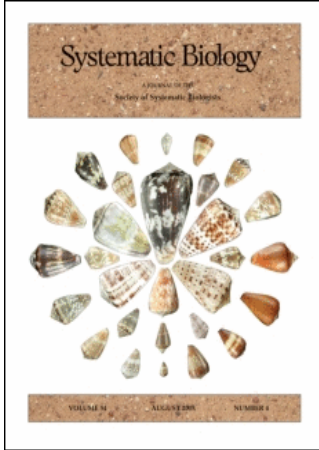


This article was downloaded by:[American Museum of Natural History]  
On: 3 July 2008  
Access Details: [subscription number 767966983]  
Publisher: Taylor & Francis  
Informa Ltd Registered in England and Wales Registered Number: 1072954  
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Systematic Biology

Publication details, including instructions for authors and subscription information:  
<http://www.informaworld.com/smpp/title~content=t713658732>

### How Many Genes Should a Systematist Sample? Conflicting Insights from a Phylogenomic Matrix Characterized by Replicated Incongruence

John Gatesy<sup>a</sup>, Rob DeSalle<sup>b</sup>, Niklas Wahlberg<sup>cd</sup>

<sup>a</sup> Department of Biology, University of California Riverside, Spieth Hall, Riverside, California, USA

<sup>b</sup> Division of Invertebrates and Molecular Systematics Laboratory, American Museum of Natural History, New York, New York, USA

<sup>c</sup> Department of Zoology, Stockholm University, Stockholm, Sweden

<sup>d</sup> Laboratory of Genetics, University of Turku, Turku, Finland

First Published on: 01 April 2007

To cite this Article: Gatesy, John, DeSalle, Rob and Wahlberg, Niklas (2007) 'How Many Genes Should a Systematist Sample? Conflicting Insights from a Phylogenomic Matrix Characterized by Replicated Incongruence', *Systematic Biology*, 56:2, 355 — 363

To link to this article: DOI: 10.1080/10635150701294733  
URL: <http://dx.doi.org/10.1080/10635150701294733>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

- Wilkinson, M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. *Mol. Biol. Evol.* 13:437–444.
- Wilkinson, M., J. A. Cotton, C. Creevey, O. Eulenstein, S. R. Harris, F.-J. Lapointe, C. Levasseur, J. O. McInerney, D. Pisani, and J. L. Thorley. 2005. The shape of supertrees to come: Tree shape related properties of fourteen supertree methods. *Syst. Biol.* 54:419–31.
- Wilkinson, M., F.-J. Lapointe, and D. J. Gower. 2003. Branch lengths and support. *Syst. Biol.* 52:127–130.
- Winkworth, R. C., D. Bryant, P. J. Lockhart, D. Havell, and V. Moulton. 2005. Biogeographic interpretation of splits graphs: Least squares optimization of branch lengths. *Syst. Biol.* 54:56–65.
- Xu, S. 2000. Phylogenetic analysis under reticulate evolution. *Mol. Biol. Evol.* 17:897–907.
- Zaretskii, K. 1965. Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Matematicheskikh Nauk* 20:90–92 [in Russian].

First submitted 9 May 2006; reviews returned 7 July 2006;

final acceptance 15 October 2006

Associate Editor: Allan Baker

*Syst. Biol.* 56(2):355–363, 2007  
Copyright © Society of Systematic Biologists  
ISSN: 1063-5157 print / 1076-836X online  
DOI: 10.1080/10635150701294733

## How Many Genes Should a Systematist Sample? Conflicting Insights from a Phylogenomic Matrix Characterized by Replicated Incongruence

JOHN GATESY,<sup>1</sup> ROB DESALLE,<sup>2</sup> AND NIKLAS WAHLBERG<sup>3</sup>

<sup>1</sup>Department of Biology, University of California Riverside, Spieth Hall, Riverside, California 92521, USA; E-mail: john.gatesy@ucr.edu

<sup>2</sup>Division of Invertebrates and Molecular Systematics Laboratory, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024, USA; E-mail: desalle@amnh.org

<sup>3</sup>Department of Zoology, Stockholm University, S-106 91, Stockholm, Sweden and Laboratory of Genetics, University of Turku, 20014 Turku, Finland; E-mail: niklas.wahlberg@utu.fi

The average size of molecular systematic data sets has grown steadily over the past 20 years. Combined phylogenetic matrices that include multiple genetic loci currently are the norm, and in many cases, rapid compilation of extremely large DNA data sets is feasible. Thus, a frequently asked question is “How many genes should a systematist sequence in order to generate a robust phylogenetic hypothesis?” This query generally has been addressed by computer simulation, where the amount of virtual DNA sequence data that can be generated is unlimited (e.g., Huelsenbeck and Hillis, 1993). Genomic data, however, provide systematists with a multitude of empirical molecular data for phylogenetic analysis, and several authors have taken advantage of this resource to examine the effects of increasing the number of genes to quantities that seemed impossible in the recent past (e.g., Cummings et al., 1995; Baptiste et al., 2002; Goremykin, 2004).

In one noteworthy study, Rokas et al. (2003) compiled a large systematic matrix of 127,026 nucleotide positions from 106 genes for 7 species of *Saccharomyces* yeast and an outgroup (*Candida albicans*). Maximum likelihood (ML) and parsimony analyses of this large data set produced congruent, well-supported results with bootstrap scores of 100% for all clades (Fig. 1a). In spite of this overwhelming support, Rokas et al. (2003) noted that there was widespread topological conflict among gene trees. Separate analyses of individual genes produced various

topologies that contradicted all nodes in the tree based on concatenation of 106 genes (Fig. 1a). Pairwise comparisons of gene trees showed extensive incongruence, and one conflicting clade, *S. kudriavzevii* + *S. bayanus*, was supported by a very large percentage of the gene trees (Fig. 1a). Replicated support for this anomalous clade was apparent in analyses of nucleotides, transversions, codons, and amino acids for a variety of systematic methods (Rokas et al., 2003; also see Holland et al., 2004, 2006; Phillips et al., 2004; Taylor and Piel, 2004; Collins et al., 2005; Gatesy et al., 2005; Ren et al., 2005; Hedtke et al., 2006).

