Published in final edited form as:
Syst Biol. 2003 February ; 52(1): 119–124.

Taxon Sampling, Bioinformatics, and Phylogenomics

Michael S. Rosenberg and **Sudhir Kumar**
Department of Biology, Arizona State University, Tempe, Arizona 85287-1501, USA

Michael S. Rosenberg: msr@asu.edu; Sudhir Kumar: s.kumar@asu.edu

Taxon sampling is often thought to be of extreme importance for phylogenetic inference, and increased sampling of taxa is commonly advocated as a solution to resolving problematic phylogenies. Another solution is to increase the number of sites (by sequencing additional genes) sampled for each taxon. In an ideal world, one would like to increase samples of both taxa and genes, but taxon sampling has not kept up with the pace of gene sampling increase because of the increasing ease and emphasis on genome sequencing.

The question of taxon sampling is necessarily driven by resource limitation. The precise scope of "sufficient" taxon sampling is always dependent on questions being addressed. If we need to know the complete phylogeny of a genus, we must sample the genus exhaustively. In experimental design, partial sampling is an issue only when certain taxa can stand as proxies for the clades to which they belong (clade-based or stratified sampling; see Hillis, 1998). In bioinformatics studies, taxon sampling is restricted by the data availability in genetic databases (database-restricted sampling). Clearly, the nature of the problem in these two research programs is different. In stratified sampling, we are interested in knowing whether to sequence more genes per species or fewer genes for a large number of species per clade. In contrast, in database-restricted sampling it is important to know whether the overall accuracy of inferred phylogenetic trees for small taxa sets is similar to that of trees inferred from larger taxa sets. We recently addressed the issue of the database-restricted sampling (Rosenberg and Kumar, 2001) and concluded that although there was a consistent decrease in error when using more taxa, the decrease was generally minor relative to the number of taxa added to the data set.

Pollock et al. (2002) challenged this conclusion by modifying our measure of the phylogenetic error. This measure, ΔE, differs from ours in that we used the difference in error between the subsampled tree [E_S] and full sampled tree [E_P], whereas Pollock et al. (2002) divided this difference by E_S to measure the relative reduction in error. ΔE plotted against the number of additional taxa in the full sampled tree (=66 minus the number of taxa in the subsample tree) shows a clear positive effect (Pollock et al., 2002: Figs. 4, 5). Unfortunately, this impressive result brings little biological benefit, as clearly shown by a scatterplot of the average number of additional branches inferred correctly in each case (Fig. 1). In no instance are there more than 1.5 additional branches reconstructed correctly, even though the number of taxa has often increased many fold. For instance, more than doubling the number of taxa only led to an average increase of 0.7 additional correct branches (points in the middle of the *x*-axis in Fig. 1). This fact was clearly noted in our original article: "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" (Rosenberg and Kumar, 2001: 10754). Therefore, although an increase in the number of taxa sampled will lead to improvement in accuracy, the improvement is minimal, particularly when we consider the amount of data (in terms of the

Associate Editor: Chris Simon

number of total nucleotides) being added. We do not advocate using fewer taxa when more are available, as is clear from the results presented by Rosenberg and Kumar (2001:10754).

