Intragenic Conflict in Phylogenomic Data Sets

Stephen A. Smith,*[†,1] Nathanael Walker-Hale,[†,2] and Joseph F. Walker[†,3]

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI
[2]Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom
[3]The Sainsbury Laboratory (SLCU), University of Cambridge, Cambridge, United Kingdom
[†]These authors contributed equally to this work.
***Corresponding author:** E-mail: eebsmith@umich.edu.
Associate editor: Keith Crandall

Abstract

Most phylogenetic analyses assume that a single evolutionary history underlies one gene. However, both biological processes and errors can cause intragenic conflict. The extent to which this conflict is present in empirical data sets is not well documented, but if common, could have far-reaching implications for phylogenetic analyses. We examined several large phylogenomic data sets from diverse taxa using a fast and simple method to identify well-supported intragenic conflict. We found conflict to be highly variable between data sets, from 1% to >92% of genes investigated. We analyzed four exemplar genes in detail and analyzed simulated data under several scenarios. Our results suggest that alignment error may be one major source of conflict, but other conflicts remain unexplained and may represent biological signal or other errors. Whether as part of data analysis pipelines or to explore biologically processes, analyses of within-gene phylogenetic signal should become common.

Key words: **intragenic conflict, phylogenomics, recombination.**

Letter

The sequencing and analysis of whole genomes, transcriptomes, and thousands of individual genes has illustrated that, throughout the tree of life, genomes are a composite of evolutionary histories. This heterogeneity is a source of biological insight (Mendes et al. 2019) as well as a source of computational and analytical complexity (Kosakovsky Pond et al. 2006a, 2006b; Boussau et al. 2013; Smith et al. 2015). Though heterogeneity is found across the genome, researchers have primarily focused on conflict among trees inferred using individual genes, in many cases limited to combined exons, and the inferred species tree.

Driven primarily by an interest in identifying recombination break points, over the past two decades, researchers have examined some sources of heterogeneous topological signal within single-gene alignments. To facilitate this, several methods have been developed (e.g., Salminen et al. 1995; Husmeier and McGuire 2003; Kosakovsky Pond et al. 2006a, 2006b; Inagaki et al. 2006; Hobolth et al. 2007; Boussau et al. 2009; Suchard et al. 2002; Ané 2011; Allman et al. 2017). Recombination plays a large role in population genetics. However, at the species level, the impact of recombination on phylogenetic inference remains debated (Edwards 2009; Lanier and Knowles, 2012; Wu et al. 2013). Scornavacca and Galtier (2017) showed that exons within the same gene may present different genealogies, a pattern confirmed by Mendes et al. (2019). Despite this proliferation of methods, many are not tractable for phylogenomic data sets and/or assume the conflict arose from a biological process.

Although biological processes can introduce phylogenetic heterogeneity within gene sequence alignments, given the data set size and automated nature of genomic analyses, systematic error may also contribute to intragenic phylogenetic conflict in empirical data sets. These alignment and/or assembly errors have become more common as data sets have increased in both taxon sampling and regions of the genome analyzed. The volume of data has made automation a requirement of any phylogenomic pipeline. Inevitably, errors make their way into alignments and, if not filtered, can lead to phylogenetic conflict and downstream errors (Song et al. 2012; Gatesy and Springer 2013; Brown et al. 2017; Walker et al. 2018).

Regardless of the source, incorrectly modeling intragenic conflict will result in inaccurate phylogenies, biased branch lengths, and erroneous selection analyses, among other errors. Importantly, how common mixed signal is within genes is still unknown. Whether due to computationally taxing methods or because researchers assume that errors will be overwhelmed by the signal from hundreds of other genes (but see Brown et al. 2017; Shen et al. 2017; Walker et al 2018), analyses of intragenic conflict are not common. Here, instead of modeling the biological processes that generate conflicting signal (e.g., recombination), we explore a rapid procedure that allows users to assess violations of the underlying assumptions of phylogenetic reconstruction (e.g., mixed signal). We evaluate this approach using simulated data to help isolate and identify key aspects of phylogenetic analyses that contribute to conflict within genes. We examine several empirical data sets using a sliding-window approach and characterize the extent to which single gene regions show evidence for conflicting histories across a broad range of phylogenetic data sets.

