QUANTIFYING PHYLOGENETIC INCONGRUENCE AND IDENTIFYING CONTRIBUTING FACTORS IN A YEAST MODEL CLADE

By

Leonidas Salichos

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

In partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

In

Biological Sciences

December 2014

Nashville, Tennessee

Approved:

Professor Antonis Rokas

Professor Patrick Abbot

Professor Brian C. O'Meara

Professor David E. McCauley

Professor Seth R. Bordenstein

To my mother, my father, my brother, my family and friends,

that mean the world to me

ACKNOWLEDGMENTS

6 years. This took a while…

First and most importantly, I want to acknowledge my advisor, Antonis, who has been an amazing mentor and a great friend. Expressing all my gratitude is not enough to describe how much I feel he has helped me all these years to arrive to this point. He believed in me, he guided me, he helped me, he motivated me, he patiently tolerated every possible new way that I could think of just to bring about a bit of embarrassment, and supported me until the end. Thank you Antoni.

I am also especially grateful for my chair and my committee members; Patrick, Brian, Dave, Seth. I would like to thank these great professors for all their support, advices, encouragement and guidance throughout these years. Especially Patrick, who has been a fantastic committee chair.

I would also like to acknowledge all the wonderful people at Vanderbilt that I had the opportunity to meet, work with, collaborate with, or be engaged in promising discussions. All the Rokas lab members, throughout these years, especially John, David, Jason, Patricia, Kris, Ioannis, Jen, Xiaofan, Abigail, Haley, Mara, Pad and Ken, Brian, and Holly, as well as people that I collaborated with, including Bethany, Chunyao, professor Jim Patton and professor Carl Johnson. Finally, I would like to acknowledge all the people that helped me get here, with a special mention to professor John Sourdis and professor Vandamme.

Saving for last, but certainly not least, and speaking of wonderful people, friends and collaborators, I would like to particularly mention a couple of dear friends that kept me going all these years through thick and thin. And lamb. And poker. This one is from the heart; Thank you Jibroni, Bonas, Drinky, Beth, Sarah and Yannis.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER V

LIST OF TABLES

CHAPTER V

CHAPTER I

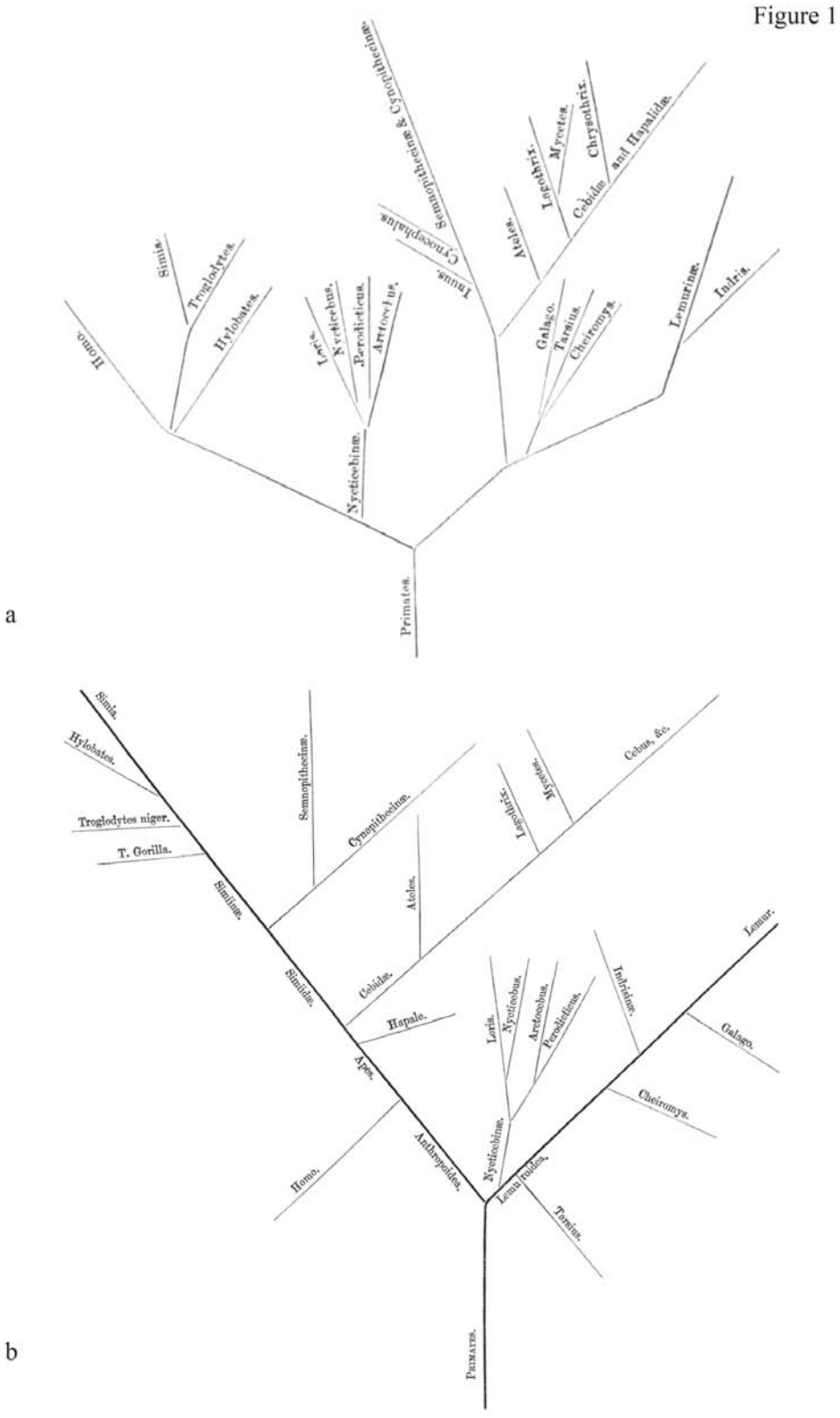
INTRODUCTION

A Brief Historical Perspective of Phylogenetic Incongruence

In his famous work, "*On the Origin of Species*", which is considered as the foundation of evolutionary biology, Charles Darwin introduces the idea of "…the great tree of life", as a metaphor that explains how all life on earth is evolutionarily related[1]. However, in his magnum opus, Darwin did not provide an actual phylogenetic tree drawn from real data, but rather a conceptual tree-like branched diagram.

One of the first phylogenetic trees from actual data was published in 1865 by St. George Mivart, which provided a reconstruction of primate evolutionary history based on the axial skeleton, or spinal column (fig. 1.1a) [2]. Interestingly, in 1867, Mivart published a second phylogeny of primates (fig. 1.1b), this time using data from the appendicular skeleton or limbs, which differed from the first topology[3]. Thus, the birth of phylogenetics goes in hand with phylogenetic incongruence. In 1870, not being able of reconciling the observed topological differences, and in line with his growing and strong opposition to Darwin's theory, Mivart writes to Darwin "…*I have really expressed no opinion as to Man's origin…Pro. Z. Soc. expresses what I believe to be the degree of resemblance as regards the spinal column only. The diagram in the Phil. Trans. expresses what I believe to be the degree of resemblance as regards the appendicular skeleton only…*" (Darwin Correspondence Project letter 7170).

Figure 1.1. The reconstruction of primate evolution history by SG Mivart in 1865 and 1867, depicting conflicting topologies based on a) axial skeleton or spinal column and b) appendicular skeleton or limbs.



Contrary to Mivart's rejection of the Darwin's theory of evolution, evolutionary biologists have been trying ever since to identify heritable traits that may reveal and resolve the evolutionary relationships between organisms. Although phylogenetics as a field has existed since Darwin, its major growth came with the molecular biology revolution and its integration with evolutionary biology. The ensuing dramatic increase of characters led several researchers to try to generate new classifications of organisms, and different schools of phylogenetic thought developed. For example Cladistics, also later known as phylogenetic systematics, originated with the work of entomologist Willi Hennig[4]. Hennig advocated for the reconstruction of evolutionary history based on the analysis of certain types of informative characters. On the other hand, Phenetics, influenced by the work of Peter Sneath and Robert Sokal[5], supported the construction of dendrograms using similarity matrices of numerous characters, without necessarily invoking an evolutionary scenario. These two schools of thought were engaged in a severe philosophical battle during the 1960s and 1970s, with cladistics eventually becoming the dominant school of thought.

In the mid 80's and 90's, the development of powerful and robust computational algorithms in the presence of numerous and continuously increasing number of phylogenetic characters- enabled the emergence of computational phylogenetics, due to their ability to account for complex evolutionary models, while providing support for the inferred topologies in an explicit statistical framework. Proposed methods for estimating phylogenetic trees like Maximum Likelihood (ML)[6] and Bayesian Inference (BI)[7], were established as dominant in the field of phylogenetics[8,9]. However, even nowadays, in the presence of extremely large data sets and complex evolutionary models, phylogenetic incongruence continues to confound evolutionary biologists by providing studies with conflicting results, across the "great Tree of Life".

