Zoologica Scripta

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Invited Review

Phylogenomic subsampling: a brief review

Scott V. Edwards

Submitted: 11 August 2016
Accepted: 11 August 2016
doi:10.1111/zsc.12210

Edwards, S.V. (2016). Phylogenomic subsampling: a brief review. — *Zoologica Scripta*, *45*, 63–74.

Phylogenomic subsampling is a method for studying the stability of phylogenetic analyses by taking random or ordered subsamples of loci and comparing phylogenetic analyses of those subsamples. The method is made possible by the large number of loci made available by application of next-generation sequencing methods to phylogenetic questions. My laboratory has used phylogenomic subsampling to investigate the consistency and stability of various methods of phylogenetic analysis of multilocus data sets, including the so-called species tree methods, which use the multispecies coalescent as a framework for interpreting gene tree heterogeneity. I was inspired to focus on this method for this symposium volume because my Harvard colleague Gonzalo Giribet had independently been using phylogenomic subsampling to explore various questions in invertebrate phylogenomics. Phylogenomic subsampling has many useful applications in phylogenomics, yet when reporting the particulars of the results of such analyses, care should be taken to focus primarily on discrepancies that achieve a high level of support by the bootstrap or other methods. Using a recently published example, I show that the methods used to summarize the results of a subsampling experiment, such as the threshold for reporting support for one or another tree or clade, can influence the perceived success or failure of concatenation or species tree methods. Single- versus double-bootstrapping is also shown to produce different subsampling results. I suggest guidelines for analysing and reporting the results of phylogenomic subsampling and suggest that it should become a routine part of phylogenetic analysis in the next-generation era.

Corresponding author: *Scott V. Edwards, Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA. E-mail: sedwards@fas.harvard.edu*

Scott V. Edwards, Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA. E-mail: sedwards@fas.harvard.edu

Introduction

In the Department of Organismic and Evolutionary Biology at Harvard, as is probably true in many university departments, we have a saying that we tend to see our colleagues more often on the road, at conferences, than at home. Although not strictly true, given the rarity with which professors assemble to hear their own colleagues discuss their research, we infrequently hear what our departmental colleagues have been up to research-wise unless at a conference. Even more surprising is the situation that arises when one realizes that one's own research has been travelling along a similar path as one's departmental colleague, only to be discovered at a far away conference. Such was my experience at the 2015 special symposium on phylogenetics assembled by *Zoologica Scripta*.

Systematics being what it is – a discipline often divided by taxon such that vertebrate phylogeneticists rarely interact with those working on invertebrates – it had been several years since I had heard my dear friend and colleague Gonzalo Giribet discuss his impressive research on phylogenomics of metazoans, including molluscs, chelicerates and many other invertebrate clades. Aside from the excitement of hearing about the heroic efforts of his laboratory to resolve various clades and higher taxa, in particular I

was struck by his discussion of his paper on phylogenomics of Arachnida (Sharma *et al.* 2014) in which he presented their use of phylogenomic subsampling to study the signal of various data sets in resolving various clades of spiders. I was particularly surprised because my own laboratory had been using phylogenomic subsampling for a variety of purposes, in particular to study the accumulation of signal for various clades with different-sized data sets and in particular to compare the behaviour of concatenation approaches with the so-called species tree approaches when analysing data sets of different sizes (Song *et al.* 2012). I had known about the challenges of matrix occupancy – the challenges of obtaining a high fraction of genes for all taxa in a phylogenomic analysis, particularly when analysing highly divergent groups – that Giribet and his colleagues had grappled with in their efforts to resolve metazoan phylogeny, and their efforts to study its consequences (e.g. Dunn *et al.* 2008; Hejnol *et al.* 2009). But in the Arachnid paper, and, as I later learned, in even earlier papers from collaborative work (Hejnol *et al.* 2009), I learned that the Giribet laboratory had begun to use subsampling routinely to test various phylogenetic hypotheses and their behaviour with different-sized data sets. The convergence of research practice that I witnessed at the *Zoologica Scripta* meeting inspired me to focus this contribution on the practice of phylogenomic subsampling. Clearly, the practice is becoming more widespread (Narechania *et al.* 2012; Simon *et al.* 2012), yet it is being used in a variety of ways and for a variety of purposes, and has shed light on different facets of phylogenomic analysis, depending on how it is deployed.

Phylogenomic subsampling

To my knowledge, phylogenomic subsampling has not yet been defined, or really even coined, and the term likely means different things to different researchers. For the present purpose, phylogenomic subsampling can be defined as a phylogenomic protocol in which loci are sampled at random to create different-sized locus-by-species matrices, with the goal of exploring the stability of a phylogenetic hypothesis (Song *et al.* 2012). Many kinds of subsampling have been employed throughout the history of phylogenetics and phylogenomics. Historically, phylogenomic subsampling emerged from the practice of subsampling characters or assessing the consequences of different trees or data perturbations to test for stability of phylogenetic hypotheses (Davis *et al.* 1993; Bremer 1994; Gatesy *et al.* 1999a,b; Giribet & Wheeler 2003). As phylogenomic data sets increased in size, researchers began exploring the effects of subsampling on the scale discussed here, with random removal of both sites and loci, albeit solely within a concatenation framework (Rokas *et al.* 2003). Bootstrapping can be considered a type of subsampling in which pseudomatrices of the same size are generated as a means of assessing the strength of support for particular phylogenetic hypotheses (Felsenstein 1985). And of course, with traditional bootstrapping, it is sites within and between concatenated genes, not entire loci, that are sampled with replacement to create the pseudomatrices. Multilocus bootstrapping has been explored by Seo (2008; see also Seo *et al.* 2005), where, for example, he compared the performance of sampling sites within loci vs. entire loci vs. sampling sites and loci in large multilocus data sets. Phylogenomic subsampling can be distinguished from these other types of sampling in its explicit focus on comparing matrices of increasing size, and in exploring the ability to recover a given clade under phylogenetic analysis of those matrices. The Random Addition Concatenation Analysis (RADICAL) approach of Narechania *et al.* (2012) captures many of the aspects of phylogenomic subsampling discussed here, albeit solely within a concatenation framework as applied thus far. Simmons *et al.* (2016) explored the effects of subsampling of previously estimated gene trees on species tree estimation, examining both random subsamples and subsamples ordered by average Robinson-Foulds (1981; RF) distance to other gene trees. They provide a number of suggestions for testing the robustness of species tree methods via sampling of different subsets of gene trees. However, their approach only subsampled gene trees and, by neglecting to subsample sites within alignments for each gene tree, ignored the uncertainty of each gene tree estimate and hence overinterpret the significance of their gene tree estimates and associated RF distances (see below). Subsampling can also be performed on taxa, and although we and others have explored this approach (Song *et al.* 2012), often called jackknifing in past studies. To our knowledge, jackknifing has not yet been scaled up in a way deserving of the term 'phylogenomic'.