Zwickl and Hillis (2002) also challenged conclusions reached by Rosenberg and Kumar (2001) by using the concept of tree diameter (the maximum distance between all pairs of taxa) to partition genes with different subsampled sets of taxa for analysis. They showed that four-taxon subsamples with a smaller tree diameter generate more accurate results than those subsamples with larger tree diameters. This result is expected because, with sequence divergence and length kept constant, the larger diameter four-taxon trees will encompass higher average divergence and would thus involve larger estimation errors. Furthermore, for the simulations involving the model tree in Figure 2a, four-taxon data sets containing sequences with larger diameters would include interordinal relationships (with many small interior branches) more frequently than would small diameter samples (see also Zwickl and Hillis, 2002: Fig. 3a). Therefore, Zwickl and Hillis’s study is an examination of the phylogenetic error at different evolutionary divergence cross sections of the phylogenetic tree specifically simulated. This and the complete absence of resource limitation (a must for any sampling study) clearly establish that Zwickl and Hillis have not evaluated either stratified or database-restricted taxon-sampling problems. Therefore, Zwickl and Hillis were not correct in stating that their results are in contradiction with our previous results (Rosenberg and Kumar, 2001). In fact, Zwickl and Hillis’s results represent another facet of statistical analysis of the same data. Also, Zwickl and Hillis took issue with our choice of a fast heuristic search used in computer simulations (Rosenberg and Kumar, 2001). We chose this strategy based on results of multiple previous studies, which showed that the most optimal tree is often more optimal than the true tree and that the fast and more exhaustive searches produce trees with comparable phylogenetic errors (Kumar, 1996; Nei et al., 1998; Takahashi and Nei, 2000). Zwickl and Hillis found that with the maximum parsimony (MP) method for the given data set, the TBR searches produced topologies that had less error than those from NNI. This result (based on a single simulation data set) seems to be in conflict with previous studies. We plan to evaluate this result more thoroughly analytically and by computer simulation in the future.

However, we extrapolated our database-restricted sampling and random sampling results to conclude that the phylogenetic trees with fewer taxa but large numbers of genes per taxon may be more accurate than those with many taxa but fewer genes (Rosenberg and Kumar, 2001). Neither Pollock et al. (2002) nor Zwickl and Hillis (2002) addressed that issue, which lies at the heart of the experimental design. Here, we tackle this issue along with biological relevance of many other assumptions made and conclusions reached by Rosenberg and Kumar (2001) that Zwickl and Hillis (2002) objected to. We show that the conclusions reached by Rosenberg and Kumar (2001) are applicable for both phyloinformatic and phylogenomic studies.

Taxon Sampling in Bioinformatics

In bioinformatics efforts, taxon sampling is directly restricted by the data available in genetic databases (e.g., GenBank). Mining these sources for data on a specific clade of interest (e.g., mammals) usually leads to extremely unbalanced data sets. A few genes may be available for dozens of taxa, but the number of available taxa decreases dramatically as the number of genes increases. We recently conducted a study to examine the relative accuracies of small trees compared with large trees by means of computer simulations in two taxon-sampling regimes: one where the sampling was biased toward taxa that were more common in GenBank (e.g., humans) and the other one where sampling was purely random (Rosenberg and Kumar, 2001). We reported a consistent decrease in error when using more taxa but found that this decrease was generally minor relative to the number of additional taxa sampled (Rosenberg and Kumar, 2001). Because the number of sites showed a larger effect on phylogenetic accuracy, we suggested using longer genes and fewer taxa rather than shorter genes for more taxa.

Zwickl and Hillis (2002) suggested that the partition metric-based method (Robinson and Foulds, 1981; Penny and Hendy, 1985) used in our study should have been normalized by considering the number of possible topologies for a given set of taxa. Without this normalization, the phylogenetic error for small trees could be underestimated. Although this normalization might be useful, the difference between the adjusted and original measurements is almost 0 if the number of taxa is greater than seven (Fig. 2; Zwickl and Hillis, 2002). Clearly, this problem affects only very small trees. Removal of all simulations sampling <10 taxa changes none of our primary conclusions (see also Zwickl and Hillis, 2002).

However, E_{adj} only takes into account the space of topological (branching pattern) configuration for a given number of taxa in a graph theoretic style; it implicitly assumes that all topologies with a certain topological difference are equally probable with respect to the optimality score. This assumption is clearly false. A measure based on the number of equally optimal trees (statistically) determined specifically for a given data set will likely be a better alternative (e.g., Kumar, 1996). In any case, we prefer computing the accuracy of phylogenetic inference in reconstructing a branch and an average of this number over all interior branches in a tree (as used in Rosenberg and Kumar, 2001), because these measures are direct and easily understood and can be computed when one conducts a large number of simulation replicates for each condition.