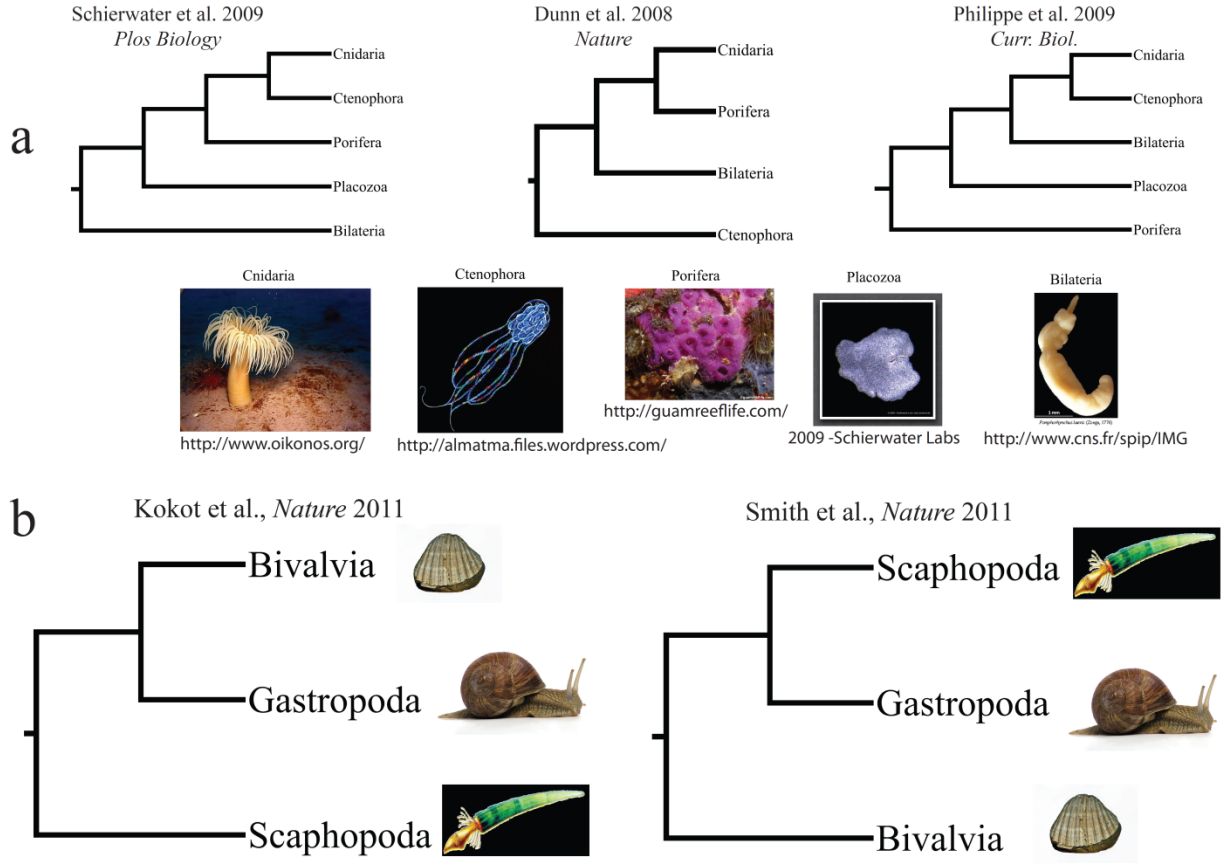
Phylogenetic Incongruence in the Modern Era Using Yeasts as a Model Clade

The modern era of phylogenetics

Advances in sequencing technologies have enabled the whole-genome sequencing of hundreds of prokaryote and eukaryote genomes providing researchers with large amounts of biological data[10], especially in the field of phylogenetics. However, using molecular data, researchers often focused their interest on resolving large taxonomic groups, resulting in the use of few genes with insufficient phylogenetic information, and consequently the inference of weakly supported topologies[11,12]. Moreover, high levels of phylogenetic incongruence were reported across very diverse clades (Primates, fruit flies, yeasts, arthropods, metazoan phyla)[13–23].

In a 2003 study, aimed at benchmarking the identification of phylogenetic incongruence, Rokas et al. revealed a high degree of phylogenetic incongruence, as well as the ability to obtain highly supported clades using concatenation, the analysis of all genes in a data set as a single supermatrix. This study, together with several others (e.g. see references [11,24]), signified the beginning of the "phylogenomic era", an era initially greeted as the "end of incongruence"[25]. Since then, studies that use concatenation approaches have become commonplace and are commonly portrayed to have resolved several vexing ancient divergences with a high degree of confidence[26–47]. Consequently, concatenation analysis became the standard approach for reconstruction of the major and deep branches of the ToL[26–28,48,49]. However, despite the progress that the advent of concatenation has brought, its use has not eliminated incongruence[26,28,34,44,48](Fig. 1.2), suggesting that it might not be as robust as confidence indices purport it to be.

Figure 1.2. The use of concatenation on datasets with hundreds of genes provide conflicting topologies in a) the evolution of mollusks and b) the evolution of early metazoa



Yeast As a Model Clade

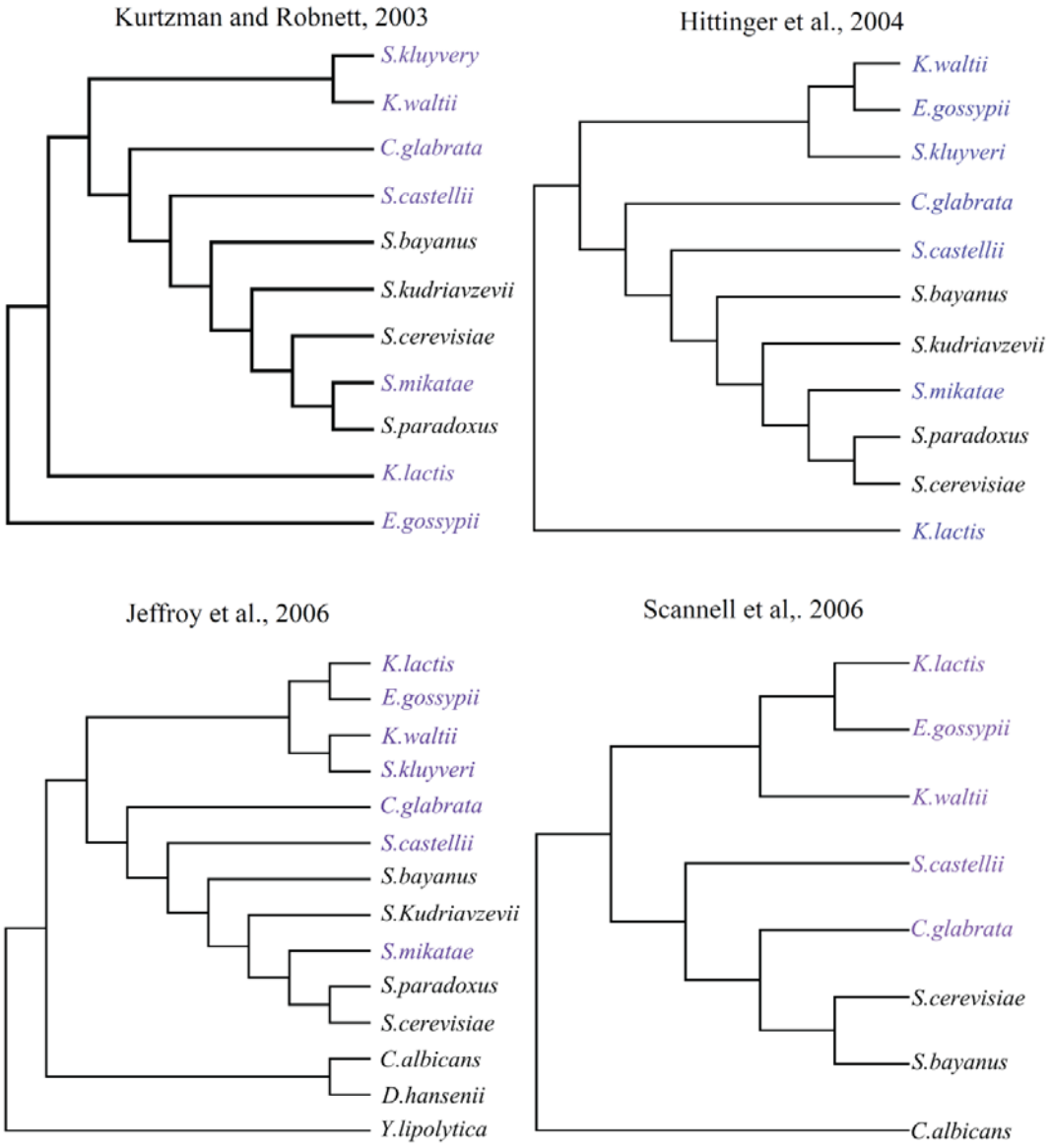
In their 2003 study, Rokas et al. constructed a matrix of 106 widely distributed orthologous genes from 8 species and showed that data sets consisting of single or a small number of concatenated genes have a significant probability of supporting conflicting topologies, although occasionally supported with high bootstrap values[17,18]. Those findings combined with re-analyses of the same data matrix[50–52] as well as with studies on other genome-scale yeast data matrices[36,53,54] suggested that the problem of incongruence in yeasts was not simply a problem of statistical efficiency and low support for different clades. Instead, the lack of accuracy in phylogenetic inference has resulted in different studies with conflicting placements of several

species[55–58]. The presence of (a) unresolved clades, (b) factors that affect the phylogenetic accuracy and information, combined with (c) the large number of fully sequenced genomes that are publicly available, has rendered the fungal class Saccharomycetes as an excellent model for the study of phylogenomics. More analytically:

a) The presence of unresolved clades

A number of studies[55–58] produced conflicting results concerning the placement of several yeast taxa including *Eremothecium gossypii, Saccharomyces castellii, Candida glabrata, Kluyveromyces lactis* and *K. waltii*. For example, while *E. gossypii* and *K. lactis* appeared to be sister taxa in studies from Jeffroy et al. (2006), Kurtzman and Robnett (2003) and Scannell et al. (2006), in Hittinger et al. (2004) the sister taxon for *E. gossypii* is *K. waltii* (fig. 1.3). At the same time, even more puzzling was the phylogenetic topology of *Saccharomyces castellii* and *Candida glabrata*, when compared to *Saccharomyces cerevisiae*. Contrary to all known studies based on molecular data, in the work of Scannell et al. (2006), *Candida glabrata* was presented as more closely related to *Saccharomyces cerevisiae* than *Saccharomyces castellii*, based on syntenic characters derived from the loss of genes after a whole genome duplication (WGD)[59]. At the same time, different topologies have been obtained as the result of using different optimality criteria and sequence data types (nucleotides vs amino acids)[17,50]. In Rokas, Williams et al study, the topology of *K. lactis* could not be resolved. Maximum likelihood analysis placed *K. lactis* as the first species to diverge, while parsimony analysis suggested an alternative placement. Both topologies were supported with a high bootstrap value. In the second case, the placement of *E. gossypii* differed significantly when nucleotide sequence data were used instead of amino acid data.