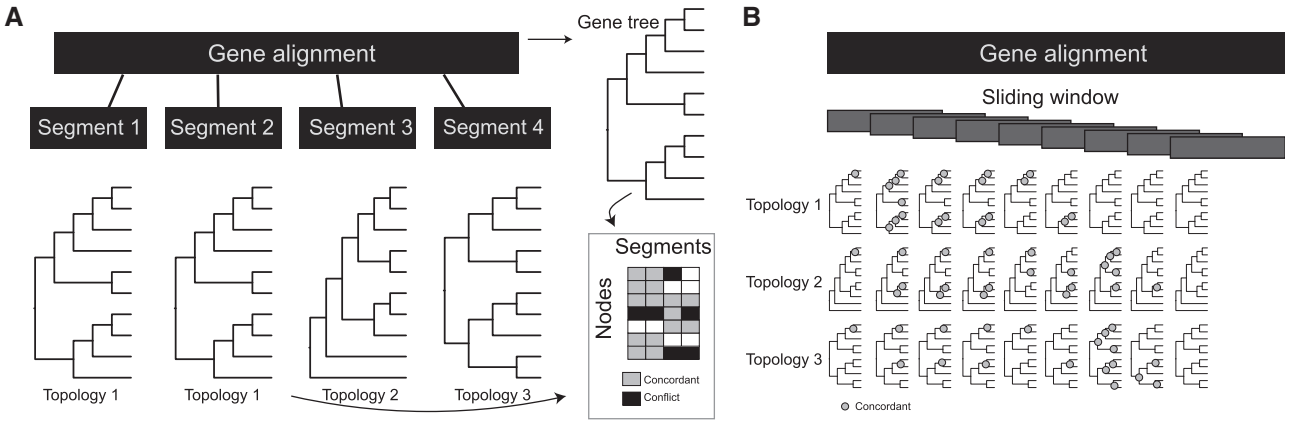

FIG 1. (A) Analysis of gene segments and comparisons between gene segment trees and the gene tree recorded in the table. Light boxes show no support in the gene segment for the node in the gene tree. Shaded boxes show support in the gene segment for the gene tree node. Dark boxes exhibit conflict between the gene tree and gene segment. These are nonoverlapping segments. There are three unique topologies found with segments 1 and 2 displaying the same topology. (*B*) Sliding window analysis for a more detailed look at where clades display well-supported conflict or concordance (shown here) between the gene tree and the segment trees. Support for clades in the three unique topologies are displayed across the sliding-window lengths. The sliding window has overlapping portions.

Materials and Methods

Empirical Data Sets

We gathered a broad sampling of nucleotide data sets across the tree of life. Many of these include a mix of exons, introns, and other genetic elements. For simplicity, when we refer to gene, we mean a locus or set of genetic elements but not necessarily a complete or protein-coding gene. This conforms to typical naming conventions such as "gene tree," which may or may not refer to the tree of a coding sequence or complete gene. The data sets examined here included those designed to analyze contentious relationships broadly across mammals: MAM1 consisting of 10,259 genes (Chen et al. 2017); MAM2 consisting of 424 genes (Song et al. 2012 as refined by Mirarab et al. 2014); and MAM3 consisting of 183 ultra-conserved element (UCE) loci (McCormack et al. 2012). We analyzed an insect data set (BUGS) focused on analyzing relationships in the Strepsiptera consisting of 4,485 genes (Niehuis et al. 2012). We examined three vertebrate data sets: VERT, consisting of 1,113 genes (Wang et al. 2013); FISH, a data set assembled to understand the relationships among ray finned fish, consisting of 1,105 genes (Hughes et al. 2018); and FROG, an data set of frogs, consisting of 95 genes (Feng et al. 2017). We analyzed three plant data sets: PLAN, a data set generated to investigate land plant evolution, consisting of 852 genes (Wickett et al. 2014); MOS (Mitochondria [M], Nuclear [N], Plastome[P]), a data set generated to analyze moss ordinal relationships, consisting of three data sets of 40 (MOSM), 105 (MOSN), and 82 (MOSP) genes (Liu et al. 2019); and CARN, a carnivorous plant data set, that consisted of 1,237 genes (Walker et al. 2017). Finally, we gathered a fungal data set (FUNG) consisting of 2,256 genes (Pizarro et al. 2018). As a result of missing data and short alignment lengths that prohibited division into at least two 1,000-bp segments (or 500 bp for data sets with very small alignments), the number of genes analyzed was often lower than the number from the original study (table 2).

Identification of Multiple Trees within a Gene

For each data set, we examined the conflicting signal inferred from segments of each gene. For most data sets we used 1,000 bp for the segment length to increase phylogenetic signal per segment and reduce gap only taxa within segments. However, for data sets with generally shorter alignments including FISH, FROGS, and MOSS, we considered a segment length of 500 bp. Phylogenetic trees were calculated for the entire gene and for each gene segment using IQ-TREE (Nguyen et al. 2015), the GTR + Γ model of molecular evolution, 1,000 ultrafast bootstrap (UFB) replicates (Hoang et al. 2018), and SH-aLRT (Guindon et al 2010) analyses. Although the UFB has been shown to be conservative with at a cutoff of 95% (Hoang et al. 2018), with any relationship whose support value is below that is inferred to have low support, our initial analyses demonstrated that this still generated support (BS $\geq$ 95%) for some poorly supported relationships. To be more conservative, we therefore calculated both SH-aLRT (with a cutoff of 80% used) and UFB. We then compared the segment trees with the maximum likelihood (ML) tree for the entire gene and gene segments found to contain strong-supported conflicting signal (UFB $\geq$ 95% and SH-aLRT $\geq$ 80%), for any given relationship were extracted to be examined using the sliding-window approach described below (fig. 1A). These analyses were conducted with the python program phynd (https://github.com/FePhyFoFum/phynd).

Simulations

We tested the sensitivity of the approach described above for type I error rate under a variety of scenarios and for positive identification of conflict when present. To do this, we simulated several alignments and analyzed them using our procedure. Each simulation was replicated 100 times.

First, to test for false-positive identification under no topological conflict, we simulated 25-taxon trees using pxbdsim with a birth rate of 1 and a death rate of 0. We then scaled the tree to a root height of 0.75 and simulated 3000 bp under JC

with INDELible (Fletcher and Yang 2009) on the same tree. We conducted a similar simulation to test the influence of taxon sampling with 50-taxon trees and a root height of 1, and another simulation to test the influence of gene length with 25 taxa and 1,500 bp. To increase model complexity and test for the impact of differing molecular models, we conducted a simulation with 25 taxa, 3,000 bp and different GTR models for first 2,000 bp and final 1,000 bp (0.6, 0.4, 0.2, 0.8, 1.2 with state freqs 0.3, 0.4, 0.1, 0.2 vs. 0.2, 0.4, 0.6, 0.8, 1.2 with state freqs 0.1, 0.2, 0.3, 0.4), simulated on the same tree. Finally, to generalize this, we conducted a simulation with 25 taxa, 3,000 bp and different, randomly parameterized GTR models for each 1,000-bp segment. We generated GTR parameters by random draws from an exponential distribution with scale = 1, with base frequencies drawn randomly from a uniform distribution and scaled to sum to 1.

