Incomplete taxon sampling is not a problem for phylogenetic inference

Michael S. Rosenberg and Sudhir Kumar*

Department of Biology, Arizona State University, Tempe, AZ 85287-1501

Edited by John C. Avise, University of Georgia, Athens, GA, and approved July 24, 2001 (received for review May 17, 2001)

A major issue in all data collection for molecular phylogenetics is taxon sampling, which refers to the use of data from only a small representative set of species for inferring higher-level evolutionary history. Insufficient taxon sampling is often cited as a significant source of error in phylogenetic studies, and consequently, acquisition of large data sets is advocated. To test this assertion, we have conducted computer simulation studies by using natural collections of evolutionary parameters—rates of evolution, species sampling, and gene lengths—determined from data available in genomic databases. A comparison of the true tree with trees constructed by using taxa subsamples and trees constructed by using all taxa shows that the amount of phylogenetic error per internal branch is similar; a result that holds true for the neighbor-joining, minimum evolution, maximum parsimony, and maximum likelihood methods. Furthermore, our results show that even though trees inferred by using progressively larger taxa subsamples of a real data set become increasingly similar to trees inferred by using the full sample, all inferred trees are equidistant from the true tree in terms of phylogenetic error per internal branch. Our results suggest that longer sequences, rather than extensive sampling, will better improve the accuracy of phylogenetic inference.

Taxon sampling refers to the process of selecting representative taxa for a phylogenetic analysis. Nonexhaustive taxon sampling occurs for a number of reasons. Data may not be available from every extant species because of constraints of time, money, or rarity. In most cases, the number of potential species increases quickly if one is interested in phylogenetic relationships above the level of genus or family. Therefore, it is impractical, if not impossible, to sample every species from clades of interest. Rather, representative species from each clade are chosen and the reconstructed phylogenetic relationships of these species are taken to represent the evolutionary history of their respective clades.

Insufficient taxon sampling is often cited as a major source of error in phylogenetic analysis (e.g., refs. 1–10). However, as expected, the value of increasing the number of sequences (species) in a data set depends on the scope of sampling (11–14). Sampling within a fully framed monophyletic group may improve phylogenetic accuracy, but sampling outside of the group pushes the most recent common ancestor of the new set of taxa back in time and may decrease accuracy (13). Random sampling of additional taxa is thought to decrease, rather than increase, phylogenetic accuracy (12–14).

One reason why increased taxon sampling is thought to improve phylogenetic resolution is that it may counteract the "long branch attraction" problem, where long, unrelated branches may group together erroneously (15, 16). Increased taxon sampling may break long branches and help reduce the average branch length throughout the tree (13, 17–19). However, computer simulation results have been equivocal about the benefit of increased taxon sampling for reducing the long branch problem (11, 12, 19–21). The importance of extensive taxon sampling is already well established for estimating evolutionary parameters (4, 22, 23) and in independent contrasts (24).

There have also been a number of empirical studies on the value of taxon sampling on phylogenetic inference (2–10). These studies typically begin with a large number of species and then examine the results of analyzing subsamples; most have concluded that phylogenetic trees reconstructed with more taxa are more accurate than those inferred from fewer taxa. These conclusions assume that the phylogeny inferred by using the largest data set available is closest to the true tree; an assumption that is not well established, because the "true tree" is not known in empirical studies. At present these studies appear to have simply demonstrated that topologies reconstructed by using larger subsamples show higher congruence with the full tree. Therefore, this problem is most readily studied by computer simulation because the "true tree" is known. However, previous simulation and theoretical studies (11, 19–21) were often not conducted by subsampling from a large tree, as mentioned above, but rather began with a small number of species and progressively added additional species to long branches in the starting cluster, keeping the subsample tree fixed.

We conducted a simulation study motivated by issues an evolutionary biologist would encounter with real data. We began with a large predetermined phylogeny (as is the case with all empirical studies, the true tree of life having been fixed via evolution) and generated data sets consisting of sampled taxa from the "known" full phylogeny. In our simulations, we examined the problem of taxon sampling by using evolutionary rates, species representations, and gene length parameters for DNA and amino acid sequences derived from molecular sequence databases. In addition, we used model trees based on actual trees published in the literature, rather than an artificial tree created from a theoretical branching process or an artificial clustering scheme, in order to make our simulations an accurate representation of the topologies and distributions of branch lengths found in real data.

Materials and Methods

We used two different simulation schemes. For the first case, we chose the 66-taxa tree representing the phylogenetic relationships among Eutherian mammals from Murphy *et al.* (ref. 1; Fig. 1). The branch lengths represent the number of substitutions per site. This tree was chosen because it revises many well established beliefs about mammalian evolution (see also ref. 25). For instance, lagomorphs had previously been found to be distantly related to sciurognath rodents in analyses of large numbers of genes for a few taxa (e.g., ref. 26), and rodents were thought to be an outgroup to artiodactyls and primates (27). Murphy *et al.* (1) place lagomorphs in a monophyletic assemblage with rodents and identify the rodents as a sister group of primates to the

EVOLUTION

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: MP, maximum parsimony; ME, minimum evolution; NJ, neighbor joining; ML, maximum likelihood.