By examining correlations between bootstrap scores and possible confounding factors, however, Rokas et al. (2003) concluded that “... none of the factors known or predicted to cause phylogenetic error could systematically account for the observed incongruence, suggesting that there may be no good predictor of the phylogenetic informativeness of genes” (p. 802). Therefore, many randomly selected genes were necessary to overwhelm conflicting signals. In this case study, very large concatenated data sets of ~20 genes were required to provide 95% bootstrap support for all nodes in the combined data tree, “substantially more genes than commonly used but a small fraction of any genome” (p. 799). Rokas et al. (2003) concluded that “These results have important implications for resolving branches of the tree of life” (p. 799) and “... important implications for many current practices in

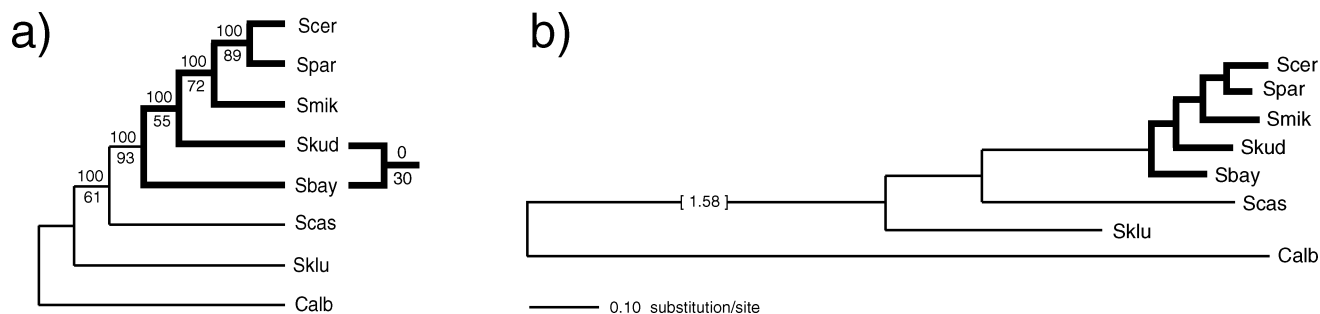


FIGURE 1. (a) The tree supported by the concatenation of 106 genes from Rokas et al. (2003). ML bootstrap scores are above internodes, and the percentage of ML gene trees that strictly supported a particular clade are indicated below internodes. The most common, conflicting clade, Skud+Sbay, also is shown. (b) Branch lengths for the optimal ML tree for the concatenation of 106 yeast genes; scale bar shows expected numbers of substitutions per site (length of outgroup branch is indicated). All phylogenetic analyses in this paper were branch and bound searches executed in PAUP\* 4.0b10 (Swofford, 2002). All ML models were chosen by likelihood ratio tests as in Rokas et al. (2003) using PAUP\* and ModelTest 3.06 (Posada and Crandall, 1998). Bootstrap analyses (Felsenstein, 1985) were as in Gatesy and Baker (2005). Scer = *Saccharomyces cerevisiae*; Spar = *S. paradoxus*; Smik = *S. mikatae*; Skud = *S. kudriavzevii*; Sbay = *S. bayanus*; Scas = *S. castellii*; Sklu = *S. kluyveri*; and Calb = *Candida albicans*.

molecular phylogenetics" (p. 802), points that were reasserted in a commentary by Gee (2003).

Specifically, if 20 or more genes generally are required to yield robust support, then most previous phylogenetic analyses are inadequate in terms of character sampling. This assertion is based on the assumption that the 8 taxa analyzed by Rokas et al. (2003) represent a typical systematic problem. Rokas et al. (2003) considered this issue, noting that "It is possible that the 8 yeast taxa we have analyzed represent a very difficult phylogenetic case, atypical of the situations found in other groups. However, the widespread occurrence of incongruence at all taxonomic levels argues strongly against such a view. Rather, we believe that this group is a representative model for key issues that researchers in phylogenetics are confronting" (p. 802). Large matrices that combine information from 20 or more gene fragments are rare (e.g., Murphy et al., 2001; Baptiste et al., 2002; Gatesy et al., 2002; Goremykin, 2004); therefore, if the test case of Rokas et al. (2003) is representative, most published molecular systematic studies are, at best, preliminary efforts.

Rokas et al. (2003) primarily used the nonparametric bootstrap (Felsenstein, 1985) to assess support and to search for correlates of incongruence in the yeast matrix. Recent reanalyses have utilized a variety of techniques to further characterize conflicting signals in the yeast data set. These approaches included Bayesian analysis (Taylor and Piel, 2004; Jeffroy et al., 2006), transversion coding (Phillips et al., 2004; Jeffroy et al., 2006), removal of rapidly evolving third codon positions (Collins et al., 2005; Jeffroy et al., 2006), partitioned Bremer support scores (Collins et al., 2005; Gatesy et al., 2005), consensus networks (Holland et al., 2004, 2006), isolation of genes with shifting base compositional biases (Collins et al., 2005), supertree bootstrapping (Burleigh et al., 2006), increased taxon sampling (Rokas and Carroll, 2005; Hedtke et al., 2006), and better fitting models of molecular evolution (Ren et al., 2005). Alternatively, several authors have suggested that reducing the number of taxa included in analysis can yield insights regarding the stabil-

ity of phylogenetic hypotheses (Lanyon, 1985; Philippe and Douzery, 1994; Siddall, 1995; Brochu, 1997; Poe, 1998; Siddall and Whiting, 1999; Holland et al., 2003).

Here, we use selected removal of taxa to explore patterns of incongruence in the yeast data set. In particular, we analyze different subsets of species to determine whether disagreements among gene trees are tempered or accentuated by altering taxonomic representation. In combination with documentation of branch lengths for individual gene trees, our subsampling results show that the set of species analyzed by Rokas et al. (2003) is not representative of most published systematic studies. We suggest that the yeast matrix does not provide a coherent, general recommendation for how many genes to sample in future molecular systematic studies. However, patterns of conflict for different subsets of species offer a very simple explanation for replication of the discrepant *S. kudriavzevii* + *S. bayanus* clade in many gene trees (Fig. 1a).

#### EXCEPTIONALLY LONG BRANCHES

Examination of the optimal topology for the concatenation of 106 genes showed a striking difference between branches that connected 5 closely related *Saccharomyces* species (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*) and branches that led to *S. castellii*, *S. kluyveri*, and the outgroup *C. albicans* (Fig. 1b; Hedtke et al., 2006; Jeffroy et al., 2006). For the ML model utilized by Rokas et al. (2003), the branches that connected to *S. castellii*, *S. kluyveri*, and *C. albicans* ranged from 0.31 to 1.58 expected substitutions per site, whereas the branches that joined the remaining, closely related *Saccharomyces* species were from 0.03 to 0.08 substitutions per site. Only 15% of the inferred nucleotide substitutions occurred on branches that linked these 5 species (Fig. 1b).