The study of the information content of phylogenomic matrices of different sizes has been used recently for many different purposes (Fig. 1). Many studies have studied the effect of matrix occupancy and missing data on their particular phylogenomic analysis, particularly when the full matrix is relatively sparse (e.g. Dunn *et al.* 2008; Hejnol *et al.* 2009; Sharma *et al.* 2014; Katz & Grant 2015). For example, there has been some debate in the literature as to whether sparse matrices, in some cases missing as much as 90% of the cells in a locus-by-species matrix, have reliable phylogenomic information content. One of the early 'shots across the bow' in this debate came from Driskell *et al.* (2004), who suggested that large matrices in which nearly 80% of the data were missing nonetheless had significant information content in resolving the tree of life when analysed in a supermatrix or concatenation framework. Several studies have gone on to show that supermatrices that are

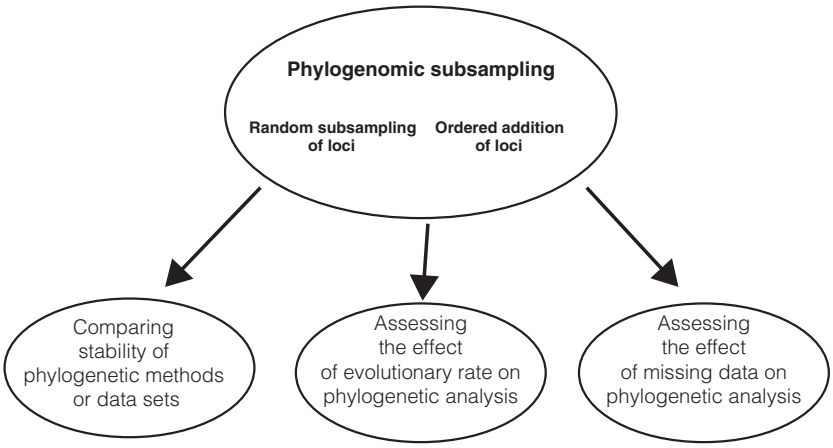

Fig. 1 Overview of uses of phylogenomic subsampling. Two types of subsampling are indicated, those which sample loci at random (the main method discussed in this paper) and those that add loci to matrices of increasing size in some ordered fashion (e.g. by increasing evolutionary rate). Phylogenomic subsampling has been used for three main purposes, as discussed in the text: to test the stability of various phylogenetic methods on matrices of different size and composition; to test the effect of differences in evolutionary rate on phylogenomic analysis; and to test the effects of missing data on phylogenomic analysis. This paper focuses on (and advocates the use of) subsampling primarily for the leftmost purpose, but acknowledges the use of ordered subsampling as well. As matrices become increasingly occupied (e.g. filled with sampled loci, as opposed to empty), the rightmost purpose will become less important.

sampled very sparsely nonetheless can possess significant phylogenetic signal (Dunn *et al.* 2008; Hejnol *et al.* 2009; Sharma *et al.* 2014). Although many of these early phylogenomic studies exhibited promising signal even when using very sparsely populated matrices, there is an undeniable trend in the field towards generating more complete matrices (Katz & Grant 2015; Sharma *et al.* 2015). In the way in which I use the term, phylogenomic subsampling is applied primarily to fully populated or near-populated matrices, that is matrices in which all or nearly all the species have been sampled for all genes. In this case, subsampling is not used to test whether missing data is influencing a particular phylogenetic analysis (Wiens 2006; Jiang *et al.* 2014). Rather it is used principally to study the stability of various clades with phylogenomic matrices of different sizes, with the prediction that clades should in general exhibit increased bootstrap or other support as matrix size increases.

It is beyond the scope of this brief review to discuss the pros and cons of different next-generation sequencing methods in generating large but strongly populated phylogenomic matrices, a topic that has been covered elsewhere (Lemmon *et al.* 2012). An informal survey of the literature and discussions with colleagues suggests that depending on the temporal depth of phylogenetic comparisons, transcriptomes can sometimes be challenging as a means of generating strongly populated matrices, particularly among deeply diverged lineages. Hybrid sequence capture methods, especially those leveraging the phylogenetic relationship and estimated ancestral sequences of loci, may well prove better at generating strongly populated matrices at deep phylogenetic depths (Lemmon & Lemmon 2013; Bragg *et al.* 2015; Brandley *et al.* 2015; Potter *et al.* 2016). The Rad-seq method has initially been applied to shallow divergences at the phylogeographic level, yet is being used more frequently for deeper divergences, albeit thus far primarily in a concatenation framework (Cruaud *et al.* 2014; Leache *et al.* 2015; Zimmer & Wen 2015). At both shallow and deep divergences, however, Rad-seq can sometimes yield patchy matrices (Edwards *et al.* 2016b). These matrices, even when reduced to fully or mostly occupied matrices, provide substantial resolution at many taxonomic levels, yet, if only for financial reasons, researchers generally prefer maximizing the size of fully occupied matrices. Improved laboratory methods, including new next-generation sequencing platforms that yield longer reads, as well as better bioinformatics pipelines, will likely contribute strongly to more fully occupied matrices for phylogenomics in the future (Bi *et al.* 2013; Jones & Good 2016).

Phylogenomic subsampling and species trees

Phylogenomic subsampling has been useful retrospectively, for studying the effects of sparsely sampled matrices on phylogenetic analysis, but will also be useful for a future in which large, fully occupied matrices are the norm. The use of phylogenomic subsampling in my laboratory was first used by my collaborators and me as a means of testing the stability of various phylogenetic methods and their consistency across different data sets

(Song *et al.* 2012). As previously stated, we knew of the extensive use of 'downsizing' character matrices to achieve use as a means of concentrating signal and reducing the effects of missing data on phylogenetic analysis. Still, and somewhat surprisingly (although perhaps not for Harvard), our use of subsampling was completely independent of its use in the Giribet laboratory, only two floors above in the Museum of Comparative Zoology Laboratories! This is all the more surprising as the Giribet laboratory began to expand its use of subsampling from exploration of patchy matrices to also studying the consistency of specific clades across data sets. We essentially converged on the approach by different routes. In our case, my laboratory arrived at subsampling after several years of developing methods for inferring phylogenetic trees, or 'species trees', that allowed the underlying gene trees to vary from locus to locus. We use a model, the multispecies coalescent (MSC) model, that provides an elegant means of assessing the likelihood of an overarching species tree in the presence of gene tree variation that is dominated by incomplete lineage sorting (ILS; Rannala & Yang 2003, 2008; Degnan & Rosenberg 2009; Liu *et al.* 2009a, 2015b). As reviewed extensively elsewhere, there is now a diverse ecosystem of species tree methods, ranging from Bayesian to likelihood to summary statistic methods, and which can handle a variety of data types, from multiple SNPs linked into single loci to multiple unlinked SNPs (Bryant *et al.* 2012; DeGiorgio *et al.* 2014; reviewed in Edwards 2016). These methods are known to vary in their efficiency and also in their ability to handle large data sets, with Bayesian methods being the desired goal, specifying the MSC completely but unable to handle large data sets, and summary statistic methods being less efficient, but still statistically consistent and able to handle data sets befitting the title 'phylogenomics'. The consistency of a species tree method is confirmed by its estimation of the correct species tree across the full parameter space of the MSC, especially in the so-called anomaly zone, a region of species tree space that generates a distribution of gene trees where the most common gene tree differs from the species tree (Degnan & Rosenberg 2006; Rosenberg 2013) – a region in which supermatrix approaches are guaranteed to be inconsistent. We view the supermatrix approach not as antithetical to the MSC but as a subset of it, a model in which all gene trees are forced to be the same and presumably identical to the species tree (Liu *et al.* 2015b; Edwards *et al.* 2016a). Our enthusiasm for species tree methods has been recently strengthened by the observation that the anomaly zone, originally just a theoretical curiosity and believed to be unlikely in nature, is strongly implied if not demonstrated in a number of empirical studies, including some unpublished studies in our laboratory (Huang & Knowles 2009; Linkem *et al.*, 2016).