The distribution of phylogenetic error based on the tree diameter (the maximum distance between all pairs of taxa) of four-taxon samples reported by Zwickl and Hillis (2002) is another facet of analysis. The results reported are expected; they are not comparable and thus not in contradiction with those of Rosenberg and Kumar (2001). We had, however, extrapolated our result to suggest that phylogenetic trees with fewer taxa but large numbers of genes per taxon may be as accurate (if not more so) than those trees with many taxa but fewer genes.

Taxon Sampling in Experimental Design

In experimental design (e.g., sequencing strategies), the approach to taxon sampling is necessarily different from that used in informatics. One is not restricted to data already available in genomic databases; we have the freedom to choose which taxa and genes to add in an optimal way. Clearly, it would be best to add as many genes and taxa as possible, but resources rarely allow this luxury. Therefore, the basic question is: Is it better to sample more genes for fewer species or fewer genes for more species? This difference in objective leads to a difference in sampling design. To evaluate this question, we present results from additional simulations conducted using the same 66-taxon tree (Fig. 2a) as presented by Rosenberg and Kumar (2001). We simulated 448 genes (100 replicates each) with evolutionary parameters estimated from actual genes in GenBank, using the Hasegawa–Kishino–Yano (HKY) model of nucleotide substitution (Hasegawa et al., 1995). Phylogenetic analysis was performed in PAUP$^{\pm}$ (Swofford, 1998), using neighbor joining (NJ) and minimum evolution (ME) methods with Tamura–Nei (Tamura and Nei, 1993) distances, unweighted MP, and maximum likelihood (ML) under the HKY model.

For each replicate we constructed data sets consisting of all 66 taxa and subsets of 45, 30, and 15 taxa. For the subsets, we used a stratified sampling approach (purposefully spreading the sampled taxa among different clades) rather than pure random sampling as in our previous study. Specifically, we constrained the sampling to contain at least one representative from each of the 14 mammalian orders present in the model tree by first choosing one taxon at random from each order. Additional taxa were chosen completely at random from all remaining taxa. By stratifying the sampling, we focused the analysis towards inferring relationships among the sampled clades (in our case, mammalian orders): although we may be sampling from the taxa in Figure 2a, we are interested in recovering the phylogeny in Figure 2b. We evaluated the effects of taxon sampling only with respect to the branches that represent these relationships (the thick branches of Fig. 2a). For each of these 10 branches, we calculated the percentage of replicates in which each branch was reconstructed correctly.

From our simulation results, we plotted the percentage of times the branch was reconstructed correctly in the full 66-taxon tree (across all genes and all replicates) against the percentage of times the branch was reconstructed correctly in a tree constructed from a subsample of taxa (Fig. 3). All inference methods showed a similar pattern: these branches were reconstructed more accurately when all 66 taxa were used than when fewer taxa were used. Furthermore, the larger taxon samples were generally more accurate than the smaller taxon samples. These results indicate that increasing the number of taxa can dramatically increase the accuracy of the relationships among the sampled clades under a stratified sampling regime when considering only the among-clades relationships.

However, the results in Figure 3 and those of Zwickl and Hillis (2002) confound the effects of taxon sampling and the increase in the number of nucleotides because, for example, a 66-taxon data set has more than three times as many nucleotides as a 15-taxon data set and would therefore require much larger sequencing effort. Is the increase in efficiency due to an increase in the number of nucleotides in data sets containing a larger number of taxa? To answer this question, we examined the effect of taxon sampling independent of the overall number of nucleotides (number of sites ± number of taxa) by subdividing the results as follows. Instead of calculating the accuracy of reconstruction over all possible genes, we subdivided the data set into genes with specific numbers of sites (we used all genes with lengths of ±100 of the target number of sites). We created four sets: 500 sites (54 genes); 1,000 sites (67 genes); 1,500 sites (29 genes); and 2,000 sites (15 genes). For each set we again calculated the percentage of replicates in which each interordinal branch was reconstructed correctly. (An interordinal branch is one that describes a relationship among mammalian orders; it does not indicate a specific taxonomic level as would, for example, an infraorder.) We then plotted these values such that each point represented approximately the same total number of sites in the data set.