*To whom reprint requests should be addressed at: Life Sciences A-371, Department of Biology, Arizona State University, Tempe, AZ 85287-1501. E-mail: s.kumar@asu.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

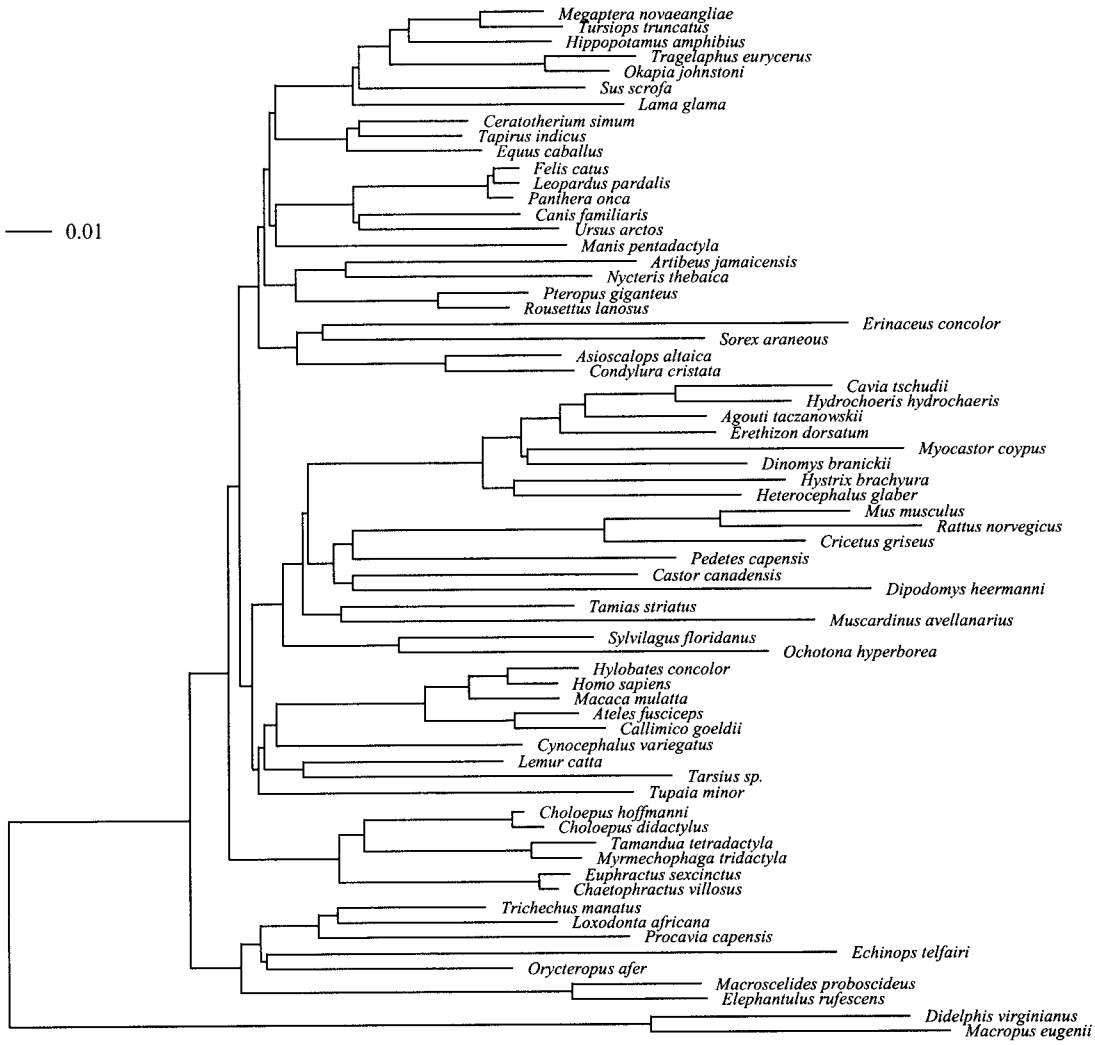

Fig. 1. Model tree used in the DNA simulations based on the Eutherian mammal tree from ref. 1. Branch lengths indicate the number of substitutions per site.

exclusion of artiodactyls. Therefore, we derived the topology of the tree in Fig. 1 by using their mammalian phylogeny. Note that Murphy *et al.*'s tree is based on a larger taxonomic sample than other studies, but has a fraction of the genes when compared with other studies that have smaller numbers of representative taxa (26, 27).

