Point of View

Syst. Biol. 55(3):522–529, 2006
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150600697358

Resolution of Phylogenetic Conflict in Large Data Sets by Increased Taxon Sampling

SHANNON M. HEDTKE,[1] TED M. TOWNSEND,[1,2] AND DAVID M. HILLIS[1]

[1]*Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas Austin, Austin, Texas 78712, USA; E-mail: s.hedtke@mail.utexas.edu (S.M.H)*
[2]*Current Address: Department of Biology and Center for Applied and Experimental Genomics, San Diego State University, San Diego, California 92182, USA*

The debate about whether phylogenetic accuracy is most efficiently increased by sampling more characters or more taxa is certainly not new (e.g., Kim, 1996; Graybeal, 1998; Poe, 1998a,b; Rannala et al., 1998; Poe and Swofford, 1999; Pollock and Bruno, 2000; Rosenburg and Kumar, 2001; Pollock et al., 2002; Zwickl and Hillis, 2002; Rosenberg and Kumar, 2003; Hillis et al., 2003). However, the recent increase of whole genomic sequences available from an assortment of distantly related taxa makes this debate highly relevant to researchers across fields of biology. Recently, Rokas et al. (2003) argued that the true species tree can be recovered despite conflicting phylogenetic signal between genes if enough genes are used in the analysis. Using the bootstrap proportion (BP) as a measure of phylogenetic accuracy, they concluded that approximately 20 genes are needed to ensure a robustly supported tree (>95% BP) for their study group of eight yeast taxa. From these empirical results, they generalized that most molecular phylogenetic studies have probably included insufficient numbers of genes to confidently resolve relationships within their respective focal groups.

This approach to measuring accuracy can be sensitive to method inconsistency, or the failure to converge on the correct tree as the data set becomes infinitely large. When a method is inconsistent, measures of support such as nonparametric bootstrapping can increase as more sequence data are added—but in support of the wrong phylogeny (Phillips et al., 2004; Collins et al., 2005; Delsuc et al., 2005). Although most methods perform well over most of tree space (Huelsenbeck, 1995; Poe, 2003), regions of inconsistency have been identified in the literature for all of the most commonly used phylogenetic methods. For example, compositional bias can affect the accuracy of minimum evolution (Phillips et al., 2004), model misspecification may affect parametric methods such as maximum likelihood (ML) (Poe, 2003; Philippe et al., 2005; Collins et al., 2005), and branch-length asymmetry can lead to inconsistency in maximum parsimony (Felsenstein, 1978; Hendy and Penny, 1989). Parsimony is particularly prone to long-branch attraction (LBA), an analytical artifact in which two taxa on long branches are incorrectly placed as sister taxa (Felsenstein, 1978; Hendy and Penny, 1989; Huelsenbeck and Hillis, 1993).

Although there are many reasons for conflicting phylogenetic signal between genes, one relevant reason could be related to method inconsistency: differing rates of evolution between genes could cause a particular method to be inconsistent for some genes and not for others. We argue that by addressing this source of conflict between genes, fewer genes may be needed to return an accurate phylogeny. One source of conflict in the Rokas et al. (2003) data set may be nonstationarity: taxa that differ from the others in their base compositional bias may be erroneously drawn together as sister taxa (Collins et al., 2005). Here, we show that an additional source of conflict between the 106 genes in the Rokas et al. data set may be branch-length asymmetry. Using simulations of 106 genes from the Rokas et al. data set on a 79-taxon yeast phylogeny, we additionally show that when genes are added to a data set, support for the wrong reconstruction can increase when there is LBA. However, when taxa are added to the analysis, support for the correct reconstruction increases, and fewer genes are needed to achieve accuracy.

LONG BRANCHES AND ROOTING THE YEAST TREE

It is instructive to place the taxa included by Rokas et al. (2003) in the context of a more intensively sampled yeast phylogeny (Fig. 1). We realigned and reanalyzed data from an eight-gene, 78-taxon study of the "*Saccharomyces* complex" (Kurtzman and Robnett, 2003), which included the ingroup taxa of Rokas et al., plus sequences for their outgroup, *Candida albicans*, obtained from GenBank (accession numbers AACQ01000295, AJ508555, X70659, AF455531, M29935, 002653, AF285261, X16377, AY497614). The tree for analysis was generated using maximum likelihood as an optimality criterion (program GARLI; D. Zwickl, University of Texas, Austin; available at

FIGURE 1. Tree topology from maximum likelihood analysis of 79 yeast. Arrows denote taxa used in Rokas et al. (2003). Four taxa used in our initial analyses are indicated by bold branches.