For methods that use information criterion to determine the existence and location of breakpoints (Kosakovsky Pond et al. 2006a, 2006b), sectional rate shifts may mislead inference because fit will be improved by separate models, which nonetheless correspond to the same topology. We therefore conducted simulations to test the sensitivity of our approach to sectional rate shifts. We generated 25 taxon trees and 3,000 bp in 1,000-bp segments, where each segment received a different randomly parameterized GTR model as above. Each segment was simulated on the same tree but scaled such that the root height was 0.5, 0.75 and 1.0, respectively.

Alignment error could induce false-positives. So, in addition to testing our method on simulated alignments, we also simulated alignments under a variety of indel parameters and realigned using MAFFT with defaults (Katoh and Standley 2013), FSA with defaults (Bradley et al. 2009) and PRANK (-iterate = 5) (Löytynoja and Goldman 2005). We conducted three sets of simulations using a negative binomial indel model following the geometric distribution with the number of successes ($r = 1$) and the probability of success ($P = 0.25$) and differing insertion and deletion rates. In each case, we generated a 25-taxon tree as above, and then simulated three 1,000-bp segments under a randomly parameterized GTR model, alongside the specified indel model. In the first set of simulations, the insertion and deletion rate per site were equal, with 0.03, 0.01, and 0.005, respectively. In the second two sets of simulations, the insertion and deletion rate differed. In the first simulation, the first segment had insertion = 0.03 and deletion = 0.04, the second 0.02 and 0.01, and the third 0.003 and 0.006. In the second simulation, the first segment had insertion = 0.04 and deletion = 0.03, the second 0.01 and 0.02, and the third 0.006 and 0.003.

We conducted simulations with conflict to test the efficacy of our method. In each, we simulated alignments under a single randomly parameterized GTR, but one segment was simulated on a separate conflicting 25-taxon tree. We conducted four simulations. In the first, we simulated 3,000 bp where the last 200 bp were simulated under a conflicting topology. In the second, we simulated 3,000 bp where the last 500 bp were simulated under a conflicting topology. In the third, we simulated 3,000 bp where the last 1,000 bp were simulated under a conflicting topology. Finally, we simulated 1,500 bp where the last 200 bp were simulated under a conflicting topology.

Specific Examples

To examine the patterns of conflict within genes in more detail, we identified four examples where the conflict identified was not a result of obvious errors (i.e., the examples are not representative of all the inferred conflicts). For each gene examined in more detail, we conducted several additional analyses. First, we calculated site-specific *lnL* for each tree constructed from the 1,000-bp segments, called the segment trees. These values were then compared with site-specific *lnL* of the maximum-likelihood tree constructed from the entire data set. The site-specific *lnL* calculates the likelihood each site has for an ML topology (Castoe et al. 2009), the likelihood of each site can then be compared between multiple topologies to identify the degree of which the likelihood supports one topology over another (Delta SS *lnL*). By performing this analysis, we were able to examine the degree to which each site supports the segment trees vs. the ML tree. This allowed us to both quantify the degree of significance a site has for an ML tree and whether the regions of genes show bias in their support across topologies. We also conducted sliding-window analyses summing the site-specific *lnL* of the segment trees for 100-bp windows every 20 bp (fig. 1*B*). This analysis was performed to help determine the significance of the conflict. For each window, we calculated the difference between the maximum and the minimum *lnL* of the window and if the difference was <0.05, we considered the window to be uninformative. Otherwise, the segment trees that were within 2 *lnL* of the maximum *lnL* were recorded for the window and concordant edges were summarized using *bp* from *gophy* and reported for each window (Edwards 1984). We considered these edges to be supported by the window. Finally, we also calculated sliding-window analyses of the base composition to determine if there were any biases in the alignments.

To compare our analyses with previously published methods, we also assessed these specific examples with GARD (Kosakovsky Pond et al. 2006a, 2006b) and phyML_multi (Boussau et al. 2009). GARD fits individual Neighbor-Joining trees to subsections of the alignment and compares the AICc (AIC with correction for small sample size) of a model allowing separate trees to that with a single tree. It then implements a genetic algorithm to search for extra breakpoints, using the same test. Finally, GARD compares the AICc of the model allowing separate trees for each subsection to the model fitting the same tree but independent branch lengths. phyML_multi uses a mixed model or phylo-HMM approach to calculate the likelihood of the alignment over multiple topologies, each inferred from a distinct segment of the alignment. It then uses the Viterbi and Forward-Backward algorithms to assign breakpoints along the alignment. For GARD, we used GTR + Γ and repeated each run in triplicate. Because the outgroup in the respective studies was rarely monophyletic for inferred segment trees, we arbitrarily selected one member of the outgroup to root each tree. Most runs were conducted in HYPHY v2.5.5, but some could not complete due to errors in likelihood calculation, and these were instead

conducted in HYPHY v2.5.8. In all analyses, the final AICc for the separate tree model was much lower than the alternative, suggesting that at least one breakpoint reflected a true incongruence. For phyML_multi, we analyzed each gene using TN93+ Γ, because GTR + Γ was not available. Following our segment trees (above), we optimized three trees for EOG2711, 2798 and ENSG00000074803, and two trees for 131. We ran phyML_multi using both the mixed-model and HMM approach, but due to issues with the python library used to process results only mixed-model results are presented here. To further explore the possible impact of alignment error in these specific examples, we realigned each gene with FSA (–nucprot) and PRANK (-iterate = 5-translate) and analyzed with GARD and phyML_multi as above. For phyML_multi, we optimized two or three trees as above for comparative purposes, even if the alignment was longer. Some phyML_multi analyses and analyses using the detailed sliding-window method (in fig 2) failed due to the significant addition of gaps in the altered alignment.