The contrast of these plots (Fig. 4) with the previous results (Fig. 3) is striking. For a constant number of sites, distance and likelihood methods reconstructed correct branches more often when there were more sites per taxon (i.e., smaller taxa samples). Parsimony showed a similar pattern, although not as universally. The differences between large numbers of taxa and shorter sequences versus few taxa and longer sequences were not large, but they were consistent. Not unexpectedly, the comparisons with more total sites tended to be more accurate than comparisons with fewer total sites (i.e., 45,000 sites versus 15,000 sites), which explains the result in Figure 3. These results speak clearly to experimental design: when resources are limited, one would appear to do better by sequencing more sites/genes per taxon than by increasing the number of taxa with shorter sequences. This finding is congruent with current genome sequencing project design, producing more genes for fewer taxa as a natural outcome of the sequencing strategies.

Conclusions

A distinction should be made between taxon sampling in informatics and in experimental design. Rosenberg and Kumar (2001) focused on the bioinformatics aspects, and we have presented here additional results for clade-based stratified taxon sampling. Zwickl and Hillis (2002) primarily dealt with error partitioning based on the maximum sequence divergence in a tree; they did not evaluate database-restricted or clade-based taxon sampling. We have also discussed the error measures used by Pollock et al. (2002) and Zwickl and Hillis (2002). The error measures used by Pollock et al. produce strong statistical correlation between taxon sample size and accuracy, even when the absolute increase (as indicated by Rosenberg and Kumar, 2001) is minimal; for example the overall improvement even after adding 36 taxa to a 30-taxon tree is less than one additional branch. Therefore, Pollock et al.'s error metric is strongly influenced by the factor used to normalize the Rosenberg and Kumar (2001) statistic. Zwickl and Hillis's (2002) correction of the metric used by Rosenberg and Kumar (2001) makes a difference only when the number of taxa is less than seven. It also makes some biologically unrealistic assumptions.

As expected, with more data (total nucleotides) we were able to reconstruct more accurate phylogenies. When the number of taxa sampled per clade was increased, the interrelationships of those clades could be inferred more accurately. However, sampling in experimental design is only relevant in the context of resource limitation; therefore, to compare apples to apples, we used the number of nucleotides per sequence (number of taxa ± sequence length) as a control. In this case, trees are more accurately reconstructed when using more sites for fewer taxa than when using more taxa for fewer sites when the total number of nucleotides is held constant in a data set. This result is stronger for distance and likelihood methods of phylogeny reconstruction but less so for parsimony. We reconstructed most of the short internal branches with a reasonably degree of accuracy (Fig. 4) with an adequate amount of data (whether taxa or sites). This result is certainly encouraging for phylogenetic reconstruction in general.

The results presented here and by in Rosenberg and Kumar (2001) provide a useful framework for analyzing the effect of taxon sampling in phyloinformatic and phylogenomic studies.

Acknowledgments

We thank B. Friedman, K. Tamura, J. Thorne, D. Hillis, D. Pollock, and two anonymous reviewers for comments on this 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.

References

Eizrik E, Murphy WJ, O'Brien SJ. Molecular dating and biogeography of the early placental mammal radiation. J Hered 2001;92:212–219. [PubMed: 11396581]

Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22:160–174. [PubMed: 3934395]

Hillis DM. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol 1998;47:3–8. [PubMed: 12064238]

Kumar S. A stepwise algorithm for finding minimum evolution trees. Mol Biol Evol 1996;13:584–593. [PubMed: 8882501]

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. Molecular phylogenetics and the origins of placental mammals. Nature 2001;409:614–618. [PubMed: 11214319]

Nei M, Kumar S, Takahashi K. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc Natl Acad Sci USA 1998;95:12390–12397. [PubMed: 9770497]

Penny D, Hendy MD. The use of tree comparison metrics. Syst Zool 1985;34:75–82.

Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol 2002;51:664–671. [PubMed: 12228008]

Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci 1981;53:131–147.

