More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy

Antonis Rokas and Sean B. Carroll

Howard Hughes Medical Institute and Laboratory of Molecular Biology, University of Wisconsin–Madison

The relative contribution of taxon number and gene number to accuracy in phylogenetic inference is a major issue in phylogenetics and of central importance to the choice of experimental strategies for the successful reconstruction of a broad sketch of the tree of life. Maximization of the number of taxa sampled is the strategy favored by most phylogeneticists, although its necessity remains the subject of debate. Vast increases in gene number are now possible due to advances in genomics, but large numbers of genes will be available for only modest numbers of taxa, raising the question of whether such genome-scale phylogenies will be robust to the addition of taxa. To examine the relative benefit of increasing taxon number or gene number to phylogenetic accuracy, we have developed an assay that utilizes the symmetric difference tree distance as a measure of phylogenetic accuracy. We have applied this assay to a genome-scale data matrix containing 106 genes from 14 yeast species. Our results show that increasing taxon number correlates with a slight decrease in phylogenetic accuracy. In contrast, increasing gene number has a significant positive effect on phylogenetic accuracy. Analyses of an additional taxon-rich data matrix from the same yeast clade show that taxon number does not have a significant effect on phylogenetic accuracy. The positive effect of gene number and the lack of effect of taxon number on phylogenetic accuracy are also corroborated by analyses of two data matrices from mammals and angiosperm plants, respectively. We conclude that, for typical data sets, the number of genes utilized may be a more important determinant of phylogenetic accuracy than taxon number.

Introduction

Phylogenetic reconstruction is fundamental to comparative biology research (Felsenstein 1985) as the phylogeneticists' conclusions (i.e., their phylogenetic inferences) become the comparative biologists' assumptions. Consequently, the generation of robust phylogenetic hypotheses and the understanding of the factors influencing accuracy in phylogenetic reconstruction are crucial to evolutionary hypothesis testing. In recent years, advances in molecular biology and genomics have led to the acquisition of genome-scale data from select model organisms (e.g., Cliften et al. 2003; Kellis et al. 2003; Dujon et al. 2004) as well as to new opportunities for the study of organismal diversity (Tautz et al. 2003). However, neither will researchers have access to full genome data from thousands of species in the near future nor will all extant taxa of the tree of life be amenable for study. Given these constraints in data availability, what is the best strategy for reconstructing a reliable broad sketch of the tree of life?

The standard approach in phylogenetics has been to reconstruct phylogenetic hypotheses by maximizing the number of taxa, utilizing one or a few genes (Baldauf et al. 2000; Peterson and Eernisse 2001; Moncalvo et al. 2002). The standard approach has been shaped by the relative technical ease of increasing the number of taxa and the desire to study as many taxa as possible (for most phylogeneticists adding taxa is far more interesting than adding genes). However, the use of small amounts of sequence data has been shown to generate phylogenetic hypotheses that are incongruent or lacking support (e.g., Satta, Klein, and Takahata 2000; Kopp and True 2002; Rokas et al. 2003*b*). To overcome the observed incongruence, many researchers have generated phylogenetic hypotheses by maximizing gene number (Bapteste et al. 2002; Blair et al. 2002; Rokas et al. 2003*b*; Wolf, Rogozin, and Koonin 2004), which typically necessitates that the number of taxa sampled is restricted. Phylogenetic reconstruction using a small number of taxa has, in turn, been criticized to be more sensitive to homoplasy (e.g., Soltis et al. 2004). A priori, it is easy to appreciate the merits inherent in each approach. Ideally, phylogenetic hypotheses should be constructed by utilizing the maximum number of genes and taxa, and currently, data sets composed of a few genes from tens to hundreds of taxa are becoming the standard (e.g., P. S. Soltis, D. E. Soltis, and Chase 1999; Kurtzman and Robnett 2003; Lutzoni et al. 2004). However, with finite resources, the question of the relative contribution of gene number per taxon and total taxon number to phylogenetic accuracy is critical.

The question of which strategy may be preferred has been examined by simulation. For example, Graybeal (1998) demonstrated that, in certain cases, if taxa are chosen specifically to break up long branches, increasing the number of taxa sampled is preferable to increasing gene number. This and similar results (Hillis 1996) have led to the conclusion that for difficult phylogenetic problems, it is generally preferable to increase the number of taxa in a data set rather than the number of genes per taxon (Swofford et al. 1996; Hillis 1998). These inferences have received support from analytical studies, which have shown that, under certain assumptions, the upper bound of the amount of sequence data required to resolve relationships in data matrices containing large taxon numbers can be surprisingly small (Erdos et al. 1999*a*, 1999*b*). However, increasing taxon number while keeping gene number constant can also lead to a decrease in accuracy, either as a result of reducing the amount of phylogenetic information available to resolve the newly added branches (Kim 1998; Bininda-Emonds et al. 2001) or by the introduction of new long branches (Poe and Swofford 1999). In recent years, the debate over which strategy is to be preferred has been renewed (e.g., Rosenberg and Kumar 2001; Pollock et al. 2002; Zwickl and Hillis 2002).

Key words: phylogenetics, taxon number, gene number, phylogenetic accuracy, tree of life, genomics.

E-mail: sbcarrol@wisc.edu.