Consistent with an estimated Precambrian (~723 Mya) divergence of *Candida albicans* from *Saccharomyces cerevisiae* (Hedges et al., 2004), the outgroup branch in the yeast tree was exceptionally long. Each site in the concatenated data set is expected to change 1.58 times on this

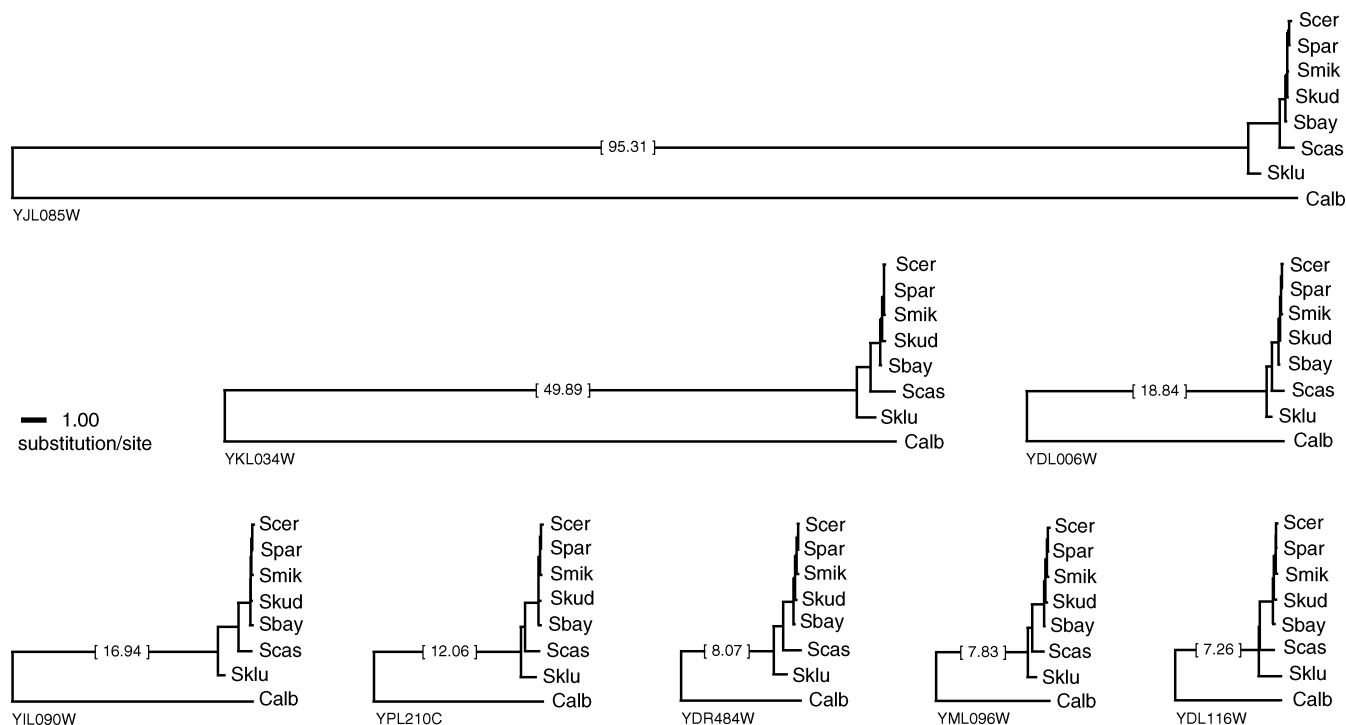


FIGURE 2. The 8 yeast genes with the longest ML branch lengths for the tree supported by the concatenation of 106 genes. The scale bar represents 1.00 expected substitution per site; branches that connect the 5 most closely related *Saccharomyces* species are tiny at this scale. The length of the longest branch in each yeast gene tree is indicated. Note that some topologies show more than one branch that is  $>1.00$  expected substitution per site. Abbreviations for yeast species are as in Figure 1.

branch according to the ML estimate (Fig. 1b). For the topology supported by the concatenation of 106 genes, 43% of the yeast genes had one branch that was  $>2.00$  expected substitutions per site, and 79% of the yeast genes had at least one branch that was  $>1.00$  substitution per site (also see Hedtke et al., 2006). For comparison, in an often cited discussion of long branches in a 28S rDNA tree of holometabolous insects, Huelsenbeck (1998) remarked that two branches in his analysis were "among the longest ever observed (approximately 1.0 substitution per site)" (p. 530). However, branches in many of the yeast gene trees dwarfed those in the insect rDNA tree and were up to 95 times longer (Fig. 2). From another perspective, the longest branches in the yeast data set exceeded those in a tree based on mitochondrial genomes from 5 animal phyla (Naylor and Brown, 1998) and also were much longer than branches in simulations designed to assess misplacement of long branches (e.g., Anderson and Swofford, 2004). Although it has been suggested that the set of species in the yeast data set represents a typical phylogenetic problem (Rokas et al., 2003), the extraordinarily long branch lengths in most yeast gene trees demonstrate that this is not the case (e.g., Fig. 2).

#### COMPLETE CONGRUENCE FOR FIVE CLOSELY RELATED *Saccharomyces* SPECIES

For the yeast data set, gene trees that included all 8 species showed many conflicts among the 5 most closely

related *Saccharomyces* species (Fig. 1a). In ML analyses, 11% of gene trees conflicted with the *S. cerevisiae* + *S. paradoxus* clade, 28% conflicted with the *S. cerevisiae* + *S. paradoxus* + *S. mikatae* clade, and 45% conflicted with the *S. cerevisiae* + *S. paradoxus* + *S. mikatae* + *S. kudriavzevii* clade. Given the moderate lengths of branches that linked *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus* (Fig. 1b), we were surprised by the widespread discrepancies among genes at this level.

To further explore differences among gene trees, we reanalyzed the 5 closely related species of *Saccharomyces* in isolation from their distant relatives, *S. castellii*, *S. kluyveri*, and *C. albicans*. Given the diversity of gene trees for all 8 taxa, we expected to find many conflicting topologies but were shocked by complete congruence among the 106 gene trees in ML analyses (Fig. 3). There are 15 possible bifurcating trees (unrooted) for a data set of 5 taxa; assuming an equal probability for each topology a priori, the chance of recovering the same tree 106 straight times is astronomically low ( $P = 3.24 \times 10^{-124}$ ). Ironically, a systematic data set that has been presented as a prime example of pervasive, inexplicable conflict among genes (Rokas et al., 2003) can be transformed, with the removal of 3 species, into a remarkably congruent data set that shows 100% agreement among 106 genes. For the set of 5 closely related *Saccharomyces* species, 20 genes were not necessary to resolve relationships; basically any gene will do (Fig. 3c). Subsets of only 600 randomly resampled nucleotides from the yeast data