Specifically, our first use of subsampling had the goal of comparing the performance of supermatrix (concatenation) and MSC methods to phylogenetic reconstruction. We followed a simulation and reporting protocol that we encourage others to follow (Song *et al.* 2012):

1. Determine a relatively well-occupied phylogenomic data matrix for a given set of taxa. Choose a specific set of branches or clades to test for stability via subsampling. This might best be done a priori but in the case of Song *et al.* (2012) we chose clades that differed strongly between supermatrix and MSC analyses.
2. According to Seo (2008), subsample columns of loci, as well as sites within loci, at random with replacement, creating pseudomatrices of different and increasing numbers of loci, for example, from an original matrix of 500 loci, one could subsample matrices of 10, 20, 50, 100, 200 and up to 500 loci. Sampling only sites within loci, but not the loci themselves, will likely yield different results that may not capture the complexities of the data (Seo 2008).
3. Each subsampling should be replicated at least 10 times within each matrix size class.
4. Build a phylogenetic tree by whatever method on each of the replicates within each size class.
5. For each tree, assess the bootstrap or other support of the clade(s) identified in step 1.
6. Summarize the distribution of support for the clade or alternative incompatible clades across replicates and pseudomatrix sizes in a form that captures the full range of support values observed. For example, a heat map is a useful way of recording the incidence of high bootstrap support for the chosen clade, or a conflicting clade, or simply equivocal support for the clade (less than some bootstrap threshold value).

In our paper on mammals (Song *et al.* 2012), we used a heat map to report the distribution of support values for our phylogenomic subsample, which is reproduced using a corrected data set (Wu *et al.* 2015) in Fig. 2. We thought this representation was useful because we assumed that replicates in which a specific clade of interest was recovered at less than some threshold value (in the case of Song *et al.* 2012, 90%) could be considered 'non-committal' with respect to that clade. However, replicates in which a clade is recovered with high support could potentially reveal situations in which such support could actually be recovered in a traditional phylogenetic analysis of that data set. Recording such events becomes even more critical as one observes the support for that clade in data sets of different sizes, or in different replicates of the same size. Our choice of a 90% bootstrap threshold was ultimately arbitrary, but we

found that it highlighted well the differences in behaviour of concatenation vs. species tree approaches, which was one of our goals.

What we observed in the mammal data set was a surprising tendency for supermatrix analyses, specifically those using RAxML, to 'flip-flop' between strong support for a clade on the one hand, and an alternative, incompatible clade on the other, *even within replicates of the same matrix size*, and even more regularly across matrices of different sizes (Fig. 2). By contrast, we found that clades estimated using phylogenetic methods employing the MSC (MP-EST is used here; Liu *et al.* 2010) generally exhibited gradually increasing bootstrap support with increasing matrix size, without evidence of jumping back and forth between strong support alternative, conflicting topologies. A similarly smooth approach to high support was found in a study on coelacanth relationships (Liang *et al.* 2013). Here we reanalyse the data of Song *et al.* (2012) using corrected data and the same analysis pipeline as in Song *et al.* (2012). We used RAXML version 7.0.4 (Stamatakis 2006) to build gene trees from concatenated data. For species tree construction, we used PHYML v. 3.0 to estimate gene trees and MP-EST v. 1.2 (Liu *et al.* 2010), and PHYBASE version 1.4 (Liu & Yu 2010) to conduct multilocus bootstrapping (see

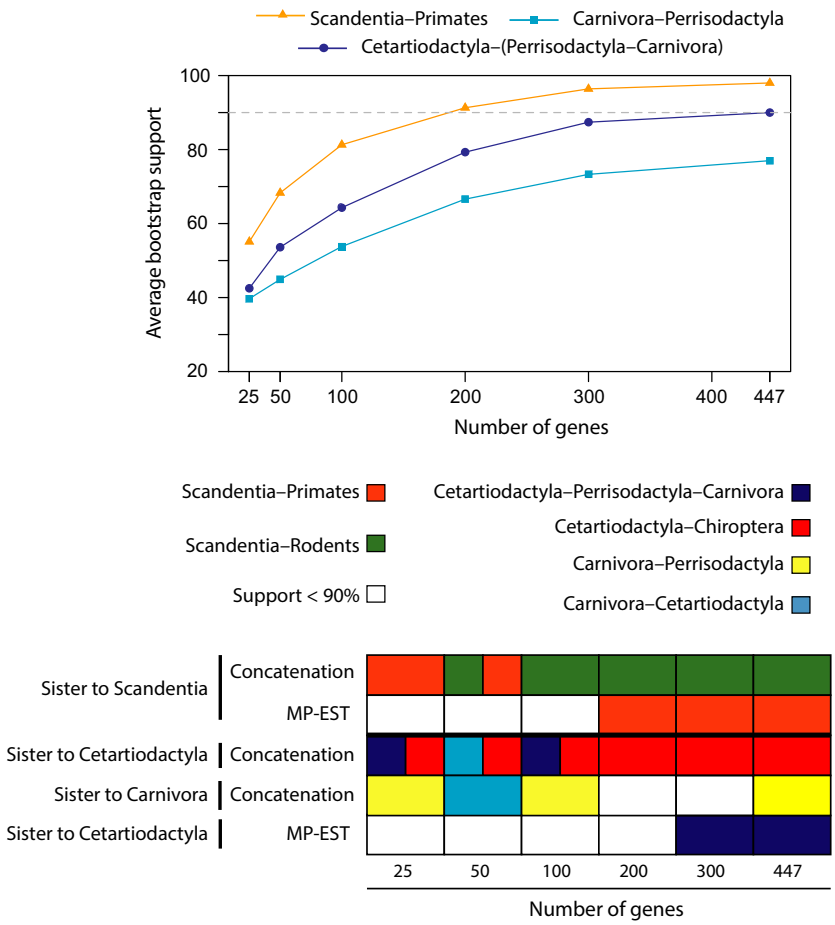

Fig. 2 Phylogenomic subsampling results for mammals in Song *et al.* (2012) using a corrected data set (Wu *et al.* 2015). This analysis uses the data of Song *et al.* (2012) but with 21 misannotated gene trees noted by Mirarab *et al.* (2014) corrected here. A typical command line for RAxML on concatenated data was as follows: raxmlHPC -f a -x 12345 -p 12345 -# 100 -m GTRGAMMA -s 447genes.phy -n 447 genesoutfile. A typical command line for using Phyml to construct gene trees for MP-EST analysis was as follows: phyml –i seq_file –mGTR. In panel A, the gradual approach to statistical significance of three clades of mammals is illustrated across subsamples of increasing size. Only averages across the 10 replicates for each subsample are shown. In panel B, each replicate of each subsample is colour-coded according to the favoured topology for that replicate and whether or not the results for that subsample achieved a bootstrap support of >90%. If a replicate did not achieve 90%, the cell is left white. A second row of results for concatenation is shown, focusing on the sister to Carnivora, because some results of this clade (e.g. at 50 gene subsamples) conflict with results for the sister to Cetartiodactyla in the row above. Other authors (Springer & Gatesy 2016) have suggested additional errors in the Song *et al.* (2012) alignments, but many of these errors and their effect on the results of Song *et al.* (2012) are disputed (Edwards *et al.* 2016a). The numbers from which these figures were drawn can be found in Table S1.