Mol. Biol. Evol. 22(5):1337–1344. 2005
doi:10.1093/molbev/msi121
Advance Access publication March 2, 2005

Noticeably absent from this debate are empirical studies of biological data sets containing both large numbers of genes and taxa. Given that models of sequence evolution often have a poor fit to real data (Goldman 1993), it is critical that inferences based on analytical and simulation studies are tested against biological data sets. Here, we have taken advantage of many available yeast genome sequences (Souciet et al. 2000; Cliften et al. 2003; Kellis et al. 2003; Dietrich et al. 2004; Jones et al. 2004; Kellis, Birren, and Lander 2004) to explore the relative merits of increasing taxon number and gene number in a genome-scale data set of several yeast species. We have tested the generality of our conclusions by investigating the effect of increasing taxon number on a taxon-rich data set from the same yeast clade and of increasing taxon number and gene number on two gene-rich/taxon-rich data matrices from mammals and angiosperms, respectively. Using symmetric difference tree distance as a metric for phylogenetic accuracy, we have found that in all cases, taxon number does not significantly correlate with phylogenetic accuracy. In contrast, we find that increasing gene number has a significant positive effect on phylogenetic accuracy and that our results may hold across the tree of life.

Materials and Methods

Ortholog Identification and Data Matrix Generation

The genomes of *Debaryomyces hansenii* and *Yarrowia lipolytica* (Dujon et al. 2004) were screened for orthologs of 106 genes from a previously published genome-scale data set of 12 taxa (*Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, *Saccharomyces bayanus*, *Saccharomyces castellii*, *Saccharomyces kluyveri*, *Candida glabrata*, *Candida albicans*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, and *Eremothecium gossypii*) (Rokas et al. 2003*b*; Hittinger, Rokas, and Carroll 2004). Ortholog identification followed the published annotation as established by synteny and Blast (Dujon et al. 2004). Individual genes were codon-aligned using ClustalW (Thompson, Higgins, and Gibson 1994) as implemented in BioEdit version 5.0.9 (Hall 1999). All gene alignments were manually edited to exclude indels and areas of uncertain alignment from further analysis. The data matrix is available from the authors upon request. Three other published data matrices from mammals, angiosperms, and yeasts (kindly provided by W. Murphy, D. Soltis, and C. Kurtzman, respectively) were also utilized (Murphy et al. 2001; Zanis et al. 2002; Kurtzman and Robnett 2003).

Testing the Effect of Taxon Number and Gene Number on Phylogenetic Accuracy

The effect of taxon and gene number on phylogenetic accuracy was estimated by the following procedure (see also Rosenberg and Kumar 2001): (1) starting from the complete data matrix containing g genes from t taxa (denoted $DM_{g/t}$; in the case of the yeast data matrix, $g =$ 106 genes and $t = 14$ taxa), the most parsimonious tree $MP_{g/t}$ is estimated. Next, x genes and y taxa (where $0 < x < g$ and $0 < y < t$) are randomly chosen to create a data matrix $DM_{x/y}$ that is a subset of $DM_{g/t}$, and the most parsimonious tree $MP_{x/y}$ for $DM_{x/y}$ is estimated. In cases where more than one most parsimonious tree is recovered, one of them is randomly chosen as the $MP_{x/y}$. Next, the $MP_{g/t}$ tree is pruned so that it contains exactly the same y taxa as $MP_{x/y}$, and the symmetric difference tree distance metric (denoted *SymDif*; Robinson and Foulds 1981) between $MP_{x/y}$ and $MP_{g/t}$ is calculated. This procedure was repeated 1,000 times for each value of x and y tested, and we tested a range of x and y values. The choice of x genes and y taxa approaches true randomness as the binomial coefficient (i.e., the number of all possible combinations) increases relative to replicate number. However, for values of x and y close to g and t, respectively, the number of possible combinations can be lower than replicate number. In such cases, instead of random sampling, all possible combinations were analyzed. To allow comparisons between topologies containing different taxon numbers (the maximum value of *SymDif* differs with respect to the number of taxa), all *SymDif* values are scaled by $2(y - 3)$, the maximum value of *SymDif* in a y taxon topology (denoted as average-scaled symmetric difference or *ASSymDif*). According to this definition of phylogenetic accuracy, a data matrix with x_1 gene and y_1 taxon dimensions is considered to be more accurate than a data matrix with x_2 gene and y_2 taxon dimensions (where $x_1 \neq x_2$ and/or $y_1 \neq y_2$), if the *ASSymDif* of the first data matrix is lower than the *ASSymDif* of the second data matrix.

All phylogenetic analyses were performed in PAUP* version 4.0b10 (Swofford 2002). The PAUP* command files for each set of replicates (including the random sampling of x genes and y taxa) were generated using Perl scripts. Significance between data matrices differing in taxon number and/or gene number was assessed by a two-tailed, paired-two-sample-for-means t-test (Sokal and Rohlf 1995) on their *ASSymDif*s.

