree but present in the Misof et al. (2014) final concatenation tree. The evidence from the literature is (subjectively) classified into three groups. *Strong*: virtually always recovered in previous molecular phylogenetic studies and not controversial based on comparative morphology *Fairly strong*: usually recovered by phylogenomic studies but either not clearly supported by morphology or sometimes another well-supported alternative exists. *Weak:* either controversial based on comparative morphology or seldom strongly supported by any analysis.

increases to 12% with the highest level of filtering (i.e., 80% threshold). Similarly, the median branch bootstrap support for RAxML (or FastTree) gene trees is 37% (29%) with no filtering but gradually increases to 48% (48%) with 80% filtering. Conversely, the number of branches with <33% support decreases from 47% (53%) with no filtering to 38% (39%) with 80% filtering (fig. 3*b*). Overall, both RAxML and FastTree gene trees improve in their bootstrap support, but the improvements are larger for FastTree.

The increased bootstrap with increased filtering can be attributed to the negative impact of fragmentary sequences on estimated gene trees (both ML and bootstrap replicates). Fragmentary sequences and their treatment as ambiguous data tend to result in long branches in an ML estimation (Lemmon et al. 2009). Consistent with this expectation, we observe that filtering out fragmentary data consistently reduces the evolutionary diameter (fig. 3*c* and supplementary fig. S12, Supplementary Material online) of gene trees, indicating that fragments may result in long branches. This reduction in evolutionary diameter coincides with increasing gene tree bootstrap support (fig. 3*c*). In the case of bootstrap replicates used to estimate support, the problem of fragmentation is exacerbated by the resampling of sites, which may leave no or few sites with any nongap characters. Finally, note that the perceived improvements in gene trees support and evolutionary diameter come at the cost of reduced taxon occupancy (fig. 3*c*).

Overall, the occupancy plot (fig. 3*a*), gene tree statistics (fig. 3*b* and *c*), and our simulations (fig. 2*a*) lead us to favor thresholds between 20% and 50%. In the rest of the discussion, whereas we will continue to discuss all thresholds, where we need only one threshold, we will use 50% as the default (we return to this choice in the discussion section).

Species Trees

With all fragments included, the ASTRAL tree inferred from FastTree gene trees fails to recover a number of relationships (fig. 4*a*): the monophyly of Hexapoda (ingroup), monophyly of Dicondylia (M), monophyly of Pterygota (L), monophyly of Neoptera (K), monophyly of Thysanoptera + Hemiptera + Psocodea + Holometabola (G), and monophyly of Neuropteroidea (D). As filtering of fragments increases (moving left to right in fig. 4*a*), the species tree improves, so that monophyly of Hexapoda (Ingroup), Dicondylia (M), Neoptera (K), and Thysanoptera + Hemiptera + Psocodea + Holometabola (G) are recovered, albeit not strongly so in the case of Neoptera and Hexapoda. ASTRAL run using FastTree gene trees never recovers the monophyly of Pterygota (L) and monophyly of Neuropteroidea (D), which have fairly strong support in the literature (table 1). Neither does it recover the Psocodea + Holometabola clade (P) or the I node found by Misof et al. (2014), which don't have strong support in prior analyses. Some expected clades, especially within