legend Fig. 2 for additional details). I used the 'sumtrees' module in Dendropy to count clade frequencies in bootstrap replicate species trees (Sukumaran & Holder 2010, 2015). Mirarab *et al.* (2014) discovered some errors in 21 of the 447 genes used by Song *et al.* (2012) to study mammal phylogeny. Wu *et al.* (2015) reanalysed the corrected data set and found that the conclusions of Song *et al.* (2012) were unchanged. However, the results of that analysis have not yet been presented in detail. Figure 2 and the Supporting Information present a reanalysis of the corrected data of Song *et al.* (2012). Surprisingly, we found that the tendency for concatenation to flip-flop with high support for conflicting topologies across different subsampling replicates was even more severe than in the original study. Strongly conflicting topologies were recovered using concatenation (RAxML) not only for small numbers of genes, but also for moderate numbers (~100).

It was at the Oslo meeting that I learned that Giribet's laboratory had observed exactly the same pattern in their studies of arachnid phylogenomics (see fig. 5 of Sharma *et al.* 2014), albeit with an important twist. Whereas our protocol specifically sampled loci (columns) at random to create pseudomatrices, Sharma *et al.* (2014) did not subsample at random, but rather added loci in order of increasing evolutionary rate to create increasingly large matrices, without replication within size classes (Narechania *et al.* 2012; Simon *et al.* 2012). They observed that for some clades (e.g. Chelicerata, Tetrapulmonata and, to a lesser extent, Acariformes+ Pseudoscorpions and Ricinulei + Xiphosura, among others), bootstrap support gradually increased as faster evolving loci were added. However, other clades, such as Ricinulei + Solifugae, Acari + Pseudoscopiones or Arachnida itself, exhibited the same type of 'flip-flopping' as we had observed in the mammal data set. Although Sharma *et al.* provide compelling evidence that the effects of evolutionary rate in their study were distinct from the effects of matrix occupancy, technically, it seems to me, given their framework, they cannot distinguish the effects of increasingly large matrices on their results from matrices increasingly dominated by faster evolving loci. To distinguish whether evolutionary rate or matrix size most strongly influence support values in a phylogenetic context, one would have to conduct a simulation and, say, for the 500-locus class of matrices, build multiple matrices including only fast or slow loci, or a mix of the two. Then one could observe the behaviour of specific clades across these data sets and potentially disentangle the effects of increasing evolutionary rate from those of matrix size (as opposed to matrix occupancy). This is precisely what Sharma *et al.* (2015) subsequently did in their updated approach in a recent paper on scorpions, where the entire matrix was partitioned into equally large tertiles of slow-, medium- and fast-evolving loci. Interestingly, that subsampling protocol made no difference in the support values they observed for that data set. Katz & Grant (2015), Sun *et al.* (2016) and others have tested the effect of removing fast *sites* from concatenated alignments. Results of such an analysis tend to vary by study: Katz & Grant (2015) found little improvement in signal with removal of fast sites. By contrast, Sun *et al.* (2016) and Xi *et al.* (2014) found very marked effects of including or excluding fast sites from alignments, the latter especially when analysed by concatenation methods.

Additionally, whereas in Song *et al.* (2012) both supermatrix and MSC methods were included and compared, Sharma *et al.* (2014) only studied the behaviour of a supermatrix approach, namely RAxML (although they did explore the effects of another concatenation approach, PhyloBayes (Lartillot *et al.* 2013), finding very similar results to RAxML). Still, the similarity in results between our two papers was striking, at least for the case of concatenation. And, appropriately, their conclusion from these analyses was a general one, namely that significant phylogenetic conflict existed in their data. For the Arachnida paper, Sharma *et al.* (2014) perhaps favoured the supermatrix approach due to the patchiness of the data set (P. Sharma, pers. comm.). By contrast, in their recent scorpion paper (Sharma *et al.* 2015), which was characterized by a much more complete character matrix, both supermatrix and species tree methods recovered the same basal topology.

Single vs. double bootstrap subsampling

Seo (2008) described several types of bootstrapping that could be applied to multilocus data and analysed their performance under concatenation and multilocus sequence data. He outlined a '2-stage' bootstrap procedure (what we call 'double bootstrapping' here) in which both loci and sites within loci are sampled with replacement to create pseudo-data sets for phylogenetic analysis (his B3 method). He compared the performance of double bootstrapping with traditional bootstrapping (his B1 method) as well as bootstrapping only across sites within loci (what we call 'single bootstrapping'; his B2 method). In concatenation or supermatrix studies, bootstrapping is usually performed without reference to the locus from which a site is sampled; even though great pains are often used to model the substitution dynamics of individual loci ('partitioning'), bootstrapping is almost always conducted while ignoring the multilocus structure of the data. Seo (2008) found that traditional bootstrapping in a concatenation framework, in which sites (columns) are sampled at random without regard to locus length or membership, produced gross over- and underestimations of support during bootstrapping, especially with smaller numbers of genes where

individual genes can dominate the signal. He also found that double bootstrapping was more successful at capturing the variation in the data than either single or traditional bootstrapping.

In the course of reanalysing the data in Song *et al.* (2012; Wu *et al.* 2015), we found some interesting differences in results depending on the type of bootstrapping employed (Fig. 3). Song *et al.* (2012) employed double bootstrapping in their original study, although when the full data sets (447 loci) were analysed, only a single replicate was analysed (instead of 10). We found that single bootstrapping generally gave lower support for the species tree for smaller data sets (<200 loci), but that the support exceeded the support yielded by double bootstrapping for the largest data sets in two of three clades tested (Fig. 3; Supporting information). In the case of the Scandentia (tree shrew)–Primate relationship that resulted from analysis of the full data set using species tree methods, single bootstrapping yielded lower support than double bootstrapping for all data sets, sometimes differing from double bootstrapping by over 40 percentage points (in the case of 300 loci). For the other two clades, the point at which single bootstrapping support exceeded that of double bootstrapping varied from 200 loci for the Cetartiodactyla–Perissodactyla–Carnivora clade to the full data set for the Cetartiodactyla–Perissodactyla clade. Single bootstrapping tests the uncertainty in individual gene trees and potential for gene tree error (Roch & Warnow 2015), with the overarching assumption that the gene tree distribution is as expected. Double bootstrapping tests both gene tree and gene distribution uncertainty and provides a more general test of the efficacy of species tree methods. However, even when studying the stability of species tree methods, single bootstrapping is sometimes used when "very large" data sets are analyzed (here, "large" is arbitrarily defined), as in Jarvis et al. (2014). Clearly, we need additional studies to understand the behaviour of single vs. double bootstrapping in phylogenomics in general and in subsampling in particular.