We simulated DNA evolution for 50 hypothetical genes for Fig. 1, each with independent evolutionary properties, using the Jukes–Cantor (28) model of nucleotide substitution [we also conducted analyses under the Hasegawa–Kishino–Yano model (29) and obtained similar results (data not shown)]. The sequence length and substitution rate were determined randomly for each gene. The sequence length was picked from a uniform distribution ranging from 500 to 3,000 (the range of sequence lengths commonly found in the literature). Because the branch lengths of the model tree (Fig. 1) already represent substitutions per site, the substitution rate for each gene represented a random multiplier of these branch lengths, picked from a gamma distribution with a gamma parameter of 1 (as observed from a data set of homologous human and mouse genes at only first and second codon positions). After simulating evolution across the full tree, a random subsample of taxa was chosen. The size of the subsample was randomly selected from a uniform distribution of 5 to 50, and the specific sampled taxa were selected randomly from the full complement of 66 taxa.

The model tree for the second set of simulations was based on an 18-taxa phylogeny of vertebrates (refs. 30–32; Fig. 2). These 18 taxa represent the most commonly found taxa in the genetic databases. There were 1,167 genes in our orthology database derived from HOVERGEN (33) with sequences for at least four of the taxa. The observed amino acid substitution rate and sequence length (100 to 2,696 sites) for each gene was used as the basis for simulating amino acid sequences on the tree by using the Poisson substitution model. Empirical substitution rates were estimated by using least-squares regression through the origin of pairwise sequence divergences and divergence times (34) for species in Fig. 2. Taxon sampling for each gene was determined by the availability of taxa for that gene in GenBank. Because some species are much more common in the database than others, this sampling is biased toward certain taxa and allows us to explore the effect of biased sampling as would be experienced by practicing biologists today. There were 100 simulation replicates for each gene for both the DNA and amino acid simulations. All simulations were conducted by using programs written by the authors.

All analyses were performed by using PAUP* Version 4.0b4a for Windows (35). Basic phylogeny reconstruction for both sets of simulations was performed by using neighbor joining (NJ), minimum evolution (ME), and maximum parsimony (MP) methods, as well as maximum likelihood (ML) for the DNA simula-