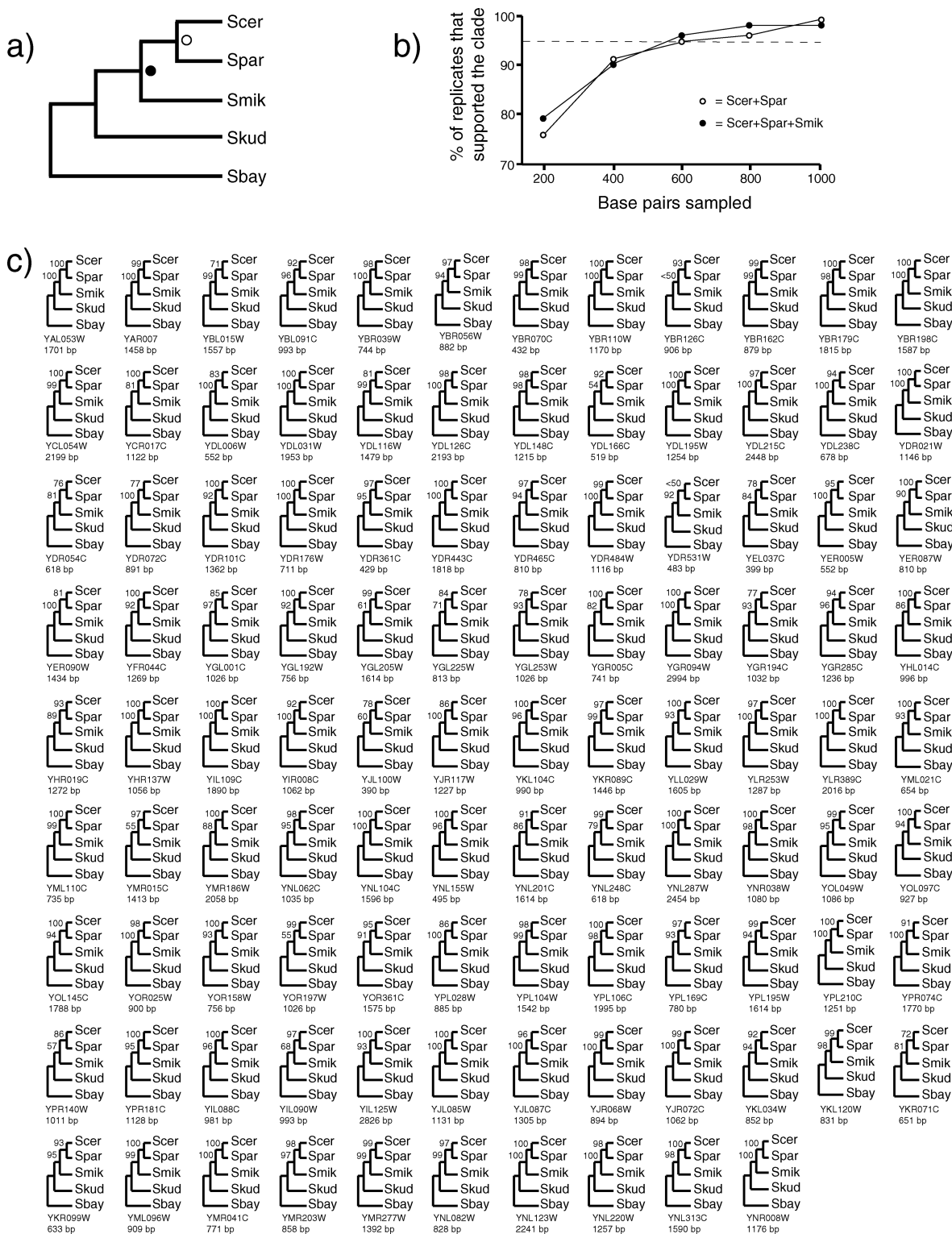


FIGURE 3. One hundred percent congruence among 106 gene trees for the 5 closely related *Saccharomyces* species. In separate ML analyses, all genes supported the same topology (a) that included Scer+Spar (white circle) and Scer+Spar+Smik (black circle). Very few randomly sampled nucleotides were required to consistently recover these clades in analyses of the 5 closely related *Saccharomyces* species (b); subsamples of only 600 nucleotides supported each clade >95% of the time (dotted line = 95%). The 106 completely congruent ML gene trees for Scer, Spar, Smik, Skud, and Sbay are shown in (c); ML bootstrap scores and the number of nucleotides for each gene are indicated. Species abbreviations are as in Figure 1. For all genes, ML models for 5 species were chosen as in Rokas et al. (2003) using PAUP\* and ModelTest 3.06. Analyses of randomly sampled sites from the yeast data set (b) were done in PAUP\* and included 500 replicates for each sample size (200, 400, 600, 800, and 1000 nucleotide sites). ML searches were branch and bound.

set were sufficient to support the 2 nodes in the 5 taxon tree in >95% of replicates, and only 200 nucleotides recovered each clade >75% of the time (Fig. 3a, b). These are very small samples of characters relative to most modern systematic studies.

#### INCONGRUENCE AMONG GENES WITH THE ADDITION OF DIVERGENT TAXA

In ML analyses of *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus*, there were no topological discrepancies among gene trees (Fig. 3c), but when more distantly related taxa (*S. castellii*, *S. kluyveri*, and *C. albicans*) were included, gene trees showed extensive conflicts regarding relationships among the 5 closely related *Saccharomyces* species. Previously, we noted such conflicts for the full complement of 8 species (Fig. 1a). Analyses of the 7 *Saccharomyces* species, excluding the outgroup *C. albicans*, also revealed widespread incongruence among genes (Rokas et al., 2003), as did analyses of 6 species (Fig. 4).

For the 6 species set composed of 5 closely related *Saccharomyces* species and *C. albicans*, the disparity in length between the outgroup branch and ingroup branches was greatest. Approximately 87% of the expected character

change was restricted to the outgroup branch (3.04 substitutions per site), and the concatenation of 106 genes supported a grouping of *S. kudriavzevii* + *S. bayanus* with an ML bootstrap score of 76% (Fig. 4). This relationship was incompatible with the *S. cerevisiae* + *S. paradoxus* + *S. mikatae* + *S. kudriavzevii* clade that had 100% bootstrap support in the analysis of 106 genes for 8 species (Fig. 1a). Thus, a systematic data set of 8 species, in which 20 genes were considered sufficient for robust phylogenetic support (Rokas et al., 2003), can be transformed, with the deletion of 2 species, into a data set of 106 genes that yields a contradictory tree; >100 concatenated genes did not provide  $\geq 95\%$  bootstrap support at all nodes for this set of 6 species (Fig. 4). Phylogenetic analyses of taxon subsamples (Figs. 3, 4) clearly show that the number of genes required to yield strong bootstrap scores is highly dependent on the particulars of a given systematic problem and suggest that long branches (Fig. 2) explain much of the incongruence among genes in the yeast data set (Taylor and Piel, 2004; Hedtke et al., 2006; Jeffroy et al., 2006).

#### HIGHLY REPLICATED INCONGRUENCE WITH THE ADDITION OF DISTANT TAXA

In ML analyses of the 5 closely related *Saccharomyces* species, all gene trees had the same 7 branches (Fig. 3a). Assignment of the root to 5 of these 7 branches will yield the incongruent *S. kudriavzevii* + *S. bayanus* clade (Fig. 5a). When the distantly related *C. albicans* was added to a matrix that included *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus*, all ML gene trees were consistent with the unrooted topology for these 5 *Saccharomyces* species (Fig. 3a), but rooting position was scattered among the 7 ingroup branches (Fig. 5b). The most common root position was on the "correct" branch (*S. bayanus*; 31 times). For the other 75 genes, the root was distributed across the remainder of the topology, and the majority of gene trees (57 of 106) supported the *S. kudriavzevii* + *S. bayanus* clade (Figs. 4 and 5b). Assuming an equal a priori probability of recovering the *S. kudriavzevii* + *S. bayanus* clade or the *S. cerevisiae* + *S. paradoxus* + *S. mikatae* + *S. kudriavzevii* clade, it would be highly unlikely for one of these groups to be supported in  $\geq 57$  of 88 gene trees, as was the case here (binomial probability of 0.007). Analogous but less extreme patterns were observed in ML gene trees for other combinations of 6 taxa, in which the 5 closely related *Saccharomyces* species were rooted with either *S. castellii* or *S. kluyveri*. *S. kudriavzevii* + *S. bayanus* was always the most common, conflicting clade (Fig. 4). Likewise, Rokas et al. (2003) documented the same pattern of replicated support for the conflicting *S. kudriavzevii* + *S. bayanus* in gene trees for the 7 *Saccharomyces* species, excluding the outgroup *C. albicans*.