TABLE 1. Categories of topological discordance between 106 individual genes and concatenated data set of Rokas et al. (2003).

Category	Maximum parsimony	Maximum likelihood
No discordance	37	41
Discordance in rooting of ingroup	13	18
Discordance in rooting of *S. cerevisiae* clade	20	22
Discordance in rooting of ingroup and *S. cerevisiae* clade	25	23
Other discordance	11	2

http://www.zo.utexas.edu/faculty/antisense/Download.html). Additional TBR swapping and branch length optimization was performed in PAUP* (Swofford, 2002).

Initial inspection of individual maximum parsimony (MP) consensus and maximum likelihood (ML) trees for each of Rokas et al.'s (2003) 106 genes reveals that a large proportion (65% for MP, 61% for ML) fail to recover the final combined-data topology (Table 1). Of the 69 MP trees inconsistent with the combined-data topology, 38 (55.1%) differ in the rooting of the ingroup. Results under ML are very similar: 41 of 65 trees (63.1%) show the ingroup rooted on *S. castellii* rather than *S. kluyveri* (Fig. 2, Table 1). Correctly rooting an ingroup is dependent on inclusion of closely related outgroup taxa (Philippe, 1997). The outgroup used by Rokas et al., *C. albicans*, is distantly related to the seven ingroup taxa, based on branch lengths estimated by ML for each individual gene. The average branch length across genes from *C. albicans* to the root node of the ingroup was 2.35 substitutions/site (range 0.35–16.82; more than 75% were greater than 1.0; we excluded two outliers estimated to have branch lengths of 48.4 and 95.3 substitutions/site, respectively). We would expect most phylogenetic methods to have trouble inferring the root when the outgroup is on such a long branch. Our 79-taxon tree (Fig. 1) suggests several potentially better single outgroups for this group. *Saccharomycodes ludwigii*, for example, is outside the focal group of Rokas et al., and has an uncorrected "p" distance of only 0.047 from *S. kluyveri*, the most basal member of Rokas et al.'s ingroup. In contrast, the uncorrected distance from *C. albicans* to *S. kluyveri* is 0.118, over twice as large.

Our 79-taxon tree (Fig. 1) further illustrates the uneven coverage of species from the *Saccharomyces* group in the Rokas et al. study. Five of Rokas et al.'s seven ingroup taxa are closely related members of the small, highly nested *S. cerevisiae* crown clade, and the other two, *S. kluyveri* and *S. castellii*, are widely spaced on the remainder of the larger tree. Of the 69 MP trees incongruent with the combined-data topology, 45 (65.2%) contain an incorrectly rooted *S. cerevisiae* clade, and in the ML case 45 of 65 trees (69.2%) show this pattern (Fig. 2, Table 1). In all these cases, if the *S. cerevisiae* clade is rooted on the branch leading to *S. bayanus*, all other relationships within the clade are congruent with the combined-data topology. We are not asserting that every case of incongruent rooting in the Rokas et al. study was directly due to method inconsistency. Some individual genes contained as few as 390 base pairs, and a few yielded trees with extensive polytomies, suggesting that insufficient character sampling probably accounts for some of the aberrant rooting. Processes such as horizontal gene transfer, convergent selection, and incomplete lineage sorting are other possibilities. However, a large proportion of conflicts between individual and concatenated gene trees involve incongruent rooting at exactly the spots predicted to be problematic due to taxon sampling and method inconsistency (Fig. 1).

Of course, the taxa chosen by Rokas et al. (2003) were not chosen randomly, but rather were the only taxa from this group of yeasts for which complete genomic sequence was available. If more species had been available, they presumably would have been included. We are therefore not faulting their choice of taxa per se, nor are we arguing with Rokas et al.'s final topology: this topology was consistent across methods and agrees with the topology we estimated using additional taxa. However, Rokas et al. (2003) argue that a large number of genes is required in a phylogenetic analysis to overcome conflicting signals between genes and reveal the true topology. Here we explore another possibility: that smaller sets of genes can be just as effective given increased attention to taxon sampling.