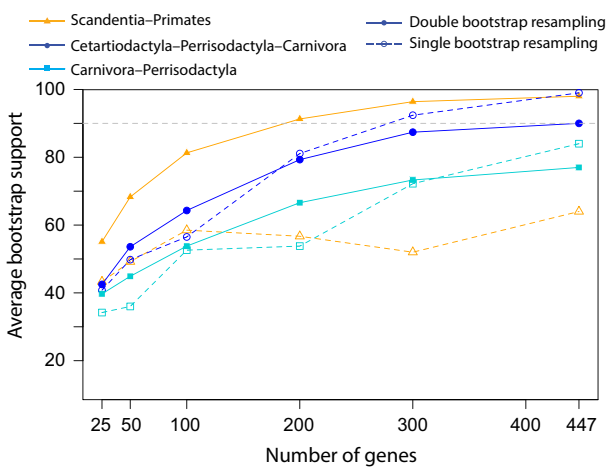

Fig. 3 Phylogenomic subsampling and the accumulation of signal in the Song *et al.* (2012) mammal data with single bootstrapping and double bootstrapping. The corrected data (Wu *et al.* 2015) were used to subsample using single bootstrapping and double bootstrapping (Seo 2008). The stability of three clades was analysed as in Fig. 2. See main text for further details. The numbers from which this figure was drawn can be found in Table S1.

Simmons *et al.* (2016) explored the effect of subsampling of gene trees on species tree estimates. Because they only subsampled gene trees and did not explore the uncertainty of each gene tree estimate via bootstrapping, their single bootstrap approach was different yet again from the methods suggested by Seo (2008). They make some useful suggestions for testing the robustness of various species tree methods, but their evaluation of species tree methods is flawed. By focusing solely on clades 'contrary' to specific reference clades, and RF distances of point estimates of both gene trees and species trees, without interrogating the uncertainty of these point estimates, they overinterpret the apparent conflict and incongruence in many of their tests. Many of the incongruences and large RF distances of gene and species trees they discovered in their analyses likely differ inconsequentially from their reference trees by branches with low bootstrap support, a prediction that double bootstrapping could confirm or reject. I believe gene tree subsampling could be a useful tool in species tree analysis but only while acknowledging that any point estimate of a gene or species tree has a variance that may render its perceived 'conflict' with or large RF distance from a reference tree less meaningful.

An example and a comment

Recently, Richart *et al.* (2016) used phylogenomic subsampling to compare the performance of MSC and supermatrix methods in another study of arachnid phylogeny (Richart *et al.* 2016). Specifically, starting from a matrix of ~672 loci for five species (three in-group and two out-group taxa), they used a phylogenomic subsampling protocol very similar to the 6-step protocol reported above, proceeding in increments of 25–100 loci. However, they diverged from our protocol in one key aspect, namely they reported in their main text the average bootstrap support for a given clade across replicates in the same matrix size class, and did not report a the ranges and distribution of bootstrap values. Their focus of interpretation was on average support values within and across replicates, whereas the focus in Song *et al.* (2012) was also on the full range of support values within and across replicates. Although seemingly a trivial difference from our protocol, I believe this led to a bias in

their interpretation of their results. Focusing exclusively on average support values within and across replicates, they observed a trend in supermatrix analyses towards 100% bootstrap support for several clades believed to be correct, whereas for the MSC methods they used (MP-EST and STAR; Liu *et al.* 2009a, 2010), the average support for these clades never reached 100%, but instead stayed generally below 80%, even for the largest matrices. They interpreted this result as evidence that supermatrix approaches were more efficient and more accurate than MSC approaches in this case, and suggested that 'our example does not support this claim' by Edwards *et al.* (2016a) that phylogenomic subsampling reveals inconsistencies in concatenation. However, in failing to discuss anything about the range of support values within and across replicates, they neglected to report a troubling signal in their data. First, I believe the heat map they produce in their Supporting Information is misleading, in so far as it categorizes trees that recover different relationships within the same replicate (e.g. 25, 50 or 100 loci) as different, regardless of the bootstrap support they achieve. Thus, for example, cells in their supplementary heat maps receive different colours when one clade receives 61% support vs. when an alternative clade receives 66% support. This protocol gives the unjustified impression that both MSC and supermatrix approaches frequently produce the 'flip-flopping' results that we observed in Song *et al.* (2012), although such a conclusion would not be justified for many of the replicates. We suggest that such discrimination is not meaningful because it portrays results as different that arguably receive no statistical support for making any statement about phylogenetic relationships. The subsampling analysis by Simmons *et al.* (2016) similarly tried to evaluate various species tree methods using RF distances of estimated trees from a reference tree but ignoring the strength of support for those estimates. In a similar vein, Richart *et al.* (2016) were surprised by the fact that the STAR method (Liu *et al.* 2009b) produced an unexpected topology for one of the 600-locus replicates, yet this topology was only supported at the level of 53%! We reject their suggestion that STAR has performed poorly in this case and suggest that such low-support results do not warrant attention in phylogenomic analyses.

When we reproduce their heat map using the colouring protocol of Song *et al.* (2012; Fig. 4), in which only cells achieving >90% support are reported, we see how the impression of flip-flopping is eliminated, with cells corresponding to MSC methods largely being coded as white, indicating lack of definitive support. Second, as observed in their supplementary material, the range of support values for supermatrix analyses is indeed much wider than for MSC analyses, and their supermatrix analyses recover