Results

Simulation Results

Simulations results are presented in table 1. Our approach had little to no false-positives in simple simulations where no conflict is present within genes. This included the situation where data were generated under two different GTR models or under three randomly parameterized GTR models. Our approach also produced no false-positives in the presence of rate shifts on the same tree within a gene. Incorporating indels in the simulation process also led to few false-positives. However, realignment of sequences simulated under indel models increased false-positive rates. MAFFT and PRANK realignments showed consistently more false-positive results with FSA realignments showing fewer false-positives. Simulations with conflict indicated that our method was sensitive in cases where conflicting signal made up a large proportion of the overall alignment, with 98% true-positives in the case where 1,000 bp were generated under a conflicting tree in a 3,000-bp alignment. Our method was less sensitive when conflicting signal made up a smaller proportion of the alignment, with only 19.5% true-positives in the case where 200 bp were generated under a conflicting tree in a 3,000-bp alignment, but 51% true-positives in the case where 200 bp were generated under conflicting tree in a 1,500-bp alignment. Generally, this demonstrated that this method was effective in identifying conflict when present and with a very low false-positive rate.

Empirical Data Sets

We analyzed 13 data sets for intragenic conflict and found variable results regarding the proportion of those genes with conflict (table 2). We noted that as taxon sampling increased, so did the inferred conflict. This is, in part, expected because as the number of taxa increases, complexity due to potential errors or biology may be assumed to increase. Nevertheless, this should be explored further. Several data sets consisted of genes that could not be analyzed due to short alignment length. We analyzed four genes in more detail to better document patterns that are not the result of obvious data assembly errors.

Invertebrate Example

We analyzed the EOG2711 gene in the BUGS data set (fig. 2A). This gene did not exhibit the pattern using the more conservative tests presented in table 2. Although several relationships disagree between the segment trees and the ML tree, the primary difference involves the placement of *Bombyx* and relatives. In the ML tree, *Bombyx* is placed in a clade with *Aedes*, *Culex*, and *Drosophila*. This is consistent with the second segment tree but conflicts with the first and third that have *Bombyx* sister to *Pediculus* and sister to *Acyrthosiphon*, respectively. In comparison, GARD inferred between 9 and 15 breakpoints. The different segment trees represent a diversity of topologies, with several conflicting in the placement of *Bombyx* and *Acyrthosiphon* among others (supplementary figs. S1–S3, Supplementary Material online). phyML_multi had optimal support for three trees over 86 segments (supplementary fig. S4, Supplementary Material online). Most segments supported a tree with a placement of *Bombyx* sister to a clade of *Aedes*, *Culex*, and *Drosophila*, whereas the other two trees placed *Bombyx* sister to *Drosophila* and differed in the placement of *Harpegnathos* and *Trilobium*, among others (supplementary fig. S4, Supplementary Material online).

Mammal Examples

We analyzed the ENSG00000074803 gene in the MAM1 data set (fig. 2B). The primary conflicts involved the placement of *Mus*, *Canis*, *Sorex*, and *Erinaceus*. The ML tree for the entire gene placed *Mus* sister to a clade of seven taxa, *Sorex* sister to *Erinaceus*, and *Canis* sister to *Ailuropda* and *Mustela*. However, the first 1,000 bp place *Mus* sister to *Canis* and the second 1,000 bp place *Sorex* sister to *Mus*. Over three runs, GARD inferred between five and eight breakpoints, with a diversity of topologies (supplementary figs. S5–S7, Supplementary Material online). Several conflicted in the placement of *Mus*, *Sorex*, *Erinaceus*, *Canis*, among others. phyML_multi had optimal support for only two of three trees over three segments, with the majority supporting a tree that placed *Erinaceus* sister to *Sorex* and *Mus* sister to *Echinops* (supplementary fig. S8, Supplementary Material online). A small number of sites supported a tree placing *Mus* sister to *Canis* and *Erinaceus* sister to *Echinops* (supplementary fig. S8, Supplementary Material online).

We analyzed the 131 gene in the MAM2 data set (fig. 2C). This gene was not flagged by the conservative tested noted above and reflects an instance where detailed analyses can identify more subtle intragenic conflict. The second segment conflicted with the ML tree in the placement of *Pan* with *Gorilla* instead of *Pan* with *Homo*. Triplicate GARD runs determined support for three trees in only one analysis, with one placing *Gorilla* similar to the ML tree, and another placing it close to *Echinops* and *Dasypus*, among other conflicts (supplementary fig. S9, Supplementary Material online). phyML_multi optimized support for two trees over 28 segments (supplementary fig. S10, Supplementary Material