Genes, Taxa, and Phylogenetic Accuracy

Previous research on the effects of taxon sampling on phylogenetic analyses of sequence data has taken four approaches: (1) comparisons of expected phylogenies based on morphology with those created using reduced versus expanded data sets (e.g., Philippe, 1997; Lin et al., 2002; Delsuc et al., 2003; Philippe et al., 2005); (2) subsampling taxa from a larger tree and comparing trees generated by the reduced taxon set to the full set (e.g., Lecointre et al., 1993; Graybeal, 1998; Poe, 1998b; Rokas et al., 2005); (3) analyzing simulated data and comparing results to the phylogeny used for simulation (e.g., Kim, 1996; Hillis, 1996, 1998; Rannala et al., 1998; Poe and Swofford, 1999; Pollock and Bruno, 2000; Pollock et al., 2002; Zwickl and Hillis, 2002; Poe, 2003; Rosenberg and Kumar, 2003); and (4) evolving organisms in the laboratory and comparing trees generated using different sampling schemes to the known, true phylogeny (Hillis et al., 1994; Cunningham et al., 1998; Poe, 1998a). All of these approaches contribute to our understanding of sampling strategies and method performance. For example, studies based on real data can examine the sensitivity of data sets to species sampling (Lecointre et al., 1993) without simplifying evolutionary processes. Studies that use experimental or simulated phylogenies can test accuracy because the true tree is known. For the purposes of this study, we treated our 79-taxon tree (Fig. 1) as the true yeast phylogeny and simulated all 106 genes from the Rokas et al. (2003) study on this tree.

Simulations were performed using Seq-Gen v. 1.2.7 (Rambaut and Grassly, 1997). Sequences were simulated using the maximum likelihood parameter estimates for the real gene under the best-fit model found by the