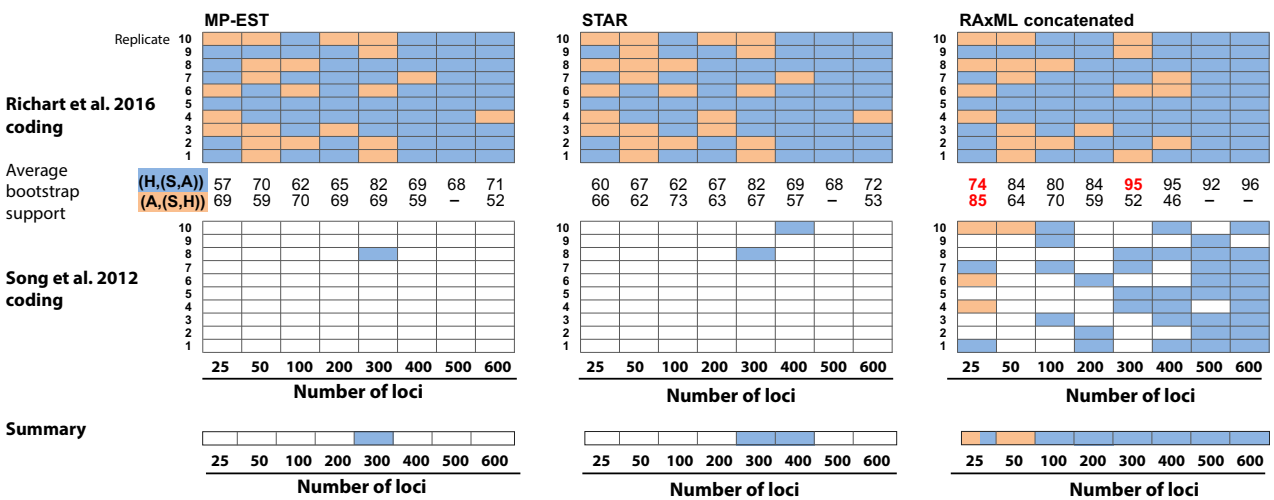

Fig. 4 Reinterpretation of phylogenomic subsampling results from Richart *et al.* (2016). On the top row is the coding of individual subsamples from Richart *et al.*'s analysis of spider phylogenomics. In the top row, cells are coloured according to which phylogeny, (H,S(A) in blue or (A,H)S) in orange, is favoured, regardless of bootstrap support for that replicate. A = Acuclavella, S = Sabacon, H = Hespernemastoma. In the lower row is the recoding of the subsampling cells according to the metric of Song *et al.* (2012), whereby cells receive no colour coding if the bootstrap support for that replicate was less than 90%. A cell receives colour coding according to the subtree favoured by that replicate if the bootstrap support was ≥90%. Clear evidence of erratic behaviour of concatenation appears in subsamples of 25 and 50 genes, whereas no such erratic behaviour is observed for species tree approaches or by concatenation for larger samples of genes. We also recalculated average bootstrap support for the 10 replicates for each subsample in Richart *et al.* (2016), reported below the top row of matrices and found three discrepancies with their calculations, indicated in red.

© 2016 Royal Swedish Academy of Sciences, **45**, s1, October 2016, pp 63–74

situations in which both a clade and an incompatible alternative are recovered with high (>90%) support within and between the 25 and 50 gene replicates. Indeed, supermatrix analyses recover such flip-flopping at a higher rate than MSC methods. Thus, I feel that their rejection of our claim stated above is not justified. Such flip-flopping of concatenation analyses, particularly when analysing small numbers of genes, was recognized in early subsampling studies (e.g. Rokas *et al.* 2003).

We agree that their analysis is broadly, but not fully, consistent with the idea that in some cases, supermatrix analyses will converge to phylogenomic certainty (high support for a given clade) more quickly than will MSC methods. This tendency has been known for several years (Edwards *et al.* 2007; Bayzid & Warnow 2013), yet it should not necessarily be taken as a 'win' for concatenation, because without deeper study one never knows whether full support for a clade represents true or spurious signal. Another example already mentioned in this context is support for Arachnida in Sharma *et al.* (2014), which received 100% bootstrap support in the partition of 500 slowest evolving genes, but whose non-monophyly also receives 100% bootstrap support as matrices surpass the 2000-locus mark. The fact that strong conflicting results are seen with concatenation only among the smaller subsamples (25 and 50 genes) may suggest a trend, similar to predictions of Seo (2008), but the reanalysis of Song *et al.* (2012) data (Fig. 2) suggests that subsamples of larger numbers of gene can still yield incongruent results.

Additionally, we have known for a while, and many simulation studies have shown, that when ILS is minimal, concatenation will recover the correct tree topology with greater efficiency than will MSC methods (Liu *et al.* 2015a; Mirarab *et al.* 2016). But such rapid approach to certainty as matrix sizes increase – you can call it 'efficiency' if you want – is not necessarily evidence of a method accurately reflecting the true rate of increase of certainty. We and many others have suggested that supermatrix approaches often unnaturally inflate support values because they represent strong violations of the MSC and of the conditional independence of loci that genomes exhibit even when ILS is low or absent (Edwards *et al.* 2007; Liu *et al.* 2010, 2015b,c; Xi *et al.* 2016). It is certainly comforting – but of course, potentially addicting – to achieve 100% support for a hypothesis that is topologically correct but should only receive 75% support given the data collected. Such a situation hardly engenders real confidence. The Richart *et al.* (2016) study also suffers from the very low number of taxa analysed; although they suggest that their results are driven primarily by biological signals, they do not rule out statistical artefacts, and, with only three in-group taxa, we agree that long-branch attraction and other artefacts are a worry.

Examining their analysis from the contrasting perspective of MSC methods, the tendency in the Richart *et al.* (2016) analysis for MSC methods to recover many clades across replicates only weakly and their failure to show increasing support across replicates as matrix size increases could be taken as a positive, because there is by necessity no 'flip-flopping' or if the topology recovered by concatenation is wrong, a conclusion that we agree is unlikely. The pattern of support for MSC analyses in their study could also mean that the assumptions of the MSC have been violated. Indeed, Richart *et al.* (2016) go to great lengths to demonstrate violations of the MSC model in their data, by exploring triplet frequencies, for example a practice that we encourage. Although we dispute their conclusion that such violations of the MSC vindicate supermatrix approaches (see Edwards *et al.* 2016a), we believe such testing of basic predictions of the MSC model are a positive step for phylogenomics (see also Tarver *et al.* 2016). In the end, their subsampling analysis also reveals the same tendency for supermatrix analyses to flip-flop as shown in Song *et al.* (2012), but their focus on average support values tends to mask the underlying erratic behaviour in some replicates. We encourage the community, as well as software focused on subsampling (e.g. Narechania *et al.* 2012) to report the full range of, as well as average for, support values in subsampled analyses.

Conclusion

Phylogenomic subsampling should become a standard practice in phylogenomics, at least until we have discovered and can potentially predict trends and patterns of support in phylogenomic data sets. For the first time in phylogenetics, we are assembling data sets that are large enough to allow subsampling in a way that robustly explores the effects of different parameters on phylogenomic results. We have seen how phylogenomic subsampling has had several uses throughout its brief history: it has been used to study the effect of missing data on phylogenomic results, to measure the effect of increasing evolutionary rates on phylogenomic results and to compare the consistency of different methods of phylogenomic analysis. Given the increasing ease of producing relatively full matrices, we believe phylogenomic subsampling has particularly important uses with regard to the latter two goals. Phylogeneticists can explore the impact of matrix size, composition and evolutionary rate using a subsampling design that controls for each of these while exploring variation in other variables. We believe that both random subsampling of phylogenomic matrices and ordered addition of loci to increasingly large matrices (Narechania *et al.* 2012; Simon *et al.* 2012) have a place in modern phylogenomics. An important use of

phylogenomic subsampling is likely as a test for the consistency of a particular phylogenetic method across subsampled matrices of different size and composition. Such a practice is undertaken under the reasonable assumption that strength of support for a given clade should increase with increasing matrix size, and should not jump erratically between high support for multiple phylogenetic hypotheses. In the pre-next-generation era, data sets of even a few tens of loci were achieved with such painful effort that the results of analysis of the largest matrix possible were usually taken as the best estimate. Now, in the next-generation era, matrices are so large that we can envision situations where smaller (perhaps more complete or less biased) matrices may actually perform better than large ones that include conspicuous gaps or a greater diversity of conflicting signals. Phylogenomic subsampling is a useful addition to other increasing practices, such as data filtering, whether based on parameters of the loci or on some *a priori* phylogenetic hypothesis (Narechania *et al.* 2012; Chen *et al.* 2015). At the very least, the writing of this brief review, and especially my fortunate attendance at the *Zoologica Scripta* meeting in Oslo, revealed a case of parallel evolution between the phylogenomic practices of vertebrate and invertebrate systematists – even those working under the same roof in Cambridge!