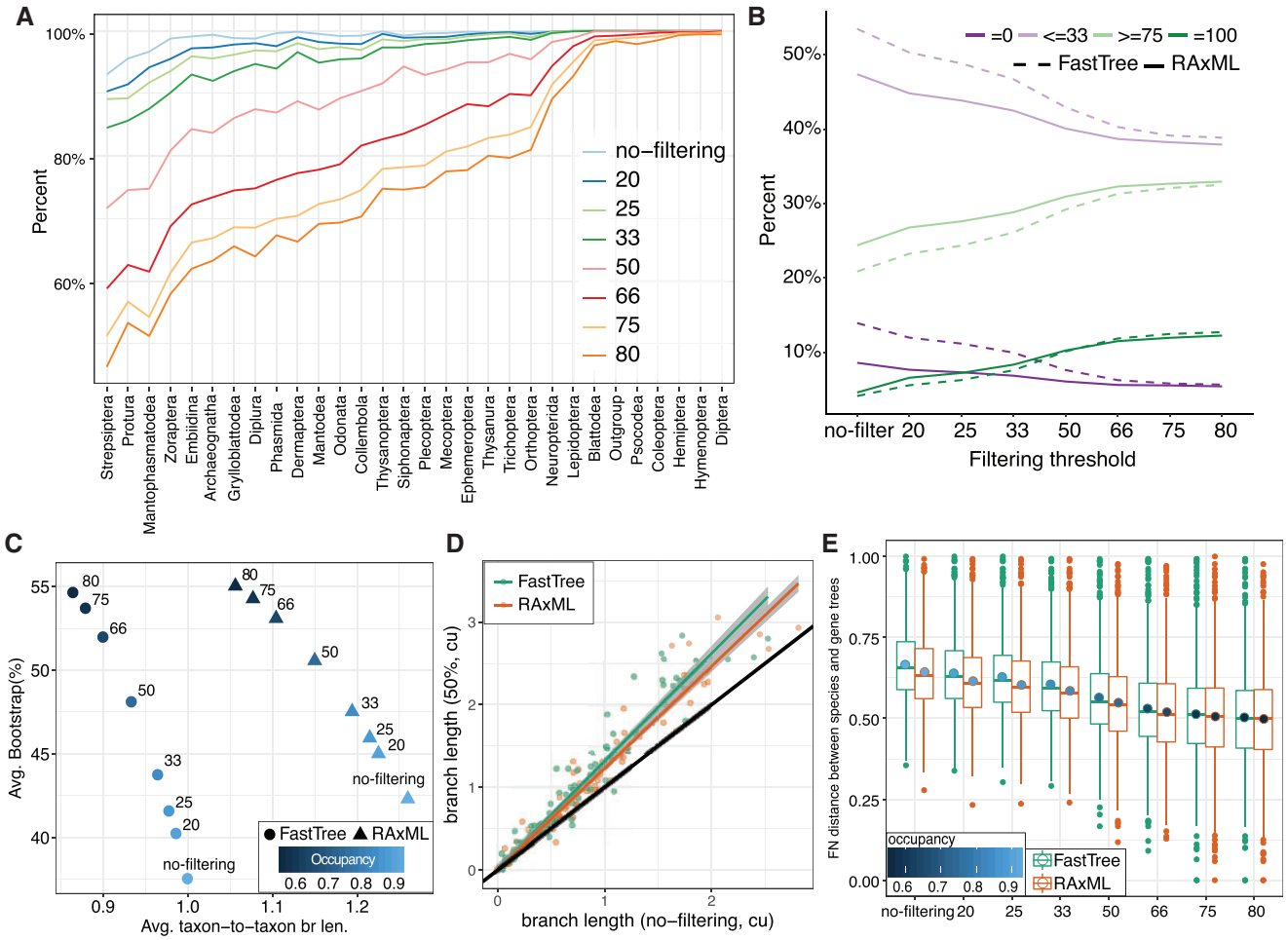

Fig. 3. Impacts of filtering on the biological insect data set. (*a*) Occupancy of major clades after filtering fragmentary data with various thresholds. (*b*) Distribution of average BS values for different filtering thresholds. We show the percent of branches that have BS value of 0%, <33%, >75% and 100%. (*c*) Average gene tree bootstrap support (y-axis) versus average (over genes) of average (over leaves) of root-to-tip distances (x-axis) with different filtering thresholds (text next to the dots). Shades represent the average occupancy. Average of maximum tip-to-tip distances (evolutionary diameter) shows similar patterns (supplementary fig. S12, Supplementary Material online). (*d*) Coalescent unit branch lengths computed by ASTRAL. Each dot corresponds to a branch and its coalescent unit length is shown when estimated from unfiltered gene trees (x-axis) or 50% filtered gene trees (y-axis). Several branches (25 for FastTree and 13 RAxML) that were not shared between the two trees are removed. A line is fitted to all the points corresponding to each method, and the dashed line shows the unity line. (*e*) Discordance of gene trees with various filtering thresholds (x-axis) versus the corresponding ASTRAL species tree. Boxplots show distributions of the proportion of species tree branches not found in gene trees.

Holometabola, always had strong support no matter how much filtering was done (supplementary fig. S13, Supplementary Material online).

In contrast, when RAxML gene trees are used, the only strong relationship that is not recovered is the monophyly of Hexapoda (fig. 4*a* and *b*). Any level of filtering would result in a monophyletic Hexapoda, showing that even for RAxML gene trees, correct handling of fragmentary data can improve the species tree topology. Moreover, even though the species tree topology inferred from RAxML gene trees is relatively robust in the presence of fragmentary data, the ASTRAL estimated branch lengths increase with filtered gene trees (fig. 3*d*). Since coalescent unit branch lengths tend to be underestimated (Mirarab et al. 2014c; Sayyari and Mirarab 2016), the increased branch lengths are likely to be more accurate. We note that our simulations corroborate that coalescent unit branch lengths are underestimated and show a strong positive correlation between gene tree accuracy and estimated branch lengths (supplementary fig. S14, Supplementary Material online).

Our final ASTRAL tree (fig. 4*b*) using RAxML gene trees and 50% filtering includes all major clades with prior support in the literature, and all but two of them (monophyly of Hexapoda, and Psocodea + Holometabola) had full support. Interestingly, Misof et al. (2014) found a clade of Embiidina, Phasmida, Grylloblattodea, and Mantophasmatodea (I) that our final tree does not recover, but this clade has little historical support; it will be interesting to see if this clade is supported by further studies.