Rosenberg MS, Kumar S. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci USA 2001;98:10751–10756. [PubMed: 11526218]

Swofford, DL. Sinauer; Sunderland, Massachusetts: 1998 p.PAUP$^{\pm}$: Phylogenetic analysis using parsimony ($^{\pm}$and other methods), version 4;

Takahashi K, Nei M. Efficiencies of fast algorthims of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol 2000;17:1251–1258. [PubMed: 10908645]

Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993;10:512–526. [PubMed: 8336541]

Zwickl DJ, Hillis DM. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 2002;51:588–598. [PubMed: 12228001]

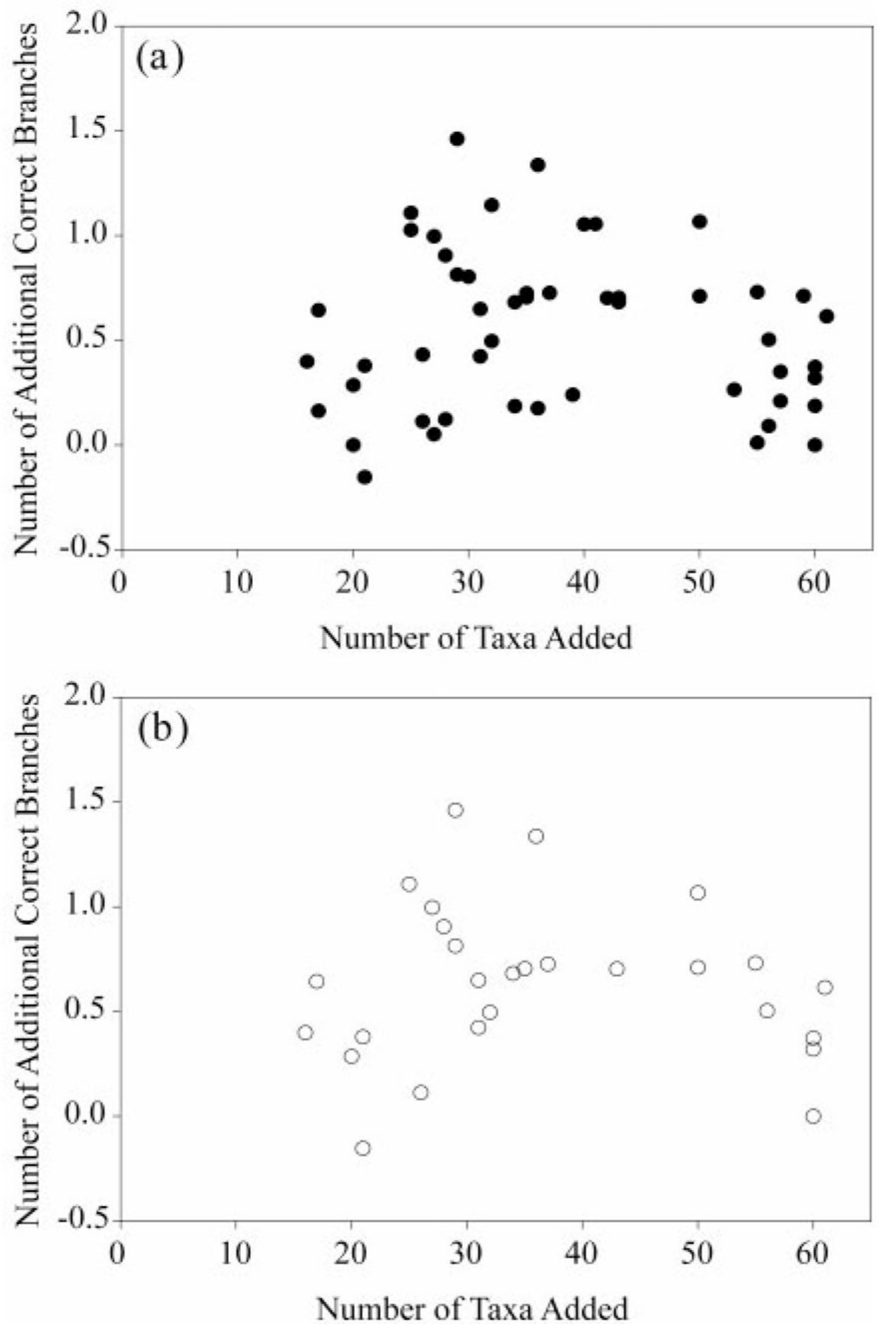

Figure 1.
Number of branches reconstructed correctly with increased taxon sampling. (a) All simulated genes. (b) Genes with rates >0.7 and >500 sites (after Pollock et al., 2002).