Acknowledgements

I thank the organizers for inviting me to the *Zoologica Scripta* symposium, Prashant Sharma, Charles Davis, Cheryl Hayashi and Zhenxiang Xi for helpful comments on the manuscript, and Liang Liu, Charles Davis, Laura Kubatko and Shaoyuan Wu for helpful discussion. Casey Richart generously provided details of the subsampling performed in Richart *et al.* (2016). Liang Liu and Shaoyuan Wu generously reanalysed the data for Figs 2 and 3, which are associated with Wu *et al.* 2015. Zhenxiang Xi helped produce and recalculate values for Fig. 4. Lily Lu generously tabulated the results of the subsampling analyses into spreadsheets. Phylogenomics work in the Edwards lab is supported by the US National Science Foundation and Harvard University.

References

Bayzid, M. S. & Warnow, T. (2013). Naive binning improves phylogenomic analyses. *Bioinformatics*, *29*, 2277–2284.

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R. & Moritz, C. (2013). Unlocking the vault: next generation museum population genomics. *Molecular Ecology*, *22*, 6018–6032.

Bragg, J. G., Potter, S., Bi, K. & Moritz, C. (2015). Exon capture phylogenomics: efficacy across scales of divergence. *Molecular Ecology Resources*, *16*, 1059–1068.

Brandley, M. C., Bragg, J. G., Singhal, S., Chapple, D. G., Jennings, C. K., Lemmon, A. R., Lemmon, E. M., Thompson, M. B. & Moritz, C. (2015). Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian *Eugongylus* group scincid lizards. *BMC Evolutionary Biology*, *15*, 1.

Bremer, K. (1994). Branch support and tree stability. *Cladistics*, *10*, 295–304.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, *29*, 1917–1932.

Chen, M. Y., Liang, D. & Zhang, P. (2015). Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. *Systematic Biology*, *64*, 1104–1120.

Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Saun, L.é., Genson, G., Dubois, E., Nidelet, S., Deuve, T. & Rasplus, J.-Y. (2014). Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution*, *31*, 1272–1274.

Davis, J. I., Frohlich, M. W. & Soreng, R. J. (1993). Cladistic characters and cladogram stability. *Systematic Botany*, *18*, 188–196.

DeGiorgio, M., Syring, J., Eckert, A. J., Liston, A., Cronn, R., Neale, D. B. & Rosenberg, N. A. (2014). An empirical evaluation of two-stage species tree inference strategies using a multilocus dataset from North American pines. *BMC Evolutionary Biology*, *14*, 67.

Degnan, J. H. & Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *Public Library of Science Genetics*, *2*, 762–768.

Degnan, J. H. & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference, and the multispecies coalescent. *Trends in Ecology & Evolution*, *24*, 332–340.

Driskell, A. C., Ane, C., Burleigh, J. G., McMahon, M. M., O'Meara, B. C. & Sanderson, M. J. (2004). Prospects for building the tree of life from large sequence databases. *Science*, *306*, 1172–1174.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q. & Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, *452*, 745–749.

Edwards, S. V. (2016). Inferring species trees. In R. Kliman (Ed.) *Encyclopedia of Evolutionary Biology* (pp. 236–244). New York: Elsevier Inc.

Edwards, S. V., Liu, L. & Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 5936–5941.

Edwards, S. V., Xi, Z. X., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B. J., Wu, S. Y., Lemmon, E. M., Lemmon, A. R., Leache, A. D., Liu, L. & Davis, C. C. (2016a). Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, *94*, 447–462.

Edwards, S. V., Potter, S., Schmitt, C. J., Bragg, J. G. & Moritz, C. (2016b). Reticulation, divergence, and the phylogeography–phylogenetics continuum. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 8025–8032.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, *39*, 783–791.

Gatesy, J., Milinkovitch, M., Waddell, V. & Stanhope, M. (1999a). Stability of cladistic relationships between Cetacea and higher-level Artiodactyl taxa. *Systematic Biology*, *48*, 6–20.

Gatesy, J., O'Grady, P. & Baker, R. H. (1999b). Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics*, *15*, 271–313.

Giribet, G. & Wheeler, W. (2003). Sensitivity analysis in cladistic analyses: Navaho rugs, stability, and their relationship to nodal support. In Abstracts of the 21st Annual Meeting of the Willi Hennig Society. pp. 148–163. Available via http://doi.wiley.com/10.1016/S0748-3007(02)00152-4.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., Martinez, P., Baguna, J., Bailly, X., Jondelius, U., Wiens, M., Muller, W. E. G., Seaver, E., Wheeler, W. C., Martindale, M. Q., Giribet, G. & Dunn, C. W. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B*, *276*, 4261–4270.

Huang, H. T. & Knowles, L. L. (2009). What is the danger of the anomaly zone for empirical phylogenetics? *Systematic Biology*, *58*, 527–536.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J. W., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M.S., Zavidovych, V., Subramanian, S., Gabaldon, T., Capella-Gutierrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W.C., Ray, D., Green, R.E., Bruford, M.W., Zhan, X.J. & Dixon, A., Li, S.B., Li, N., Huang, Y.H., Derryberry, E.P., Bertelsen, M.F., Sheldon, F.H., Brumfield, R.T., Mello, C.V., Lovell, P.V., Wirthlin, M., Schneider, M.P.C., Prosdocimi, F., Samaniego, J.A., Velazquez, A.M.V., Alfaro-Nunez, A., Campos, P.F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D.M., Zhou, Q., Perelman, P., Driskell, A.C., Shapiro, B., Xiong, Z.J., Zeng, Y.L., Liu, S.P., Li, Z.Y., Liu, B.H., Wu, K., Xiao, J., Yinqi, X., Zheng, Q.M., Zhang, Y., Yang, H.M., Wang, J., Smeds, L., Rheindt, F.E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F.K., Jonsson, K.A., Johnson, W., Koepfli, K.P., O'Brien, S., Haussler, D., Ryder, O.A., Rahbek, C., Willerslev, E., Graves, G.R. & Glenn, T.C., McCormack, J., Burt, D., Ellegren, H., Alstrom, P., Edwards, S.V., Stamatakis, A., Mindell, D.P., Cracraft, J., Braun, E.L., Warnow, T., Jun, W., Gilbert, M.T.P. & Zhang, G.J. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, *346*, 1320–1331.