Results

The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy in a 106-Gene, 14-Taxon Phylogeny of Yeast Species

We have utilized publicly available genomic data on 14 yeast species to examine the effect of increasing gene number and taxon number on phylogenetic accuracy. As gene number increases, *ASSymDif* decreases, irrespective of taxon number (fig. 1). For example, data matrices of 3 genes and 11 taxa show an *ASSymDif* of 0.28, whereas data matrices of 5, 8, 20, and 50 genes for the same number of taxa exhibit progressively smaller *ASSymDif*s (0.23, 0.20, 0.17, and 0.17 respectively). This decrease in *ASSymDif* among data matrices differing in gene number is statistically significant (table 1), with the exception of the difference in *ASSymDif*s between data matrices with taxon numbers 20 and 50 (table 1). Therefore, as gene number in a data matrix increases, *ASSymDif* values decrease and phylogenetic accuracy increases. In contrast, increasing taxon number results in slight increases in *ASSymDif* (fig. 1) and hence a reduction in phylogenetic accuracy. For example, data matrices of 20 genes and 6 taxa show an *ASSymDif* of 0.11, whereas data matrices of 7, 8, 9, 10, and 11 taxa for

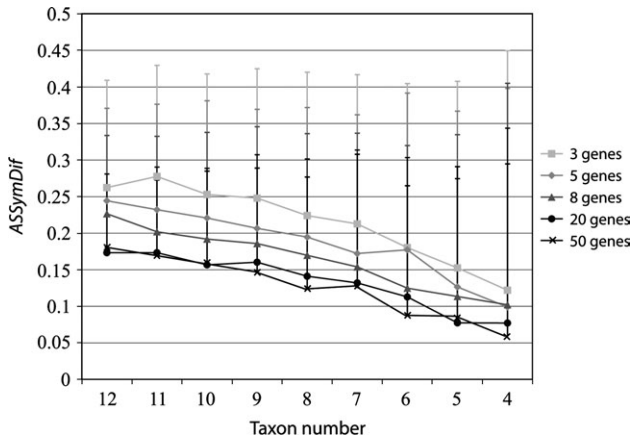

FIG. 1.—The relative contribution of gene number and taxon number to phylogenetic accuracy in a 106-gene, 14-taxon data matrix of yeast species. The effect of taxon number on phylogenetic accuracy is significant, albeit weakly; increasing taxon number (abscissa) leads to higher values of *ASSymDif* (ordinate) and hence a decrease in phylogenetic accuracy. In contrast, increasing gene number is significantly and positively correlated with phylogenetic accuracy; increasing gene number generates decreasing values of *ASSymDif*. Variation in gene number is represented by lines and symbols of different shades of gray. For example, the 3-gene data points are drawn as light gray squares connected by lines of the same shade, whereas the 20-gene data points are drawn as dark gray circles connected by dark gray lines. Each data point is the mean value of 1,000 replicates. Error bars denote one standard deviation from the mean. Standard deviations associated with different gene numbers sampled are drawn in the corresponding gray line scales. In each replicate, the selection of taxa and genes was random.

the same number of genes exhibit increasingly higher *ASSymDif*s (0.13, 0.14, 0.16, 0.16, and 0.17, respectively). The increases in *ASSymDif* are significant in most pairwise comparisons between data matrices differing in taxon number by a value of 3 or more (table 1). The only observed trend with respect to taxon number in figure 1 is the decrease in the variance of *ASSymDif* in response to an increase in taxon number.

Separating the Effect of Taxon Number and Gene Number on Phylogenetic Accuracy

In the experimental design underlying the data shown in figure 1, both variables (gene number and taxon number) were allowed to vary. In order to distinguish the effect of each variable on phylogenetic accuracy and to verify that the effects of gene number and taxon number are due to the variables themselves and not their interaction, we performed the same analyses, allowing one variable to vary freely while holding the other variable constant. The results are shown in figure 2. As gene number increases (fig. 2*A*), while holding the taxon set for each taxon number constant, *ASSymDif*s decrease. In contrast, an increase in taxon number (fig. 2*B*), while holding the gene set for each gene number constant, does not affect *ASSymDif* values. These results are in agreement with those of figure 1; phylogenetic accuracy is positively correlated with increasing gene number and not correlated (or slightly negatively correlated) with taxon number.

Increasing Taxon Density Does Not Improve Accuracy

The potential shortcomings with the experimental design shown in figure 1 are that the taxon number analyzed is relatively small compared to numbers typically found in phylogenetic studies and that the gene number is very large. The 14 taxa utilized in this study are members of a much larger clade, raising the question of whether taxon number has a significant effect only when a larger number of taxa are added. To test whether the dimensions of our experimental design limited our ability to detect the effect of increasing taxon number on phylogenetic accuracy, we also analyzed a recently published data set composed of 8 genes from 76 taxa (Kurtzman and Robnett 2003). To test whether there was any

Table 1
Significance Tests Between Subsamples of the 106-Gene, 14-Taxon Yeast Data Matrix Differing in Taxon Number and Gene Number

Taxon Number	4	5	6	7	8	9	10	11	12
4		NS	NS	*	*	*	*	*	**
5	0.0283		NS	**	*	**	**	**	**
6	0.0123	0.0451		NS	NS	NS	*	*	*
7	0.0006	0.0002	0.0416		NS	*	*	*	*
8	0.0007	0.0006	0.0027	0.0662		*	*	*	*
9	0.0003	0.0002	0.0014	0.0005	0.0012		NS	NS	NS
10	0.0004	9×10^{-5}	0.0008	0.0009	0.0013	0.0915		NS	NS
11	0.0004	0.0002	0.0006	0.0004	0.0007	0.0022	0.0055		NS
12	0.0001	2×10^{-5}	0.0005	0.0008	0.0008	0.0092	0.0071	0.3827	
Gene number	3	5	8	20	50				
3		**	**	**	**				
5	0.0002		*	**	**				
8	2×10^{-5}	0.0019		**	**				
20	7×10^{-7}	6×10^{-6}	6×10^{-5}		NS				
50	1×10^{-7}	4×10^{-6}	4×10^{-7}	0.1204					

