Genome-scale approaches to resolving incongruence in molecular phylogenies

Antonis Rokas*, Barry L. Williams*, Nicole King & Sean B. Carroll

Howard Hughes Medical Institute, Laboratory of Molecular Biology, R. M. Bock Laboratories, University of Wisconsin-Madison, 1525 Linden Drive, Madison, Wisconsin 53706, USA

* These authors contributed equally to this work

One of the most pervasive challenges in molecular phylogenetics is the incongruence between phylogenies obtained using different data sets, such as individual genes. To systematically investigate the degree of incongruence, and potential methods for resolving it, we screened the genome sequences of eight yeast species and selected 106 widely distributed orthologous genes for phylogenetic analyses, singly and by concatenation. Our results suggest that data sets consisting of single or a small number of concatenated genes have a significant probability of supporting conflicting topologies. By contrast, analyses of the entire data set of concatenated genes yielded a single, fully resolved species tree with maximum support. Comparable results were obtained with a concatenation of a minimum of 20 genes; substantially more genes than commonly used but a small fraction of any genome. These results have important implications for resolving branches of the tree of life.

Understanding the historical relationships between living organisms has been one of the principal goals of evolutionary research. Molecular phylogenetic data are instrumental in research on the history of life[1–3], the polarity of phenotypic and developmental evolution[4], and on the diversity of living organisms[5]. Despite tremendous progress in recent years, phylogenetic reconstruction involves many challenges that create uncertainty with respect to the true historical associations of the taxa analysed. One of the most notable difficulties is the widespread occurrence of incongruence between alternative phylogenies generated from single-gene data sets. Incongruence occurs at all taxonomic levels, from phylogenies of closely related species[6,7] to relationships between major classes[8,9] or phyla and higher taxonomic groups[10–12].

Both analytical and biological factors may cause incongruence[10,13]. Analytical factors affecting phylogenetic reconstruction include the choice of optimality criterion[14], limited data availability[15,16], taxon sampling[17] and specific assumptions in the modelling of sequence evolution[18]. Biological processes such as the action of natural selection or genetic drift[19–22] may cause the history of the genes under analysis to obscure the history of the taxa. The large number of potential explanations for the presence of incongruence in molecular phylogenetic analyses makes decisions on how to handle conflict in larger sets of molecular data difficult[23]. For example, two genes with different evolutionary histories (for example, owing to hybridization or horizontal transfer) for a particular taxonomic group will by definition be incongruent while still depicting true histories[20]. Data sets composed of genes showing heterogeneity in mode of sequence evolution may also compound bias rather than resolve the true history[24]. Furthermore, because current tests are not always reliable[25,26], it has been difficult to estimate incongruence. To overcome the effect of analytical and biological factors by increasing the signal-to-noise ratio, many researchers have attempted to address difficult phylogenetic questions by analysis of concatenated data sets[1,27–29]. However, phylogenetic analyses of different sets of concatenated genes do not always converge on the same tree[8,9], and some studies have yielded results at odds with widely accepted phylogenies[30]. Although theory suggests that a number of factors (such as gene number, sequence length, optimality criterion and rate of evolution) may influence phylogenetic reconstruction[14,15,31], the effect of these factors has not been systematically explored with large data sets derived from biological sequences. Recent progress in the genomics of the yeast genus *Saccharomyces*[32–34] has presented an unprecedented opportunity to evaluate these issues in eukaryotic phylogenetics.

Screen for and phylogenetic analysis of orthologous genes

Genome sequence data have been obtained for seven *Saccharomyces* species (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii* and *S. kluyveri*)[32–34] as well as for the outgroup fungus *Candida albicans*. The genomes of all eight species were screened for orthologous genes to serve as phylogenetic markers (see Methods). Previous work[35–37] has suggested the occurrence of a genome duplication in the evolutionary history of the *Saccharomyces* yeasts, so our gene selection criteria relied on synteny to establish orthology (see Methods). We retained 106 genes, which are distributed throughout the *S. cerevisiae* genome on all 16 chromosomes and comprise a total length of 127,026 nucleotides (42,342 amino acids), corresponding to roughly 1% of the genomic sequence and 2% of the predicted genes. Phylogenetic reconstructions were performed in three ways: maximum likelihood (ML) analysis of the nucleotide data, and maximum parsimony (MP) analysis of both the nucleotide and the amino acid data. Because the study comprised eight taxa, we used a search strategy in each analysis that was guaranteed to find the most parsimonious or most likely tree with respect to each gene (see Methods).

Single-gene phylogenies reveal extensive incongruence

Analyses of the 106 genes resulted in more than 20 alternative ML or MP trees (see Methods). To assess the degree of support for each of these trees, 100 bootstrap replicates were generated and summarized as 50% majority-rule consensus trees. Several of the many strongly supported (defined here as a bootstrap value >70%) alternative phylogenies are shown in Fig. 1a–f. For example, some genes recovered strong support for various alternative placements of species within the *sensu stricto* group (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii* and *S. bayanus*; Fig. 1a, c–e), whereas other genes strongly supported alternative placements of *S. castellii* and *S. kluyveri* relative to the outgroup (Fig. 1a, b, f).