Figure 1.3. Conflicting topologies between 4 phylogenetic studies in yeast. Taxa with various topologies are shown in blue



b) The presence of factors that affect the accuracy and phylogenetic information

Several different analyses on the yeast clade have shown the existence of factors that affect phylogenetic inference[17,50]. Exploring those factors using the 8-taxon data matrix of 106 gene alignments, Rokas et al. (2003) tested whether clade support could be explained by or correlated

with 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. Interestingly, their results showed a significant correlation between clade support and the number of variable sites, the number of parsimony -informative sites and gene size for some of the branches. Subsequent analyses of the data matrix by Phillips et al. (2004), further established the presence of systematic error in the yeast phylogeny. By using minimum evolution as their optimality criterion, the authors inferred a different topology. Moreover, this topology was mainly attributed to the existence of nucleotide compositional bias as the recoding of nucleotides to purines and pyrimidines rendered the original phylogeny. Similarly, Collins et al. demonstrated that the use of stationary genes in their dataset provided on average more accurate results[51], while other groups demonstrated how the influence of long branch attraction frequently resulted in misplacements for some of the eight taxa in some single gene analyses[52,60].

c) The increasing availability of fully sequenced genomes

The original data matrix constructed by Rokas et al. (2003) used data from the 8 then available yeast genomes. Currently, more than 20 yeast genomes have been fully sequenced, which enables the construction of much bigger data matrices from many more taxa. Moreover, the recent development of databases such as the Yeast Genome Order Browser (YGOB)[61] and Candida Gene Order Browser (CGOB)[62] provides valuable information concerning the identification and validation of orthologous groups of genes from many of those genomes, rendering this data set a model dataset for future functional and phylogenetic analyses.

What is the Source for Phylogenetic Incongruence?

In general, the reasons for observing phylogenetic incongruent data sets may be characterized as either analytical or biological. In biological reasons potential incongruence in a dataset may exist by genes that have different histories than their respective species and stem from historical events. This type of incongruence includes events such as partial or whole genome duplication, introgression, lineage sorting of ancestral polymorphisms or horizontal transfer[63].

In analytical reasons, we find two main types of error that may explain the presence of incongruence; sampling and systematic[64]. Sampling error arises when the sample is not representative of the whole population. Factors that may increase sampling error are the number and appropriateness of included genes, and the phylogenetic information of the inferred alignment[64]. Systematic error may result from a misspecification of the selected evolutionary model[50,65]. Example of factors that contribute to systematic error are base composition and branch length[17,64].

One factor that deserves special mention is ortholog determination, because its accurate determination is fundamental to evolutionary analyses. The identification of orthologs is not always straightforward because genetic (e.g., gene duplications and losses) and population-level (e.g., hybridization and speciation) events can yield complex gene histories[66,67]. For example, gene duplication, especially when it affects many genes as in the case of whole genome duplications and is followed by extensive gene loss, can generate large numbers of single-copy paralogs, which complicate ortholog determination[66–70]. The difficulty in accurately determining orthology, combined with the utility of orthology in many different applications and disciplines and the need for high-throughput pipelines for prediction, have led to the development of several different algorithms for ortholog-based prediction[71].

In chapter II, I present an evaluation of different graph-based algorithms that define orthology between genes in a phylogenetic dataset, including the performance of clustering Reciprocal Best Hit (cRBH), a clustering algorithm for reciprocal best hit ortholog identification that I developed using custom perl scripts. Analysis of 4 algorithms showed that cRBH algorithm outperformed all other three algorithms in almost all of my comparisons. Even though all algorithms seem to deal well with paralogy in most data sets, their performance seems to decrease dramatically in data sets with high levels of paralogy, especially when the orthologous genes have been lost. In chapter IV, where I construct a data matrix from 23 yeast species some of which underwent a whole genome duplication[59], the construction of orthogroups was accomplished by retrieving information from two high quality databases –YGOB[61] and CGOB[62]– where orthology is determined based on syntenic information, sequence similarity and manual curation. This, enabled the construction of a premium dataset of orthogroups that is essentially free of paralogy. However, in constructing data matrices from Metazoans and Vertebrates, lineages for which high quality databases of syntenic orthologs are lacking, I was able to apply my cRBH algorithm to infer orthology.

Measuring data incongruence

Handling and quantifying incongruent data sets has confounded systematists since the beginning of evolutionary biology[72–74]. In general, methods that have been developed for the quantification of incongruence can be classified into two main categories: a) methods that identify character-based incongruence[75–83] and b) methods that calculate incongruence between trees[84–86]. In the first group, the identification of incongruence is achieved by examining how well the data set fits a given phylogenetic tree. In contrast, the second group of methods attempt to calculate the

difference between two distinct trees[87]. In general, even though several of these methods are extremely useful, in practice, they tend to lack generality, as they depend either on a particular optimality criterion[77,78,81] or clade support measure[76,85].

Given the increasing number of available genes for phylogenetic analysis, an interesting group of tree-based methods for measuring incongruence and summarizing conflict are consensus methods[88]. Since each internode in a phylogenetic tree can be represented by a bipartition of two sets of taxa, a set of trees can be potentially also summarized into a consensus tree by including only those bipartitions that are "representative of the set". A very popular and widely-used particular form of consensus methods is the Majority Rule Consensus (MRC) tree method which summarizes the shared bipartitions across all trees in a set, in order to provide a single tree with a value for each internode that corresponds to either the number or percentage of individual phylogenetic trees. However, although very useful, this value does not differentiate between the presence of a strong secondary conflicting signal on the specific internode or simply phylogenetic noise. For example, when a MRC tree reports that 51 out of 100 phylogenetic trees contain a specific bipartition, whether the rest of the bipartitions strongly support a secondary signal remains unknown.

In chapter III, I present four novel measures based on information theory and Shannon's entropy to quantify phylogenetic incongruence. Each internal branch (or internode) in a phylogenetic tree can also be represented by the bipartition of two disjoint sets of taxa (partitions). Consequently, using the prevalence of conflicting bipartitions, I calculate the level of support for each internode as well as the level of conflict. Specifically, Internode Certainty (IC) and Internode Certainty All (ICA) measure the level of certainty for a specific internode either by selecting either the two most prevalent conflicting bipartitions (IC) or all prevalent conflicting bipartitions (ICA),

respectively. Furthermore, the sum of all IC or ICA values on a given phylogeny provides a level of support and conflict for that phylogeny, which is captured by the measures Tree Certainty (TC) and Tree Certainty All (TCA), respectively.

Available High-Profile Practices That Decrease Phylogenetic Incongruence

To reduce data incongruence and improve phylogenetic inference, different phylogenomic studies have relied upon the use of several practices; these include the removal of rogue (unstable or fast evolving) taxa[26,48,89], the trimming and exclusion of ambiguous columns from the gene alignments[17,26,44,49], the use of only the slow-evolving and highly conserved genes[26,41,48], the use of 'good-marker' genes identified based on whether these genes recover internodes that are widely considered as known[27], or finally the use of certain types of characters that are thought to be more informative, such as conserved amino acid (aa) substitutions[90] or indels[91]. Although their effect and magnitude of impact has not been systematically evaluated, these -highly popular- practices are being generously applied, despite different empirical and simulation studies that have argued for their utility[92–94].

In chapter IV, by analyzing a dataset of 1,070 high-quality orthologous groups from 23 yeast genomes, as well as two additional data sets of 1,086 orthogroups from 18 vertebrates species and 225 groups from 21 metazoan species, I show that selecting genes with strong phylogenetic signal reduces incongruence and allows the more accurate reconstruction of ancient divergences. Additionally, using IC and TC (as presented in chapter III) I demonstrate that widely used methods that intend to reduce incongruence, have little or no significant effect on the yeast phylogeny. Finally, I propose two novel methods that dramatically decrease the level of incongruence in the dataset. However, even with achieving a significant decrease of

incongruence, I was unable to resolve certain very short internodes at the base of the yeast species phylogeny, suggesting that conflict in the genes' phylogenetic signal is strong or that phylogenetic signal for these internodes has been almost lost. Perhaps one of the most surprising results of my thesis was that the 1,070 inferred gene trees differed with the species phylogeny, as well as with each other.

Evaluating Phylogenetic Properties and Functional Factors that Influence Phylogenetic Incongruence

Despite significant efforts in accurately reconstructing the tree of life[95], the phylogeny of different evolutionary clades still remains unresolved[14–16,18,20,21,49]. Contradicting results on whether more genes or more taxa need to be included in order to minimize phylogenetic incongruence[52,92,96–101], has opened the discussion of which genes should be considered appropriate and informative. In my comparison of 1,070 yeast gene trees against the species phylogeny, I discovered great differences among the gene trees, as well as between the gene trees and the species phylogeny. This begged the question what are the factors that drive these large amounts of incongruence. As mentioned previously (see "*Yeast as a model clade*", section '*b'*), in their 2003 study, Rokas et al. (2003) tested whether the clade support could be explained by or correlated with 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.

In chapter V, I explore a set of different functional factors (including the percentage and variance of GC content in genes, percentage of variable sites, branch length, number of physical and genetic interactions, level of gene expression, codon adaptation and codon bias[102]) together with a set of phylogenetic gene properties (including Tree Certainty[103], Average Bootstrap Support[104],

Robinson-Foulds[105] mean distance(mRF)and Robinson-Foulds gene variance) in order to examine to which degree these factors or properties are driving phylogenetic incongruence. Rokas et al., showed a correlation between clade support and many of these factors. In my new data set, I explored the impact of several of the factors tested by Rokas et al. as well as new ones on a data set that included ten times as many orthogroups and three times as many taxa. Moreover, using data mining and statistical techniques such as regression, principal component analysis and neural networks, I measure the predictability of each gene's phylogenetic behavior based on its functional factors. Overall, I show that approximately 15-20% of gene-tree incongruence can be directly attributed to gene factors like the percentage of GC content, codon bias, codon adaptation, percentage of variable sites, and distorts the gene's topology away from the species phylogeny. However, even though these functional factors may provide extremely useful insights on the evolutionary behavior of genes, their impact on reducing data incongruence appears to be small, especially for resolving short internodes at the base of the phylogeny. Thus, selecting genes based on their phylogenetic properties such as gene Tree Certainty, average bootstrap support or mean Robinson Foulds distance (the best-performing measure), remains the best way to select genes for phylogenetic inference.