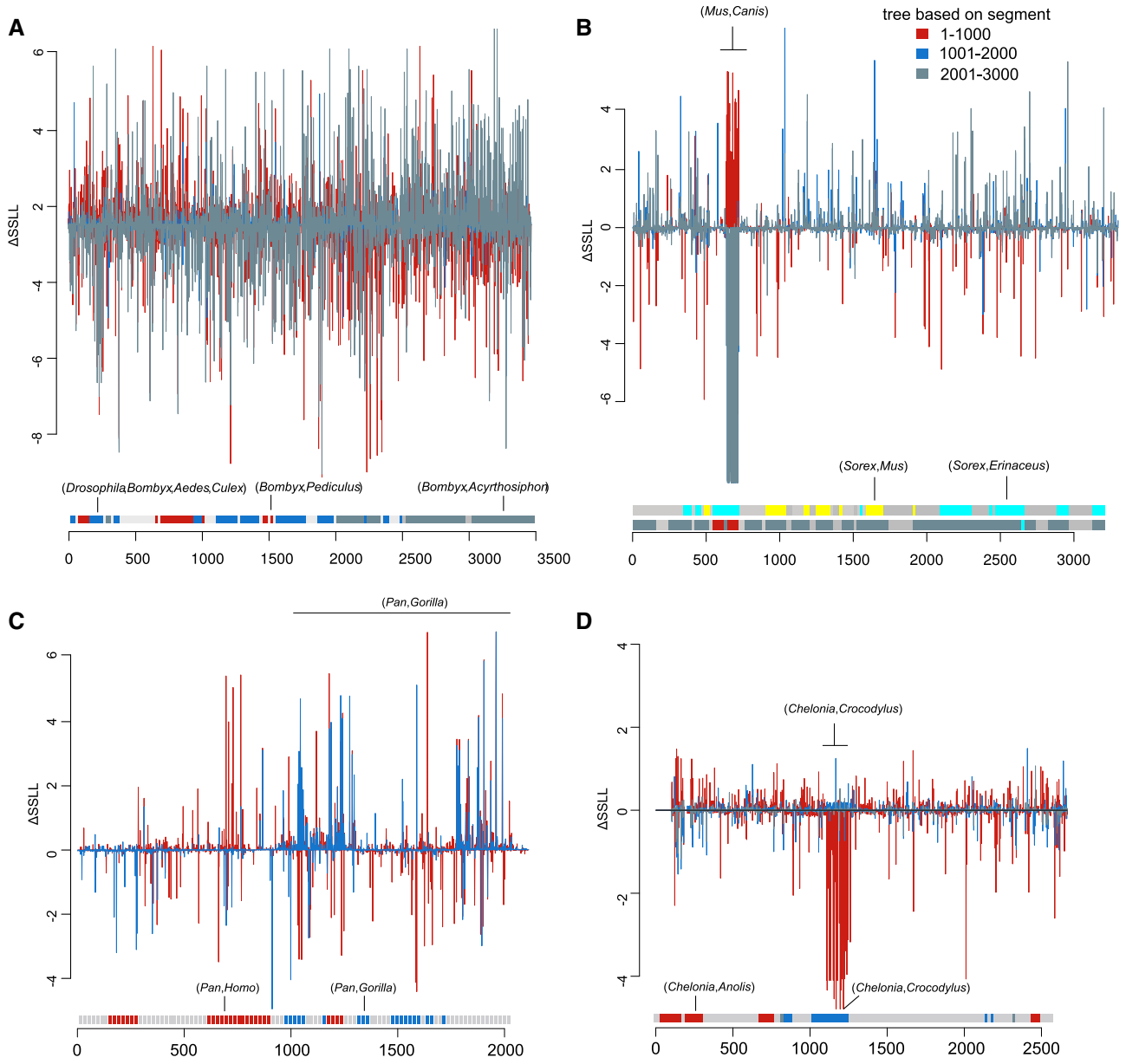

Fig 2. Detailed analysis of within-gene conflict. Delta site-specific *lnL* (ΔSSLL) for each site and the set of segment trees estimated for each data set are shown. Shades represent the ΔSSLL for the tree estimated from the denoted segments. Major deviations for contiguous segments are denoted in *B*, *C*, *D*. Below each plot are the results for the sliding window analysis (using 100-bp segments every 20 bp). Major clades that differ in these segments are shown above the shade in abbreviated Newick. Light gray denotes that there was no strong support for any clade in that gene section. For (*B*) only one sliding-window analysis was conducted; however, we show how the support changes through the alignment for two focal clades (top includes *Sorex*, whereas the bottom shows the shift to *Mus*+*Canis*). (*A*) *EOG2711* gene from the BUGS data set. (*B*) *ENSG00000074803* gene in the MAM1 data set. (*C*) *131* gene in the MAM2 data set. (*D*) *2798* gene in the VERT data set.

online). These trees similarly conflicted with one another and the ML tree in primate relationships, with the first placing *Pan* sister to *Homo* as in the ML tree, and the second placing *Pan*, *Gorilla* and *Homo* in a grade leading to *Callithrix* and *Pongo*, among other conflicts.

Vertebrate Example

We analyzed the 2,798 gene in the VERT data set (fig. 2D). The primary conflicts between the trees based on 1,000-bp segments and the ML analyses involved the placement of crocodile and sea turtle. The ML analyses supported sea turtle (*Chelonia mydas*) as sister to soft-shell turtle (*Pelodiscus sinensis*). However, the second 1,000-bp segment supported sea turtle as sister to crocodile. The dramatic site likelihood shift at ~1,100–1,300 bp (fig. 2D) was not associated with any notable homology or alignment problem. GARD inferred between 17 and 19 segments with topologies varying in the placement of sea turtle, soft-shell turtle, and *Anolis* (supplementary figs. S11–S13, Supplementary Material online). phyML_multi had optimal support for three trees over 10 segments, with most sites supporting one of two trees placing sea turtle sister to *Anolis*, and a small number of sites supporting a topology sea turtle sister to crocodile (supplementary fig. S14, Supplementary Material online).

Table 1. Results from Simulation Analyses of Differing Alignment Lengths Simulated under Different Models with Different Within-Gene Conflicts.

Type	Description	% Genes with Conflicts
No conflict	25 t 3,000 bp JC	0
	50 t 3,000 bp JC	0
	25 t 1,500 bp JC	0
	25 t 3,000 bp diff. GTR	0
	25 t 3,000 bp random GTR	1
Rate shift	25 t 3,000 bp random GTR diff. t height	0
Indel	25 t 3,000 bp random GTR indel equal	1
	25 t 3,000 bp random GTR indel equal: MAFFT	10 (10% bad ML[a])
	25 t 3,000 bp random GTR indel equal: FSA	0 (2% bad ML)
	25 t 3,000 bp random GTR indel equal: PRANK	6 (10% bad ML[a])
	25 t 3,000 bp random GTR indel diff. 1	1
	25 t 3,000 bp random GTR indel diff. 1: MAFFT	11 (18% bad ML[a])
	25 t 3,000 bp random GTR indel diff. 1: FSA	1 (6% bad ML[a])
	25 t 3,000 bp random GTR indel diff. 1: PRANK	18 (19% bad ML[a])
	25 t 3,000 bp random GTR indel diff. 2	2
	25 t 3,000 bp random GTR indel diff. 2: MAFFT	12 (14% bad ML[a])
	25 t 3,000 bp random GTR indel diff. 2: FSA	2 (2% bad ML)
	25 t 3,000 bp random GTR indel diff. 2: PRANK	18 (11% bad ML[a])
Conflict	25 t 3,000 bp random GTR 200 bp conf.	19.5
	25 t 3,000 bp random GTR 500 bp conf.	68.4
	25 t 3,000 bp random GTR 1,000 bp conf.	98
	25 t 1,500 bp random GTR 200 bp conf.	51