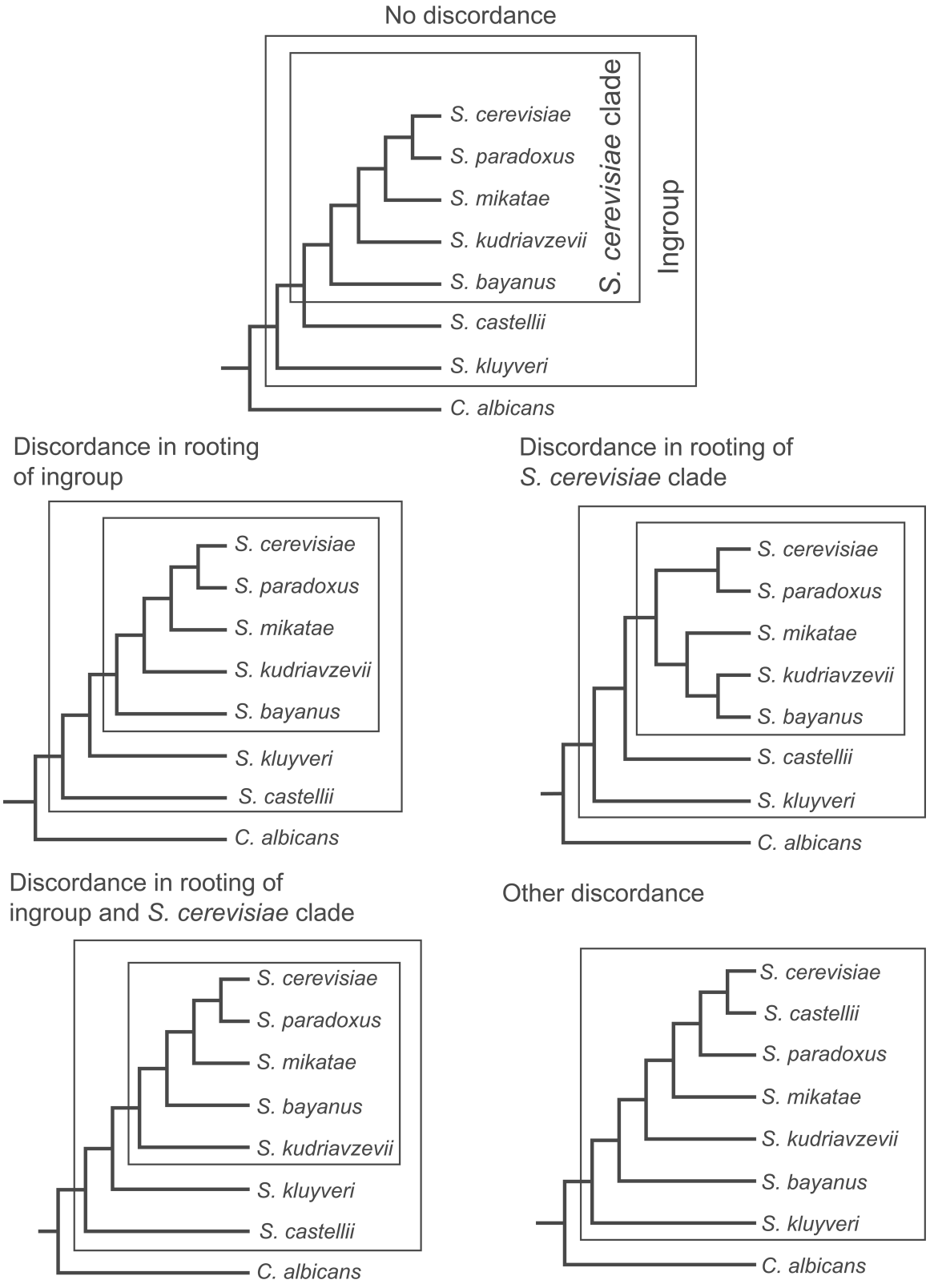

FIGURE 2. Examples of topological discordance in the Rokas et al. (2003) data set.

hierarchical likelihood ratio test using ModelTest v. 3.06 (Posada and Crandall, 1998). Each gene was simulated on the topology estimated for our 79-taxon tree, with branch lengths scaled individually for each gene to account for differences in relative substitution rates. Branch lengths were scaled by plotting the branch lengths of each eight-taxon tree based on the individual gene versus the branch lengths for the eight-taxon tree based on all genes concatenated, and using the best-fit line to estimate the expected branch length for each gene on the larger phylogeny.

Unfortunately, we could not test the impact of increased taxon sampling on the relationships that differed between genes in the Rokas et al. (2003) data set. In our 79-taxon tree, the length of the branch leading to *C. albicans* from the ingroup of the eight-taxon subsample (0.416) is much shorter than that estimated from the eight-taxon data set alone (1.554). As a result, the branch length between *C. albicans* and the ingroup for individual gene trees is also shorter in the simulated data set compared to the actual data set. Parsimony analyses of individual simulated genes did not result in conflicting relationships between *S. kluyveri, S. castellii,* and the crown group of the remaining five taxa.

Therefore, to examine the effect of taxon sampling on this data set, we selected a taxon quartet we suspected would be prone to LBA: *S. dairenensis, Kluyveromyces blattae, S. kluyveri,* and *C. albicans* (Fig. 1). For each of 10 replicates, we randomly selected 25 genes and 36 taxa to add to the analysis. Each replicate began with the four taxa selected above. We performed a parsimony analysis on these four taxa using one randomly chosen gene, added another randomly chosen gene and analyzed those two genes, and repeated up through 25 total genes. Next, we took that same set of 25 random genes, and added one taxon at a time to our selected taxon quartet, such that analyses were performed on four through 40 taxa for one through 25 genes (9000 total data sets). We used PAUP* (Swofford, 2002) to perform heuristic searches using parsimony as the optimality criterion, with TBR branch swapping, ten replicates, and random sequence addition. We ran 100 bootstrap pseudoreplicates and recorded the proportion of trees supporting each reconstruction for the initial taxon quartet. The optimality criterion and number of replicates were chosen to make our results comparable to those of Rokas et al. (2003).

For the initial four-taxon tree of *S. dairenesis, K. blattae, S. kluyveri,* and *C. albicans*, none of the 10 replicates had a bootstrap proportion (BP) >70 for the correct reconstruction, no matter how many genes were added to the analysis (Fig. 3). This is expected because we specifically selected a taxon quartet difficult to reconstruct. However, as genes are added to the analysis, the average BP for the *correct* reconstruction of the relationships between the four taxa decreases (Fig. 3)—in other words, bootstrap