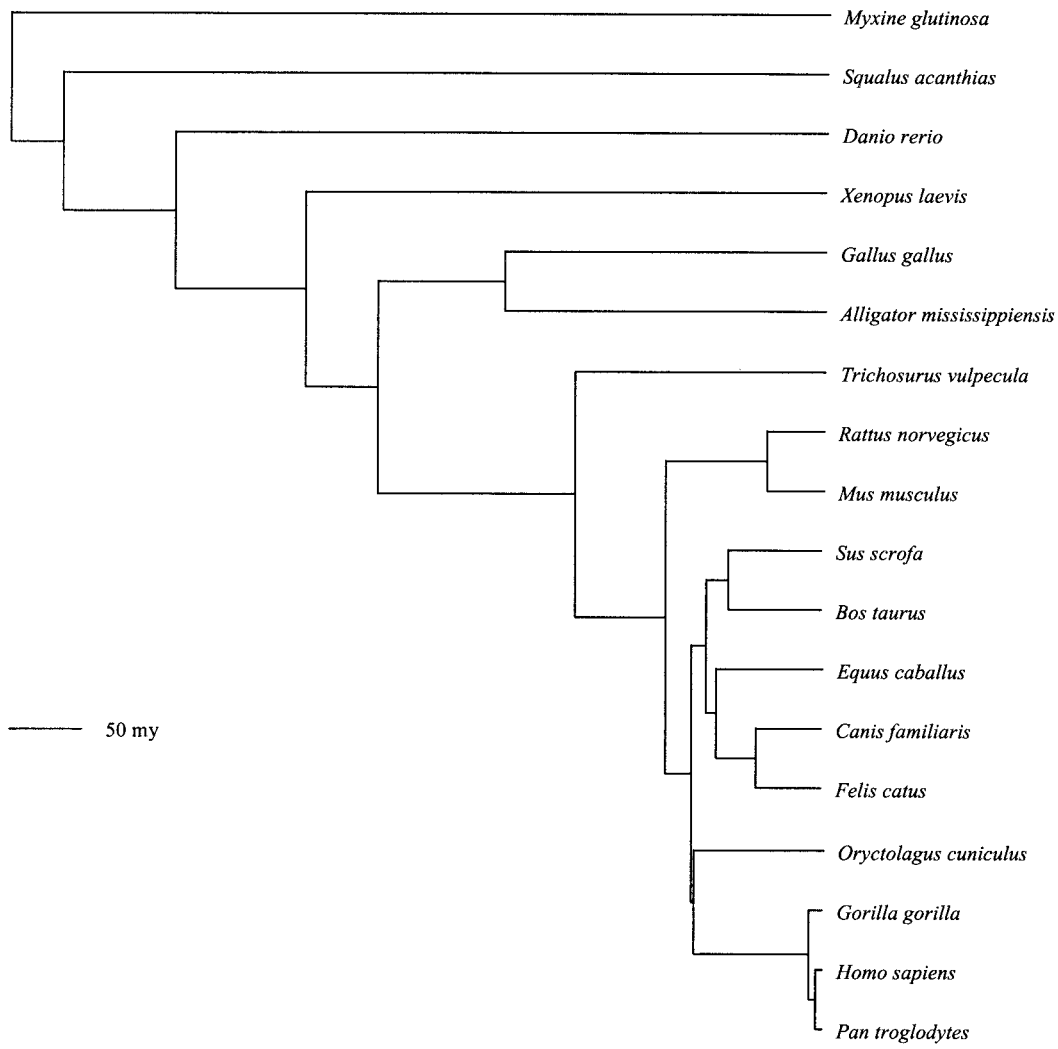

Fig. 2. Model tree used in the amino acid simulations based on a known vertebrate phylogeny. Branch lengths indicate time in millions of years.

tions only. Distances for NJ and ME were calculated under the JC model. In ME, MP, and ML a single heuristic search was performed with Nearest-Neighbor-Interchange branch swapping. For ME and ML, the NJ tree was used as the starting tree for the heuristic search; for MP, a stepwise addition procedure was used. A more exhaustive, time consuming search is not necessary because it is clear that it does not improve phylogenetic accuracy (36–39). The maximum number of trees that could be saved during the heuristic search procedures was set to 10,000 (most of the searches never came close to reaching this limit). When multiple trees were found under the ME, MP, and ML procedures, a majority rule consensus tree (retaining all compatible clades even under 50% frequency of occurrence with the LE50 option in PAUP*) was used to create a single resultant tree for each analysis. The resultant tree was then compared with the true (model) tree and the topological distance, d_T (40, 41), was recorded. This distance is twice the number of interior branches at which the two trees being compared differ. For subsample analyses, the topological distance between the true tree and the inferred subsample tree was computed by using the pruned true tree that contained only the subsampled taxa. Tree distance is not directly comparable among trees with different numbers of taxa, because d_T directly depends on the number of taxa (two unrooted trees of four taxa trees have a maximum d_T of 2, whereas two 66-taxa trees have a maximum d_T of 126). Therefore, we normalize the d_T value and define the phylogenetic error per internal branch, $E = d_T/2(n-3)$, where n is the number of taxa, and $n-3$ is the number of internal branches in a bifurcating tree. E ranges from 0 to 1, indicating that the proportion of internal branches inferred incorrectly. Other scaling metrics (e.g., scaling by the number of taxa) led to similar conclusions.

Phylogeny reconstruction was performed under a number of scenarios for each simulation. First, all of the sequences for all genes were concatenated into a single data set; the error between these inferred trees and the true tree is designated E_{concat}. Second, each gene was analyzed individually (gene-by-gene analysis) with all taxa included. In this case, we compute phylogenetic error (E_G) by directly comparing the true tree with the inferred full tree for the given gene. Third, each gene was analyzed individually with only the subsampled taxa for that gene included. In this case, the phylogenetic error (E_S) was computed by comparing the inferred subsample tree and true tree pruned to contain only the taxa in the subsample (Fig. 3).

Results and Discussion

DNA Sequence Evolution. In the DNA simulations, the 50 simulated genes consisted of a total of 79,410 sites (an average of 1,588 sites per gene). The average number of taxa subsampled was 28. The results for ME, NJ, MP, and ML were quite similar.

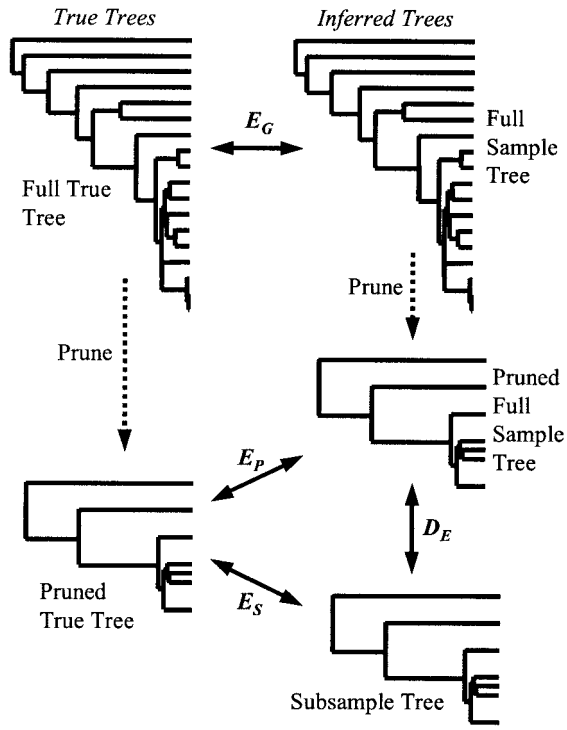

Fig. 3. Diagrammatic representation of the relationships among the true and inferred trees and error statistics. E_G is the phylogenetic error between the true tree and the full sample inferred tree; E_S is the phylogenetic error between the true tree and the subsample inferred tree; E_P is the phylogenetic error between the true tree and the full sample inferred tree, pruned to contain only the subsampled taxa; and D_E is the phylogenetic difference between the inferred subsample tree and the inferred full sample tree.