REFERENCES

1. Darwin C. *Origin of Species*. (Comfort R, ed.). John Murray; 1859:424. doi:10.1038/475424a.

2. Mivart S. Contributions towards a more complete knowledge of the axial skeleton in the primates. *Proc Zool Soc London*. 1865;33:545-592.

3. Mivart S. On the appendicular skeleton of the primates. *Philos Trans R Soc London*. 1867;157:299-429.

4. Hennig W. Phylogenetic Systematics. *Annu Rev Entomol*. 1965;10(1):97-116. doi:10.1146/annurev.en.10.010165.000525.

5. Sneath PHA, Sokal RR. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*.; 1973:573. doi:citeulike-article-id:2347143.

6. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 1981;17(6):368-376. doi:10.1007/BF01734359.

7. Rannala B, Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol*. 1996;43(3):304-311. doi:10.1007/BF02338839.

8. Siddall ME. Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris zone. *Cladistics*. 1998;14:209-220. doi:10.1006/clad.1998.0063.

9. Swofford DL, Peter J Waddell PJ, John P Huelsenbeck JP, Foster PG, Paul O Lewis PO JSR. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst Biol*. 2001;50(4):525-539.

10. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res*. 2006;34(Database issue):D332-D334.

11. Bapteste E, Brinkmann H, Lee JA, et al. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proc Natl Acad Sci U S A*. 2002;99(3):1414-1419.

12. Berbee ML, Carmean DA, Winka K. Ribosomal DNA and resolution of branching order among the ascomycota: how many nucleotides are enough? *Mol Phylogenet Evol*. 2000;17(3):337-344. doi:10.1006/mpev.2000.0835.

13. Satta Y, Klein J, Takahata N. DNA archives and our nearest relative: the trichotomy problem revisited. *Mol Phylogenet Evol*. 2000;14(2):259-275.

14. Giribet G, Edgecombe GD, Wheeler WC. Arthropod phylogeny based on eight molecular loci and morphology. *Nature*. 2001;413(6852):157-161. doi:10.1038/35093097.

15. Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W. Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature*. 2001;413(6852):154-157.

16. Kopp A, True JR. Phylogeny of the Oriental Drosophila melanogaster species group: a multilocus reconstruction. *Syst Biol*. 2002;51(5):786-805. doi:10.1080/10635150290102410.

17. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*. 2003;425(6960):798-804. doi:10.1038/nature02053.

18. Rokas A, Carroll SB. Bushes in the tree of life. *PLoS Biol*. 2006;4(11):1899-1904.

19. Delsuc F, Brinkmann H, Chourrout D, Philippe H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*. 2006;439(7079):965-968.

20. Marlétaz Ferdinaz, Elise Martin, Yvan Perez, Daniel Papillon, Xavier Caubit, Christopher J. Lowe, Bob Freeman, Laurent Fasano, Carole Dossat, Patrick Wincker, Jean Weissenbach YLP. Chaetognath phylogenomics: a protostome with deuterostome-like development. *Curr Biol*. 2006;16(15).

21. Matus DQ, Copley RR, Dunn CW, et al. Broad taxon and gene sampling indicate that chaetognaths are protostomes. *Curr Biol*. 2006;16(15).

22. Pollard DA, Iyer VN, Moses AM, Eisen MB. Widespread discordance of gene trees with species tree in drosophila: Evidence for incomplete lineage sorting. *PLoS Genet*. 2006;2(10):1634-1647.

23. Rokas A, Chatzimanolis S. From gene-scale to genome-scale phylogenetics: the data flood in, but the challenges remain. *Methods Mol Biol*. 2008;422:1-12. doi:10.1007/978-1-59745-581-7_1.

24. Lerat E, Daubin V, Moran NA. From Gene Trees to Organismal Phylogeny in Prokaryotes:The Case of the gamma-Proteobacteria. *PLoSBiol*. 2003;1(1544-9173):E19. doi:10.1371/journal.pbio.0000019.

25. Gee H. Evolution: ending incongruence. *Nature*. 2003;425(6960):782. doi:10.1038/425782a.

26. Philippe H, Derelle R, Lopez P, et al. Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Curr Biol*. 2009;19(8):706-712.

27. Regier JC, Shultz JW, Zwick A, et al. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*. 2010;463(7284):1079-1083.

28. Schierwater B, Eitel M, Jakob W, et al. Concatenated analysis sheds light on early metazoan evolution and fuels a modern “urmetazoon” hypothesis. *PLoS Biol*. 2009;7(1).

29. Burleigh JG, Bansal MS, Eulenstein O, Hartmann S, Wehe A, Vision TJ. Genome-scale phylogenetics: Inferring the plant tree of life from 18,896 gene trees. *Syst Biol*. 2011;60(2):117-125.

30. Decker JE, Pires JC, Conant GC, et al. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proc Natl Acad Sci U S A.* 2009;106(44):18644-18649.

31. Hackett SJ, Kimball RT, Reddy S, et al. A phylogenomic study of birds reveals their evolutionary history. *Science*. 2008;320(5884):1763-1768. doi:10.1126/science.1157704.

32. Hampl V, Hug L, Leigh JW, et al. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci U S A.* 2009;106(10):3859-3864.

33. Jacobsen F, Friedman NR, Omland KE. Congruence between nuclear and mitochondrial DNA: Combination of multiple nuclear introns resolves a well-supported phylogeny of New World orioles (Icterus). *Mol Phylogenet Evol*. 2010;56(1):419-427.

34. Kocot KM, Cannon JT, Todt C, et al. Phylogenomics reveals deep molluscan relationships. *Nature*. 2011;477(7365):452-456. doi:10.1038/nature10382.

35. Lehtonen S. Towards resolving the complete fern tree of life. *PLoS One*. 2011;6(10).

36. Medina EM, Jones GW, Fitzpatrick DA. Reconstructing the fungal tree of life using phylogenomics and a preliminary investigation of the distribution of yeast prion-like proteins in the fungal kingdom. *J Mol Evol*. 2011;73(3-4):116-133.

37. Philippe H, Telford MJ. Large-scale sequencing and the new animal phylogeny. *Trends Ecol Evol*. 2006;21(11):614-620.

38. Pratt RC, Gibb GC, Morgan-Richards M, Phillips MJ, Hendy MD, Penny D. Toward resolving deep neoaves phylogeny: Data, signal enhancement, and priors. *Mol Biol Evol*. 2009;26(2):313-326.

39. Ragsdale EJ, Baldwin JG. Resolving phylogenetic incongruence to articulate homology and phenotypic evolution: a case study from Nematoda. *Proc Biol Sci*. 2010;277(1686):1299-1307.

40. Regier JC, Shultz JW, Ganley ARD, et al. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst Biol*. 2008;57(6):920-938.

41. Rodriguez-Ezpeleta N, Brinkmann H, Burger G, et al. Toward Resolving the Eukaryotic Tree: The Phylogenetic Positions of Jakobids and Cercozoans. *Curr Biol*. 2007;17(16):1420-1425.

42. Savard J, Tautz D, Richards S, et al. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res*. 2006;16(11):1334-1338.

43. Simon S, Strauss S, Von Haeseler A, Hadrys H. A phylogenomic approach to resolve the basal pterygote divergence. *Mol Biol Evol*. 2009;26(12):2719-2730.

44. Smith SA, Wilson NG, Goetz FE, et al. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*. 2011;480(7377):364-367.

45. Wiens JJ, Hutter CR, Mulcahy DG, et al. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. *Biol Lett*. 2012.

46. Zhou X, Xu S, Xu J, Chen B, Zhou K, Yang G. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the laurasiatherian mammals. *Syst Biol*. 2012;61(1):150-164.

47. Zou X-H, Zhang F-M, Zhang J-G, et al. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol*. 2008;9(3):R49.

48. Dunn CW, Hejnol A, Matus DQ, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. 2008;452(7188):745-749.

49. Rokas A, Krüger D, Carroll SB. Animal evolution and the molecular signature of radiations compressed in time. *Science*. 2005;310(5756):1933-1938.

50. Phillips MJ, Delsuc F, Penny D. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol*. 2004;21(7):1455-1458.

51. Collins TM, Fedrigo O, Naylor GJP. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. *Syst Biol*. 2005;54(3):493-500. doi:10.1080/10635150590947339.

52. Hedtke SM, Townsend TM, Hillis DM. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol*. 2006;55(3):522-529. doi:10.1080/10635150600697358.

53. Fitzpatrick DA, Logue ME, Stajich JE, Butler G. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol*. 2006;6:99.

54. Hess J, Goldman N. Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: Yeasts revisited. *PLoS One*. 2011;6(8).

55. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*. 2006;440(7082):341-345. doi:10.1038/nature04562.

56. Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? *Trends Genet*. 2006;22(4):225-231.

57. Hittinger CT, Rokas A, Carroll SB. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci U S A*. 2004;101(39):14144-14149.

58. Kurtzman CP, Robnett CJ. Phylogenetic relationships among yeasts of the "Saccharomyces complex" determined from multigene sequence analyses. *FEMS Yeast Res*. 2003;3(4):417-432.

59. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*. 1997;387(6634):708-713. doi:10.1038/42711.

60. Gatesy J DR and WN. How Many Genes Should a Systematist Sample? Conflicting Insights from a Phylogenomic Matrix Characterized by Replicated Incongruence. *Syst Biol*. 2006;56(2):355-363.