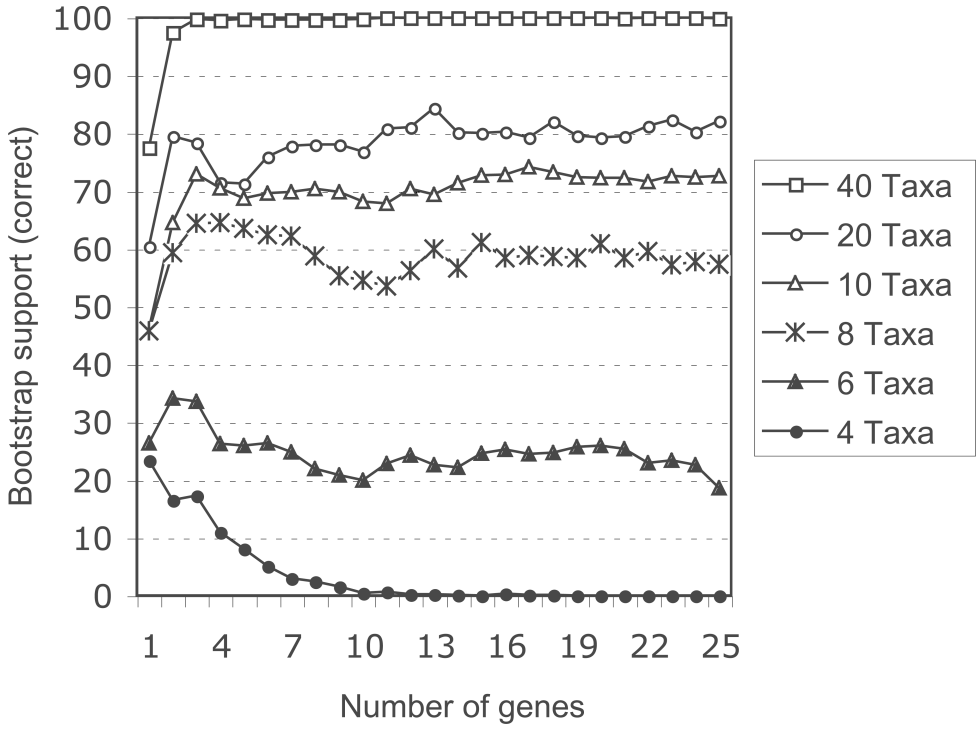

FIGURE 3. Average bootstrap support for the correct phylogenetic reconstruction of the four-taxon quartet ((*S. dairenensis, K. blattae*) (*S. kluyveri, C. albicans*)) over 10 simulated runs. When taxon sampling is poor, the average bootstrap value for the correct reconstruction goes down as more genes are added. Once taxon sampling is sufficient, the average bootstrap value increases as genes are added. Results with intermediate number of taxa and variances for bootstrap support values are available in Appendix 1 (available at http://systematicbiology.org).

support gets stronger for the *wrong* reconstruction as genes are added. Conversely, as the number of taxa randomly added to the analysis increases, the average BP for the correct reconstruction increases dramatically (Fig. 3). Thus, if we had gone into the analysis without knowing the relationships among taxa, we would not know whether a high BP represents confidence in the correct or incorrect reconstruction.

The minimum number of taxa required to obtain increased BP for the correct reconstruction ranges widely between replicates, from 6 to 22, because taxa are added randomly with respect to their relationships. This approach is intended to mimic reality, in which phylogenetic relationships between taxa are not known a priori. However, examination of each individual run reveals that the point at which adding genes begins to increase, rather than decrease, the BP for the correct reconstruction is only after one or both long branches have been broken by the addition of a new taxon. That is, taxa added to the analyses that break the internal branch of our taxon quartet, or break the relatively short branches leading to *S. dairenensis* or *S. kluyveri*, did not increase accuracy of reconstruction unless one or both of the long branches had already been broken.

The number of genes required to achieve BP greater than 95 for the correct reconstruction of our four-taxon quartet also varied between replicates (Table 2). Analyzing even as many as 25 genes did not ensure accurate reconstruction in all replicates if fewer than 26 taxa were included. A BP of at least 95 for the correct reconstruction was achieved in 90% of replicates when as few as three genes were used in the analysis, as long as taxon sampling was greater than 27 (Table 2). This is far fewer than the 20 genes suggested by Rokas et al. (2003) for their eight-taxon data set.