Of the 10,395 possible bifurcating topologies for all 8 species, the *S. kudriavzevii* + *S. bayanus* bipartition is found in only 9% of all trees. However, ML analyses of 8 species showed that *S. kudriavzevii* + *S. bayanus* was recovered 32 times in 106 gene trees (Fig. 1a);

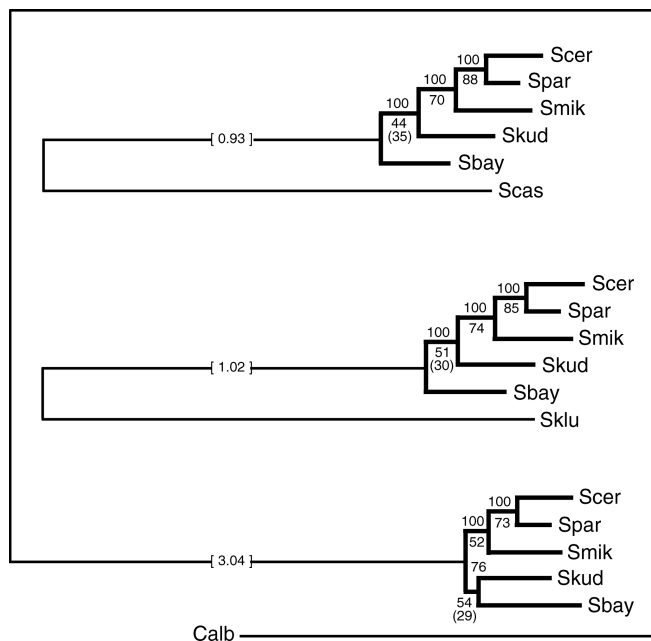


FIGURE 4. Trees supported by the concatenation of 106 genes in ML analyses of 6 species. Note the very long outgroup branches; expected substitutions per site for the outgroup branches are indicated. In the topology rooted with Calb, the conflicting Skud+Sbay clade was supported. ML bootstrap scores are above internodes, and the percentage of ML gene trees that strictly supported a particular clade are indicated below internodes. For the top and middle trees, clades in parentheses indicate the percentage of times that the Skud+Sbay clade was supported; for the bottom tree, the number in parentheses is the percentage of gene trees that supported the Scer+Spar+Smik+Skud clade. Species abbreviations are as in Figure 1. For each gene and for the concatenation of 106 genes, ML models for each set of 6 species were chosen as in Rokas et al. (2003) using PAUP\* and ModelTest 3.06.

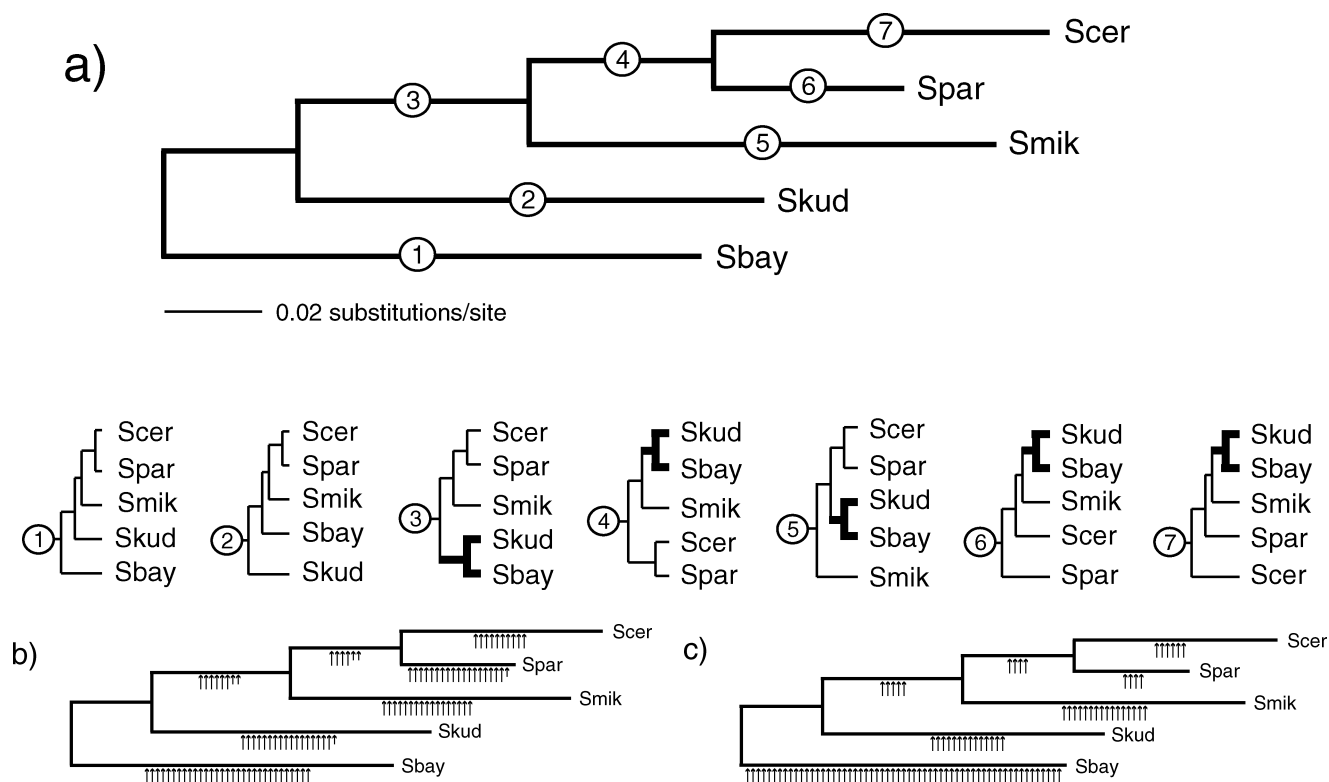


FIGURE 5. (a) ML tree supported by concatenation of 106 genes for the 5 closely related *Saccharomyces* species, with possible root placements shown. Five of the 7 roots yield the conflicting Skud+Sbay clade. (b) Placement of the root in 106 ML gene trees for all 8 yeast species (Scer, Spar, Smik, Skud, Sbay, and Calb). (c) Placement of the root in 106 ML gene trees for all 8 yeast species (Scer, Spar, Smik, Skud, Sbay, and Calb). Arrows on branches indicate the number of times a particular branch was rooted by Calb (b) or by Calb, Scas, and Sklu (c). For 3 genes, there were 2 optimal rootings; truncated arrows on branches in (b) indicate that in 1 of the 2 best trees, the root was assigned to that branch. Species abbreviations are as in Figure 1.