61. Byrne KP, Wolfe KH. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*. 2005;15(10):1456-1461.

62. Fitzpatrick DA, O'Gaora P, Byrne KP, Butler G. Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser. *BMC Genomics*. 2010;11:290.

63. Baum DA. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon*. 2006;56(2):417-426. Available at: papers2://publication/uuid/51000653-3ADB-4080-8AEA-E54898ADE024.

64. Swofford DL et al. Phylogenetic inference. In: Hillis, David M., Craig Moritz BKM, ed. *Molecular Systematics*. 2nd ed. Sinauer Associates, Inc; 1996:407-514.

65. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: from genes to genomes and back. *J Mol Biol*. 1998;283(4):707-725. doi:10.1006/jmbi.1998.2144.

66. Koonin E V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 2005;39:309-338. doi:10.1146/annurev.genet.39.073003.114725.

67. Mindell DP, Meyer A. Homology evolving. *Trends Ecol Evol*. 2001;16(8):434-440.

68. Tatusov RL, Koonin E V, Lipman DJ. A genomic perspective on protein families. *Science*. 1997;278(5338):631-637. doi:10.1126/science.278.5338.631.

69. Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol*. 2002;321(1):7-20.

70. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13(9):2178-2189.

71. Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*. 2008;24(11):539-551.

72. Cunningham CW. Can three incongruence tests predict when data should be combined? *Mol Biol Evol*. 1997;14(7):733-740.

73. Huelsenbeck JP, Bull JJ, Cunningham CW. Combining data in phylogenetic analysis. *Trends Ecol Evol*. 1996;11(4):152-158. doi:10.1016/0169-5347(96)10006-9.

74. Bull J. J., J. P. Huelsenbeck, Clifford W. Cunningham DLS and PJW. Partitioning and Combining Data in Phylogenetic Analysis. *Syst Biol*. 1993;42(3):384-397.

75. Goldman N, Anderson JP, Rodrigo AG. Likelihood-based tests of topologies in phylogenetics. *Syst Biol*. 2000;49(4):652-670.

76. Shimodaira H, Hasegawa M. Letter to the Editor Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol*. 1999;16(8):1114-1116. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21178118.

77. Baker RH, DeSalle R. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Syst Biol*. 1997;46(4):654-673. doi:10.1093/sysbio/46.4.654.

78. Farris JS, Kallersjo M, Kluge AG BC. Testing significance of incongruence. *Cladistics*. 1994;10:315-319.

79. Faith DP. Cladistic permutation tests for monophyly and nonmonophyly. *Syst Zool*. 1991;40:366–375.

80. Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol*. 1989;29(2):170-179.

81. Templeton AR. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution (N Y)*. 1983;37(2):221-244. doi:10.2307/2408332.

82. Le Quesne WJ. A method of selection of characters in numerical taxonomy. *Syst Zool*. 1969;18:201-205.

83. Wilson EO. A consistency test for phylogenies based on contemporaneous species. *Syst Zool*. 1965;14:214–220.

84. Thorley JL WM. Testing the phylogenetic stability of early tetrapods. *J Theor Biol*. 1999;200:343-344.

85. Rodrigo AG, Kelly-Borges M, Bergquist PR, Bergquist PL. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *New Zeal J Bot*. 1993;31(3):257-268. doi:10.1080/0028825X.1993.10419503.

86. Thorley JL, Page RD. RadCon: phylogenetic tree comparison and consensus. *Bioinformatics*. 2000;16(5):486-487. doi:10.1093/bioinformatics/16.5.486.

87. Planet PJ. Tree disagreement: Measuring and testing incongruence in phylogenies. *J Biomed Inform*. 2006;39(1 SPEC. ISS.):86-102.

88. Bryant D. A classification of consensus methods for phylogenetics. *DIMACS Ser Discret Math Theor Comput Sci*. 2003;61:163-184.

89. Rokas A, Carroll SB. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol*. 2005;22(5):1337-1344.

90. Rogozin IB, Wolf YI, Carmel L, Koonin E V. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. *Mol Biol Evol*. 2007;24(4):1080-1090.

91. Belinky F, Cohen O, Huchon D. Large-scale parsimony analysis of metazoan indels in protein-coding genes. *Mol Biol Evol*. 2010;27(2):441-451.

92. Poe S, Swofford DL. Taxon sampling revisited. *Nature*. 1999;398(6725):299-300. doi:10.1038/18592.

93. Philippe H, Lopez P, Brinkmann H, et al. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc Biol Sci*. 2000;267(1449):1213-1221.

94. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*. 2007;56(4):564-577. doi:10.1080/10635150701472164.

95. Cracraft J, Donoghue MJ. *Assembling the Tree of Life*. (Cracraft J, Donoghue MJ, eds.). Oxford University Press; 2004:592. doi:10.1016/S0169-5347(97)01242-1.

96. Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. Is sparse taxon sampling a problem for phylogenetic inference? *Syst Biol*. 2003;52(1):124-126.

97. Kim J. Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Syst Biol*. 1998;47(1):43-60. doi:10.1080/106351598261021.

98. Rannala B, Huelsenbeck JP, Yang Z, Nielsen R. Taxon sampling and the accuracy of large phylogenies. *Syst Biol*. 1998;47(4):702-710.

99. Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. Increased taxon sampling is advantageous for phylogenetic inference. *Syst Biol*. 2002;51(4):664-671.

100. Rosenberg MS, Kumar S. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc Natl Acad Sci U S A*. 2001;98(19):10751-10756.

101. Rosenberg MS, Kumar S. Taxon sampling, bioinformatics, and phylogenomics. *Syst Biol*. 2003;52(1):119-124.

102. Sharp PM, Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 1987;15(3):1281-1295.

103. Salichos L, Stamatakis A, Rokas A. Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Mol Biol Evol*. 2014;31(5):1261-71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24509691.

104. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*. 2013;497(7449):327-31. doi:10.1038/nature12130.

105. Robinson DF, Foulds LR. *Comparison of Phylogenetic Trees*.; 1981:131-147.

CHAPTER II

EVALUATING ORTHOLOG PREDICTION ALGORITHMS IN A YEAST MODEL CLADE

Leonidas Salichos[1] and Antonis Rokas[1]
[1]*Department of Biological Sciences, Vanderbilt University, VU Station B #35-1634, Nashville, TN, 37235, United States of America*

This chapter is published in *Plos One, April 13, 2011*

ABSTRACT

Background

Accurate identification of orthologs is crucial for evolutionary studies and for functional annotation. Several algorithms have been developed for ortholog delineation, but so far, manually curated genome-scale biological databases of orthologous genes for algorithm evaluation have been lacking. We evaluated four popular ortholog prediction algorithms (MULTIPARANOID; and ORTHOMCL; RBH: Reciprocal Best Hit; RSD: Reciprocal Smallest Distance; the last two extended into clustering algorithms cRBH and cRSD, respectively, so that they can predict orthologs across multiple taxa) against a set of 2,723 groups of high-quality curated orthologs from 6 Saccharomycete yeasts in the Yeast Gene Order Browser.

Results

Examination of SENSITIVITY [TP/(TP+FN)], SPECIFICITY [TN/(TN+FP)], and ACCURACY[(TP+TN)/(TP+TN+FP+FN)] across a broad parameter range showed that CRBH was the most accurate and specific algorithm, whereas ORTHOMCL was the most sensitive. Evaluation of the algorithms across a varying number of species showed that CRBH had the highest ACCURACYand lowest FALSE DISCOVERY RATE [FP/(FP+TP)], followed by CRSD. Of the six species in our set, three descended from an ancestor that underwent whole genome duplication. Subsequent differential duplicate loss events in the three descendants resulted in distinct classes of gene loss patterns, including cases where the genes retained in the three descendants are paralogs, constituting ‘traps’ for ortholog prediction algorithms. We found that the FALSE DISCOVERY RATE of all algorithms dramatically increased in these traps.

Conclusions

These results suggest that simple algorithms, like CRBH, may be better ortholog predictors than more complex ones (e.g., ORTHOMCL and MULTIPARANOID) for evolutionary and

functional genomics studies where the objective is the accurate inference of single-copy orthologs (e.g., molecular phylogenetics), but that all algorithms fail to accurately predict orthologs when paralogy is rampant.

INTRODUCTION

Orthologous genes are homologs that originated by speciation events, whereas paralogs are homologs that originated by gene duplication events [1]. Accurate determination of orthologs and paralogs is fundamental to molecular evolution analyses, the first step in any comparative molecular biology study, and incredibly useful for functional prediction and annotation [2], [3],[4], [5], [6]. However, identifying orthologs and distinguishing them from paralogs is not always straightforward because genetic (e.g., gene duplications and losses) and population-level (e.g., hybridization and speciation) events can yield complex gene histories [2], [7].