We argue, however, that the values for the appropriate number of genes or taxa for phylogenetic analysis are not generalizable. Three genes may be the correct number for this particular phylogenetic problem, but a different sequence set, different taxonomic group, or different method of analysis may require more than three genes to clearly resolve historical relationships, particularly if those genes support conflicting phylogenies due to hybridization events or convergent selection (e.g., Bull et al., 1997). Alternatively, for simpler problems, far fewer genes and far fewer taxa may be needed. For example, we ran six replicates using the taxon quartet *S. cerevisiae*, *C. castellii*, *K. yarrowii*, and *Zygosaccharomyces bisporus*. We expect this to be much easier to resolve based on their phylogenetic positions and relative branch lengths (Fig. 1). With only four taxa, two randomly chosen genes were sufficient to get 100% BP for the correct reconstruction of this taxon quartet (Appendix 2; available at http://systematicbiology.org). BP for the correct reconstruction for the more difficult taxon quartet of *S. dairenensis*, *Z. bisporus*, *K. blattae*, and *C. albicans* did not get above 95 until after 37 taxa and 7 genes were added (Appendix 2; available at http://systematicbiology.org). Additional quartets suspected of long-branch attraction demonstrated qualitatively similar results, although the number of taxa required varied (Appendix 2; available at http://systematicbiology.org).

TABLE 2. Number of taxa and simulated genes necessary to achieve bootstrap proportion (BP) of at least 95 for the *correct* reconstruction of the four-taxon statement (*S. dairenensis, K. blattae*) (*S. kluyveri, C. albicans*). When less than 26 taxa are used in the analysis, at least 1 out of 10 runs fails to achieve a BP of greater than 95. When less than 22 taxa are used, 25 genes are insufficient to return an accurate phylogeny in 9/10 replicates (indicated in the table with an x).

Number of taxa	No. of runs in which BP <95 using 25 genes	Range across runs of No. of genes needed to reach BP >95	No. of genes needed to reach BP >95 in 90% of runs
4	All 10	Not applicable	x
5	9	12	x
6	9	15	x
7	9	16	x
8	5	2–18	x
9	4	2–15	x
10	4	2–16	x
11	4	2–15	x
12	4	2–19	x
13	5	2–12	x
14	4	2–24	x
15	3	2–23	x
16	3	2–11	x
17	3	2–16	x
18	4	2–14	x
19	4	1–14	x
20	4	1–13	x
21	2	1–19	x
22	1	1–15	15
23	1	1–14	14
24–25	1	1–8	8
26	0	1–11	4
27	0	1–10	2
28	0	1–3	2
29	0	1–3	2
30	0	1–3	3
31	0	1–3	2
32	0	1–4	2
33	0	1–3	2
34	0	1–4	3
35	0	1–3	2
36	0	1–3	2
37	0	1–3	2
38	0	1–3	2
39	0	1–3	3
40	0	1–3	3

Finally, the addition of a single or a few taxa will not necessarily increase accuracy. In 3 of our 10 replicates, there were instances in which adding a taxon decreased phylogenetic accuracy, no matter how many genes were added (results similar to Poe and Swofford, 1999; Poe, 2003; Rokas and Carroll, 2005). Importantly, once sufficient additional taxa were also included, this trend reversed itself. Although we only examined the effect of taxon addition on one bipartition, we expect the same trend to hold across the tree as a whole (Hillis, 1996; Zwickl and Hillis, 2002).

We acknowledge that simulated data do not capture all the complexities of real evolutionary processes, and that empirical data sets may require more sequence data than suggested here. In addition, there are regions of tree space where adding taxa will not increase accuracy, but adding more characters will (Poe and Swofford, 1999).

Nevertheless, the phylogenetic conflict represented here is typical of genomic-scale data sets derived from model organisms, which are more likely to suffer from limited taxon sampling. In these cases, improved accuracy from increased taxon sampling is clear.

CONCLUSIONS

No particular number of genes or taxa will guarantee that phylogenetic reconstruction is accurate, even if bootstrap support for that reconstruction is high. If conflicting signals between genes are due to method inconsistency, adding more genes may lead to increasing support for the incorrect phylogenetic reconstruction. In such cases, increasing taxon representation may improve accuracy more than does increasing gene number. If we incorporate our understanding of sources of inconsistency into study design, resulting phylogenies are more likely to be representative of evolutionary history.