NOTE.—Significance between the *ASSymDif* values of data matrices differing in taxon number was assessed for all pairwise comparisons by a two-tailed *t*-test (Sokal and Rohlf 1995); the *ASSymDif* values were paired so that only values generated by identical gene numbers were compared. Significance between the *ASSymDif* values of data matrices differing in gene number was assessed in the same way, with the *ASSymDif* values paired so that only values generated by identical taxon numbers were compared. The *P* values of all pairwise comparisons are shown on the lower diagonal. The level of significance of the *P* values following a Bonferroni correction for multiple comparisons is shown on the upper diagonal. NS, corrected *P* value $<$ 95% significant; *, 95% significant $<$ corrected *P* value $<$ 99% significant; **, corrected *P* value $>$ 99% significant.

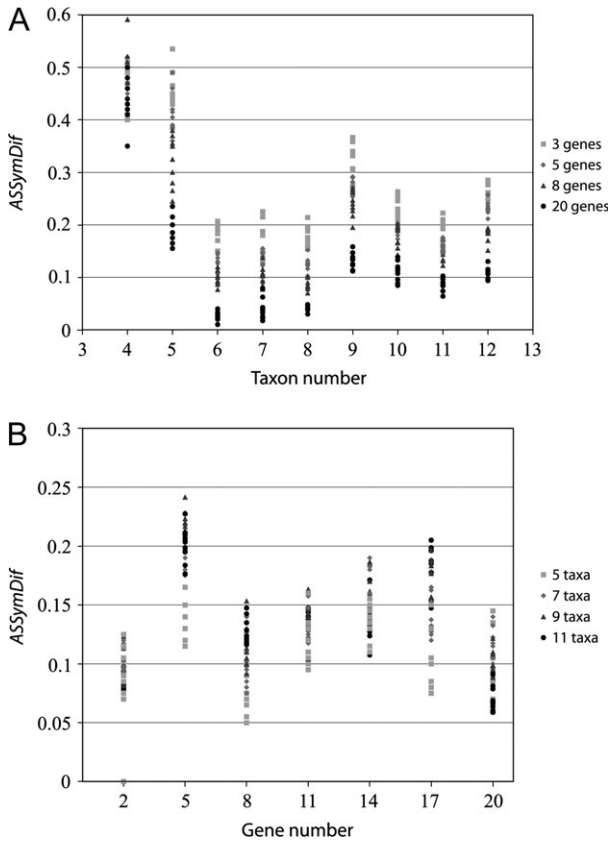

FIG. 2.—Separating the effect of taxon number and gene number on phylogenetic accuracy. (*A*) As gene number increases while the taxon set for each taxon number (abscissa) is held fixed, *ASSymDif* values (ordinate) decrease and hence, phylogenetic accuracy increases. Variation in gene number is represented by symbols of different shades of gray, e.g., the 3-gene data points are drawn as light gray squares, whereas the 20-gene data points are drawn as dark gray circles. (*B*) Increases in taxon number while the gene set for each gene number (abscissa) is held fixed are not correlated with *ASSymDif* values (ordinate) and phylogenetic accuracy. Variation in taxon number is represented by variation in symbol shading. For example, the 5-taxon data points are drawn as light gray squares, whereas the 11-taxon data points are drawn as dark gray circles. Ten replicates were performed for each combination of taxon number and gene number. Each data point is the mean value of 100 replicates.

heterogeneity among different gene partitions, the same procedure was repeated for three gene partitions. The three partitions were mitochondrial genes (small-subunit ribosomal DNA and cytochrome oxidase II), rDNA array genes (5.8S, 18S, and 26S rDNA and internal transcribed spacer), and protein-coding genes (actin, elongation factor-1α, and RNA polymerase II). The results show that *ASSymDif*s do not differ significantly with respect to the taxon number utilized (table 2), although a trend of slight decreasing *ASSymDif* values is observed in response to increasing taxon number when all genes are utilized (fig. 3). In contrast, the different gene partitions in this data matrix exhibit dramatically different *ASSymDif*s (fig. 3), a finding in agreement with previous results on the inaccuracy of phylogenetic hypotheses based on a few genes (Rokas et al. 2003*b*). In conclusion, these results also suggest that phylogenetic reconstruction is not strongly influenced by taxon number.

Table 2
Significance Tests Between Subsamples of the 8-Gene, 76-Taxon Yeast Data Matrix Differing in Taxon Number and Gene Partitions

Taxon Number	10	20	30	40	50	60	70
10		NS	NS	NS	NS	NS	NS
20	0.0168		NS	NS	NS	NS	NS
30	0.0908	0.4427		NS	NS	NS	NS
40	0.2926	0.1021	0.0198		NS	NS	NS
50	0.5481	0.0720	0.0341	0.0719		NS	NS
60	0.7642	0.0549	0.0363	0.0673	0.0733		NS
70	0.9289	0.0701	0.0677	0.1286	0.2037	0.4590	

Gene partitions	mtDNA	nuDNA	riDNA	All genes
mtDNA		**	**	**
nuDNA	9×10^{-6}		**	**
riDNA	8×10^{-8}	2×10^{-8}		**
All genes	8×10^{-11}	9×10^{-8}	2×10^{-6}	

NOTE.—Significance tests were performed as described in the note to table 1.