The 106 genes analysed were selected without consideration of their function, and their phylogenetic performance has not been evaluated previously. Genes more commonly used in molecular systematics are often chosen for both historical (for example, ease of amplification, previous analyses of related taxa) and analytical (for example, copy number, rate of evolution) reasons[38]. It is possible

that commonly used genes may provide better resolution of the same set of taxa. We repeated ML and MP analyses of nucleotide sequence on a sample of six commonly used genes (actin, hsp70, β-tubulin, RNA polymerase II, elongation factor 1-α and 18S rDNA), and also obtained alternative and often strongly supported phylogenies (Fig. 1g–l). Specifically, actin and RNA polymerase II (Fig. 1g, j) each support the same tree as in Fig. 1a, whereas hsp70 (Fig. 1h) supports the topology in Fig. 1b. β-Tubulin (Fig. 1i), 18S rDNA (Fig. 1l) and elongation factor 1-α (Fig. 1k) each support, albeit weakly at some branches, trees not represented in Fig. 1a–f. Therefore, analyses of both commonly used genes and the set of 106 orthologous genes provide strong support for several alternative phylogenies and fail to indicate which gene tree(s) might represent the actual species tree.

The alternative phylogenies could have resulted from a number of different scenarios: (1) most genes could have weakly supported most phylogenies and strongly supported only a few alternative trees; (2) most genes could have strongly supported one phylogeny and a few genes strongly supported only a small number of alternatives; (3) there could have been some combination of these scenarios so that each branch among alternative phylogenies had either weak or strong support depending on the gene. To distinguish between these possibilities, we identified all of the branches recovered during the single-gene analyses, and recorded each bootstrap value with respect to the gene and method of analysis (see Supplementary Information). Eight branches were shared by all three analyses with multiple instances of bootstrap values >50%; hence, we focus on these eight branches in our analyses.

For each branch, we observed a full range of bootstrap values (Fig. 2c). Only two branches (branches 1 and 4 in Fig. 2) were supported with high bootstrap values by a majority of genes. The remaining six branches were supported by bootstrap values that range evenly from 0 to 100 (Fig. 2c). Two summary statistics of these distributions, the mean bootstrap value and the percentage of genes supporting each branch (Fig. 2c), also indicate weak support overall for each of these six branches (Fig. 2c). Notably, bootstrap values for branches 3 versus 6 and among branches 5, 7 and 8 exhibit significant and strong negative correlations so that a low bootstrap value for one branch corresponds to a high bootstrap value for an alternative branch (Pearson correlation coefficients and *P*-values are: −0.684 and <0.001, −0.734 and <0.001, and −0.607 and <0.001 for branches 3 versus 6, 5 versus 7, and 5 versus 8, respectively; values are for ML analysis; similar results were obtained from the remaining two MP analyses (data not shown)). Although the values of some other pairs of branches were also correlated, these were the only pairs of branches exhibiting a very strong negative correlation across all three analyses. This negative correlation is indicative of the pervasive direct conflict among genes with respect to these branches. In summary, the support for a given branch was strongly dependent on the gene analysed.

Although the trees recovered from single-gene analyses are incongruent in that they exhibit topological differences, the degree of conflict among trees could be relatively minor. We examined the degree of incongruence between the trees recovered by determining the number of taxa that would need to be removed in order to make two trees congruent (see Methods). These values were computed for all possible pairwise comparisons among the 106 50% majority-rule consensus trees generated within each of the three analyses. The distribution of values shows that only a small proportion of all pairwise comparisons are in total agreement (<15%), whereas most of the trees (>50%) require that at least two of the eight taxa are removed before they become congruent, regardless of the method of analysis (Fig. 3). These results were consistent with other metrics of incongruence (see Methods and Supplementary Information) and indicate extensive incongruence among the 106 data sets.

Factors potentially influencing incongruence between trees

We mitigated some potential sources of incongruence by analysing a genome-wide sample of orthologous genes. There are a number of