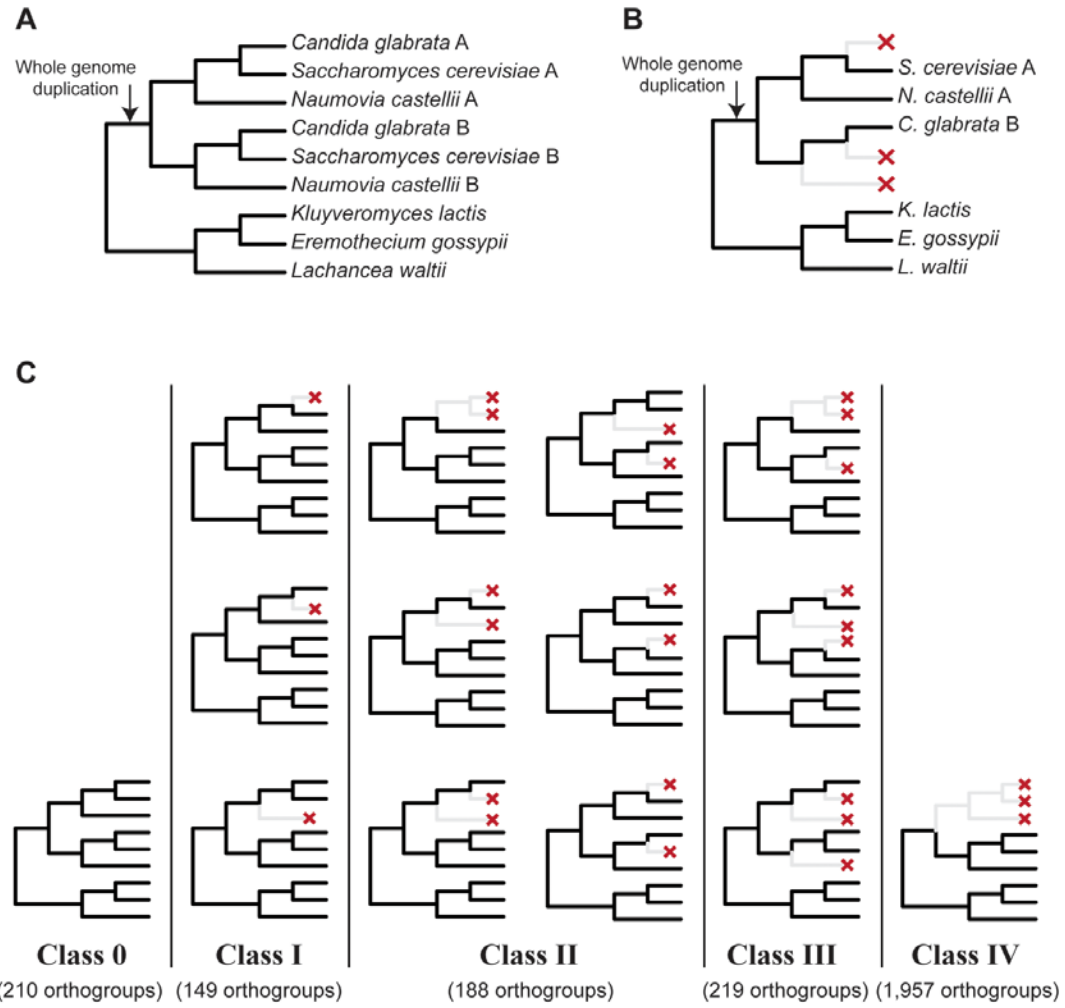
The difficulty in accurately determining orthology, the utility of orthology in many different applications and disciplines, and the abundance of genomic data necessitating high-throughput pipelines for prediction, have led to the development of several different types of ortholog prediction algorithms [8]. For example, a number of graph-based algorithms use similarity searches, such as BLAST [9], to predict groups of orthologous genes (orthogroups), either in pairwise (between two taxa) or clustering (between multiple taxa) fashion [3], [6], [10], [11],[12], [13], [14], [15], [16], [17]. In contrast, tree-based algorithms predict orthogroups using explicit phylogenetic criteria [18], [19], [20], [21], [22], [23]. Although all these different types of ortholog prediction algorithms are widely used, studies that evaluate ortholog prediction algorithm performance for molecular phylogenetic purposes are not

available. Furthermore, large-scale studies that evaluate the relative performance of a wide variety of different ortholog prediction algorithms have yielded contradictory results [10], [24],[25], [26]. For example, whereas Alexeyenko and co-workers [10] found that the graph-based MULTIPARANOID clustering algorithm produced the fewest errors, a different analysis showed that ORTHOMCL, another graph-based clustering algorithm, had the best balance of SENSITIVITY and SPECIFICITY [27]. In contrast, Hulsen and co-workers [24] found that the INPARANOID pairwise algorithm outperformed ORTHOMCL in predictions of orthologous gene pairs. Furthermore, Altenhoff and Dessimoz [25] found that the graph-based OMA clustering algorithm [16] had the highest SPECIFICITY (together with the homolog prediction algorithm HOMOLOGENE [28]), and that certain tree-based algorithms were occasionally outperformed by graph-based pairwise algorithms. Unfortunately, several differences in algorithm design make many of the above comparisons hard to interpret. For example, it is unclear how to interpret comparisons between pairwise and clustering ortholog prediction algorithms (e.g.,[24]), or between algorithms that predict orthologs and paralogs (e.g., [25]), or how the results should be interpreted when the objective is not functional prediction but phylogenetic inference (e.g., [24]).

One potential explanation for these contradictory results might be that each one of the efforts to evaluate ortholog prediction algorithms makes assumptions likely to be violated [10], [24], [25],[27]. For example, several studies evaluated algorithms using functional similarity as a proxy for orthology [24], [25], whereas others evaluated algorithms against sets of orthologs identified by phylogenetic analysis [10], [25]. However, orthologous genes are not always functionally similar [2], and single-gene phylogenies frequently yield erroneous results [29], [30].

The contradictory results in studies of ortholog prediction algorithm performance and the range of evaluation approaches developed suggest that there is a clear need for reliable reference genome-scale ortholog databases. One such high-quality reference database of homologous gene groups is the Yeast Gene Order Browser (YGOB) [31]. The YGOB is an excellent reference dataset for evaluating different ortholog prediction algorithms (e.g., [19], [32]) for two reasons. First, it contains genomes of varying evolutionary distances, and the homology of several thousand of their genes has been accurately annotated through sequence similarity, phylogeny, and synteny conservation data [31], [33]. Second, approximately 100 million years ago, a subset of species in the clade underwent a single round of whole genome duplication (WGD) (Figure 2.1A) [34]. Subsequent differential loss of gene duplicates originating from the WGD event resulted in groups of different gene retention pattern where in some cases the duplicates retained are paralogs [35] (Figure 2.1B), constituting 'traps' for ortholog prediction algorithms (e.g., Class III gene retention patterns in Figure 2. 1C). Importantly, the YGOB database contains accurate ortholog annotations from species that predate and postdate the WGD event, as well as an accurate annotation of hundreds of such 'trap groups', allowing us to compare algorithm performance against orthogroup sets that are much more challenging to decipher.

Figure 2.1. The generation of the five distinct classes of gene loss patterns following the yeast whole genome duplication (WGD).
(**A**) Approximately 100 million years ago, the common ancestor of *S. cerevisiae*, *C. glabrata*, and *N. castellii* underwent WGD, resulting in the doubling of chromosomes. Segments that correspond to the two chromosome sets are known as tracks A and B. (**B**) An example of how the loss of paralogs from different tracks, if undetected, can generate an incorrect species tree. In the example, *C. glabrata* has lost a paralog from track A, whereas *S. cerevisiae* and *N. castellii* have lost paralogs from track B, 'trapping' ortholog prediction algorithms in incorrectly grouping the three post-WGD paralogs in an orthogroup. (**C**) In the aftermath of WGD, extensive loss of paralogs within homologous gene groups resulted in different gene loss patterns, known as classes 0 – IV [35]. Class 0 consists of groups that have not lost any paralogs. Groups in classes I and II have lost one and two paralogs, respectively. Finally, all groups in classes III and IV have lost three paralogs, however, all paralogs lost in class IV groups were on the same track (A or B).



Here, we evaluated the performance of four commonly used ortholog prediction algorithms – MULTIPARANOID [10], ORTHOMCL [3], RBH [4], [6], [12], [13], and RSD [14] in predicting orthogroups in six yeast proteomes by comparing their results against reference orthogroups retrieved from the YGOB database. To ensure that we evaluated all algorithms for their performance in detecting orthogroups across *multiple* species, we extended RBH and RSD into clustering algorithms (CRBH and CRSD, respectively). We selected these four algorithms among the several different ones available [8], based on their popularity, availability as standalone algorithms, and that they are not tree-based, which allows their implementation for downstream molecular phylogenetic analyses. We assessed the performance of each algorithm under a range of parameters and conditions, including in 'traps', as well as using varying numbers of species. We found that CRBH almost always outperformed all other algorithms, suggesting that simpler algorithms may often perform better than more complex ones in identifying orthologs across species, but that the FALSE DISCOVERY RATE of all algorithms was dramatically increased when groups of paralogs stemming from the WGD event were examined.

METHODS

The Test Dataset

The test dataset consists of 31,012 proteins from the proteomes of the following six Saccharomycete yeasts: *Saccharomyces cerevisiae*, *Candida glabrata* (also known as*Nakaseomyces glabrata* [36]), *Naumovia castellii* (also known as *Saccharomyces castellii*[36]), *Lachancea waltii* (also known as *Kluyveromyces waltii* [36]), *Eremothecium gossypii* (also known as *Ashbya gossypii* [36]), and *Kluyveromcyes*

lactis [37], [38], [39], [40], [41]. A common ancestor of three of these six yeast species (*S. cerevisiae, C. glabrata,* and *N. castellii*) underwent a single round of WGD (Figure 2.1A) [34]. Although the quality of annotations differs between the six species included in this study [31], it is unlikely to influence significantly our results. This is so because in our analyses we test all four algorithms on exactly the same data, and we have no reason to think that annotation quality differences would differentially affect the performance of ortholog prediction algorithms in our study.

Constructing 'Gold Groups', a Reference Set of Orthogroups

The Yeast Genome Order Browser (YGOB) database is a manually curated homolog database of Saccharomycete proteins [31] from species that predate the WGD event (*K. lactis, L. waltii*and *E. gossypii*) as well as from species that postdate the WGD event (*S. cerevisiae, C. glabrata,* and *N. castellii*). Thus, for every chromosomal segment in the three pre-WGD species (*L. waltii*, *E. gossypii*, and *K. lactis*), assuming no loss, there are two corresponding chromosomal segments (known as track A and B) in the three post-WGD species. As a result, each homologous gene group in the YGOB database, assuming no gene loss, contains a single ortholog from each pre-WGD species, and two paralogs from each post-WGD species, one from track A and one from track B.

To construct a reference dataset of orthogroups deprived of paralogy we first retrieved all 2,723 annotated homologous gene groups from the YGOB (note that this set is a fraction of the total set of true orthogroups) and split each group into two subgroups. The first subgroup contained all ortholog genes from pre-WGD species together with all orthologs from post-WGD species found on track A, whereas the second subgroup contained the same orthologous genes from pre-WGD

species together with all orthologs from post-WGD species found on track B. To avoid the double counting of orthologs from pre-WGD species in our assessment of ortholog predictions, we evaluated each prediction only against the subgroup that had the best match. We used these orthogroups, from here on referred to as ‘gold groups’, as the reference set to evaluate the performance of ortholog prediction algorithms.