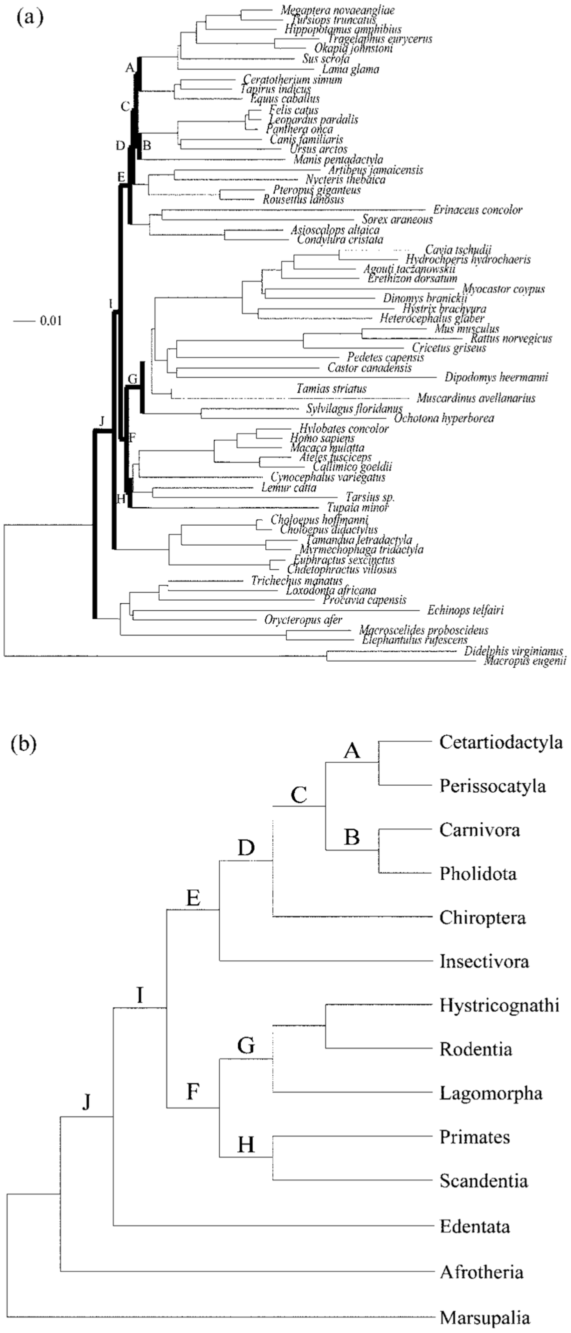

Figure 2.

Model tree for the simulations based on the Eutherian mammal tree from Murphy et al. (2001) and Eizrik et al. (2001). (a) Full 66-taxon tree; interordinal relationships are represented by thick branches designated with letters. (b) Phylogenetic relationships of the mammalian orders present in the model tree.