Table 1 lists the number of sampled taxa, number of sites, substitution rate, and results of the ME analysis for each gene; the overall results for all methods are summarized in Table 2. In this set of analyses it is clear that the ML method reconstructs the trees most accurately, followed by MP, then NJ and ME.

For the concatenated data set, the mean phylogenetic error (E_{concat}) was 2–3%, with the true tree inferred in about 17% of the replicates. In any case, the inferred tree contained only one or two incorrect partitions. For individual genes, the error (E_G) varied tremendously by gene, from 0.04 to 0.72 with a mean of 0.17 for the ME method. In general the true tree was never recovered by using individual genes. The variation among genes is due to a combination of both mutation rate and number of sites. Although both were important factors, the number of sites seems to have been more critical. For ME, the correlation between number of sites and E_G was -0.552, whereas the correlation of substitution rate and E_G was -0.326. As would be expected, there is no correlation ($r = 0.023$) between number of sites and substitution rate. For trees inferred with a subsample of taxa, the mean error (E_S) was 0.19. This value is very similar to that obtained from trees inferred by using the full set of taxa (cf. $E_G = 0.17$, Table 1). The correlation between the number of taxa and E_S ($r = -0.055$) was 10-fold lower than the correlation with sequence length.

To construct a direct comparison between the phylogenetic relationships of the subsampled taxa as obtained by using all taxa and only the subsample taxa, we pruned the inferred trees obtained by using all of the taxa to contain just the subsampled taxa and determined the phylogenetic error between this pruned inferred tree and the true tree (E_P). The mean phylogenetic error for ME was 0.16, which is again quite similar to E_G and E_S (Table 2). On average, more than doubling the number of taxa increased the percent of correct branches by only 2–3% (with an average subsample of 28, this increase represents less than a single branch). Note that even though E_S is greater than E_G and E_P for very small subsamples (<10 taxa), the difference in phylogenetic error is usually much smaller than one branch per tree. Therefore, use of only a fraction of taxa provides practically indistinguishable results. The similarities among E_G, E_S, and E_P persist

Table 1. Results of minimum evolution phylogenetic analysis for the DNA simulations

Number of taxa	Number of sites	Rate	E_G	E_S	E_P
24	202	1.094	0.36	0.39	0.36
34	211	1.156	0.35	0.40	0.36
41	217	1.068	0.35	0.41	0.38
23	308	0.853	0.29	0.36	0.33
30	343	0.216	0.48	0.48	0.47
32	360	0.063	0.72	0.77	0.77
37	566	2.204	0.17	0.20	0.16
36	660	0.522	0.20	0.24	0.22
6	684	2.300	0.16	0.15	0.07
7	688	0.326	0.25	0.43	0.29
37	706	1.486	0.15	0.15	0.12
38	741	0.873	0.15	0.20	0.18
50	853	1.689	0.13	0.13	0.12
46	889	0.087	0.44	0.47	0.47
30	889	0.920	0.13	0.17	0.12
25	941	0.394	0.17	0.24	0.19
5	1044	1.182	0.10	0.23	0.03
41	1145	1.517	0.09	0.11	0.08
10	1196	0.102	0.34	0.27	0.26
35	1262	6.523	0.21	0.16	0.14
31	1300	0.259	0.18	0.24	0.22
29	1370	1.092	0.09	0.07	0.04
34	1375	1.066	0.08	0.08	0.07
23	1378	0.869	0.09	0.16	0.13
45	1384	4.895	0.14	0.11	0.11
49	1488	1.539	0.07	0.08	0.06
39	1597	2.770	0.08	0.09	0.07
6	1692	1.119	0.07	0.00	0.00
26	1850	0.638	0.07	0.10	0.06
39	1859	0.080	0.29	0.28	0.28
35	2084	2.222	0.06	0.05	0.04
10	2085	1.004	0.06	0.07	0.01
38	2126	0.390	0.08	0.11	0.11
32	2163	2.222	0.06	0.10	0.08
16	2230	4.268	0.07	0.13	0.05
9	2244	0.158	0.15	0.10	0.07
40	2285	1.834	0.05	0.05	0.05
31	2463	1.163	0.05	0.08	0.06
45	2533	2.297	0.04	0.06	0.05
40	2561	0.189	0.12	0.13	0.12
9	2588	0.034	0.40	0.61	0.56
13	2604	0.690	0.05	0.05	0.03
46	2643	2.284	0.04	0.05	0.04
11	2652	0.041	0.36	0.37	0.37
16	2688	0.838	0.05	0.10	0.05
6	2742	2.259	0.04	0.09	0.00
6	2749	0.631	0.05	0.05	0.00
27	2877	0.556	0.05	0.03	0.02
11	2919	1.997	0.04	0.10	0.02
49	2976	0.166	0.11	0.12	0.11

Tree distances are averaged over 100 simulations. The variables are described in the text and Fig. 3. Each row represents a simulated gene, sorted by number of sites.