Ortholog Prediction Algorithms Tested

The MULTIPARANOID algorithm [10] is an extension of the graph-based INPARANOID clustering algorithm [11], [42] for identifying orthologs and inparalogs across multiple species. INPARANOID uses bi-directional best BLAST [9], [43] to identify putative orthologs and a clustering algorithm to identify their inparalogs. To do so, INPARANOID assumes that any sequences from the same species that are more similar to the predicted ortholog than to any sequence from other species are inparalogs [11], [42]. MULTIPARANOID generates multi-species orthogroups by merging all pairwise INPARANOID predictions, while minimizing the number of internal conflicts. Furthermore, the algorithm uses a ‘cut-off’ parameter based on the distance of candidate inparalogs to the predicted target ortholog to filter out weakly supported candidates. MULTIPARANOID was obtained from http://multiparanoid.sbc.su.se and INPARANOID (version 3beta) was obtained upon request from inparanoid@sbc.su.se.

The ORTHOMCL algorithm also builds upon the INPARANOID algorithm [11], [42] by using the Markov Cluster (MCL) algorithm for predicting orthogroups across multiple species based on their sequence similarity information [3]. The algorithm uses an ‘inflation rate’ parameter, to

regulate the ‘tightness’ of the predicted orthogroups. ORTHOMCL (version 1.4) was obtained from http://orthomcl.org/common/downloads/software/v1.4/.

The Reciprocal Best Hit (RBH) algorithm [4], [6], [12], [13] relies on BLAST [9], [43] to identify pairwise orthologs between two species. According to the RBH algorithm, two proteins *X* and *Y*from species *x* and *y*, respectively, are considered orthologs if protein *X* is the best BLAST hit for protein *Y* and protein *Y* is the best BLAST hit for protein *X*. We integrated a ‘filtering’ parameter *r* that enabled us to avoid constructing orthogroups that contained distant homologs by considering the degree by which the two proteins differed in sequence length or BLAST alignment [44], [45]. Thus, putative orthogroups are retained if:

$$r \leq \frac{\text{BLAST length or sequence length of putative ortholog A}}{\text{BLAST length or sequence length of putative ortholog B}} \leq \frac{1}{r},$$

where $0 < r < 1$.

From the above equation, it follows that *r* values close to 1 are likely to filter out a larger number of putative orthologs, whereas *r* values close to 0 are likely to include all putative orthologs. The default mode of the algorithm does not use the filtering parameter *r*.

The Reciprocal Smallest Distance (RSD) algorithm [14] generates global sequence alignments for a small number of top BLAST hits against a query gene *X* from species *x*. RSD then calculates the maximum likelihood evolutionary distance between *X* and its top BLAST hits, identifying the gene with the smallest evolutionary distance from *X* (e.g., gene *Y* from species*y*). If the RSD search using gene *Y* from species *y* as the query also identifies gene *X* from species *x* as its closest relative, then proteins *X* and *Y* are considered orthologs [14], [15]. In RSD, the user can modify the shape parameter *a* of the gamma distribution, a key determinant of the estimated evolutionary distance between genes. The RSD algorithm was obtained fromhttp://roundup.hms.harvard.edu/site/.

Extending the Pairwise RBH and RSD Algorithms into Clustering Algorithms cRBH and cRSD

To directly compare the clustering performance of all four ortholog prediction algorithms we extended the pairwise algorithms RBH and RSD into clustering algorithms CRBH and CRSD, respectively. CRBH and CRSD construct orthogroups from more than two species as follows (see also [46]). Considering all pairwise BLAST similarity searches for genes *A, B, C,…, N-1, N* from species *a, b, c,…, n-1, n* to form an orthologous gene group, gene *B* must be the reciprocal best hit to gene *A*, gene *C* the reciprocal best hit to gene *B* or gene *A*, …, and gene*N* the reciprocal best hit to any gene ∈[*A, B, C,…, N-1*]. In cases such as when gene *A* from species *a* is the reciprocal best hit to gene *B* from species *b* and to gene *C1* from species *c*, but gene *B* is the reciprocal best hit to gene *C2* from species *c*, the algorithm drops species *c*from the orthogroup.

Evaluating the Performance of Ortholog Predictions

We used a BLASTP cut-off *E*-value of ≤ 1e-5 in all orthogroup predictions made with all four algorithms. We run the MULTIPARANOID algorithm using a range of cut-off parameter values (cut-off = {0.0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}; 0.0 is the default value), the ORTHOMCL algorithm using a range of inflation rate parameter values (inflation rate = {0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 5, 7.5, 10.0, 100.0}; 1.5 is the default value), the CRBH algorithm by ranging the values assigned to the filtering parameter *r* (*r* = {no *r*, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}; no *r* is the default option), and the CRSD algorithm by ranging the values of the shape parameter *a* (*a* = {0.1, 0.4, 0.5, 0.6, 0.7, 1.0, 1.5, 2.0, 2.5, 5.0}; 0.5 is the default value). For each algorithm and its range of parameter values, we calculated

its ACCURACY,SENSITIVITY, SPECIFICITY, and FALSE DISCOVERY RATE using the following equations: $$\text{ACCURACY} = \frac{\text{True Positives } (TP) + \text{True Negatives } (TN)}{\text{True Positives } (TP) + \text{True Negatives } (TN) + \text{False Positives } (FP) + \text{False Negatives } (FN)}$$

$$\text{SENSITIVITY} = \frac{TP}{TP + FN}$$

$$\text{SPECIFICITY} = \frac{TN}{TN + FP}$$

$$\text{FALSEDISCOVERYRATE (FDR)} = \frac{FP}{FP + TP}$$

Finally, we graphically plotted

the RECEIVER OPERATING CHARACTERISTIC (ROC curve) of

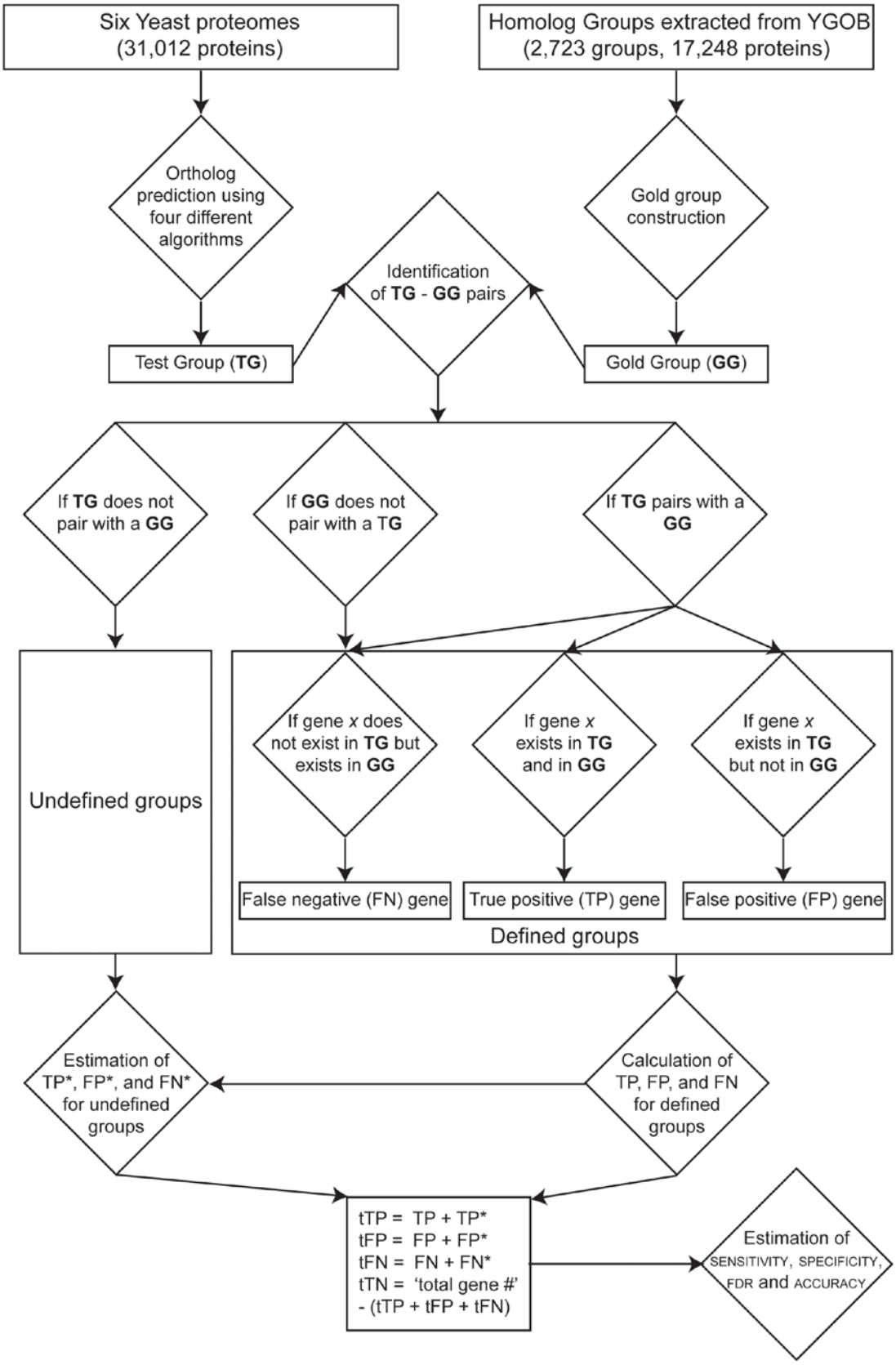
SENSITIVITY versus (1 − SPECIFICITY).