previous parsimony, ML, and Bayesian results showed this same pattern of replicated incongruence whether nucleotides, transversions, codons, or amino acids were analyzed (Rokas et al., 2003; Phillips et al., 2004; Taylor and Piel, 2004; Collins et al., 2005; Gatesy et al., 2005; Ren et al., 2005; Burleigh et al., 2006; Holland et al., 2004, 2006). Repeated recovery of the incongruent *S. kudriavzevii* + *S. bayanus* clade in ~30% of our ML gene trees strongly suggested an underlying bias. Once again, all ML gene trees for 8 species were compatible with relationships in the unrooted tree for the 5 closely related *Saccharomyces* species (Fig. 3a), but different placements of the 3 long branch taxa (Fig. 6a, b) resulted in many gene trees that supported the *S. kudriavzevii* + *S. bayanus* clade (Figs. 1a and 5c). As in the analyses of 6 taxa (Fig. 4), replicated support for the conflicting *S. kudriavzevii* + *S. bayanus* grouping was due to erratic rooting of the uniformly supported, pectinate topology for *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus* (Fig. 3a) by the very distantly related *S. castellii*, *S. kluyveri*, and *C. albicans* (Figs. 4 to 6; for discussion of distant outgroups see Wheeler, 1990; Huelsenbeck et al., 2002; Holland et al., 2003; Anderson and Swofford, 2004; Bergsten, 2005; Susko et al., 2005; Goloboff and Pol, 2005; Hedke et al., 2006).

#### HOW MANY GENES ARE ENOUGH?

Rokas et al. (2003) suggested that their analyses of 106 genes from 8 species had important implications for resolving the tree of life. In particular, they argued that 20 or more genes might be required to garner robust support for phylogenetic relationships. This assertion is based on two critical assumptions: (1) There are no good predictors for the utility of different genes, and (2) the 8 species in their data set represent a typical phylogenetic problem. Recent reanalyses of the yeast data set have contested both of these assumptions. Phillips et al. (2004) noted that differences in G-C content might explain non-historical signals in the yeast matrix. Subsequently, Collins et al. (2005) found that shifts in base composition were most prominent at third codon positions (also see Jeffroy, 2006). When Collins et al. (2005) resampled genes with stationary base compositions, only 10 genes were required to record high bootstrap percentages for relationships supported by the concatenated data set. By contrast, 23 genes characterized by large shifts in base composition were necessary to yield the same level of support (Collins et al., 2005). In a study focused on Bayesian support measures, Taylor and Piel (2004) found that, "Overall the external/internal branch length ratios were greater for trees that were incongruent

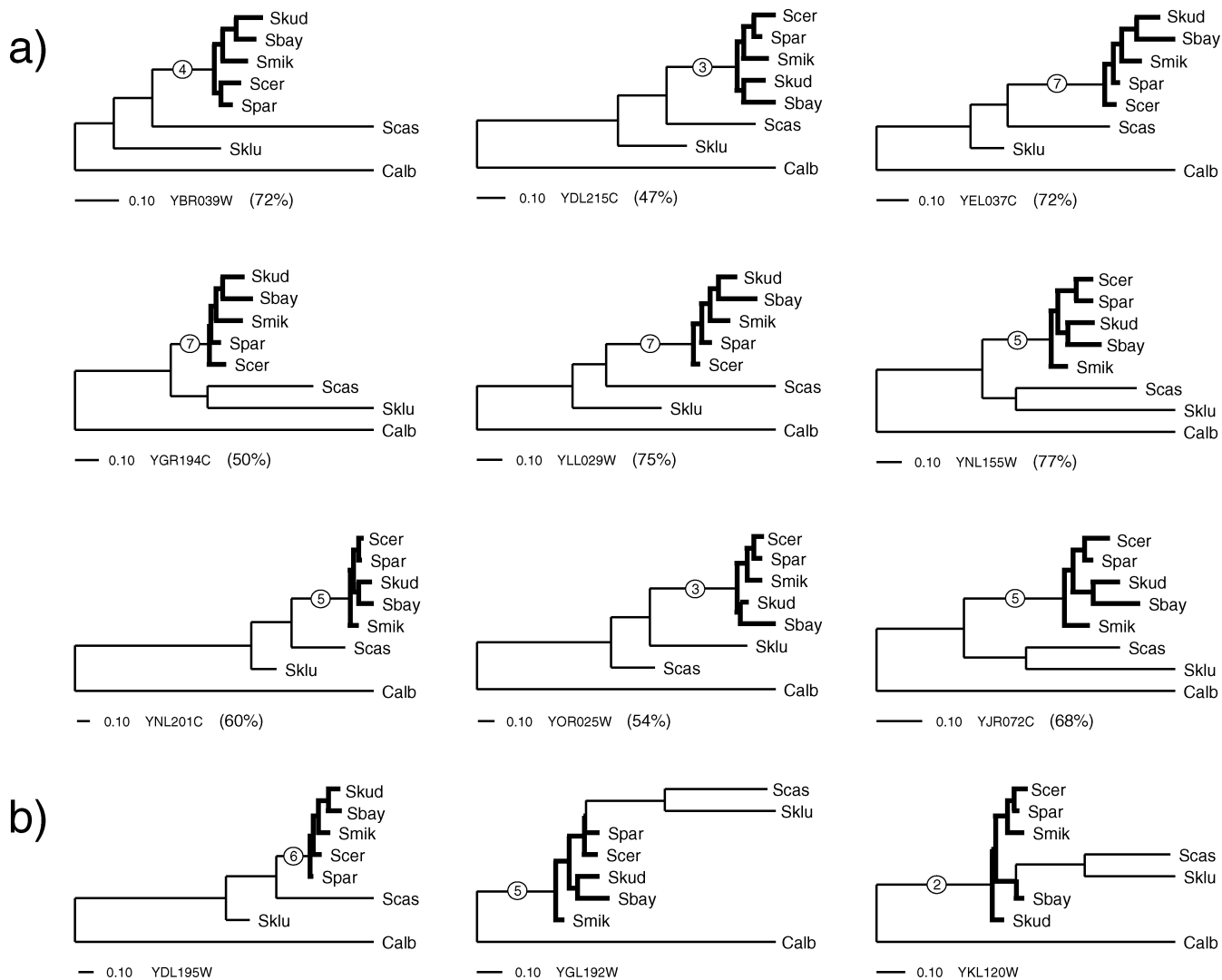


FIGURE 6. Optimal ML topologies for 12 yeast genes. All gene trees are consistent with the unrooted topology for the 5 closely related *Saccharomyces* species (Fig. 3); branches that connect these 5 species are shown as thick lines. Numbers in circles indicate rooting position according to Figure 5. The 9 genes in (a) all supported the Skud+Sbay clade (bootstrap support for Skud+Sbay is shown in parentheses). When these 9 genes were combined, Skud+Sbay was not supported, and the topology favored by the concatenation of 106 genes was optimal (see Gatesy and Baker, 2005). The three gene trees in (b) support alternative topologies. Two of these genes did not support monophyly of the 5 closely related *Saccharomyces* species but were compatible with the unrooted tree for these species (Fig. 3a).

with the reference tree [our Fig. 1a, b]...” (p. 1536), a result that was statistically significant. In sum, the contention that there are no good predictors of phylogenetic utility for particular genes does not seem to hold for this phylogenomic data set.