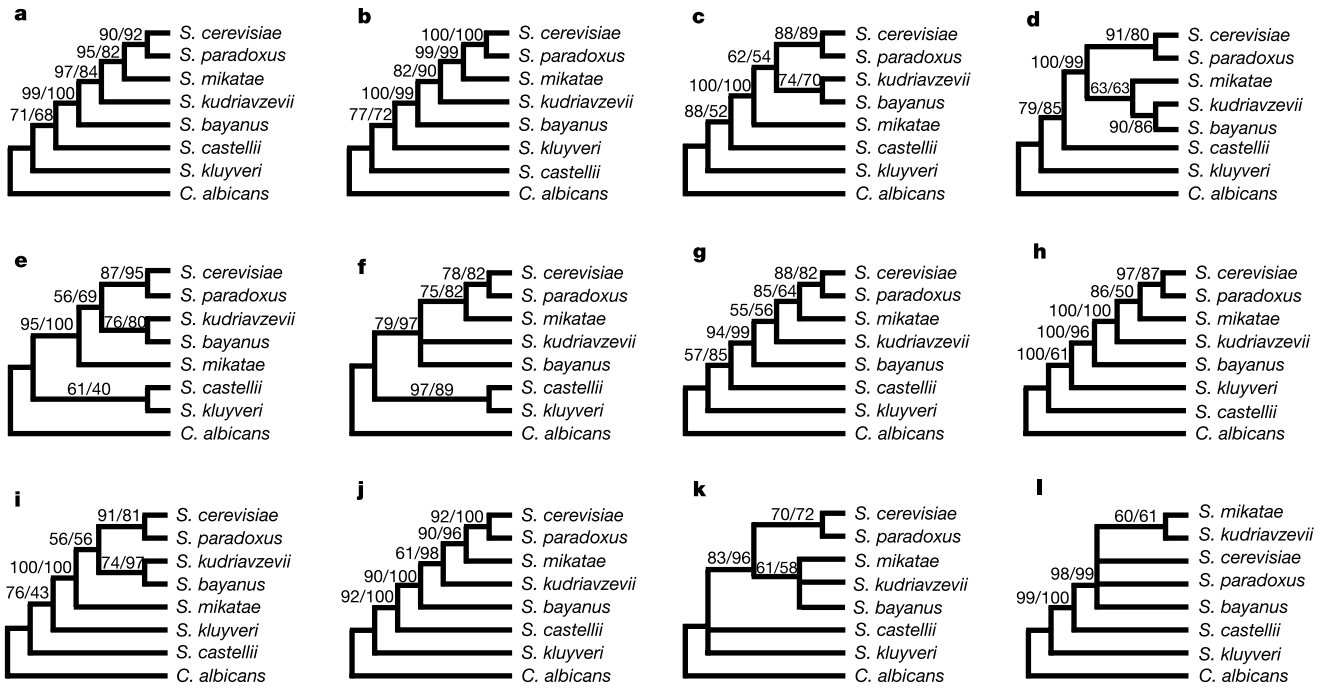

Figure 1 Single-gene data sets generate multiple, robustly supported alternative topologies. Representative alternative trees recovered from analyses of nucleotide data of 106 selected single genes and six commonly used genes are shown. The trees are the 50% majority-rule consensus trees from the genes YBL091C (**a**), YDL031W (**b**), YER005W (**c**), YGL001C (**d**), YNL155W (**e**) and YOL097C (**f**), as well as those from the commonly used genes actin (**g**), hsp70 (**h**), β-tubulin (**i**), RNA polymerase II (**j**) elongation factor 1-α (**k**) and 18S rDNA (**l**). Numbers above branches indicate bootstrap values (ML on nucleotides/MP on nucleotides).

additional analytical and biological factors—such as outgroup choice, number of variable sites and rate of evolution—that may lead to incongruence between single-gene phylogenies[10,13]. To test whether the outgroup accounted for the incongruence between phylogenies, we repeated all of the analyses without the outgroup *C. albicans*. We found no change in the distribution of bootstrap values (correlations among pairwise comparisons of each distribution for the remaining branches were significant with $P < 0.05$) or in the degree of incongruence between the remaining branches (Supplementary Information). We also examined whether support for each branch was explained by the number of variable sites, number of parsimony-informative sites, gene size, rate of evolution, nucleotide composition, base compositional bias, genome location, or gene ontology (Table 1; see also Supplementary Information). Number of variable sites, number of parsimony-informative sites and gene size were significantly correlated with bootstrap values for some branches, although they accounted for only a small amount of the total variation in each case (Table 1; see also Supplementary Information). With a single exception (branch 4 was correlated with the rate of evolution for the ML analysis; Table 1), none of the remaining variables was correlated with bootstrap values for any branch (Table 1; see also Supplementary Information). In summary, there were no identifiable parameters that could systematically account for or predict the performance of single genes.

Concatenation of single genes yields a single tree

Although we do not know the cause(s) of incongruence between single-gene phylogenies, the critical question is how the pervasive incongruence between single trees might be overcome to arrive at the actual species tree. Although many potential options exist, we explored the effect of concatenating single genes into one large data set[1,27,39]. Remarkably, all three methods of analysis of the concatenated sequences yielded a single tree with 100% bootstrap values at every branch (Fig. 4). Furthermore, all alternative topologies generated among the single-gene analyses were rejected (Templeton test, $P < 0.001$ for each of three analyses). Thus, even though the individual genes examined supported alternative trees, the concatenated data exclusively supported a single tree. This level of support for a single tree with five internal branches is, to our knowledge, unprecedented; we conclude that it accurately represents the historical relationships of these eight yeast taxa and will be referred to hereafter as their species tree. The maximum support for a single topology regardless of method of analysis is strongly suggestive of the power of large data sets in overcoming the incongruence present in single-gene analyses.

How much data are sufficient to recover the species tree?

The concatenated data recovered a tree with maximum support on all branches, despite divergent levels of support for each branch among single-gene analyses. This raises the question: at what size