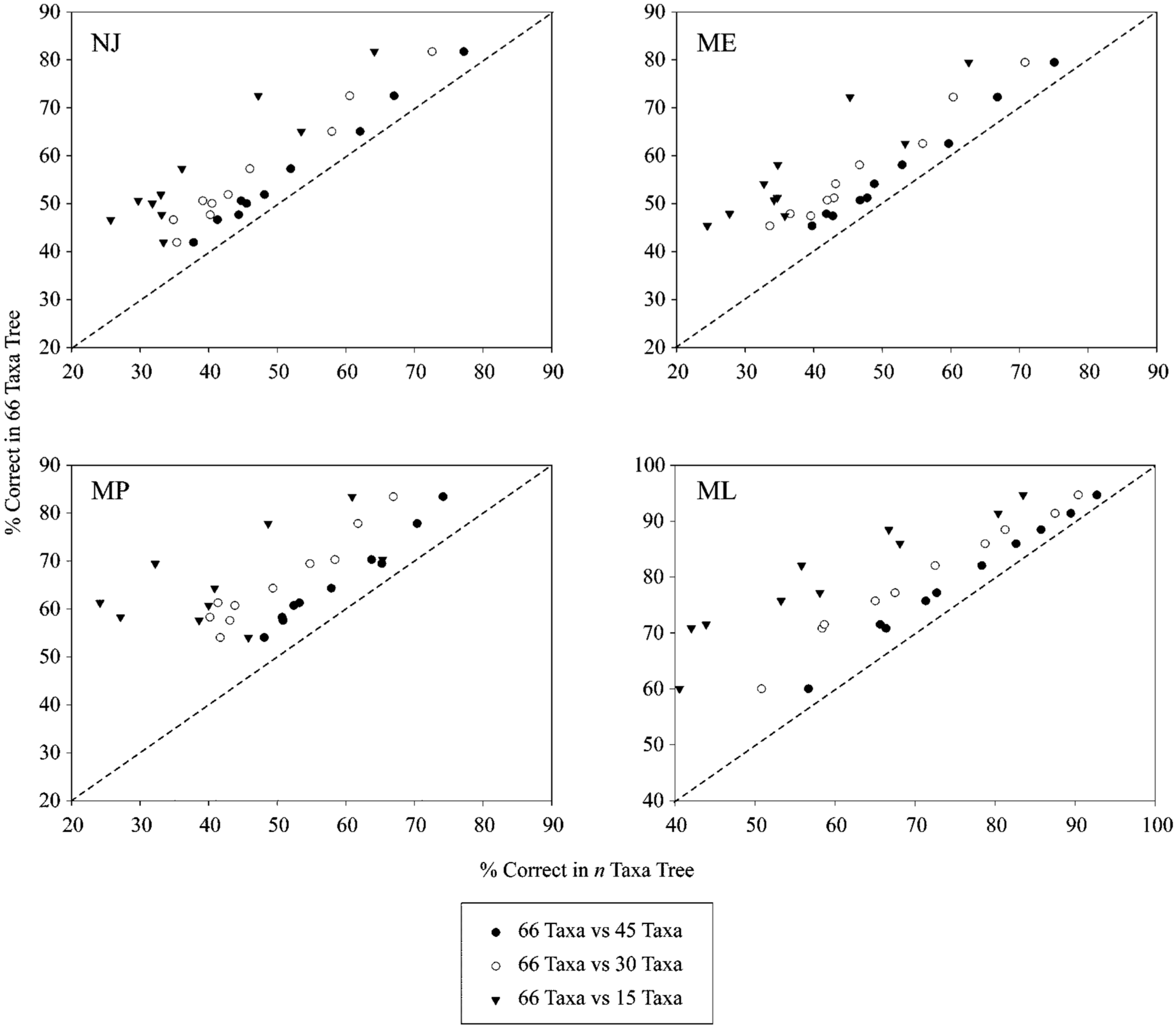

Figure 3.

Plot of the percentage of times the interordinal branches were reconstructed correctly in 66-taxon trees versus *n*-taxon trees, where *n* = 15, 30, and 45. These values are for all genes and all replicates. The dotted lines indicate a 1:1 relationship. Analyses were conducted with PAUP$^{\pm}$ using neighbor joining (NJ), minimum evolution (ME), unweighted maximum parsimony (MP), and maximum likelihood (ML) models.