For any given study, how can an investigator know whether it is better to add more characters or add more taxa to a phylogenetic analysis? High support values for individual clades indicate that sufficient characters have been collected to converge on a robust result. Unfortunately, the well-supported result may be wrong, particularly if small trees with long branches are being estimated. This outcome appears to be especially likely when intensively sampled genomes have been selected across relatively few, distantly related species—as with model organisms. In such cases, any slight systematic bias can become magnified and misinterpreted as phylogenetic signal. High bootstrap or other support values are almost guaranteed with genome-sized character sets: the analyses will tend to converge on some answer, even if the answer has more to do with biases in the analysis than phylogenetic history. Therefore, it is important to investigate possible sources of systematic bias, such as long-branch attraction or model misspecification. Simulation studies can help determine the likelihood of long-branch attraction problems in these situations and suggest where additional taxon sampling should occur.

ACKNOWLEDGEMENTS

We thank A. Rokas and C. Kurtzman for providing the two yeast data sets used in this study. For comments and suggestions, we would like to thank T. Collins, two anonymous reviewers, and members of the University of Texas IGERT phylogenetics discussion group and Jim Bull lab group, particularly J. Brown, D. Cannatella, W. Harcombe, T. Heath, R. Heineman, M. Mahoney, R. Timme, J. Wagner, and D. Zwickl. Computational support for this research and a graduate research fellowship for SMH was provided by an NSF IGERT grant in Computational Phylogenetics and Applications to Biology awarded to the University of Texas, Austin. TMT was supported by an NSF Bioinformatics Postdoctoral Fellowship (DBI 0204451).

REFERENCES

Bull, J. J., M. R. Badgett, H. A. Wichman, J. P. Huelsenbeck, D. M. Hillis, A. Gulati, C. Ho, and I. J. Molineux. 1997. Exceptional convergent evolution in a virus. Genetics 147:1497–1507.

Collins, T. M., O. Fedrigo, and G. J. Naylor. 2005. Choosing the best genes for the job: The case for stationary genes in genome-scale phylogenetics. Syst. Biol. 54:493–500.

Cunningham, C. W., H. Zhu, and D. M. Hillis. 1998. Best-fit maximum likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52:978–987.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Delsuc, F., M. J. Phillips, and D. Penny. 2003. Comment on "Hexapod origins: Monophyletic or paraphyletic?" Science 301:1482.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47:9–17.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

Hillis, D. M. 1996. Inferring complex phylogenies. Nature 383:130–131.

Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3–8.

Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham. 1994. Application and accuracy of molecular phylogenies. Science 264:671–677.

Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124–126.

Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

Kim, J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45:363–374.

Kurtzman, C. P., and C. J. Robnett. 2003. Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. FEMS Yeast Res. 3:417–432.

Lecointre, G., H. Philippe, H. L. V. Le, and H. Le Guyader. 1993. Species sampling has a major impact on phylogenetic inference. Mol. Phyl. Evol. 2:205–224.

Lin, Y.-H., P. A. McLenachan, A. R. Gore, M. J. Phillips, R. Ota, M. D. Hendy, and D. Penny. 2002. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol. Biol. Evol. 19:2060–2070.

Philippe, H. 1997. Rodent monophyly: Pitfalls of molecular phylogenies. J. Mol. Evol. 45:712–715.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. Mol. Biol. Evol. 22:1246–1253.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Poe, S. 1998a. The effect of taxonomic sampling on accuracy of phylogeny estimation: Test case of a known phylogeny. Mol. Biol. Evol. 15:1086–1090.

Poe, S. 1998b. Sensitivity of phylogeny estimation to taxonomic sampling. Syst. Biol. 47:18–31.

Poe, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52:423–428.

Poe, S., and D. L. Swofford. 1999. Taxon sampling revisited. Nature 398:299–300.

Pollock, D. D., and W. J. Bruno. 2000. Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. Mol. Biol. Evol. 17:1854–1858.

Pollock, D. D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.

Posada, D., and K. A. Crandall. 1998. ModelTest: Testing the model of DNA substitution. Bioinformatics 14:817–818.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Rannala, B., J. P. Huelsenbeck, Z. Yang, and R. Nielsen. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702–710.

Rokas, A., and S. B. Carroll. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22:1337–1344.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Rosenberg, M. S., and S. Kumar. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98:10751–10756.

Rosenberg, M. S., and S. Kumar. 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52:119–124.

Swofford, D. L. 2002. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

First submitted 12 August 2005; reviews returned 5 December 2005; final acceptance 7 January 2006

Associate Editor: Tim Collins