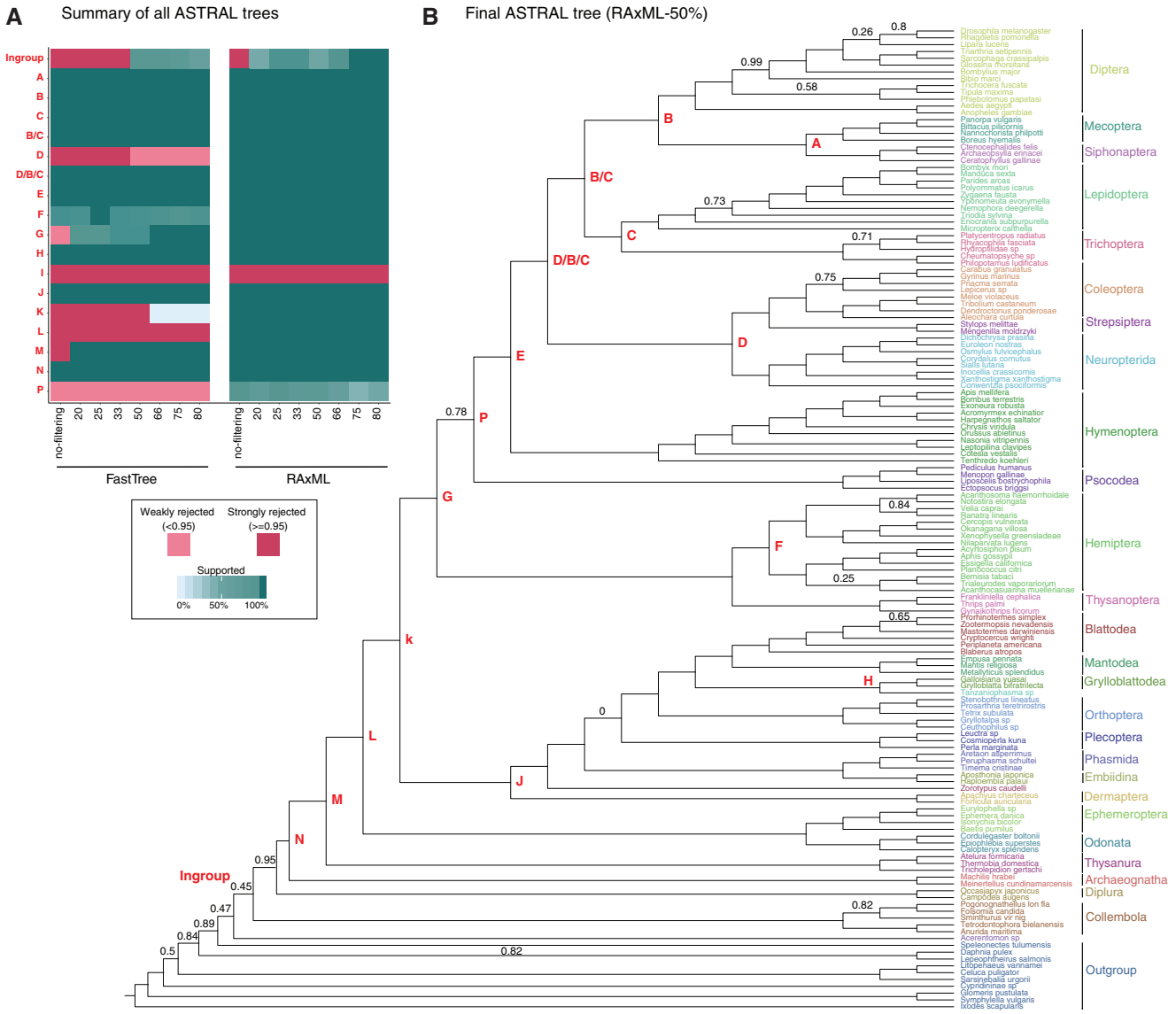

Fig. 4. ASTRAL species trees. (*a*) Recovery of important clades in ASTRAL species trees with various filtering thresholds (see supplementary figs. S4–S11, Supplementary Material online for trees), represented by columns. Rows show important clades A to P (table 1). The blue-green indicates monophyly of a clade, and the spectrum of blue to green colors show ASTRAL localPP support values (Sayyari and Mirarab 2016). Red, indicates strong or weak rejection of a clade. Weak rejection is defined as a clade that is absent from a tree but is compatible with the tree if branches below 95% support are contracted. See supplementary figure S13, Supplementary Material online for more clades. (*b*) The ASTRAL species tree using 1478 RAxML gene trees with the 50% filtering threshold.

Gene Tree Species Tree Discordance

As a final evidence that gene tree accuracy has improved, we demonstrate that after filtering, both RAxML and FastTree gene trees show reduced discordance. Overall, the amount of gene tree discordance with the species tree reduces substantially as we increase filtering, especially up to the 66% threshold (fig. 3*e*). These reductions are the reason for the reduced coalescent unit branch lengths (fig. 3*d*). Similarly, the overall ASTRAL quartet score (proportion of gene tree quartets found in the species tree) increases as more filtering is applied (supplementary table S2, Supplementary Material online).

To further break down patterns of discordance, we compare support for major clades (orders plus clades shown in table 1) in our gene trees before and after filtering (fig. 5). Before filtering, many of the insect orders receive surprisingly little support in our gene trees. For example, with RAxML (or FastTree) gene trees, 13 (or 14) out of 26 orders are recovered in less than half of the gene trees, and only six orders are recovered in at least three-quarters of the gene trees. Moreover, most gene trees have low support and cannot strongly reject or support the monophyly of these orders. However, the 50%-filtered gene trees show strong support for most orders. Only seven orders are recovered in fewer than half of these gene trees, and the number of orders supported by at least three-quarter of the genes increases to 13 and 11, respectively, for RAxML and FastTree. As an example, before filtering, 35% (28%) of RAxML (FastTree) genes recovered Lepidoptera as a monophyletic clade, and only 15% (9%) have high bootstrap support, whereas, after filtering, 64% (62%) recover it and 47% (44%) have high