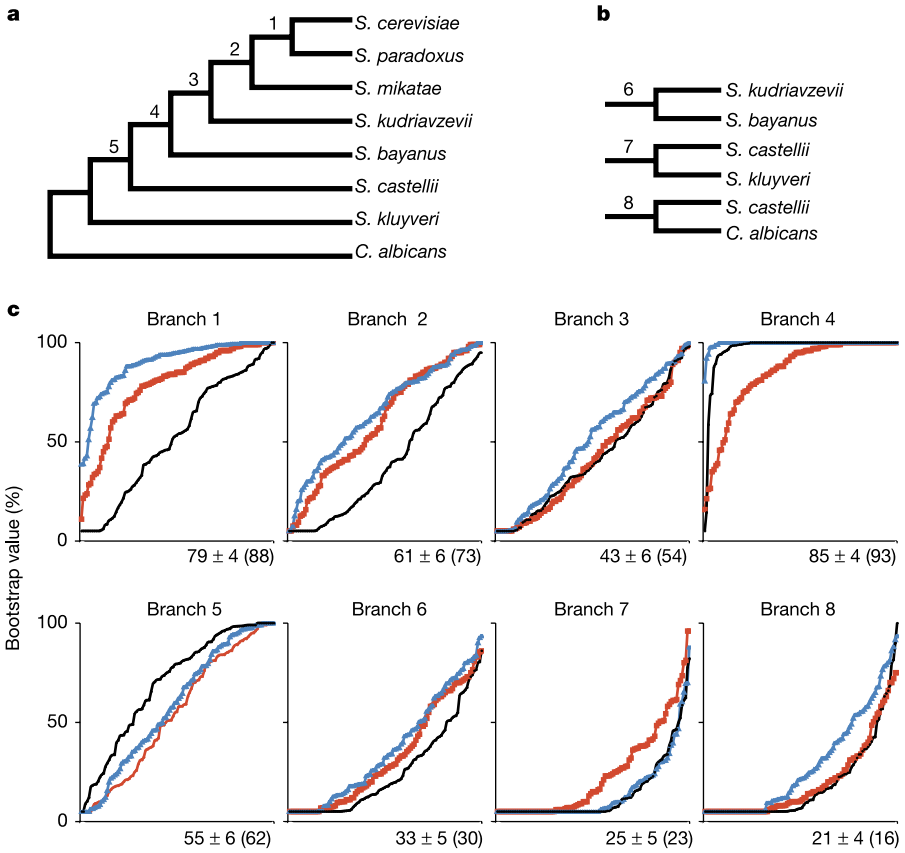

Figure 2 The distribution of bootstrap values for the eight prevalent branches recovered from 106 single-gene analyses highlights the pervasive conflict among single-gene analyses. **a**, Majority-rule consensus tree of the 106 ML trees derived from single-gene analyses. Across all analyses, there were eight commonly observed branches; the five branches in the consensus tree (numbers 1–5; **a**) and the three branches (numbers 6–8) shown in **b**. **c**, For each of the eight branches, the ranked distribution of per cent bootstrap values recovered from the three analyses of 106 genes is shown. Results from ML (blue) and MP (red) analyses of nucleotide data sets, and MP analyses of amino acid data sets (black), are shown. For each branch, the mean bootstrap value and 95% confidence intervals from the ML analyses and the percentage of ML trees supporting this branch (in parentheses) are indicated below each graph. Although the ranked distributions of bootstrap values from the three analyses are remarkably similar for most branches, on a gene-by-gene basis there is no tight correspondence between bootstrap values from ML and MP analyses (see Supplementary Information).

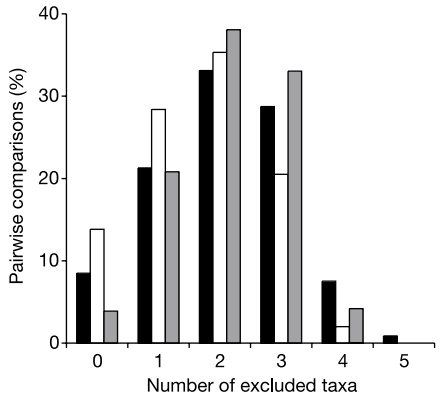

Figure 3 Extensive incongruence between trees derived from the 106 individual-gene data sets. Pairwise comparisons between 50% majority-rule consensus trees from 106 single-gene ML analyses of nucleotide data (black bars), MP analyses of nucleotide data (white bars), and MP analyses of amino acid data (grey bars) were categorized on the basis of the minimum number of taxa that need to be removed for two trees to reach congruence (*x* axis). For each of the analyses, the majority of pairwise comparisons require the removal of two or more taxa before congruence is attained. Similar results are obtained when the ML and MP trees from the analyses of the 106 genes are used (data not shown).

did the data set arrive at the species tree? To estimate the minimum number of genes required to support the species tree (Fig. 4), we randomly re-sampled and concatenated variable numbers of genes from the 106-gene data set (Fig. 5; see also Supplementary Information). Although the number of genes required to achieve a mean bootstrap value >70% across all branches of the species tree was only three concatenated genes, the variance in this estimate was high (Fig. 5, right panels). The results show that in order to achieve a mean bootstrap value of at least 70% with a 95% confidence interval, eight concatenated genes were required. For a mean bootstrap value of at least 95% with a confidence interval of 95%, the number of concatenated genes rose to 20. These results show that the number of genes sufficient to support all branches of the species tree ranged from a minimum of 8 to 20 (Fig. 5, right panels), depending on the threshold of statistical support required.

Because we observed a high variance in bootstrap values from small numbers of concatenated genes, we sought to explore the underlying source of this variance. It has been suggested that nucleotides within a given gene do not evolve independently, thus potentially influencing the phylogenetic accuracy of single genes[16]. To test this hypothesis, we used a variable length bootstrap procedure in which a subset of orthologous nucleotides was randomly re-sampled from the total data set with replacement. The results show that with only 3,000 nucleotides, all five branches were supported with a mean bootstrap value >70% and a confidence interval of 95% (Fig. 5, left panels; see also Supplementary Information). With 8,000 nucleotides, the mean bootstrap value rose to >95% with a confidence interval of 95%. The average size of a gene in our data set was 1,198 base pairs, so 3,000 randomly selected nucleotides correspond to less than three concatenated genes. Importantly, random nucleotides had much lower variance in bootstrap values when compared with corresponding numbers of concatenated genes. For example, with 3,000 nucleotides the mean bootstrap value (±95% confidence interval) was 81.03 ± 1.08 and 91.06 ± 0.83 for branches 3 and 5, respectively, whereas with three concatenated genes the corresponding bootstrap values were 74.36 ± 11.69 and 70.98 ± 18.48 (Fig. 5). The lower bootstrap value and much higher variance for concatenated genes relative to randomly re-sampled nucleotides is consistent with the hypothesis that nucleotides within genes have not evolved independently[16,30,40].