Table 2. Summary of the results from the DNA simulations

	E_{concat}	E_G	E_S	E_P	r_{rate}	r_{sites}
ME	0.02	0.17	0.19	0.16	−0.326	−0.552
MP	0.03	0.12	0.16	0.11	−0.105	−0.618
NJ	0.03	0.17	0.18	0.16	−0.332	−0.557
ML	—	0.10	0.12	0.09	−0.319	−0.599

The variables are described in the text and Fig. 3. Values are averaged over 100 simulations; the values for E_G, E_S, and E_P are also averaged across 50 genes. r_{rate} is the correlation of E_G with substitution rate; r_{sites} is the correlation of E_G with number of sites. The concatenated analyses were not performed by using ML method because of the excessive time required for their completion.

in simulations using more complex models of nucleotide substitution—e.g., the Hasegawa–Kishino–Yano (29) model (results not shown)—and are thus not a function of using a simple substitution model.

Amino Acid Sequence Evolution. In the amino acid simulations, the 1,167 simulated genes consisted of a total of 464,990 sites (an average of 398.5 sites per gene). The average number of taxa subsampled was 4.7, following the presence of species available for different genes in GenBank. ME, MP, and NJ gave similar results, which are summarized in Table 3. In these analyses, MP was the most accurate, followed by NJ and ME.

For the concatenated data set, we were always able to recover the correct tree with all of the data. For ME, the mean error for individual genes (E_G) ranged from 0.02 to 0.84, with a mean of 0.18. Again, the number of sites seems to have been more critical than the substitution rate. The correlation between number of sites and E_G was −0.459, whereas the correlation of substitution rate and E_G was −0.356.

The mean E_S was 0.10, which is lower than the 0.18 for the full complement of taxa (E_G). At first glance, it seems to imply that the subsample trees (those containing fewer taxa) were more accurate than the trees containing all of the taxa! Clearly, this observation is unexpected and contrary to the often assumed benefit of increased taxon sampling. As was done for DNA simulations, we test the validity of this result by comparing the amount of phylogenetic error in inferring phylogenetic relationships of the subsampled taxa as obtained by using all taxa and only the subsample taxa (E_P, Fig. 3). E_P was almost identical to E_S. The differences of E_S and E_G are explained by the fact that most of the errors in the trees inferred by using all taxa were those in the most basal three or four taxa (Fig. 2), which were almost always absent from the taxon-sampled trees (because of the nature of the current gene sequence databases). Therefore, the similarity of the E_S and E_P values make it clear that the subsample and full taxon set analyses were able to reconstruct trees with equivalent accuracy even though the average number of taxa in the subsample was less than five.

Taxon Sampling Versus Phylogenetic Signal. In general, the above results indicate that incomplete taxon sampling has a much smaller effect on the accuracy of a phylogeny as compared with the number of sites and substitution rates. It is therefore unwarranted to simply dismiss undesired or unexpected phylogenetic relationships obtained by using small numbers of taxa as the result of poor taxon sampling. Poor character sampling with weak phylogenetic signal is more likely to be the cause. In our study, taxon sampling had similar effects on phylogeny reconstruction for all of the major reconstruction methods. When the signal was strong, all of the methods reproduced the correct tree; when the signal was weak, none of them did. The one major difference among the different reconstruction methods in this study is their relationship with substitution rate and sequence length. All methods showed a stronger correlation between reconstruction accuracy and the number of sites than between accuracy and the substitution rate (see also ref. 37). This effect was strongest in MP, which consistently showed both a higher effect of number of sites and a lower effect of rate than any other method (Tables 2 and 3).

Table 3. Summary of the results from the amino acid simulations

	E_{concat}	E_G	E_S	E_P	r_{rate}	r_{sites}
ME	0.00	0.18	0.10	0.08	−0.356	−0.459
MP	0.00	0.12	0.07	0.05	−0.222	−0.518
NJ	0.00	0.18	0.09	0.07	−0.385	−0.445

The variables are described in the text and Fig. 3. Values are averaged over 100 simulations; the values for E_G, E_S, and E_P are also averaged across 1,167 genes. r_{rate} is the correlation of E_G with substitution rate; r_{sites} is the correlation of E_G with the number of sites.