A Positive Effect of Gene Number and Lack of Effect of Taxon Number May Be General Features of Multigene Data Matrices

An important question raised by these results for this group of yeast species concerns their relevance to other branches of the tree of life. To examine this question, we conducted the same analysis to two published data matrices, one from mammals (Murphy et al. 2001) and one from angiosperms (Zanis et al. 2002). These data matrices were chosen for two reasons. First, they contain both large numbers of taxa and genes, thus allowing testing of the effect of both taxon number and gene number on phylogenetic accuracy. The mammalian data matrix is composed of sequence data from 20 genes for 42 species, and the angiosperm data matrix is composed of sequence data from 10 genes for 16 species. Second, these data matrices cover a broad range of taxonomic diversity because each one is derived from a different eukaryotic kingdom.

Analyses of the mammalian and angiosperm data matrices reveal the same trends as those seen in the yeast data matrices (fig. 4). Increasing taxon number does not have any statistically significant effect on *ASSymDif* values in the mammal data matrix (table 3), whereas increasing gene number is significantly correlated with decreasing *ASSymDif* values and increasing phylogenetic accuracy. Increasing taxon number is correlated with a slight but nonsignificant increase in *ASSymDif* values in the angiosperm data matrix (table 4). In contrast, increasing gene number leads to a significant reduction in *ASSymDif* values. Both the mammal and angiosperm data matrices exhibit an increase in variance as taxon number decreases. The effect of gene number and taxon number on *ASSymDif* values from both data matrices is very similar to the effect observed by the two variables on the two yeast data matrices (figs. 1, 3, and 4 and tables 1–4). These results suggest that the positive effect of gene number on phylogenetic accuracy, the lack of effect of taxon number on phylogenetic accuracy, and the increased variance in phylogenetic accuracy associated with smaller taxon numbers in yeast

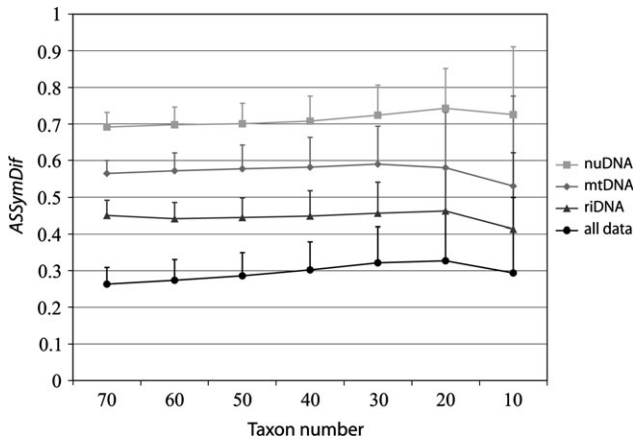

FIG. 3.—The effect of increasing taxon number on phylogenetic accuracy in a study of a data matrix comprising 8 genes and 76 taxa of yeast species (Kurtzman and Robnett 2003). Increasing taxon number (abscissa) does not correlate with *ASSymDif* values (ordinate) or phylogenetic accuracy in the analyses of mitochondrial (mtDNA), nuclear (nuDNA), and ribosomal (riDNA) DNA data partitions. A slight decrease in *ASSymDif* values and hence a slight increase in phylogenetic accuracy in response to increasing taxon number is observed when all data partitions (all data) are included in the analysis. Analyses of the different data partitions are represented by lines and symbols of different shades of gray. For example, the nuDNA data points are drawn as light gray squares connected by lines of the same shade, whereas the data points representing the analyses using all the genes (all data) are drawn as dark gray circles connected by dark gray lines.