Following Taylor and Piel (2004), Ren et al. (2005), Hedtke et al. (2006), Jeffroy et al. (2006), and Holland et al. (2006) noted the presence of exceptionally long branches and argued that a high level of divergence and associated branch length inequalities (e.g., Fig. 2) were determinants of conflict among genes in the yeast data set. Here, we extended these arguments and concluded that the 8 species in the yeast data set do not represent a “typical” phylogenetic problem. The tree based on the concatenated matrix of 106 genes showed great disparities

in branch lengths (Fig. 1b), but individual gene trees had some truly extraordinary branches that were up to 95.31 expected substitutions per site (Fig. 2). This saturation of nucleotide substitution does not represent a typical phylogenetic problem; many systematists acknowledge that this degree of divergence is a very difficult problem (Felsenstein, 1978; Hendy and Penny, 1989; Wheeler, 1990; Huelsenbeck, 1998; Pol and Siddall, 2001; Holland et al., 2003; Anderson and Swofford, 2004; Bergsten, 2005; Susko et al., 2005). *S. castellii*, *S. kluyveri*, and *C. albicans* are very genetically distant from each other and from the 5 most closely related *Saccharomyces* species (Figs. 1, 2, 4, and 6). Therefore, it was not surprising that there were wholesale conflicts among gene trees in parsimony, Bayesian, and ML analyses (Taylor and Piel, 2004; Hedtke

et al., 2006). In fact, the 3 divergent taxa, which also were characterized by the largest shifts in base composition (Collins et al., 2005), accounted for *all* conflicts among genes in ML analyses.

Because of extensive incongruence, Rokas et al. (2003) found that 20 randomly sampled genes from the yeast matrix were required for a robustly supported tree of 8 species, but this result has no generality. Even within this 106 gene matrix, it is clear that some systematic problems are much more difficult to solve relative to others. For the 5 closely related *Saccharomyces* species, one gene might be sufficient (Fig. 3c). ML analyses of individual genes produced the same tree 106 straight times, and sets of 600 randomly sampled nucleotides consistently supported this topology (Fig. 3b). By contrast, in ML analyses of these 5 species plus *C. albicans*, 106 concatenated genes (127,026 nucleotides) apparently were insufficient; the optimal ML tree (Fig. 4) contradicted the best tree for all 8 yeast species, a topology that was thought to show an unprecedented level of support (Fig. 1a, b; Gee, 2003; Rokas et al., 2003).

Clearly, the quantity of genes that is required to robustly resolve relationships will be dependent on the specifics of the phylogenetic problem at hand (Cummings and Meyer, 2005; Hedtke et al., 2006), as well as a particular researcher's definition of "adequate support" (e.g., Satta et al., 2000; Zander, 2001; Siddall, 2002; Grant and Kluge, 2003; Soltis et al., 2004; Taylor and Piel, 2004; Jeffroy et al., 2006). For easy phylogenetic problems where divergence among taxa is not great and internodes are moderately long, a single gene might provide high bootstrap support (e.g., Fig. 3). However, even in this situation, sequencing 2 or more genes may be justified, given that tightly linked nucleotides do not necessarily provide independent evidence for phylogenetic relationships (Doyle, 1992). For cases where exceptionally long branches are apparent, even 106 genes might not be enough (e.g., Fig. 4). When faced with extreme branch lengths like these, increased taxonomic sampling (e.g., Zwickl and Hillis, 2002), evidence from the fossil record (e.g., Brochu, 1997), or a set of more slowly evolving genes (e.g., Springer et al., 2001) with stationary base frequencies (e.g., Collins et al., 2005) may be required. Educated guesses can be made, but in the end, the amount of character data needed to arrive at a stable, well-supported phylogenetic hypothesis can only be quantified by adding new data to existing data and then reassessing the results.

#### ITERATED CONFLICT IN PHYLOGENOMIC MATRICES

Our reanalyses provided a very simple explanation for replicated support of the conflicting *S. kudriavzevii* + *S. bayanus* clade in many yeast gene trees (Fig. 1a). Misrooting of a stable topology for 5 close relatives (Fig. 3a) by 3 genetically distant taxa (Fig. 4) can account for this iterated pattern. Because the topology for *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus* was pectinate (Fig. 3a), erratic placement of the root (Fig. 5) repeatedly yielded the discrepant *S. kudri-*

*avzevii* + *S. bayanus* clade. In the most extreme case that we examined (a data set that included 6 of 8 species in the yeast matrix), the "wrong clade" was preferred over the "right clade" in the majority of ML gene trees (57 of 106 = 54%) and in the concatenated analysis of 106 genes (Fig. 4).

This result represents a cautionary tale for phylogenomic studies, in which >100 genes from relatively few taxa may be sampled, and where congruence among individual gene trees has been used to assess support (e.g., Rokas et al., 2003; Holland et al., 2004, 2006; Burleigh et al., 2006). Previously, many authors have argued that large concatenations of genes can provide strong, but spurious, bootstrap support because of model misspecification, inadequate taxon sampling, or both (e.g., Philippe and Douzery, 1994; Naylor and Brown, 1998; Holland et al., 2004, 2006; Phillips et al., 2004; Soltis et al., 2004; Stefanovic et al., 2004; Hedtke et al., 2006; Jeffroy et al., 2006). Here, we documented an exceptional pattern of replicated conflict in which a consensus derived from separate analyses of >100 genes failed to give the right result; nearly twice as many gene trees favored the wrong grouping of *S. kudriavzevii* + *S. bayanus* over the right *S. cerevisiae* + *S. paradoxus* + *S. mikatae* + *S. kudriavzevii* clade (Fig. 4). In comparison to concatenation of genes, it might be expected that partitioned phylogenetic analyses of individual genes should be less prone to highly supported but spurious results. Unfortunately, this is not always the case (Fig. 4), and a simple compilation of many genes for very few taxa (Rokas et al., 2003; Rokas and Carroll, 2004) cannot be trusted as a general solution for "ending incongruence" (Gee, 2003).