Table 4. Summary of the empirical study of Eutherian mammal genes

	D_E			D_E Ratio		
Gene	$n = 15$	$n = 30$	$n = 45$	15/30	30/45	15/45
12s	0.73	0.47	0.33	1.56	1.43	2.22
16s	0.79	0.55	0.42	1.42	1.32	1.87
adora3	0.63	0.35	0.22	1.78	1.60	2.86
adrb2	0.58	0.34	0.24	1.71	1.44	2.45
app	0.58	0.33	0.22	1.76	1.49	2.63
atp7a	0.57	0.30	0.20	1.90	1.50	2.85
bdnf	0.68	0.48	0.33	1.41	1.46	2.06
bmi1	0.52	0.37	0.22	1.42	1.66	2.35
cb1	0.58	0.32	0.17	1.80	1.86	3.35
crem	0.74	0.45	0.30	1.64	1.53	2.51
edg1	0.65	0.38	0.23	1.71	1.69	2.90
plcb4	0.52	0.25	0.17	2.10	1.44	3.02
pnoc	0.56	0.30	0.17	1.86	1.78	3.31
rag1	0.54	0.34	0.20	1.57	1.69	2.64
rag2	0.61	0.37	0.23	1.66	1.62	2.69
tyr	0.50	0.27	0.14	1.86	1.89	3.52
zfx	0.58	0.36	0.23	1.59	1.57	2.50

D_E is the phylogenetic difference measured between trees containing n taxa and trees containing all 66 taxa. Values are averaged across 500 subsamples. The D_E Ratio is the ratio between D_E values based on different n.

Empirical Versus Simulation Studies. Our simulation results appear to conflict with empirical studies that have reported improved performance with increased taxon sampling. We examined these patterns empirically with the raw data from the study that produced the Eutherian mammal tree (1). For each of the 17 genes (Table 4), we inferred the phylogeny for all taxa by using NJ. We then created 500 random subsamples consisting of 15, 30, and 45 taxa each. Each subsample was constrained to contain at least one species from each of 13 mammalian orders present in the tree (assuming data were available for the order). These subsamples were analyzed with NJ, and the phylogenetic difference per internal branch was determined between the results of the subsampled taxa and the pruned results from the full taxa (D_E, Fig. 3).

D_E declines as the sample size increases (Table 4). However, this does not indicate that trees with larger numbers of taxa are more accurate with respect to the true tree than those with fewer taxa, because this comparison does not involve the true tree (Fig. 3). The results in Table 4 merely show that trees based on similar numbers of taxa (e.g., 66 and 45) tend to be more similar than those based on dissimilar numbers of taxa (e.g., 66 and 15).

Because we know the true tree in computer simulations, we

Table 5. Comparison of subsampled data in the DNA simulations

	E_S	E_P	D_E	r_n
ME	0.19	0.16	0.14	−0.200
MP	0.16	0.11	0.09	−0.172
NJ	0.18	0.16	0.15	−0.229
ML	0.12	0.09	0.05	−0.324

Values are averaged over 100 simulations for each of 50 genes. r_n is the correlation between D_E and the number of subsampled taxa.

previously calculated the error between the true tree and the reconstructed trees (E_S and E_P). These errors were found to be largely independent of the size of the taxon sample. When we calculate D_E for our simulations, we find a negative correlation between it and subsample size (Table 5)—i.e., as the subsample size gets larger, the topological differences among the full and subsample inferred trees gets smaller. This is the identical pattern found for the empirical results in Table 4. In addition, D_E is only slightly smaller than E_S and E_P (Table 5). Therefore, there is as much topological difference between full and subsample trees as is observed between these trees and the true tree. This indicates that the similarity of E_S and E_p, regardless of the number of taxa, is not due to identical phylogenetic inference. The use of different sample sizes (number of taxa) may lead to different phylogenetic inferences; however, the error associated with these estimates is largely independent of the sample size.

This result has interesting implications for the difference in phylogenetic position of rabbits, rodents, and primates in the Eutherian phylogeny as obtained in studies based on small numbers of taxa and large numbers of genes versus those with more extensive taxon sampling but much fewer genes (1, 25, 26, 34, 42). If the model tree in Fig. 2 is indeed true, then our study indicates that taxon sampling does not explain the discrepancy. Otherwise, the correct phylogenetic relationships of these three groups are yet to be determined. Increasing the number of genes for the large taxon sample of Murphy *et al.* (1) and Madsen *et al.* (25) is likely to resolve this issue. In general, our results do not provide evidence in favor of adding taxa to problematic phylogenies; instead, using more genes with longer sequences would be a better use of time and resources.

We thank Sudhindra Gadagkar, Mark Miller, Tom Dowling, Mike Douglas, Koichiro Tamura, and two anonymous reviewers for providing useful comments on earlier versions of the manuscript. This research was supported by grants from the National Science Foundation (DBI-9983133), the National Institute of Health (HG-02096), and the Burroughs Wellcome Fund (BWI 1001311) (to S.K.), and National Science Foundation Grant IBN-9977063 (to James L. Collins).

1. Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A. & O'Brien, S. J. (2001) *Nature (London)* **409,** 614–618.
2. Omland, K. E., Lanyon, S. M. & Fritz, S. J. (1999) *Mol. Phylogenet. Evol.* **12,** 224–239.
3. Yoder, A. D. & Irwin, J. A. (1999) *Cladistics* **15,** 351–361.
4. Saunders, M. A. & Edwards, S. V. (2000) *J. Mol. Evol.* **51,** 97–109.
5. van Tuinen, M., Sibley, C. G. & Hedges, S. B. (2000) *Mol. Biol. Evol.* **17,** 451–457.
6. De Rijk, P., Van de Peer, Y., Van den Broeck, I. & De Wachter, R. (1995) *J. Mol. Evol.* **41,** 366–375.
7. Lecointre, G., Philippe, H., Lê, H. L. V. & Le Guyader, H. (1993) *Mol. Phylogenet. Evol.* **2,** 205–224.
8. Poe, S. (1998) *Syst. Biol.* **47,** 18–31.
9. Soltis, P. S., Soltis, D. E., Wolf, P. G., Nickrent, D. L., Chaw, S.-M. & Chapman, R. L. (1999) *Mol. Biol. Evol.* **16,** 1774–1784.
10. Johnson, K. P. (2001) *Syst. Biol.* **50,** 128–136.
11. Kim, J. (1996) *Syst. Biol.* **45,** 363–374.
12. Kim, J. (1998) *Syst. Biol.* **47,** 43–60.
13. Rannala, B., Huelsenbeck, J. P., Yang, Z. & Nielsen, R. (1998) *Syst. Biol.* **47,** 702–710.
14. Hillis, D. M. (1998) *Syst. Biol.* **47,** 3–8.
15. Felsenstein, J. (1978) *Syst. Zool.* **27,** 401–410.
16. Nei, M. (1996) *Annu. Rev. Genet.* **30,** 371–403.
17. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996) in *Molecular Systematics*, eds. Hillis, D. M., Moritz, C. & Mable, B. K. (Sinauer, Sunderland, MA), pp. 407–514.
18. Hillis, D. M. (1996) *Nature (London)* **383,** 130–131.
19. Hendy, M. D. & Penny, D. (1989) *Syst. Zool.* **38,** 297–309.
20. Graybeal, A. (1998) *Syst. Biol.* **47,** 9–17.
21. Poe, S. & Swofford, D. L. (1999) *Nature (London)* **398,** 299–300.
22. Robinson, M., Gouy, M., Gautier, C. & Mouchiroud, D. (1998) *Mol. Biol. Evol.* **15,** 1091–1098.
23. Sullivan, J., Swofford, D. L. & Naylor, G. J. P. (1999) *Mol. Biol. Evol.* **16,** 1347–1356.
24. Ackerly, D. D. (2000) *Evolution* **54,** 1480–1492.
25. Madsen, O., Scally, M., Douady, C. J., Kao, D. J., DeBry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., De Jong, W. W. & Springer, M. S. (2001) *Nature (London)* **409,** 610–614.
26. Graur, D., Duret, L. & Gouy, M. (1996) *Nature (London)* **379,** 333–335.
27. Easteal, S., Collet, C. & Betty, D. (1995) *The Mammalian Molecular Clock* (Landes, Austin, TX).
28. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), pp. 21–132.
29. Hasegawa, M., Kishino, H. & Yano, T. (1985) *J. Mol. Evol.* **22,** 160–174.
30. Tudge, C. (2000) *The Variety of Life* (Oxford Univ. Press, New York).
31. Hedges, S. B. (2001) in *Major Events in Early Vertebrate Evolution: Palaeontology, Phylogeny, Genetics and Development*, ed. Ahlberg, P. E. (Taylor and Francis, London), pp. 119–134.
32. Hedges, S. B. & Kumar, S. (1999) *Science* **285,** 2031a.
33. Duret, L., Mouchiroud, D. & Gouy, M. (1994) *Nucleic Acids Res.* **22,** 2360–2365.
34. Kumar, S. & Hedges, S. B. (1998) *Nature (London)* **392,** 917–920.
35. Swofford, D. L. (2000) PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) (Sinauer, Sunderland, MA).
36. Nei, M., Kumar, S. & Takahashi, K. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12390–12397.
37. Kumar, S. & Gadagkar, S. R. (2000) *J. Mol. Evol.* **51,** 544–553.
38. Rosenberg, M. S. & Kumar, S. (2001) *Mol. Biol. Evol.* **18,** 1823–1827.
39. Kumar, S. (1996) *Mol. Biol. Evol.* **13,** 584–593.
40. Robinson, D. F. & Foulds, L. R. (1981) *Math. Biosci.* **53,** 131–147.
41. Penny, D. & Hendy, M. D. (1985) *Syst. Zool.* **34,** 75–82.
42. Hedges, S. B., Parker, P. H., Sibley, C. G. & Kumar, S. (1996) *Nature (London)* **381,** 226–229.