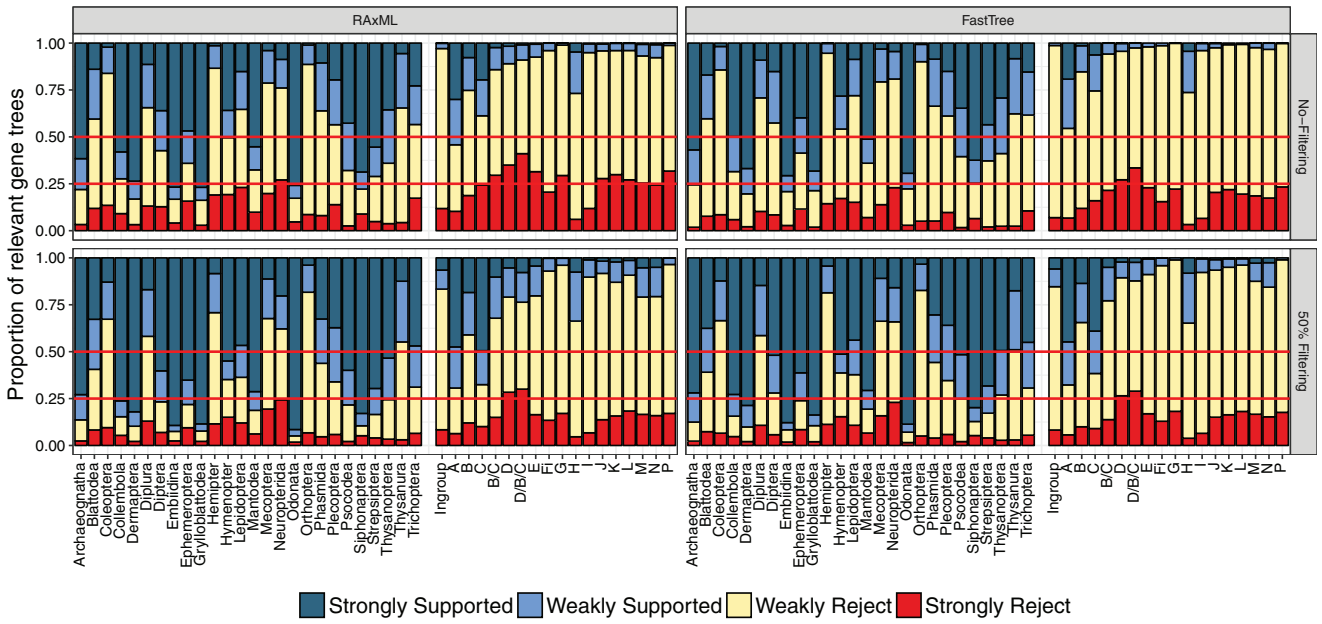

Fig. 5. Distribution of clade supports for no-filtering and 50%-filtered gene trees. The proportion of (*a*) RAxML and (*b*) FastTree gene trees supporting or rejecting each clade is shown. Branches with bootstrap support value 75% and higher (lower) are considered highly (weakly) supported; "Weakly Reject" refers to gene trees that don't recover a clade but are compatible with it once low support branches are contracted; "Strongly Reject" are clades that remain incompatible even after contracting low support branches. Two horizontal lines mark 50% and 75% support marks.

support. Although Lepidoptera has one of the biggest changes, patterns across all orders consistently point to reduced discordance. Major clades (other than orders) also have increased support and substantial reduction in highly supported discordance.

Discussion

Impacts of Fragments

We showed that fragmentary data can have substantial negative impacts on gene trees and consequently species trees estimated in a summary method pipeline. These results build on previous studies of weak phylogenetic signal, the resulting high gene tree error (or uncertainty), and species tree error (Patel et al. 2013; Salichos and Rokas 2013; Mirarab et al. 2014c; Xu and Yang 2016). It is important to note that fragmentation not only weakens the signal, but it may also create biases in an ML analysis (Lemmon et al. 2009).

The harmful impact of fragmentation was previously observed by Hosner et al. (2016). Our results corroborate their observation. However, we propose a very different solution. Unlike their solution of removing the entire gene, which can lead to loss of otherwise useful signal and perhaps a non-random sample of genes (Huang et al. 2016), we propose removing specific problematic taxa. Importantly, we observed that trading off decreased taxon occupancy with decreased levels of fragmentation (type I vs. type II in the terminology of Hosner et al. [2016]) is beneficial, but only to a point; excessive filtering can also impact the accuracy of the species tree by creating missing data in gene trees. The amount of improvement depended on the number of genes, and filtering did not seem useful when only a small number of genes was available. The reductions in gene tree error were substantial (e.g., from 0.39 to 0.28); improvements in species tree topological accuracy may be considered small in magnitude (0.01–0.02 NRF), but we note that the error is reduced by a quarter of the original error, and that, these improvements come at no extra cost. We further note the improvements in species tree branch lengths.

Consistent with the literature (Hovmöller et al. 2013; Huang and Knowles 2016; Xi et al. 2016), our results indicate that summary methods are somewhat robust to missing data, but we also show that this robustness has limits as seen by Hosner et al. (2016). In the context of a single maximum likelihood analysis, Wiens (2006) observed that the absence of enough data, and not the presence of missing data per se, can cause inaccuracy. Importantly, we do not filter entire genes because they miss some taxa or because they have some fragmentary sequences. Prior research suggests aggressive filtering of entire genes with missing data can be harmful (Huang et al. 2016; Streicher et al. 2016). Our results do not conflict with those studies and our filtering approach is in fact motivated by their observations.

Finally, we observe that the number of genes has the strongest effect on the species tree error. Therefore, removing genes is not desirable. Instead, when possible, increasing the number of genes may improve the species tree topological accuracy even in the presence of fragments.

Filtering Threshold

The best choice of the filtering thresholds will always depend on the data set. However, our analyses suggest a possible way forward for systematists. Since ML tools such as FastTree can easily compute many hundreds of gene trees quickly and

relatively accurately, one can examine different thresholds empirically. By changing the threshold, reestimating gene trees, and computing occupancy (q), average gene tree support (p), and evolutionary diameter (d) (e.g., as in fig. 3*b* and *c*), analysts can look for thresholds that reduce occupancy minimally while increasing support or decreasing long branches substantially. This involves making trade-offs and the best way of making such trade-offs requires further analyses.

Some simple rule-of-thumbs could be designed, and supplementary table S3, Supplementary Material online gives several rules and applies them to the insect data sets. For example, we can simply use the threshold that maximizes pq/d. Another simple rule is using pq, which would emphasize occupancy and support equally. Based on the belief, backed by the literature, that that reduced occupancy is less damaging than high gene tree error, we can use p^2q to weight bootstrap support more than occupancy. Finally, one can pick the threshold that gives the highest p given that q is above a threshold (say, 70%). On the insect data set, these rules selected thresholds between 20% and 50% for RAxML gene trees (supplementary table S3, Supplementary Material online). Since the 50% threshold was chosen the most often, we chose to use it as our default threshold. On the simulated data set, for FastTree gene trees, where we could perform bootstrapping, several of our rules (e.g., p^2q and $\frac{pq}{1-p}$) tend to select thresholds between 20% and 50% (supplementary fig. S15, Supplementary Material online); these match the optimal thresholds for FastTree simulations (fig. 2).

Even though these rules seem to pick reasonable thresholds on our insect data, whether any of them performs well on a wide range of data sets remains unclear and require future studies. Moreover, filtering in general, and the use of bootstrap support in particular, could always add a bias, and thus, we do not suggest using thresholds that are varied from gene to gene.

Finally, if some relationships are judged very strong by prior evidence, detecting whether one recovers them may also prove useful, though this strategy should be used judiciously to avoid confirmation bias. In absence of extensive analysis, thresholds 25–50% seemed reasonable in our simulated and empirical analyses and may prove useful as a default for other analyses.