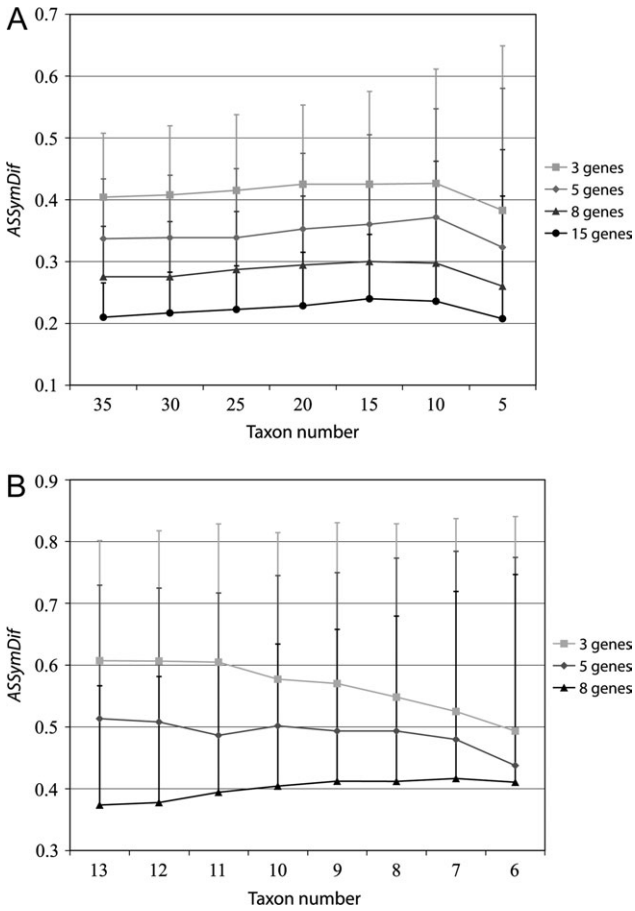

FIG. 4.—The positive effect of gene number and the lack of effect of taxon number on phylogenetic accuracy may be general features of multigene, multitaxon data sets across the tree of life. (*A*) Increasing gene number significantly correlates with increasing phylogenetic accuracy, and increasing taxon number does not significantly correlate with phylogenetic accuracy in a 20-gene, 42-taxon data matrix of mammalian taxa (Murphy et al. 2001). (*B*) Increasing gene number significantly correlates with increasing phylogenetic accuracy, whereas increasing taxon number does not significantly correlate with phylogenetic accuracy in a 10-gene, 16-taxon data matrix of angiosperm taxa (Zanis et al. 2002). In both panels, variation in gene number is represented by lines and symbols of different shades of gray. For example, the 3-gene data points are drawn as light gray squares connected by lines of the same shade, whereas the 8-gene data points are drawn as dark gray triangles connected by dark gray lines.

species may represent general properties of typical data matrices.

Discussion

We have investigated the effect of gene number and taxon number on phylogenetic accuracy in data matrices representing clades from three different kingdoms. In all clades studied, gene number was significantly and positively correlated with phylogenetic accuracy, whereas taxon number did not significantly correlate with phylogenetic accuracy. These results directly pertain to the ongoing debates regarding the choice of experimental strategies to assemble a broad sketch of the tree of life. However, the identification of the optimal experimental strategy for inferring the evolutionary history of any given clade is a complex problem, likely to be influenced by many parameters. The focus of our experimental design concentrated on the relative contribution of two variables (taxon number and gene number) to a specific measure of phylogenetic accuracy (*ASSymDif*). There are important caveats associated with the adoption of any of these parameters (*ASSymDif*, gene number, or taxon number) as the sole criterion for designing experimental strategies. Here, we discuss these caveats in turn.

Tree Distance Metrics as Measures of Phylogenetic Accuracy

In this study, estimation of phylogenetic accuracy is accomplished by measuring the tree distance between topologies obtained by the analysis of subsampled data matrices against topologies obtained by the analysis of the complete data matrices. Unfortunately, there is no way of knowing whether the topology obtained by analysis of the complete data matrix is the true one. Therefore, this index of phylogenetic accuracy may be considered an index of the sensitivity of phylogenetic analyses to variations in gene and taxon number, bearing no relevance to phylogenetic accuracy. An assumption behind the observed increase in gene number in phylogenetic studies (and many times the same assumption governs the increase in taxon number) is that more genes and taxa will lead to greater accuracy in phylogenetic estimation. Therefore, if the goal in increasing gene number and taxon number in a phylogenetic study is the achievement of better phylogenetic estimates, the index we used does provide a measure of

Table 3
Significance Tests Between Subsamples of a 20-Gene, 42-Taxon Mammalian Data Matrix

Taxon Number	5	10	15	20	25	30	35
5		NS	**	NS	NS	NS	NS
10	0.00277		NS	NS	NS	NS	NS
15	0.00037	0.69512		NS	NS	*	*
20	0.0056	0.15303	0.08228		NS	NS	**
25	0.01391	0.05037	0.00933	0.01162		NS	NS
30	0.01486	0.0063	0.00096	0.00272	0.08877		NS
35	0.04545	0.00305	0.00112	0.00048	0.03685	0.14304	
Gene number	3	5	8	15			
3		**	**	**			
5	3.8×10^{-7}		**	**			
8	6.3×10^{-11}	3.6×10^{-7}		**			
15	5.8×10^{-10}	6.3×10^{-9}	4.4×10^{-8}				

NOTE.—Significance tests were performed as described in the note to table 1.

phylogenetic accuracy, at least in the sense adopted by researchers performing multigene analyses.

Utilization of the symmetric difference tree distance does not allow observation of the scaling of phylogenetic accuracy for each internal branch. This has important consequences for how these results should be interpreted. For example, a decrease in tree distance (and an increase in phylogenetic accuracy) may be obtained by increasing success in resolving specific internal branches of a topology while leaving other internal branches of the topology unaffected (e.g., increasing success at resolving shallow internal branches, while still failing to resolve deep internal branches, Bininda-Emonds et al. 2001).

A Plethora of Genes Will Not Be a Panacea for Phylogenetic Inference

Increasing gene number positively correlates with an increase in phylogenetic accuracy in all data matrices examined in this study. Nevertheless, the availability of multiple unlinked genes will not be a panacea for phylogenetic inference. In any given phylogenetic problem, accuracy will be influenced by a large number of parameters (distance between successive lineage-splitting events, differences in the rate of evolution among lineages, etc.), many of which may be specific to the phylogeny at hand or difficult to identify a priori (Bininda-Emonds et al. 2001; Sanderson and Shaffer 2002; Rokas et al. 2003*a*). For example, increasing gene number may have no effect if care is not taken in matching the level of sequence variation of the genes selected to the phylogenetic depth to be resolved (e.g., Rokas et al. 2002). Furthermore, if the topologies of some genes differ from the history of the taxa (Maddison 1997), increasing gene number may not lead to increasing phylogenetic accuracy.