The results demonstrate that concatenation of a sufficient number of randomly selected genes overwhelms conflicting signals present in different genes. It is important to determine whether a consistent bias present in a subset of genes is also overcome by concatenation. To test this possibility, we concatenated genes with bootstrap values >50% for each of the alternative branches 6, 7 and 8 (Fig. 2). Each analysis yielded a majority-rule consensus tree with >90% bootstrap values at most branches, but none of the trees was congruent with that recovered from concatenation of the full data set (Fig. 6). For example, the six genes that most strongly supported *S. castellii* and *S. kluyveri* as sister taxa recovered these species as sister taxa with 100% bootstrap values when concatenated (Fig. 6b). All three trees in Fig. 6 also rejected all alternative topologies generated by single-gene and concatenated analyses (Templeton test, $P < 0.001$ for each of the three trees in Fig. 6). Hence, when biased genes present in the data set were concatenated, strong and misleading support for an alternative species tree was obtained. Previous studies have shown that the concatenation of genes that share some bias (for example, mitochondrial genes) can produce strong support for the incorrect phylogeny[30]. Thus, concatenation of a large number of unlinked genes is clearly the superior strategy.

Implications for resolution of phylogenies

Our results show that there is widespread incongruence between phylogenies recovered from individual genes. Therefore reliance on single or a small number of genes has a significant probability of supporting incorrect relationships for the eight yeast taxa. Perhaps surprisingly, none of the factors known or predicted to cause phylogenetic error[41] could systematically account for the observed

Table 1 **Regressions of bootstrap values on analytical factors**

	Factor							
	Variable sites		Informative sites‡		Branch lengths§		(G + C)%	
Branch*	r^2	*P*-value†	r^2	*P*-value	r^2	*P*-value	r^2	*P*-value
1	0.135	<0.001	0.091	0.002	0.040	0.041	<0.001	0.873
2	0.204	<0.001	0.120	<0.001	0.014	0.228	0.005	0.481
3	0.150	<0.001	0.115	<0.001	0.016	0.197	0.009	0.321
4	0.135	<0.001	0.042	0.035	0.110	0.001	0.022	0.134
5	0.036	0.051	0.073	0.005	0.016	0.200	0.004	0.513
6	0.052	0.018	0.099	0.001	0.002	0.640	0.006	0.432
7	0.072	0.005	0.122	<0.001	0.006	0.413	0.001	0.735
8	0.007	0.392	0.003	0.557	0.001	0.775	<0.001	0.869

*Branch numbers correspond to Fig. 2 and are provided for ML analyses only (see Supplementary Information for the remainder of the analyses).
†Significant values are <0.006 after Bonferroni correction for tests between eight branches.
‡Shown for MP analyses of the nucleotide data only.
§Branch lengths estimated using ML.

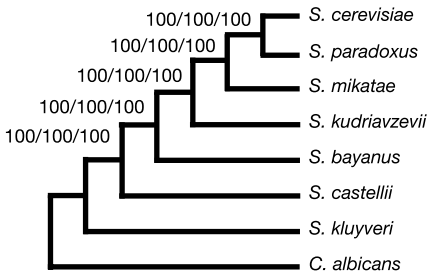

Figure 4 Phylogenetic analyses of the concatenated data set composed of 106 genes yield maximum support for a single tree, irrespective of method and type of character evaluated. Numbers above branches indicate bootstrap values (ML on nucleotides/MP on nucleotides/MP on amino acids).