RAxML versus FastTree

Although the choice of the ML method for inferring gene trees was not the focus of our study, our simulation analyses showed a clear advantage in using RAxML versus FastTree. Moreover, on the biological data set, using RAxML (with automatic model selection) rather than FastTree (with a fixed model) led to further improvements in the species tree, recovering the monophyly of Neuropteroidea (node D) and Pterygota (node L), both of which have strong evidence from the literature. It is also interesting that the support for Neoptera (Node K) and the tentative sister-group relationship of Psocodea and Holometabola found by Misof et al.'s concatenation results (node P) were increased using RAxML gene trees. Overall, simulations and real data indicate that not only gene trees are less accurate when estimated using FastTree, but also, the ASTRAL species trees inferred from FastTree gene trees are less accurate than those inferred from RAxML gene trees. Interestingly, FastTree gene trees consistently have reduced branch lengths compared with RAxML trees on the biological data set (fig. 3*c*), perhaps because of FastTree's extensive use of the minimum evolution criteria (in addition to maximum likelihood). Finally, we note that FastTree does not allow for extensive model selection (for proteins), a fact that on biological (but not simulated) data sets could further contribute to its inaccuracies.

A previous independent simulation study by (Liu et al. 2011) had concluded that the two methods are essentially identical in terms of accuracy. The opposing conclusion drawn by (Liu et al. 2011) and our study may be related to simulation conditions. Our study considers conditions that include short branches prone to ILS but includes no alignment error; in contrast, (Liu et al. 2011) use data sets originally simulated to study alignment accuracy and include very divergent sequences (at least 50% average p-distance between sequences). To our knowledge, ours is the first simulation study to show that RAxML gene trees are more accurate, and we believe, the results should discourage analyses that rely solely on FastTree. Although many practitioners have perhaps already suspected that the much slower RAxML algorithm is more accurate under some conditions, the results shown here provide direct comparative evidence.

Despite the difference in accuracy, impacts of fragmentation had broadly similar patterns, regardless of the gene tree method used. Therefore, we believe for exploratory analyses of a data set, the use of FastTree is justified whereas final analyses used to infer the species tree are more reliable when based on the RAxML gene trees. Future work should test if using Bayesian methods for estimating gene trees would similarly improve the species tree accuracy.

Insect Phylogeny

As the debate between concatenation and summary methods pipelines continues (Springer and Gatesy 2014; Xi et al. 2014; Simmons and Gatesy 2015; Edwards et al. 2016), we note that Misof et al. (2014) had only used concatenation in their analyses. Our final ASTRAL tree using RAxML gene trees is highly congruent with the concatenation tree of Misof et al. (2014). This result has several implications for insect phylogeny. Overall, it supports nearly all of the results of their concatenation analysis with respect to the major events in insect evolution, using a different analysis strategy. Where their final tree showed relationships with weak support, generally ours did as well, indicating that some results may require further effort to resolve with confidence. It is likely that with the generation of phylogenomic data with lower fragmentation (e.g., using full genomes instead of transcriptomes), gene tree summary methods will be able to improve upon the results of both studies.

Methodological Limitations

It is important to note that even in the final RAxML gene trees, extensive gene tree discordance remains, and some of

the discordance is highly supported (fig. 5). The presence of highly supported discordance can in principle favor summary methods over concatenation. However, we note that our analyses that included fragments produced results that strongly conflicted with strong evidence from the prior literature. Thus, the choice is not only between concatenation and summary methods but more broadly about choosing data generation methods and tailoring the analysis pipeline to the data. Summary methods can only produce good results when provided with good gene trees and removing fragmentary data and other sources of the error are essential to that goal. Our paper demonstrated negative impacts of fragmentary data and suboptimal gene tree estimation methods. However, several other sources of error were not addressed.

Even after filtering, the proportion of genes that fail to recover major insect orders remains arguably high. It is likely that gene tree error persists even after filtering. One major cause of the remaining discordance is likely the lack of strong signal in gene trees. In addition to insufficient signal, our models of sequence evolution are likely violated in many ways, especially when we consider 400+ My of evolution, as we did here. Factors that include convergent effects of strong selection or unexpectedly high sensitivity to individual sites (Shen et al. 2017) may lead to systematic biases. Finally, even when the gene tree discordance is real, it may be due to factors other than ILS, including incorrect detection of orthology. Future work should explore improved scalable methods of dealing with these difficulties.

Materials and Methods

Analysis Pipeline for Insect Data Set

We used the amino acid sequence data provided by Misof et al. (2014) as "supplementary 7."

Filtering Strategy

In real data, gaps can appear for two reasons: insertions and deletions (as inferred by an alignment algorithm) and missing data. Our goal is to filter out sequences that are fragmentary (partially sequenced or assembled) but we don't wish to remove sequences due only to indels. Defining what is a fragmentary sequence is complicated by the presence of gappy sites. Also, very gappy sites increase running time but provide little signal to the maximum likelihood analyses, which treat them as missing data (and not as indel signal). To address both issues, before identifying fragmentary sequences, we first remove extremely gappy sites, defined as those with >90% gaps. Although this filtering can remove photogenically informative indels, we note that indels are not incorporated in models of sequence evolution used in our gene tree estimation tools. We then remove species that have <20% (1/5), 25% (1/4), 33% (1/3), 50% (1/2), 66% (2/3), 75% (3/4), or 80% (4/5) amino-acids (i.e., characters other than gaps). In order to filter sequences, we use a tool called seqtools, implemented as part of the PASTA (Mirarab et al. 2014b) package. After filtering sequence, we reestimate gene trees, but we keep the same alignment. In order to track the occupancy and bootstrap support, we use in-house scripts, available online https://github.com/esayyari/discoVista.

Gene Trees and Species Trees

After each round of filtering, gene trees are estimated using FastTree2 (Price et al. 2010) using its default amino acid substitution model, which is JTT (Jones et al. 1992) or RAxML (Stamatakis 2014) with the automatic amino acid model selection.