A question raised by the results of this study is why the increase in gene number per taxon has a larger effect than an increase in taxon number. From a theoretical standpoint, important determinants of phylogenetic accuracy (e.g., number of informative sites) are expected to be more dependent on gene number than on taxon number (Swofford et al. 1996). However, there is a theoretical upper limit in the amount of history preserved by gene sequences to resolve polytomous splits (Bininda-Emonds et al. 2001; Mossel 2001; Sober and Steel 2002). If there is no phylogenetic signal preserved in gene sequences, it follows that increasing gene number will not be helpful (nor will anything else for that matter) in reconstructing evolutionary history (e.g., Rodrigo et al. 1994; Sober and Steel 2002; Rokas et al. 2003*a*).

Table 4
Significance Tests Between Subsamples of a 10-Gene, 16-Taxon Angiosperm Data Matrix

Taxon Number	6	7	8	9	10	11	12	13
6		NS	NS	NS	NS	NS	NS	NS
7	0.1306		NS	NS	NS	NS	NS	NS
8	0.1729	0.3207		NS	NS	NS	NS	NS
9	0.1822	0.3353	0.4104		NS	NS	NS	NS
10	0.2256	0.3844	0.4504	0.6920		NS	NS	NS
11	0.3209	0.5532	0.6895	0.8575	0.9543		NS	NS
12	0.3654	0.5685	0.6788	0.8213	0.8705	0.8646		NS
13	0.3771	0.5743	0.6801	0.8158	0.8564	0.8566	0.8255	
Gene number	3	5	8					
3		**	**					
5	5×10^{-5}		**					
8	6×10^{-5}	0.0002						

NOTE.—Significance tests were performed as described in the note to table 1.

Taxon Number May Not Be Important for Phylogenetic Accuracy but Other Properties of the Sampled Taxa Are

These results suggest that taxon number does not affect phylogenetic accuracy. While taxon number may not be a contributing factor to accuracy in phylogenetic inference, other properties of the sampled taxa may turn out to be key determinants. A factor known to generate instability in phylogenetic inference is the use of specific taxa (Thorley and Wilkinson 1999; Sanderson and Shaffer 2002). For example, higher evolutionary rates in a subset of the taxa included in a data matrix (often these taxa are identified through their long branches) can lead, through the attraction of these long branches, to failure in phylogenetic reconstruction (Felsenstein 1978; Hendy and Penny 1989). Long-branched taxa have been identified in several clades (Carmean and Crespi 1995; Philippe 2000; D. E. Soltis and P. S. Soltis 2004), but their frequency, cause, and distribution across the tree of life are largely unknown. Simulation studies have shown that problems associated with long branches can be dramatically alleviated by the addition of specific taxa (Graybeal 1998; Hillis 1998; but see also Poe and Swofford 1999). Further, studies with biological sequences have suggested that understanding patterns of covarion evolution is important for understanding and predicting the effects of taxon addition in specific cases. For example, Lockhart et al. (2000) report observations on five eubacterial data sets, suggesting that the addition of taxa with distributions of variable sites that deviate from the clade to which they belong increases phylogenetic instability. Another factor that may be associated with phylogenetic accuracy is the density of taxon sampling in a clade. In a very interesting case study of the effect of density on phylogenetic accuracy using simulations, it was shown that a phylogeny with 10% coverage out of a total of 200 species can be more accurate than a phylogeny with 10% coverage out of a total of 2,000 species, despite the fact that in the latter case, the number of taxa utilized is an order of magnitude greater (Rannala et al. 1998).

Conclusions

These results are most relevant to the objective of resolving critical branches of the eukaryotic tree where data matrices composed of a few genes have generated phylogenetic hypotheses either lacking power or in conflict with each other, such as the tree of animals (Rokas et al. 2003*a*). In such cases, increasing gene number, irrespective of the number of taxa used, may be a prerequisite for improving phylogenetic accuracy (Cummings, Otto, and Wakeley 1995; Rokas et al. 2003*b*). As the number of genomes being sequenced is rapidly increasing, many researchers are taking advantage of the enormous amount of new data available to examine long-standing phylogenetic problems. A question frequently posed is whether the phylogenies will prove robust to the addition of more taxa. Based on these results, there is reason to expect that these phylogenies will generally be robust.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online (www.mbe.oupjournals.org).

Acknowledgments

We thank P. Lockhart, D. Krueger, and B. Williams and three anonymous referees for extensive comments on previous versions of the manuscript, B. Larget for help with the experimental design, and W. Murphy, D. Soltis, and C. Kurtzman for kindly sharing data matrices. A.R. is a Human Frontier Science Program Long-Term Fellow and S.B.C. an investigator of the Howard Hughes Medical Institute.

Literature Cited

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**:972–977.

Bapteste, E., H. Brinkmann, J. A. Lee et al. (11 co-authors). 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Natl. Acad. Sci. USA **99**:1414–1419.

Bininda-Emonds, O. R., S. G. Brady, J. Kim, and M. J. Sanderson. 2001. Scaling of accuracy in extremely large phylogenetic trees. Pac. Symp. Biocomput. 547–558.

Blair, J. E., K. Ikeo, T. Gojobori, and S. B. Hedges. 2002. The evolutionary position of nematodes. BMC Evol. Biol. **2**:7.

Carmean, D., and B. J. Crespi. 1995. Do long branches attract flies? Nature **373**:666.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science **301**:71–76.