#### ACKNOWLEDGEMENTS

We thank R. Baker, T. Collins, A. de Queiroz, J. Garb, C. Hayashi, M. R. McGowen, R. Page, and two anonymous reviewers for comments on different versions of the manuscript. J. Gatesy was supported by NSF (USA) DEB-0212572, DEB-0213171, and EAR-0228629; R. DeSalle was supported by the Lewis B. and Dorothy Cullman Program in Molecular Systematics at the American Museum of Natural History and by NSF (USA) DBI-0421604; N. Wahlberg was supported by the Swedish Research Council 621-2004-2853. G. Naylor provided alignments of animal mitochondrial genomes. A. Rokas provided published multiple sequence alignments and supporting materials that made the present study possible.

#### REFERENCES

- Anderson, F. E., and D. L. Swofford. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Mol. Phylogenet. Evol.* 33:440-451.
- Baptiste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Duruflé, T. Gaasterland, P. Lopez, M. Müller, and H. Philippe. 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.
- Bergsten, J. 2005. A review of long-branch attraction. *Cladistics* 21:163-193.
- Brochu, C. 1997. Morphology, fossils, divergence timing, and the phylogenetic relationships of *Gavialis*. *Syst. Biol.* 46:479-522.
- Burleigh, J. G., A. C. Driskell, and M. J. Sanderson. 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Syst. Biol.* 55:426-440.

- Collins, T. M., O. Fedrigo, and G. J. P. Naylor. 2005. Choosing the best genes for the job: The case for stationary genes in genome-scale phylogenies. *Syst. Biol.* 54:493–500.
- Cummings, M. P., and A. Meyer. 2005. Magic bullets and golden rules: Data sampling in molecular phylogenetics. *Zoology* 108:329–336.
- 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.
- Doyle, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. *Syst. Bot.* 17:144–163.
- Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* 27:401–410.
- Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791.
- Gatesy, J., and R. H. Baker. 2005. Hidden likelihood support in genomic data: Can forty-five wrongs make a right? *Syst. Biol.* 54:483–492.
- Gatesy, J., C. Matthee, R. DeSalle, and C. Hayashi. 2002. Resolution of a supertree/supermatrix paradox. *Syst. Biol.* 51:652–664.
- Gee, H. 2003. Ending incongruence. *Nature* 425:782.
- Goloboff, P. A., and D. Pol. 2005. Parsimony and Bayesian phylogenetics. Pages 148–159 in *Parsimony, phylogeny, and genomics* (V. A. Albert, ed.). Oxford University Press, Oxford, UK.
- Goremykin, V. V. 2004. The chloroplast genome of *Nymphaea alba*: Whole-genome analyses and the problem of identifying the most basal angiosperm. *Mol. Biol. Evol.* 21:1445–1454.
- Grant, T., and A. G. Kluge. 2003. Data exploration in phylogenetic inference: Scientific, heuristic, or neither. *Cladistics* 19:379–418.
- Hedges, S. B., J. E. Blair, M. L. Venturi, and J. L. Shoe. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.* 4:2.
- Hedtke, S. M., T. M. Townsend, and D. M. Hillis. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* 55:522–529.
- Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38:297–309.
- Holland, B. R., K. T. Huber, V. Moulton, and P. J. Lockhart. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. *Mol. Biol. Evol.* 21:1459–1461.
- Holland, B. R., L. S. Jermini, and V. Moulton. 2006. Improved consensus network techniques for genome-scale phylogeny. *Mol. Biol. Evol.* 23:848–855.
- Holland, B. R., D. Penny, and M. D. Hendy. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—A simulation study. *Syst. Biol.* 52:229–238.
- Huelsenbeck, J. P. 1998. Systematic bias in phylogenetic analysis: Is the Strepsiptera problem solved? *Syst. Biol.* 47:519–537.
- Huelsenbeck, J. P., J. P. Bollback, and A. M. Levine. 2002. Inferring the root of a phylogenetic tree. *Syst. Biol.* 51:32–43.
- Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* 42:247–264.
- Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: The beginning of incongruence. *Trends Genet.* 22:225–231.
- Lanyon, S. 1985. Detecting internal inconsistencies in distance data. *Syst. Zool.* 34:397–403.
- Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* 294:2348–2351.
- Naylor, G. J. P., and W. M. Brown. 1998. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* 47:61–76.
- Philippe, H., and E. Douzery. 1994. The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationship. *J. Mamm. Evol.* 2:133–152.
- Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21:1455–1458.
- Poe, S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. *Syst. Biol.* 47:18–31.
- Pol, D., and M. E. Siddall. 2001. Biases in maximum likelihood and parsimony: A simulation approach to a 10-taxon case. *Cladistics* 17:266–281.
- Posada, D., and K. Crandall. 1998. ModelTest: Testing the model of DNA substitution. *Bioinformatics* 14:817–818.
- Ren, F., H. Tanaka, and Z. Yang. 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Syst. Biol.* 54:808–818.
- Rokas, A., and S. B. Carroll. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol. Biol. Evol.* 22:1337–1344.
- Rokas, A., B. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.
- Satta, Y., J. Klein, and N. Takahata. 2000. DNA archives and our nearest relative: The trichotomy problem revisited. *Mol. Phylogenet. Evol.* 14:259–275.
- Siddall, M. 1995. Another monophyly index: Revisiting the jackknife. *Cladistics* 11:33–56.
- Siddall, M., and M. Whiting. 1999. Long-branch abstractions. *Cladistics* 15:9–24.
- Siddall, M. E. 2002. Measures of support. Pages 80–101 in *Methods and tools in biosciences and medicine: Techniques in molecular systematics and evolution* (R. DeSalle, G. Giribet, and W. Wheeler, eds.). Birkhäuser Verlag, Basel, Switzerland.
- Soltis, D. E., V. A. Albert, V. Savolainen, K. Hilu, Y.-L. Qiu, M. W. Chase, J. S. Farris, S. Stefanovic, D. W. Rice, J. D. Palmer, and P. S. Soltis. 2004. Genome-scale data, angiosperm relationships, and 'ending congruence': A cautionary tale in phylogenetics. *Trends Plant Sci.* 9:477–483.
- Springer, M. S., R. W. DeBry, C. Douady, H. M. Amrine, O. Madsen, W. W. de Jong, and M. J. Stanhope. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. *Mol. Biol. Evol.* 18:132–143.
- Stefanovic, S., D. W. Rice, and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol. Biol.* 4:35.
- Susko, E., M. Spencer, and A. J. Roger. 2005. Biases in phylogenetic estimation can be caused by random sequence segments. *J. Mol. Evol.* 61:351–359.
- Swofford, D. L. 2002. PAUP\*. Phylogenetic analysis using parsimony (\*and other methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.
- Taylor, D. J., and W. H. Piel. 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. *Mol. Biol. Evol.* 21:1534–1537.
- Wheeler, W. C. 1990. Nucleic acid sequence phylogeny and random outgroups. *Cladistics* 6:363–367.
- Zander, R. H. 2001. A conditional probability of reconstruction measure for internal cladogram branches. *Syst. Biol.* 50:425–437.
- Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* 51:588–598.

First submitted 28 April 2006; reviews returned 7 July 2006;

final acceptance 19 October 2006

Associate Editor: Tim Collins