To infer our bestML gene trees, we use RAxML (Stamatakis 2014), version 8.2.9 with ten runs of inference using different starting trees. Unlike FastTree, RAxML implements many protein substitution models and it can find the best scoring protein-coding substitution model (Price et al. 2010; Stamatakis 2014). We used RAxML's automatic model selection approach; numbers of genes with various models are shown in supplementary table S4, Supplementary Material online. When several species have identical sequences for a gene, we keep only one of them (i.e., remove redundant ones) in our RAxML runs and add the removed species back to the final inferred gene tree as a polytomy.

For performing gene tree bootstrapping using FastTree, we first generate bootstrap sequences using RAxML and then run FastTree on those to estimate the bootstrapped gene trees. We then draw those bootstrap gene trees on ML gene tree branches using the newick utility (Junier and Zdobnov 2010). For RAxML gene trees, we use the rapid bootstrapping option on reduced sequences (after removing identical sequences). After gene tree estimations, we add back the identical species and draw these bootstrap gene trees on the best ML gene trees (RAxML) following the same procedure using the newick utility.

We use ASTRAL-II to estimate the species trees summarizing gene trees with at least four taxa left after filtering.

Simulation Procedure

We use one model condition of a previously simulated data set from Mirarab and Warnow (2015) with 100 ingroup taxa and one outgroup. For each of the 50 replicates in this data set, Simphy (Mallo et al. 2016) was used to simulate a species tree according to the Yule model, and then 1,000 gene trees were simulated using the MSC model which captures ILS. The data set has moderate levels of ILS; the average distance between true gene trees and true species trees is 0.33. We subsampled genes to create three different data sets with 50, 200, or 1,000 genes. DNA sequences of varying length (supplementary fig. S16, Supplementary Material online) were simulated down the gene trees using Indelible (Fletcher and Yang 2009) with GTR parameters and stationary distributions estimated from published biological data sets, as detailed by Mirarab and Warnow (2015). Note that simulated sequences did not include any indels and thus were already aligned. Mirarab and Warnow (2015) suggested removing two replicates that include almost no phylogenetic signal, and we use the same strategy, leaving us with 48 replicates. This creates our unfiltered base data set.

Adding Fragmentation

We add fragmentation to our complete simulated data set using a procedure that seeks to emulate patterns of fragmentation in the insect biological data set. 1) For each replicate, we order species in the biological data set and the simulated data set with respect to the tip-to-root distances. 2) We randomly select 100 of the biological species and map them to the simulated species with the same position in the order. The main outgroup (*Ixodes scapularis*) in the biological and simulated data sets always map to each other. 3) For each replicate in the simulated data set, we randomly sample (with replacement) 1,000 genes in the insect data sets that have at least 101 species, including the main outgroup. 4) For each species in each simulated gene, we compute the portion of gap sites in the corresponding gene alignment for the corresponding species in the biological data, and remove the same portion of sites in the simulated data set at random positions. When a species is missing from a gene in the biological data set, we use the same species from another randomly chosen gene.

Filtering Fragments

Although our simulated data do not include indels, injected fragments can create sites that are almost entirely gaps; these sites increase running time but include minimal signal. We, therefore, remove sites with >90% gaps, removing between 0.0% and 2.0% (median: 0.1%) of the total number of characters in all sequences. We then remove from each gene any species that has less than a certain fraction (e.g., 10% – 80%) of the full gene. For example, at 10%, we remove only sequences that have 90% or more gaps.

Gene Trees and Species Trees

For each threshold, after filtering, we estimate gene trees using both RAxML (Stamatakis 2014) version 8.2.9 with two starting trees and FastTree (Price et al. 2010) version 2.1.9 Double precision using the GTR + Γ model of sequence evolution (Tavaré 1986). We infer the species tree using ASTRAL-II (Mirarab and Warnow 2015) version 4.11.1, which is a commonly used summary method. We build species trees using all 1,000 genes or using randomly chosen subsets of 200 or 50 genes.

Statistical Tests

All *P* values reported are computed using the Analysis of variance (ANOVA) tests. For impacts on gene trees, we use the gene tree method (RAxML vs. FastTree) and filtering thresholds as independent variables. For species tree, to study the impact of presence/absence of fragments, we only include species trees of *orig-seq* and *no-filtering*, and use a binary variable to encode it and use another variable for the number of genes. To study the impact of filtering with sufficiently small thresholds, we restrict the data to those with up to 33% filtering and we use the filtering threshold as a numerical independent variable.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank Bernhard Misof, Karl Kjer, and Karen Meusemann for providing the insect data set and additional explanations as to how it was constructed. This work was supported by the National Science Foundation grant IIS-1565862 to E.S. and S.M. Computations were performed on the San Diego Supercomputer Center (SDSC) through XSEDE allocations, which is supported by the NSF grant ACI-1053575. J.B.W. would like to thank the Center for Advanced Studies, University of Illinois, for funding release time.

References

Bayzid MS, Warnow T. 2012. Estimating optimal species trees from incomplete gene trees under deep coalescence. *J Comput Biol*. 19(6): 591–605.

Bayzid MS, Mirarab S, Boussau B, Warnow T, Cellinese N. 2015. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLoS One* 10(6): e0129183.

Beutel RG, Kristensen NP, Pohl H. 2009. Resolving insect phylogeny: the significance of cephalic structures of the nannomecoptera in understanding endopterygote relationships. *Arthropod Struct Dev*. 38(5): 427–460.

Beutel RG, Friedrich F, Hörnschemeyer T, Pohl H, Hünefeld F, Beckmann F, Meier R, Misof B, Whiting MF, Vilhelmsen L. 2011. Morphological and molecular evidence converge upon a robust phylogeny of the megadiverse holometabola. *Cladistics* 27(4): 341–355.

Beutel RG, Friedrich F, Yang X-K, Ge S-Q. 2014. Insect morphology and phylogeny: a textbook for students of entomology. Berlin: Walter de Gruyter.