Cummings, M. P., S. P. Otto, and J. Wakeley. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. **12**:814–822.

Dietrich, F. S., S. Voegeli, S. Brachat et al. (17 co-authors). 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science **304**:304–307.

Dujon, B., D. Sherman, G. Fischer et al. (67 co-authors). 2004. Genome evolution in yeasts. Nature **430**:35–44.

Erdos, P. L., M. A. Steel, L. A. Szekely, and T. J. Warnow. 1999*a*. A few logs suffice to build (almost) all trees (I). Random Struct. Algorithms **14**:153–184.

———. 1999*b*. A few logs suffice to build (almost) all trees: Part II. Theor. Comput. Sci. **221**:77–118.

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

———. 1985. Phylogenies and the comparative method. Am. Nat. **125**:1–15.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

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

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41**:95–98.

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.

———. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. **47**:3–8.

Hittinger, C. T., A. Rokas, and S. B. Carroll. 2004. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. Proc. Natl. Acad. Sci. USA **101**:14144–14149.

Jones, T., N. A. Federspiel, H. Chibana et al. (12 co-authors). 2004. The diploid genome sequence of *Candida albicans*. Proc. Natl. Acad. Sci. USA **101**:7329–7334.

Kellis, M., B. W. Birren, and E. S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature **428**:617–624.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423**:241–254.

Kim, J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. **47**:43–60.

Kopp, A., and J. R. True. 2002. Phylogeny of the oriental *Drosophila melanogaster* species group: a multilocus reconstruction. Syst. Biol. **51**:786–805.

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.

Lockhart, P. J., D. Huson, U. Maier, M. J. Fraunholz, Y. Van de Peer, A. C. Barbrook, C. J. Howe, and M. A. Steel. 2000. How molecules evolve in eubacteria. Mol. Biol. Evol. **17**:835–838.

Lutzoni, F., F. Kauff, C. J. Cox et al. (47 co-authors). 2004. Assembling the fungal tree of life: progress, classification and evolution of subcellular traits. Am. J. Bot. **91**:1446–1480.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. **46**:523–536.

Moncalvo, J. M., R. Vilgalys, S. A. Redhead et al. (14 co-authors). 2002. One hundred and seventeen clades of euagarics. Mol. Phylogenet. Evol. **23**:357–400.

Mossel, E. 2001. Reconstruction on trees: beating the second eigenvalue. Ann. Appl. Probab. **11**:285–300.

Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. Nature **409**:614–618.

Peterson, K. J., and D. J. Eernisse. 2001. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol. Dev. **3**:170–205.

Philippe, H. 2000. Long branch attraction and protist phylogeny. Protist **151**:307–316.

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

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.

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

Robinson, D. R., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

Rodrigo, A. G., P. R. Bergquist, P. L. Bergquist, and P. R. Reeves. 1994. Are sponges animals? An investigation into the vagaries of phylogenetic inference. Pp. 47–54 *in* R. W. M. van Soest, T. M. G. van Kempen, and J. C. Braekman, eds. Sponges in time and space. Balkema, Rotterdam, The Netherlands.

Rokas, A., N. King, J. Finnerty, and S. B. Carroll. 2003*a*. Conflicting phylogenetic signals at the base of the metazoan tree. Evol. Dev. **5**:346–359.

Rokas, A., J. A. A. Nylander, F. Ronquist, and G. N. Stone. 2002. A maximum likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera: Cynipidae); implications for insect phylogenetic studies. Mol. Phylogenet. Evol. **22**:206–219.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003*b*. 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.

Sanderson, M. J., and H. B. Shaffer. 2002. Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. **33**: 49–72.

Satta, Y., J. Klein, and N. Takahata. 2000. DNA archives and our nearest relative: the trichotomy problem revisited. Mol. Phylogenet. Evol. **14**:259–275.

Sober, E., and M. Steel. 2002. Testing the hypothesis of common ancestry. J. Theor. Biol. **218**:395–408.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. Freeman, New York.

Soltis, D. E., V. A. Albert, V. Savolainen et al. (11 co-authors). 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci. **9**:477–483.

Soltis, D. E., and P. S. Soltis. 2004. Amborella not a "basal angiosperm"? Not so fast. Am. J. Bot. **91**:997–1001.

Soltis, P. S., D. E. Soltis, and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature **402**:402–404.

Souciet, J., M. Aigle, F. Artiguenave et al. (24 co-authors). 2000. Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. FEBS Lett. **487**:3–12.

Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hillis, C. Moritz, and B. K. Mable, eds. Molecular systematics. Sinauer Associates, Sunderland, Mass.

Tautz, D., P. Arctander, A. Minelli, R. H. Thomas, and A. P. Vogler. 2003. A plea for DNA taxonomy. Trends Ecol. Evol. **18**:70–74.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Thorley, J. L., and M. Wilkinson. 1999. Testing the phylogenetic stability of early tetrapods. J. Theor. Biol. **200**:343–344.

Wolf, Y. I., I. B. Rogozin, and E. V. Koonin. 2004. Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. **14**:29–36.

Zanis, M. J., D. E. Soltis, P. S. Soltis, S. Mathews, and M. J. Donoghue. 2002. The root of the angiosperms revisited. Proc. Natl. Acad. Sci. USA **99**:6848–6853.

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

Peter Lockhart, Associate Editor

Accepted February 23, 2005