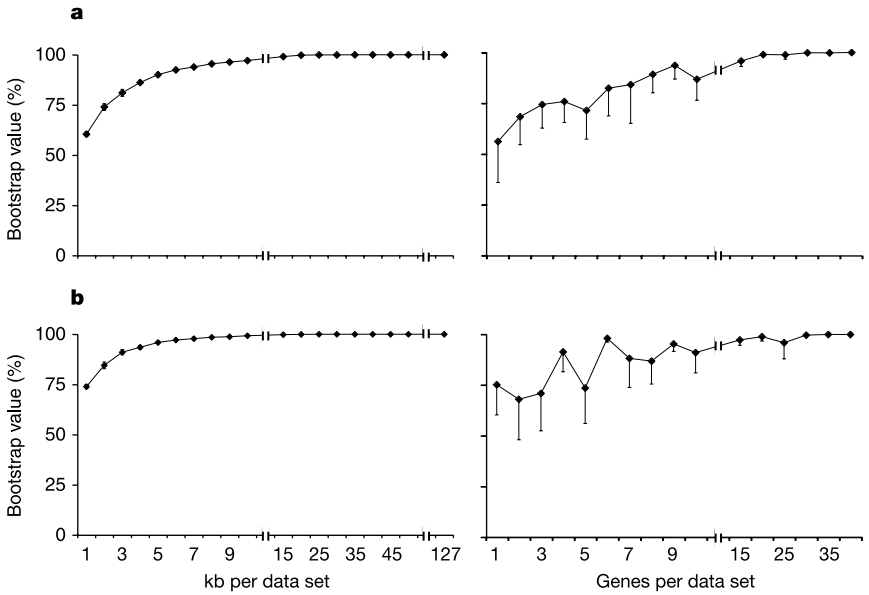

Figure 5 A minimum of 20 genes is required to recover >95% bootstrap values for each branch of the species tree. **a**, **b**, The bootstrap values for branches 3 (**a**) and 5 (**b**) were constructed from the concatenation of randomly re-sampled orthologous nucleotides (left) or random subsets of genes (right) (for the remaining branches see Supplementary Information). Each data point represents the mean and minus one confidence interval of ten replicates (confidence interval values for left panels are not clearly visible due to almost complete overlap with the mean). The species tree is recovered with robust support (>95% bootstrap values in all branches at 95% confidence interval) by analyses of a minimum of 20 concatenated genes. This same result is obtained using only 8,000 randomly selected orthologous nucleotides, indicating that nucleotides within genes evolve non-independently. The variation associated with the confidence intervals in **b** (for example, for branch 5 the data point for 25 genes has a larger confidence interval than that for 20 genes) is due to the small number of replicates. Broken lines in each graph represent a change in the linear values along the *x* axis. All analyses were performed using MP.

incongruence, suggesting that there may be no good predictor of the phylogenetic informativeness of genes. However, regardless of the source of incongruence, concatenation of a sufficient number of unlinked genes (≥20 genes in this study) yields the species tree with remarkable support.

These findings have important implications for many current practices in molecular phylogenetics. A strict interpretation of our data suggests that analyses based on single or a small number of genes provide insufficient evidence for establishing or refuting phylogenetic hypotheses. Furthermore, concatenations of small numbers of genes that provide support for specific branches leads to amplification of support for that branch, even if the branch is in favour of an incorrect topology. For example, a recent phylogenetic study of a concatenation of eight commonly used genes for 75 species belonging to the '*Saccharomyces* complex' found a bootstrap value of 69% in support of a sister group relationship between *S. paradoxus* and *S. mikatae*[42], a finding in sharp contrast to our results. This may be accounted for by both the number and non-independence of the set of genes examined. The unreliability of single-gene data sets (or data sets composed of linked genes, such as genes from the mitochondrial genome) stems from the fact that each gene is shaped by a unique set of functional constraints through evolution. Phylogenetic algorithms are sensitive to such constraints[30] and we show that with genome-wide sampling of independently evolving genes such problems can be avoided. Of course, the possibility exists that one or a few 'lucky' genes[39] may correctly reconstruct the phylogeny of a particular taxonomic group but, in the absence of other data, the confidence in the conclusion will be weak (Fig. 5).

It is only through analyses of a larger amount of sequence data that confidence in the proposed phylogenetic reconstruction can be obtained. In this phylogenetic analysis of eight yeast taxa, concatenated data sets of 20 genes are sufficient to provide very strong (>95%) support for the species tree. It is possible that the eight yeast taxa we have analysed represent a very difficult phylogenetic case, atypical of the situations found in other groups. However, the

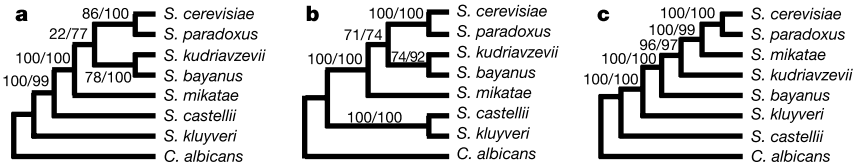

Figure 6 Concatenation of genes supporting alternative branches leads to further amplification of bias. **a**–**c**, Concatenation of the ten single genes showing the highest bootstrap values for a sister-group relationship between *S. kudriavzevii* and *S. bayanus* (**a**), the six data sets showing bootstrap values >50% for a sister-group relationship between *S. castellii* and *S. kluyveri* (**b**), or the ten data sets showing the highest bootstrap values for *S. castellii* being the most distant member of the ingroup (**c**) generates tree topologies that are inconsistent with the species tree recovered from concatenation of 106 genes (Fig. 4). These results may be explained either by a systematic bias imposed by analytical factors (for example, variable rates of nucleotide site evolution) misleading phylogenetic algorithms to artefactually group certain taxa, or by the action of biological factors (for example, certain genes having different genealogical histories from the species owing to hybridization). Numbers above branches indicate bootstrap values (ML on nucleotides/MP on nucleotides).

widespread occurrence of incongruence at all taxonomic levels argues strongly against such a view[6,8–10,21]. Rather, we believe that this group is a representative model for key issues that researchers in phylogenetics are confronting.

Of course, in other cases the amount of sequence information needed to resolve specific relationships will be dependent on the particular phylogenetic history under examination. For example, branches depicting speciation events separated by long time intervals may be resolved with a small amount of data (perhaps branches 1 and 4 in our analyses), whereas branches depicting speciation events separated by shorter intervals may be much harder to resolve. In addition, phylogenetic reconstruction may be complicated by factors such as taxon sampling[17], variable rates of nucleotide site evolution[43], hybridization[44], horizontal transfer[45] and lineage sorting of ancestral polymorphisms[21]. These factors may have an as yet unpredictable effect on the amount of sequence data required for accurate phylogenetic reconstruction of other taxonomic groups or of larger numbers of taxa. Larger concatenated data sets have been very powerful and persuasive in testing specific relationships (such as monophyly of amoebae[39] and rejection of Ecdysozoa[46]). The results of our study suggest that use of genome-wide data sets may provide unprecedented power not only in testing specific phylogenetic hypotheses but also in precise reconstruction of the historical associations of all the taxa analysed. □