Bitsch C, Bitsch J. 2004. Phylogenetic relationships of basal hexapods among the mandibulate arthropods: a cladistic analysis based on comparative morphological characters. *Zool Scripta* 33(6): 511–550.

Boussau B, Walton Z, Delgado JA, Collantes F, Beani L, Stewart IJ, Cameron SA, Whitfield JB, Johnston JS, Holland PWH, et al. 2014. Strepsiptera, phylogenomics and the long branch attraction problem. *PLoS One* 9(10): e107709.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, Roychoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol*. 29(8): 1917–1932.

Cameron SL, Barker SC, Whiting MF. 2006. Mitochondrial genomics and the new insect order mantophasmatodea. *Mol Phylogenet Evol*. 38(1): 274–279.

Chifman J, Kubatko LS. 2014. Quartet inference from snp data under the coalescent model. *Bioinformatics* 30(23): 3317–3324.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*. 24(6): 332–340.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63(1): 1–19.

Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, Zhong B, Wu S, Lemmon EM, Lemmon AR, et al. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Mol Phylogenet Evol*. 94: 447–462.

Engel MS, Grimaldi DA. 2004. New light shed on the oldest insect. *Nature* 427(6975): 627–630.

Fletcher W, Yang Z. 2009. Indelible: a flexible simulator of biological sequence evolution. *Mol Biol. Evol*. 26(8): 1879–1888.