NOTE.—The percentage of replicates where the ML tree for the whole alignment was incorrectly inferred are indicated by "bad ML."
[a]Not all these replicates overlapped with cases where multiple trees were detected.

Table 2. Results from Analyses of Individual Data Sets Including the Number of Taxa and Number of Genes in the Original Study, the Number of Genes Long Enough to Analyze, and the Proportion of Those Analyzed with Conflict.

Data Set	No. of Taxa	No. of Regions	No. of Analyzed Regions	% With Conflict
MAM1 (Chen et al. 2017)	22	10,259	3,666	6.5
MAM2 (Song et al. 2012)	37	424	293	27.7
MAM3 (McCormack et al. 2012)	29	183	1	100
VERT (Wang et al. 2013)	12	1,113	551	2.7
BUGS (Niehuis et al. 2012)	13	4,485	467	1.9
FISH (Hughes et al. 2018)	303	1105	118	92.4
FROG (Feng et al. 2017)	164	95	40	80
FUNG (Pizarro et al. 2018)	51	2,256	1,750	37.2
CARN (Walker et al. 2017)	13	1,237	343	0.6
MOSM (Liu et al. 2019)	134	40	11	100
MOSN (Liu et al. 2019)	134	105	81	85.2
MOSP (Liu et al. 2019)	134	82	23	87
PLAN (Wickett et al. 2014)	103	852	160	16.3

For all examples, analysis of realignments continued to exhibit breakpoints featuring similar conflicts with the ML tree (supplementary figs. S15–45, Supplementary Material online). For 131, all three GARD runs detected multiple breakpoints in both FSA and PRANK realignments. For GARD, the positions at which breakpoints were inferred were relatively consistent, considering changes in alignment length and the noted stochasticity between different runs. For phyML_multi, realignment could induce important differences in the results. For example, in the FSA realignment of EOG2711, most sites supported a tree with implausibly long branch lengths featuring *Bombyx* sister to *Drosophila* (supplementary fig. S21, Supplementary Material online), whereas in the PRANK realignment most sites supported a tree featuring *Bombyx* as sister to *Aedes* and *Culex* (supplementary fig. S22, Supplementary Material online). For 2798, realignment with FSA led to inference of three segments supporting only two of the three possible trees (supplementary fig. S44, Supplementary Material online), whereas PRANK realignment showed similar numbers of inferred segments (12 vs. 10) but dissimilar positions (supplementary fig. S45, Supplementary Material online).

Discussion

We have demonstrated that intragenic conflict can be common in empirical data sets. Some of the data sets analyzed here overlap in taxon sampling but vary greatly in the frequency of intragenic conflict (e.g., MAM1 and MAM2) suggesting that although biological processes may play a role in generating conflict (Mendes et al. 2019), nonbiological errors are likely to be a major source of incongruence. For example, MAM2 has been thoroughly analyzed to uncover that errors such as issues with homology exist within the data set assembly (Springer and Gatesy 2018). Importantly, all the

phylogenetic methods from the original publications for each data set in table 2 assumed a single topology underlying the alignment. Our analyses suggest that for many gene regions, this would represent model misspecification, as there are several well-supported topologies underlying many genes that may mislead analyses on these data. Several of the taxon-rich data sets, including FISH and FROG, had very high levels of intragenic conflict across the genes we analyzed. This points to systematic errors in data set assembly, extensive recombination, or other biological processes that may introduce conflict. We note that, although there was a tendency for taxon-rich data sets to exhibit these issues, the PLAN data set did not exhibit a particularly high rate of conflicts. Nevertheless, we expect that increasing taxon sampling may potentially result in more intragenic conflict for biological or nonbiological reasons as increasing sampling will necessarily increase biological heterogeneity (e.g., the probability of sampling a recombination event) and the potential for error. Additionally, longer alignments would be expected to harbor more recombination breakpoints. Our analyses filtered data sets for regions that could produce at least two gene segments and so we could potentially bias upward our estimates of within-gene conflict. However, we did not notice a trend of data sets with longer gene regions to exhibit more conflict overall, and much smaller genes are expected to yield poor-quality phylogenetic inference. Nonetheless, the addition of both more taxa and more sites will increase the potential for biological and systematic within-gene conflict.