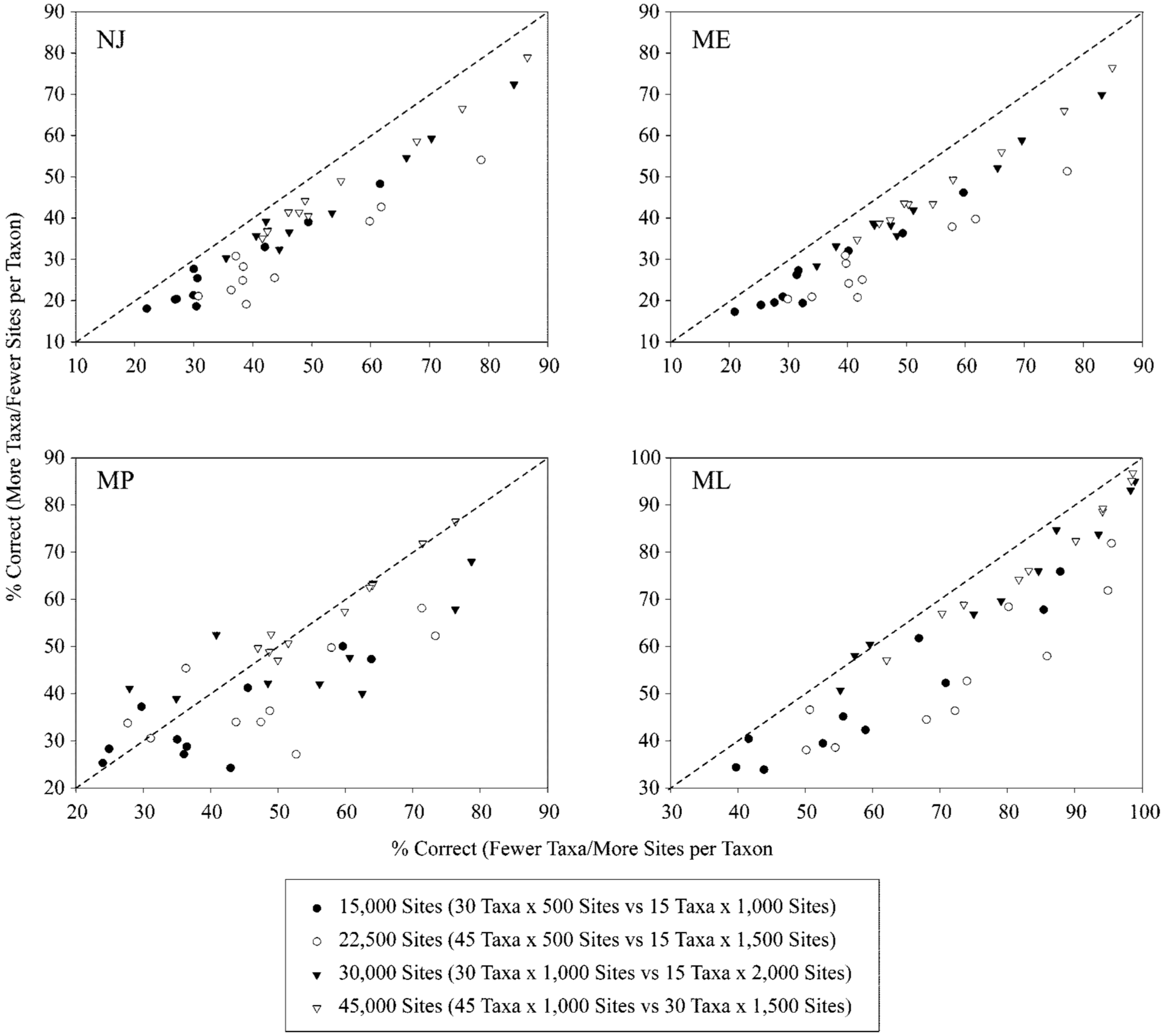

Figure 4.

Plot of the percentage of times interordinal branches were reconstructed correctly when the total number of bases was held constant. In each comparison, the data set with fewer taxa (and more sites per taxon) is always plotted on the *x*-axis. The dotted lines indicate a 1:1 relationship. Analyses were conducted with PAUP$^{\pm}$ using neighbor joining (NJ), minimum evolution (ME), unweighted maximum parsimony (MP), and maximum likelihood (ML) models.