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol Phylogenet Evol*. 80: 231–266.
Grimaldi D, Engel MS. 2005. Evolution of the insects. Cambridge: Cambridge University Press.
Grimaldi DA. 2010. 400 million years on six legs: on the origin and early evolution of hexapoda. *Arthropod Struct Dev*. 39(2): 191–203.
Hasenfuss I. 2002. A possible evolutionary pathway to insect flight starting from lepismatid organization. *J Zool Syst Evol Res*. 40(2): 65–81.
Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol*. 27(3): 570–580.
Hennig W, Hennig W, Hennig W, Zoologist G, Hennig W. 1969. Die stammesgeschichte der insekten. Germany: Krammer, Frankfurt am Main.
Heraty J, Ronquist F, Carpenter JM, Hawks D, Schulmeister S, Dowling AP, Murray D, Munro J, Wheeler WC, Schiff N, Sharkey M. 2011. Evolution of the hymenopteran megaradiation. *Mol Phylogenet Evol*. 60(1): 73–88.
Hosner PA, Faircloth BC, Glenn TC, Braun EL, Kimball RT. 2016. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (aves: Galliformes). *Mol Biol Evol*. 33(4): 1110–1125.
Hovmöller R, Knowles LL, Kubatko LS. 2013. Effects of missing data on species tree estimation under the coalescent. *Mol Phylogenet Evol*. 69(3): 1057–1062.
Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-Shehbaz I, Edger PP, Pires JC, Tan D-Y. 2016. Resolution of brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol Biol Evol*. 33(2): 394–412.
Huang H, Knowles LL. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of rad sequences. *Syst Biol*. 65(3): 357–365.
Ishiwata K, Sasaki G, Ogawa J, Miyata T, Su Z-H. 2011. Phylogenetic relationships among insect orders based on three nuclear protein-coding gene sequences. *Mol Phylogenet Evol*. 58(2): 169–180.
Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215): 1320–1331.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* 8(3): 275–282.
Junier T, Zdobnov EM. 2010. The newick utilities: high-throughput phylogenetic tree processing in the unix shell. *Bioinformatics* 26(13): 1669.
Kjer KM, Carle FL, Litman J, Ware J. 2006. A molecular phylogeny of hexapoda. *Arthropod Syst Phylogeny* 64(1): 35–44.
Klass K-D. 2009. A critical review of current data and hypotheses on hexapod phylogeny. *Proc Arthropod Embryol Soc Jpn* 43:3–22.
Kristensen NP. 1975. The phylogeny of hexapod orders. A critical review of recent accounts. *Journal of Zoological Systematics and Evolutionary Research*. 13(1): 1–44.
Kristensen NP. 1991. Phylogeny of extant hexapods. Melbourne: Melbourne University Publishing. p. 126–140.
Kristensen NP. 1999. Phylogeny of endopterygote insects, the most successful lineage of living organisms. *Eur J Entomol*. 96(3): 237–253.
Kristensen NP, Scoble MJ, Karsholt OLE. 2007. Lepidoptera phylogeny and systematics: the state of inventorying moth and butterfly diversity. *Zootaxa* 1668(699): e747.
Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. *Systc Biol*. 58(1): 130–145.
Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol Evol Syst*. 44(1): 99–121.
Letsch H, Simon S. 2013. Insect phylogenomics: new insights on the relationships of lower neopteran orders (polyneoptera). *Syst Entomol*. 38(4): 783–793.
Liu K, Linder CR, Warnow T, Wu R. 2011. Raxml and fasttree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One* 6(11): e27731.
Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 24(21): 2542–2543.
Liu L, Yu L. 2011. Estimating species trees from unrooted gene trees. *Syst Biol*. 60(5): 661–667.
Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. *Syst Biol*. 58(5): 468–477.
Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol*. 10(1): 302.
Liu L, Xi Z, Davis CC. 2015. Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. *Mol Biol Evol*. 32(3): 791–805.
Maddison WP, Wiens JJ. 1997. Gene trees in species trees. *Syst Biol*. 46(3): 523–536.
Mallo D, De Oliveira Martins L, Posada D. 2016. SimPhy: phylogenomic simulation of gene, locus and species trees. *Syst Biol*. 65(2): syv082.
McKenna DD, Farrell BD. 2010. 9-genes reinforce the phylogeny of holometabola and yield alternate views on the phylogenetic placement of strepsiptera. *PLoS One* 5(7): e11887.
Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kück P, Ebersberger I, Walzl M, Pass G, Breuers S, et al. 2010. A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol*. 27(11): 2451–2464.
Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31(12): i44–i52.
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014a. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30(17): i541–i548.
Mirarab S, Nguyen N, Warnow T. 2014b. PASTA: ultra-large multiple sequence alignment. *Res Comput Mol Biol*. 22(5):177–191.
Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014c. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346(6215): 1250463–1250463.
Mirarab S, Bayzid MS, Warnow T. 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst Biol*. 65(3): 366–380.
Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A. 2007. Towards an 18s phylogeny of hexapods: accounting for group-specific character covariance in optimized mixed nucleotide/doublet models. *Zoology* 110(5): 409–429.
Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, Niehuis O, Petersen M. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346(6210): 763–767.
Nagarajan N, Pop M. 2013. Sequence assembly demystified. *Nat Rev Genet*. 14(3): 157–167.
Nguyen N, Mirarab S, Kumar K, Warnow T. 2015. Ultra-large alignments using phylogeny-aware profiles. *Genome Biol*. 16(1): 124.
Niehuis O, Hartig G, Grath S, Pohl H, Lehmann J, Tafer H, Donath A, Krauss V, Eisenhardt C, Hertel J, et al. 2012. Genomic and morphological evidence converge to resolve the enigma of strepsiptera. *Curr Biol*. 22(14): 1309–1313.
Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol*. 5(5): 568–583.
Patel S, Kimball R, Braun E. 2013. Error in phylogenetic estimation for bushes in the tree of life. *Phylogenet Evol Biol*. 01(2): 2.
Philippe H, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol*. 21(9): 1740–1752.
Price MN, Dehal PS, Arkin AP. 2010. FastTree-2 approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3): e9490.
Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (aves) using targeted next-generationDNAsequencing. *Nature* 526(7574): 569–573.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes usingDNAsequences from multiple loci. *Genetics* 164(4): 1645–1656.
Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463(7284): 1079–1083.
Richards S, Murali SC. 2015. Best practices in insect genome sequencing: what works and what doesn't. *Curr Opin Insect Sci*. 7: 1–7.
Roch S, Steel M. 2014. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor Popul Biol*. 100: 56–62.
Roch S, Warnow T. 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Syst Biol*. 64(4): 663–676.
Rouse GW, Wilson NG, Carvajal JI, Vrijenhoek RC. 2016. New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. *Nature* 530(7588): 94–97.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449): 327–331.
Savard J, Tautz D, Richards S, Weinstock GM, Gibbs RA, Werren JH, Tettelin H, Lercher MJ. 2006. Phylogenomic analysis reveals bees and wasps (hymenoptera) at the base of the radiation of holometabolous insects. *Genome Res*. 16(11): 1334–1338.
Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol Biol Evol*. 33(7): 1654–1668.
Shen XX, Hittinger CT, Rokas A. 2017. Studies can be driven by a handful of genes. *Nature* 1(April): 1–10.
Simmons MP. 2012. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. *Cladistics* 28(2): 208–222.
Simmons MP. 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian mcmc phylogenetic analyses. *Mol Phylogenet Evol*. 80(1): 267–280.
Simmons MP, Gatesy J. 2015. Coalescence vs. concatenation: sophisticated analyses vs. first principles applied to rooting the angiosperms. *Mol Phylogenet Evol*. 91: 98–122.
Springer MS, Gatesy J. 2014. Land plant origins and coalescence confusion. *Trends Plant Sci*. 19(5): 267–269.
Springer MS, Gatesy J. 2016. The gene tree delusion. *Mol Phylogenet. Evol*. 94(Part A): 1–33.
Stamatakis A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9): 1312.
Streicher JW, Schulte JA, Wiens JJ. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? an empirical study in iguanian lizards. *Syst Biol*. 65(1): 128–145.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis ofDNAsequences. *Lect Math Life Sci*. 17: 57–86.
Terry MD, Whiting MF. 2005. Mantophasmatodea and phylogeny of the lower neopterous insects. *Cladistics* 21(3): 240–257.
Trautwein MD, Wiegmann BM, Beutel R, Kjer KM, Yeates DK. 2012. Advances in insect phylogeny at the dawn of the postgenomic era. *Annu Rev Entomol*. 57(1): 449–468.
Vachaspati P, Warnow T. 2015. Astrid: accurate species trees from internode distances. *BMC Genomics* 16(Suppl 10): S3.
Wheat CW, Wahlberg N. 2013. Phylogenomic insights into the cambrian explosion, the colonization of land and the evolution of flight in arthropoda. *Syst Biol*. 62(1): 93–109.
Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter EJ, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A*. 111(45): 4859–4868.
Wiegmann BM, Trautwein MD, Kim J-W, Cassel BK, Bertone MA, Winterton SL, Yeates DK. 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC Biol*. 7: 34.
Wiens JJ. 2006. Missing data and the design of phylogenetic analyses. *J Biomed Inform*. 39(1):34–42.
Wiens JJ, Morrill MC. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst Biol*. 60(5): 719–731.
Wipfler B, Machida R, Müller B,Beutel RG. 2011. On the head morphology of grylloblattodea (insecta) and the systematic position of the order, with a new nomenclature for the head muscles of dicondylia. *Syst Entomol*. 36(2): 241–266.
Xi Z, Liu L, Rest JS, Davis CC. 2014. Coalescent versus concatenation methods and the placement of amborella as sister to water lilies. *Syst Biol*. 63(6): 919–932.
Xi Z, Liu L, Davis CC. 2016. The impact of missing data on species tree estimation. *Mol Biol Evol*. 33(3): 838–860.
Xu B, Yang Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics* 204(4): 1353–1368.
Yoshizawa K. 2011. Monophyletic polyneoptera recovered by wing base structure. *Syst Entomol*. 36(3): 377–394.
Zimmermann T, Mirarab S, Warnow T. 2014. Bbca: improving the scalability of *beast using random binning. *BMC Genomics* 15(Suppl 6): S11.