Nonbiological errors may arise from alignment inaccuracies, homology issues, and errors in data set assembly perhaps exacerbated for large data sets assembled using significant automation. However, although errors may be a major source of intragenic conflict, the data examined in figure 2 did not exhibit obvious errors (2798 from the VERT data set could perhaps plausibly include misannotated sequence, but it is difficult to confirm this without a more in-depth examination of the soft-shell turtle genome), and breakpoints were detected with multiple methods even after re-alignment with different algorithms. Due to the differences in the goals of different methods, it is difficult to assess the concordance between approaches. By virtue of using segment trees to examine site-specific log-likelihoods as a metric for phylogenetic signal, our in-depth examination is similar to several published approaches (e.g., Likewind; Archibald and Roger 2002a, 2002b; Inagaki et al. 2006) including phyML_multi, and in some cases our approach and phyML_multi yielded nearly identical findings (e.g., support for a tree featuring *Mus*+*Canis* in supplementary fig. S8, Supplementary Material online, between positions 500 and 700 in the alignment is similar to the pattern displayed in fig. 2). However, in other cases, phyML_multi supported many more breakpoints. GARD also yielded concordant findings, for example, trees 2, 3, and 3 in supplementary figures S5–S7, Supplementary Material online, all of which derive from a segment located between positions 500 and 700 in the alignment and all of which featured the (*Mus*, *Canis*) relationship, as above. Nonetheless, GARD commonly inferred many more breakpoints supporting a wide diversity of topologies. Because GARD commonly results in inferring trees from very small subsections of the alignment, only a very small number of sites could have dramatic influence on the inferred topology, which might increase sensitivity to error. Regardless, signals of multiple trees were still supported by all methods and all alignments examined. It is still unclear what the source of the conflict is in those specific examples. However, the conflict we detected, without obvious systematic error, contributes to a growing body of literature that is discovering intragenic conflict due to biological processes such as recombination (e.g., Scornavacca and Galtier 2016; Mendes et al. 2019). Without having generated the original data sets, including assembly or alignment, it would be very difficult to untangle what the source of conflict was in each case. However, to ensure that biological conclusions were not the result of noise and error, it would be important to determine whether intragenic heterogeneity was being properly accounted for in these empirical data sets.

The importance of accurate alignment for tree inference is well understood (Ogden and Rosenberg 2006). The results we present here suggest that alignment error impacts not only ML topology estimates but can also induce intragenic conflict. Specifically, in our simulations, realignment of simulated data containing indels led to increased false-positives. Realignment with a progressive algorithm (MAFFT), which is known to over-align (Katoh and Standley 2016; Vialle et al. 2018), produced greater numbers of false-positives. By contrast, FSA, which typically under-aligns (Bradley et al. 2009), led to lower rates of false-positives. This inference is supported by analysis of empirical data sets. For example, the CARN data set from Walker et al. (2017) was aligned with PRANK and cleaned for column occupancy in the original study and showed comparatively lower rates of within-gene conflict. In our simulations, PRANK admittedly induced relatively high rates of false-positives particular in cases with asymmetric insertion and deletion rates. However, it is still expected to under-align relative to MAFFT (Vialle et al. 2018; Nute et al. 2019), suggesting that the aforementioned alignment filtering may serve to diminish false-positives. Conversely, although realignment with more accurate aligners reduced false-positive occurrence in simulations, multiple trees were still inferred for examples when realigning, suggesting true signals of multiple trees that are not due to alignment errors. Despite the overall simplicity of our simulation process, our rate shift experiments and changes in indel rates between segments should mimic some of the more complex dynamics occurring in, for example, genes encoding multi-domain proteins. Whether due to biological processes or errors in data set assembly, intragenic conflict should not be ignored. Intragenic conflict can drive false-positives in positive selection analyses (Anisimova et al. 2003), influence branch length estimation, and bias phylogenetic reconstruction (e.g., Schierup and Hein 2000). These and other errors are to be expected as intragenic conflict violates typical phylogenetic models that assume a single tree for the length of the alignment. Some have suggested that recombination may have a minimal impact on species tree inference (Lanier and Knowles 2012). However, others have noted that only a few conflicting sites can drive tree inference (Shen et al.

2017). Furthermore, our results have implications for analyses that assume that a gene region, regardless of data type, is the most meaningful phylogenetic unit (e.g., summary coalescent analyses). In the presence of intragenic conflict, an entire gene may not be a meaningful unit for phylogenetic analysis.

What should practitioners do when confronted with within-gene heterogeneity? In some cases, topologies and breakpoints may be unambiguously supported by multiple methods, suggesting a plausible biological source of conflict. However, for many alignments and data sets, complex conflicting signals may be uncovered. Genes for which the assumption of within-gene homogeneity is strongly violated may need to be filtered from data sets prior to further inference, and our method provides a means to identify such genes. However, we resist making exact recommendations considering the uncertainty surrounding the impact of violating these assumptions (Lanier and Knowles 2012). Undoubtedly, the decision whether or not to filter genes will depend on proportion of the gene impacted and the extent of conflict between within-gene trees. Nevertheless, when entire data sets exhibit widespread intragenic conflict, it warrants closer inspection of the underlying data or analyses leading to multiple sequence alignment to ensure that errors are not being introduced at some early step.

Our method presents one way to highlight general problems with data set assembly and to identify important molecular evolutionary patterns. Importantly, the method presented here does not assume a particular source of conflict. Nevertheless, if a data set exhibits significant intragenic conflict, it warrants further investigation regardless of the source. Biological sources of intragenic conflict can provide important information about molecular evolutionary processes, but to better understand these processes, we need to ensure that we identify biological conflict and not error. Disentangling the sources of intragenic conflict will lead to cleaner data sets, more robust species tree inferences, and a greater understanding of the molecular evolutionary processes shaping genomes.

Supplementary Material

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

Acknowledgments

We would like to thank James Pease, Jeremy Beaulieu, Jonathan Chang, and Caroline Parins-Fukuchi for helpful comments. We would also like to the associate editor, Sidonie Bellot, and three anonymous reviewers for comments that greatly improved the manuscript. N.W.H. was supported by a Woolf Fisher Cambridge Scholarship. J.F.W. was supported by University of Michigan Rackham Pre-doctoral fellowship. S.A.S. was supported by a University of Michigan MICDE pilot grant and NSF DEB 1917146.

References

Allman ES, Kubatko LS, Rhodes JA. 2017. Split scores: a tool to quantify phylogenetic signal in genome-scale data. *Syst Biol*. 66(4):620–636.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. *Genome Biol Evol*. 3:246–258.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164(3):1229–1236.