Jiang, W., Chen, S. Y., Wang, H., Li, D. Z. & Wiens, J. J. (2014). Should genes with missing data be excluded from phylogenetic analyses? *Molecular Phylogenetics and Evolution*, *80*, 308–318.

Jones, M. R. & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*, 185–202.

Katz, L. A. & Grant, J. R. (2015). Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. *Systematic Biology*, *64*, 406–415.

Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, *62*, 611–615.

Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. N. M. & Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology*, *64*, 1032–1047.

Lemmon, E. M. & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 99–121.

Lemmon, A. R., Emme, S. A. & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, *61*, 727–744.

Liang, D., Shen, X. X. & Zhang, P. (2013). One thousand two hundred ninety nuclear genes from a genome-wide survey support lungfishes as the sister group of tetrapods. *Molecular Biology and Evolution*, *30*, 1803–1807.

Linkem, C. W., Minin, V. N. & Leaché, A. D. (2016). Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). *Systematic Biology*, *65*, 465–477.

Liu, L. & Yu, L. (2010). Phybase: an R package for species tree analysis. *Bioinformatics*, *26*, 962–963.

Liu, L., Yu, L., Pearl, D. K. & Edwards, S. V. (2009a). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, *58*, 468–477.

Liu, L., Yu, L., Kubatko, L., Pearl, D. K. & Edwards, S. V. (2009b). Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, *53*, 320–328.

Liu, L., Yu, L. & Edwards, S. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, *10*, 302.

Liu, L., Wu, S. & Yu, L. (2015a). Coalescent methods for estimating species trees from phylogenomic data. *Journal of Systematics and Evolution*, *53*, 380–390.

Liu, L., Xi, Z., Wu, S., Davis, C. C. & Edwards, S. V. (2015b). Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences*, *1360*, 36–53.

Liu, L., Xi, Z. X. & Davis, C. C. (2015c). Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. *Molecular Biology and Evolution*, *32*, 791–805.

Mirarab, S., Bayzid, M. S., Boussau, B. & Warnow, T. (2014). Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, *346*, 1250463. doi: 10.1126/science.1250463.

Mirarab, S., Bayzid, M. S. & Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, *65*, 366–380.

Narechania, A., Baker, R. H., Sit, R., Kolokotronis, S.-O. & DeSalle, R. P. P. (2012). Random addition concatenation analysis: a novel approach to the exploration of phylogenomic signal reveals strong agreement between core and shell genomic partitions in the cyanobacteria. *Genome Biology and Evolution*, *4*, 30–43.

Potter, S., Bragg, J., Peter, B. M., Bi, K. & Moritz, C. (2016). Phylogenomics at the tips: inferring lineages and their

demographic history in a tropical lizard, *Carlia amax*. *Molecular Ecology*, *25*, 1367–1380. in press.
Rannala, B. & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, *164*, 1645–1656.
Rannala, B. & Yang, Z. H. (2008). Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics*, *9*, 217–231.
Richart, C. H., Hayashi, C. Y. & Hedin, M. (2016). Phylogenomic analyses resolve an ancient trichotomy at the base of Ischyropsalidoidea (Arachnida, Opiliones) despite high levels of gene tree conflict and unequal minority resolution frequencies. *Molecular Phylogenetics and Evolution*, *95*, 171–182.
Robinson, D. F. & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*, 131–147.
Roch, S. & Warnow, T. (2015). On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic Biology*, *64*, 663–676.
Rokas, A., Williams, B., King, N. & Carroll, S. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, *425*, 798–804.
Rosenberg, N. A. (2013). Discordance of species trees with their most likely gene trees: a unifying principle. *Molecular Biology and Evolution*, *30*, 2709–2713.
Seo, T. K. (2008). Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution*, *25*, 960–971.
Seo, T. K., Kishino, H. & Thorne, J. L. (2005). Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 4436–4441.
Sharma, P. K., Stefan, T., Alicia, P.-P. R., Gonz, V. L., Hormiga, G., Wheeler, W. C. & Giribet, G. (2014). Phylogenomic interrogation of arachnida reveals systemic conflicts phylogenetic signal. *Molecular Biology and Evolution*, *31*, 2963–2984.
Sharma, P. P., Fernandez, R., Esposito, L. A., Gonzalez-Santillan, E. & Monod, L. (2015). Phylogenomic resolution of scorpions reveals multilevel discordance with morphological phylogenetic signal. *Proceedings of the Royal Society B*, *282*, 20142953. doi: 10.1098/rspb.2014.2953
Simmons, M. P., Sloan, D. B. & Gatesy, J. (2016). The effects of subsampling gene trees on coalescent methods applied to ancient divergences. *Molecular Phylogenetics and Evolution*, *97*, 76–89.
Simon, S., Narechania, A., DeSalle, R. & Hadrys, H. (2012). Insect phylogenomics: exploring the source of incongruence using new transcriptomic data. *Genome Biology and Evolution*, *4*, 1295–1309.
Song, S., Liu, L., Edwards, S. V. & Wu, S. Y. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 14942–14947.
Springer, M. S. & Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, *94*, 1–33.
Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*, 2688–2690.
Sukumaran, J. & Holder, M. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, *26*, 1569–1571.
Sukumaran, J. & Holder, M. (2015). SumTrees: phylogenetic tree summarization, version 4.0.0. https://github.com/jeetsukumaran/DendroPy
Sun, L., Fang, L., Zhang, Z., Chang, X., Penny, D. & Zhong, B. (2016). Chloroplast phylogenomic inference of green algae relationships. *Scientific Reports*, *6*, 20528.
Tarver, J. E., dos Reis, M., Mirarab, S., Moran, R. J., Parker, S., O'Reilly, J. E., King, B. L., O'Connell, M. J., Asher, R. J., Warnow, T., Peterson, K. J., Donoghue, P. C. J. & Pisani, D. (2016). The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biology and Evolution*, *8*, 330–344.
Wiens, J. J. (2006). Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics*, *39*, 34–42.
Wu, S., Liu, L. & Edwards, S. V. (2015). Correction for Song et al., Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, E6079 Available via http://www.pnas.org/lookup/doi/10.1073/pnas.1518753112.
Xi, Z., Liu, L., Rest, J. S. & Davis, C. C. (2014). Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. *Systematic Biology*, *63*, 919–932.
Xi, Z. X., Liu, L. & Davis, C. C. (2016). The impact of missing data on species tree estimation. *Molecular Biology and Evolution*, *33*, 838–860.
Zimmer, E. A. & Wen, J. (2015). Using nuclear gene data for plant phylogenetics: progress and prospects II. Next-gen approaches. *Journal of Systematics and Evolution*, *53*, 371–379.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Bootstrap values of specific clades from Figs. 2 and 3 for concatenation trees and species trees (MP-EST) generated during 10 replicates of subsampling. Two types of subsampling were used: single bootstrap (for concatenation and MP-EST) and double bootstrap (MP-EST). For convenience, the bootstrap values greater than or equal to 90% are highlighted in pink.