Methods

Data set collection

Genes spaced approximately every 40 kilobases (kb) in the *S. cerevisiae* genome were examined and retained for analysis if they met the following criteria: (1) had homologous sequence in each of the eight species; (2) had at least two homologous flanking syntenic genes; and (3) could be aligned over most of the protein. The species of the *sensu strictu* group are thought to have undergone a whole-genome duplication[35–37], so determination of orthology was based on synteny. For *S. paradoxus*, *S. mikatae* and *S. bayanus*, synteny relative to *S. cerevisiae* was determined by visual inspection using the synteny viewer option in the *Saccharomyces* Genome Database (http://www.yeastgenome.org/). For *S. kudriavzevii*, *S. castellii* and *S. kluyveri*, synteny was determined by tblastx comparisons of the *S. cerevisiae* gene of interest, plus each of the two flanking genes. Genes were considered as orthologues if the gene of interest and at least one of the two flanking genes had significant tblastx scores (*e*-value <0.001) and were found in close proximity to each other in all genomes. For *C. albicans*, genes were retained if they were the best reciprocal tblastx hits between both *S. cerevisiae* and *C. albicans*. The genome databases for *S. kudriavzevii*, *S. castellii* and *S. kluyveri* contain draft quality sequence providing 2–3 times coverage over 85–95% of each genome[34]. Therefore, some cases of segmental duplications may have been missed; such cases would be rare and have had little, if any, influence on our results. The selection of commonly used genes (actin, hsp70, β-tubulin, RNA polymerase II, elongation factor 1-α and 18S rDNA) involved searching each genome for the single best tblastx score (in cases of paralogy, the hit with the highest tblastx score was considered as the orthologue). Many other commonly used genes were not analysed because a copy was absent in at least one of the genome databases. Data sets are available from the authors on request.

Phylogenetic analysis

Individual genes were aligned by codon using ClustalW[47] as implemented in Bioedit version 5.0.9 (ref. 48). All gene alignments were manually edited to exclude indels and areas of uncertain alignment from further analysis. After this step, on average 76% of the sequence of each gene was retained. All phylogenetic analyses were performed using PAUP* version 4.0b10 (ref. 49). Each nucleotide data set was analysed under the optimality criteria of maximum likelihood (ML) and maximum parsimony (MP), whereas the corresponding amino acid data sets were analysed using only MP. In both ML and MP analyses, tree space was searched using the branch-and-bound algorithm, which guarantees to find the optimal tree(s). Tree reliability under both optimality criteria was assessed using non-parametric bootstrap re-sampling of 100 replicates; for a given data set, bootstrap re-sampling for both ML and MP analyses started from the same seed number to eliminate variance in the re-sampling procedure between the two optimality criteria. Under MP all characters were equally weighted whereas under ML the model of sequence evolution was optimized using likelihood ratio tests as implemented in Modeltest version 3.06 (ref. 50). Analyses of concatenated data sets were performed using MP (1,000 bootstrap replicates) and, in a subset of cases, ML (100 bootstrap replicates). Random re-sampling of individual nucleotide sites from the complete 127-kb data set was performed using the variable-length-bootstrap option in PAUP*[49] and random gene re-sampling using a random number generator. Incongruence was tested using Templeton's significantly less parsimonious test, and tree distances for all pairwise comparisons among the 106 trees were calculated using two metrics (the agreement subtree metric and the symmetric-difference metric) as implemented in PAUP*[49]. Data sets and trees are available from the authors on request.

Received 4 June; accepted 15 September 2003; doi:10.1038/nature02053.

1. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290,** 972–977 (2000).
2. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. Universal trees based on large combined protein sequence data sets. *Nature Genet.* **28,** 281–285 (2001).
3. Aguinaldo, A. M. A. *et al.* Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387,** 489–493 (1997).
4. Knoll, A. H. & Carroll, S. B. Early animal evolution: emerging views from comparative biology and geology. *Science* **284,** 2129–2137 (1999).
5. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276,** 734–740 (1997).
6. Kopp, A. & True, J. R. Phylogeny of the Oriental *Drosophila melanogaster* species group: a multilocus reconstruction. *Syst. Biol.* **51,** 786–805 (2002).
7. Mason-Gamer, R. J. & Kellogg, E. A. Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). *Syst. Biol.* **45,** 522–543 (1996).
8. Giribet, G., Edgecombe, G. D. & Wheeler, W. C. Arthropod phylogeny based on eight molecular loci and morphology. *Nature* **413,** 157–161 (2001).
9. Hwang, U. W., Friedrich, M., Tautz, D., Park, C. J. & Kim, W. Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature* **413,** 154–157 (2001).
10. Rokas, A., King, N., Finnerty, J. & Carroll, S. B. Conflicting phylogenetic signals at the base of the metazoan tree. *Evol. Dev.* **5,** 346–359 (2003).
11. Loytynoja, A. & Milinkovitch, M. C. Molecular phylogenetic analyses of the mitochondrial ADP–ATP carriers: the Plantae/Fungi/Metazoa trichotomy revisited. *Proc. Natl Acad. Sci. USA* **98,** 10202–10207 (2001).
12. Baldauf, S. L. & Palmer, J. D. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl Acad. Sci. USA* **90,** 11558–11562 (1993).
13. Wendel, J. F. & Doyle, J. J. in *Molecular Systematics of Plants II: DNA Sequencing* (eds Soltis, D. E., Soltis, P. S. & Doyle, J. J.) 265–296 (Kluwer, Boston, Massachusetts, 1998).
14. Huelsenbeck, J. P. Performance of phylogenetic methods in simulation. *Syst. Biol.* **44,** 17–48 (1995).
15. Philippe, H., Chenuil, A. & Adoutte, A. Can the cambrian explosion be inferred through molecular phylogeny? *Development* (Suppl.) 15–25 (1994).
16. Cummings, M. P., Otto, S. P. & Wakeley, J. Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* **12,** 814–822 (1995).
17. Graybeal, A. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47,** 9–17 (1998).
18. Yang, Z., Goldman, N. & Friday, A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11,** 316–324 (1994).
19. Martin, A. P. & Burg, T. M. Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. *Syst. Biol.* **51,** 570–587 (2002).
20. Maddison, W. P. Gene trees in species trees. *Syst. Biol.* **46,** 523–536 (1997).
21. Satta, Y., Klein, J. & Takahata, N. DNA archives and our nearest relative: the trichotomy problem revisited. *Mol. Phylog. Evol.* **14,** 259–275 (2000).
22. Rieseberg, L. H., Whitton, J. & Linder, C. R. Molecular marker incongruence in plant hybrid zones and in phylogenetic trees. *Acta Bot. Neerland.* **45,** 243–262 (1996).
23. Huelsenbeck, J. P., Bull, J. J. & Cunningham, C. W. Combining data in phylogenetic analysis. *Trends Ecol. Evol.* **11,** 152–158 (1996).
24. Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L. & Waddell, P. J. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* **42,** 384–397 (1993).
25. Sullivan, J. Combining data with different distributions of among-site rate variation. *Syst. Biol.* **45,** 375–380 (1996).
26. Cunningham, C. W. Can three incongruence tests predict when data should be combined? *Mol. Biol. Evol.* **14,** 733–740 (1997).
27. Soltis, P. S., Soltis, D. E. & Chase, M. W. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402,** 402–404 (1999).
28. Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409,** 614–618 (2001).
29. Moreira, D., Le Guyader, H. & Philippe, H. The origin of red algae and the evolution of chloroplasts. *Nature* **405,** 69–72 (2000).
30. Naylor, G. J. P. & Brown, W. M. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* **47,** 61–76 (1998).
31. Hillis, D. M. Inferring complex phylogenies. *Nature* **383,** 130–131 (1996).
32. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
33. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546 (1996) 563–567.
34. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301,** 71–76 (2003).
35. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).
36. Wong, S., Butler, G. & Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl Acad. Sci. USA* **99,** 9272–9277 (2002).
37. Langkjaer, R. B., Cliften, P. F., Johnston, M. & Piskur, J. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* **421,** 848–852 (2003).
38. Hillis, D. M., Moritz, C. & Mable, B. K. (eds) *Molecular Systematics* (Sinauer, Sunderland, Massachusetts, 1996).
39. Bapteste, E. *et al.* 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 (2002).
40. Averof, M., Rokas, A., Wolfe, K. H. & Sharp, P. M. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287,** 1283–1286 (2000).
41. Sanderson, M. J. & Shaffer, H. B. Troubleshooting molecular phylogenetic analyses. *Annu. Rev. Ecol. Syst.* **33,** 49–72 (2002).
42. Kurtzman, C. P. & Robnett, C. J. Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. *FEMS Yeast Res.* **3,** 417–432 (2003).
43. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19,** 1–7 (2002).

44. Naumov, G. I., James, S. A., Naumova, E. S., Louis, E. J. & Roberts, I. N. Three new species in the *Saccharomyces sensu stricto* complex: *Saccharomyces cariocanus*, *Saccharomyces kudriavzevii* and *Saccharomyces mikatae*. *Int. J. Syst. Evol. Microbiol.* **50,** 1931–1942 (2000).
45. Bergthorsson, U., Adams, K. L., Thomason, B. & Palmer, J. D. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* **424,** 197–201 (2003).
46. Blair, J. E., Ikeo, K., Gojobori, T. & Hedges, S. B. The evolutionary position of nematodes. *BMC Evol. Biol.* **2,** 7 (2002).
47. Thompson, J. D., Higgins, D. G. & Gibson, T. J. Clustal-W—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 (1994).
48. Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41,** 95–98 (1999).
49. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) (Version 4.0b10)* (Sinauer, Sunderland, Massachusetts, 2002).
50. Posada, D. & Crandall, K. A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14,** 817–818 (1998).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We are grateful to P. Cliften, M. Johnston and the Washington University Genome Sequencing Center for access to genome sequence data for *S. kudriavzevii*, *S. castellii* and *S. kluyveri*; the staff of the *Saccharomyces* Genome Database (http://www.yeastgenome.org/) for access to genome sequence data for *S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus*; and the Stanford Genome Technology Center website (http://www-sequence.stanford.edu/group/candida) for access to sequence data for *C. albicans*. We thank D. Baum, B. Hersh, C. Hittinger and K. Johnson for useful comments on the manuscript, D. Baum and members of the Carroll laboratory for useful discussions on phylogenetics, and D. Lautenschleger for computer support. A.R. is a Human Frontier Science Program long-term fellow, B.L.W. and N.K. are NIH post-doctoral fellows, and S.B.C. is an investigator of the Howard Hughes Medical Institute. This work was funded by the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.B.C. (sbcarrol@wisc.edu).