Archibald JM, Roger AJ. 2002a. Gene Conversion and the Evolution of Euryarchaeal Chaperonins: A Maximum Likelihood-Based Method for Detecting Conflicting Phylogenetic Signals. *J Mol Evol*. 55(2):232–245.

Archibald JM, Roger AJ. 2002b. Gene duplication and gene conversion shape the evolution of archaeal chaperonins. *J Mol Biol*. 316(5):1041–1050.

Boussau B, Guéguen L, Gouy M. 2009. A mixture model and a hidden markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. *Evol Bioinform*. 5:EBO.S2242–79.

Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res*. 23(2):323–330.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast statistical alignment. *PLoS Comput Biol*. 5(5):e1000392.

Brown JM, Thomson RC. 2016. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Syst Biol*. 66(4):517–530.

Brown JW, Walker JF, Smith SA. 2017. Phyx: phylogenetic tools for Unix. *Bioinformatics* 33(12):1886–1888.

Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD, Hillis DM. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci U S A*. 106(22):8986–8991.

Chen M-Y, Liang D, Zhang P. 2017. Phylogenomic resolution of the phylogeny of Laurasiatherian mammals: exploring phylogenetic signals within coding and noncoding sequences. *Genome Biol Evol*. 9(8):1998–2012.

Edwards AWF. 1884. Likelihood. CUP Archive.

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

Feng Y-J, Blackburn DC, Liang D, Hillis DM, Wake DB, Cannatella DC, Zhang P. 2017. Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous–Paleogene boundary. *Proc Natl Acad Sci U S A*. 114(29):E5864–E5870.

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

Gatesy J, Springer MS. 2013. Concatenation versus coalescence versus "concatalescence". *Proc Natl Acad Sci U S A*. 110(13):E1179–E1179.

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

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 35(2):518–522.

Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLOS Genet*. 3(2):e7.

Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, Arcila D, Betancur-R R, Li C, Becker L, et al. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proc Natl Acad Sci U S A*. 115(24):6249–6254.

Husmeier D, McGuire G. 2003. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Mol Biol Evol*. 20(3):315–337.

Inagaki Y, Susko E, Roger A J. 2006. Recombination between elongation factor 1 genes from distantly related archaeal lineages. *Proc Natl Acad Sci U S A*. 103(12):4528–4533.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: improvements in Performance and Usability. *Mol Biol Evol*. 30(4):772–780.

Katoh K, Standley DM. 2016. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* 32(13):1933–1942.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006a. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol*. 23(10):1891–1901.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006b. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22(24):3096–3098.

Lanier HC, Knowles LL. 2012. Is recombination a problem for species-tree analyses? *Syst Biol*. 61(4):691–701.

Liu Y, Johnson MG, Cox CJ, Medina R, Devos N, Vanderpoorten A, Hedenäs L, Bell NE, Shevock JR, Aguero B, et al. 2019. Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. *Nat Commun*. 10(1):1485.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A*. 102(30):10557–10562.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res*. 22(4):746–754.

Mendes FK, Livera AP, Hahn MW. 2019. The perils of intralocus recombination for inferences of molecular convergence. *Phil Trans R Soc B*. 374(1777):20180244.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30(17):i541–i548.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.

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.

Nute M, Saleh E, Warnow T. 2019. Evaluating statistical multiple sequence alignment in comparison to other alignment methods on protein data sets. *Syst Biol*. 68(3):396–411.

Ogden TH, Rosenberg MS. 2006. Multiple sequence alignment accuracy and phylogenetic inference. *Syst Biol*. 55(2):314–328.

Pizarro D, Divakar PK, Grewe F, Leavitt SD, Huang J-P, Dal Grande F, Schmitt I, Wedin M, Crespo A, Lumbsch HT. 2018. Phylogenomic analysis of 2556 single-copy protein-coding genes resolves most evolutionary relationships for the major clades in the most diverse group of lichen-forming fungi. *Fungal Divers*. 92(1):31–41.

Salminen MO, Carr JK, Burke DS, McCutchan FE. 1995. Identification of Breakpoints in Intergenotypic Recombinants of HIV Type 1 by Bootscanning. *AIDS Res Hum Retrov*. 11(11):1423–1425.

Schierup MH, Hein J. 2000. Consequences of recombination ontraditional phylogenetic analysis. *Genetics* 156(2):879–891.

Scornavacca C, Galtier N. 2016. Incomplete lineage sorting in mammalian phylogenomics. *Syst Biol*. 66(1):112–120.

Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol*. 1(5):0126.

Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol Biol*. 15(1):150.

Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci U S A*. 109(37):14942–14947.

Springer MS, Gatesy J. 2018. On the importance of homology in the age of phylogenomics. *Syst Biodivers*. 16(3):210–228.

Suchard MA, Weiss RE, Dorman KS, Sinsheimer JS. 2002. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. *Syst Biol*. 51(5):715–728.

Vialle RA, Tamuri AU, Goldman N. 2018. Alignment modulates ancestral sequence reconstruction accuracy. *Mol Biol Evol*. 35(7):1783–1797.

Walker JF, Brown JW, Smith SA. 2018. Analyzing contentious relationships and outlier genes in phylogenomics. *Syst Biol*. 67(5):916–924.

Walker JF, Yang Y, Moore MJ, Mikenas J, Timoneda A, Brockington SF, Smith SA. 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *Am J Bot*. 104(6):858–867.

Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, Li C, White S, Xiong Z, Fang D, et al. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat Genet*. 45(6):701–706.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, 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):E4859–E4868.

Wu S, Song S, Liu L, Edwards SV. 2013. Reply to Gatesy and Springer: the multispecies coalescent model can effectively handle recombination and gene tree heterogeneity. *Proc Natl Acad Sci U S A*. 110(13):E1180–E1180.