The Evaluation Pipeline for Test Orthologous Genes and Orthogroups

We evaluated the ability of each ortholog algorithm to predict orthogroups by comparing their predictions against the reference gold groups. According to our evaluation pipeline (Figure 2.2 and Text S2.1), each predicted orthogroup was first compared against the set of gold groups to identify, if any, its corresponding gold group. If a test group shared at least two genes with a reference gold group, the test group was characterized as a ‘defined’ test group. In all other cases, the test group was considered ‘undefined’.

Figure 2.2. The pipeline used to evaluate the performance of the ortholog prediction algorithms.

The pipeline evaluates algorithm performance by comparing their predictions on six yeast proteomes against a high-quality reference set of orthologs (gold groups) constructed from the YGOB [31]. The pipeline first compares each test group against the set of gold groups. If the test group matches with a corresponding gold group, the test group is characterized as ‘defined’ and the two groups are further compared on a gene-by-gene basis. If there is no match, the test group is characterized as ‘undefined’. For the ‘defined’ groups, genes present in both the test and the gold groups are considered true positives (TP), whereas genes present only in the test group or only in the gold group are considered as false positive (FP) and false negative (FN), respectively. From the TP, FP, and FN values for all ‘defined’ groups we then estimated the true positives (TP*), false positives (FP*), and false negatives (FN*) for the ‘undefined’ set of groups. Finally, by adding the values obtained from the analysis of ‘defined’ and ‘undefined’ groups we calculated the total number of true positive (tTP), false positive (tFP), false negative (tFN), and true negative (tTN) genes for all test groups, and used them to estimate each algorithm's SENSITIVITY, SPECIFICITY,ACCURACY and FALSE DISCOVERY RATE (See Methods and Text S2.1).


Six Yeast proteomes
(31,012 proteins)
Homolog Groups extracted from YGOB
(2,723 groups, 17,248 proteins)
Ortholog prediction using four different algorithms
Gold group construction
Identification of TG - GG pairs
Test Group (TG)
Gold Group (GG)
If TG does not pair with a GG
If GG does not pair with a TG
If TG pairs with a GG
Undefined groups
If gene x does not exist in TG but exists in GG
If gene x exists in TG and in GG
If gene x exists in TG but not in GG
False negative (FN) gene
True positive (TP) gene
False positive (FP) gene
Defined groups
Estimation of TP*, FP*, and FN* for undefined groups
Calculation of TP, FP, and FN for defined groups
tTP = TP + TP*
tFP = FP + FP*
tFN = FN + FN*
tTN = 'total gene #' - (tTP + tFP + tFN)
Estimation of SENSITIVITY, SPECIFICITY, FDR and ACCURACY

For the defined orthogroups, we considered all genes shared between the test group and its corresponding gold group as true positive (TP), and any genes in the test group that did not also belong to the gold group as false positive (FP) (Figure 2.2 and Text S2.1). FP genes could belong to a different gold group or to be absent from the set of corresponding gold groups. Finally, we considered all those genes present in gold groups that did not belong to any test groups as false negative (FN).

Given that the number of reference gold groups is much smaller than the total number of true orthogroups in our dataset, we expect that a significant number of test orthogroups will not have corresponding gold groups, and hence will be undefined. Because we wanted to calculate values that were representative for the entire dataset, we estimated the number of true positive (TP*), false positive (FP*), and false negative (FN*) for the undefined orthogroups by multiplying the number of TP, FP, and FN calculated from the defined groups with the ratio of the number of undefined genes on the number of defined genes (Figure 2.2 and Text S2.1). For example, TP* is the product of the TP value multiplied by the ratio of the number of undefined genes on the number of defined genes. Finally, by calculating the total number of true positive (tTP = TP + TP*), false positive (tFP = FP + FP*), and false negative (tFN = FN + FN*) genes, we were able to estimate the number of total true negative genes (tTN = total number of genes – tTP – tFP – tFN) in our dataset (Figure 2.2 and Text S2.1).

To ensure that the calculated TP, FP, and FN values for proteins that belonged to ‘defined’ groups were also representative of the remainder of the proteins (i.e., those that belong to the ‘undefined’ groups) (Figure 2.2), we tested whether *S. cerevisiae* genes that belong to ‘defined’ and ‘undefined’ groups differed significantly in evolutionary rate (measured by the *dN/dS* ratio), number of paralogs in genome, and codon adaptation index. We obtained the data for

evolutionary rate and codon adaptation index calculations from the study of Wall *et al.* [47]. We calculated the number of *S. cerevisiae* paralogs per protein using BLASTP [9]. To evaluate whether the evolutionary and functional properties of genes that belong to the 'defined' and 'undefined' groups were statistically significant, we performed a two-tailed t-test (assuming unequal variance and unequal sample size) [48].

Evaluating Algorithm Performance for Varying Numbers of Species

To evaluate the performance of each algorithm across varying numbers of species, we examined all possible combinations for three, four, and five yeast proteomes and calculated each algorithm's ACCURACY and FDR. All algorithms were run using the parameter values that yielded the highest ACCURACY in orthogroup prediction on the six yeast proteomes dataset.

Evaluating Algorithm Performance against Different Classes of Gene Loss Events

Our reference dataset contains orthogroup classes where some of the homologs retained are paralogs. To investigate how each algorithm performed in these 'trap groups', we divided the 2,723 gold groups into the five classes described by Scannell *et al.* [35] (Figure 2.1C) and calculated the ACCURACY and FDR for each algorithm. All algorithms were run using the parameter values that yielded the highest ACCURACY in orthogroup prediction on the six yeast proteomes dataset.

RESULTS

We evaluated the performance of four different algorithms (MULTIPARANOID, ORTHOMCL,CRBH and CRSD) in predicting orthogroups against a manually curated, high-

quality database of ortholog groups (gold groups), by

estimating SENSITIVITY, SPECIFICITY, ACCURACY and FDR across different parameter values, using a varying number of species and across different gene loss classes (Figures 2.3, 2.4, 2.5, 2.6 and Table S2.1). *S. cerevisiae* genes that belong to 'defined' and 'undefined' groups did not differ significantly in evolutionary rate, number of paralogs in genome, and codon adaptation index (all *p*-values for all measures across all algorithms were larger than 0.05). Thus, the 'defined' and 'undefined' orthogroups do not differ significantly. Therefore, our estimation of the number of true positive (TP*), false positive (FP*), and false negative (FN*) for the undefined orthogroups based on the number of TP, FP, and FN calculated from the defined groups seems to be valid and our results should be representative of the entire population of orthogroups present in the six yeast genomes under study.

Figure 2.3. The ACCURACY **and** RECEIVER OPERATING CHARACTERISTIC **(**ROC**) curve for each ortholog prediction algorithm across a range of parameter values.**
(**A**) The ACCURACY [(TP + TN)/(TP + TN + FP + FN)] of each ortholog prediction algorithm (shown on the Y-axis) is plotted against the range of algorithm-specific parameter values (shown on the X-axis). Values for MULTIPARANOID are for the 'cut-off' parameter, values for ORTHOMCL are for the 'inflation rate' parameter, values for CRBH are for the 'filtering parameter *r*', and values for CRSD are for the 'shape parameter *a*'. (**B**) The ROC curve for each ortholog prediction algorithm shows SENSITIVITY [TP/(TP + FN)] (on the Y-axis) plotted against 1 – SPECIFICITY [1 – (TN/(TN + FP))] (on the X-axis). Optimal values and distributions reside on the top left of the graph. All values depicted in the graphs are shown in Table S2.1.

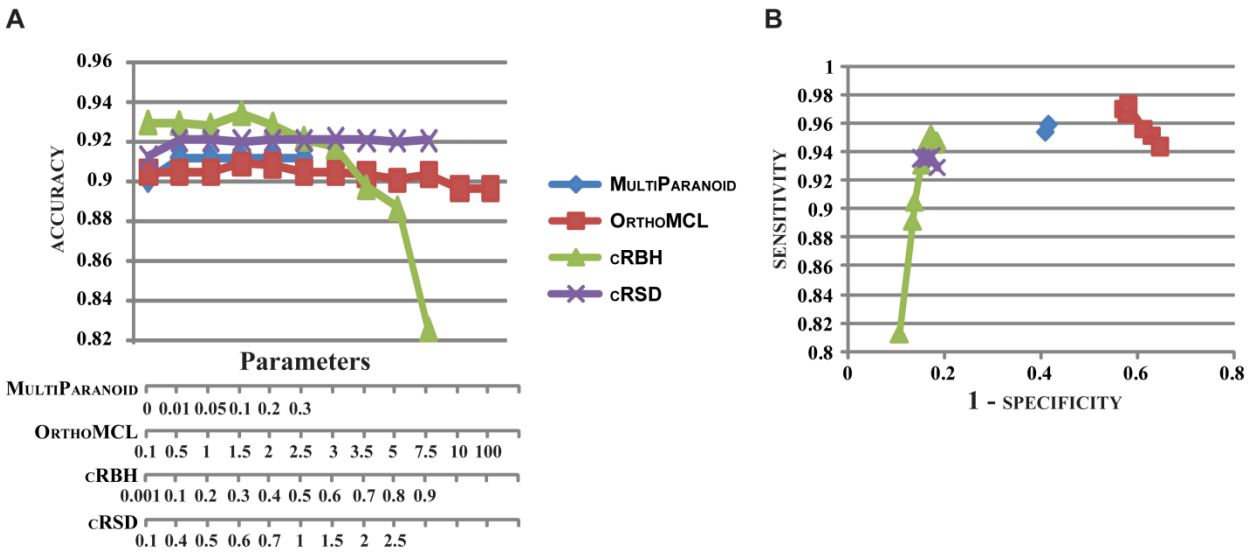


Figure 2.4. The ACCURACY **and** FDR **of ortholog prediction algorithms using varying numbers of species.**
(**A**) The ACCURACY of ortholog prediction algorithms (shown on the Y-axis) is plotted against varying numbers of species (shown on the X-axis). (**B**) The FDR of ortholog prediction algorithms (shown on the Y-axis) is plotted against varying numbers of species (shown on the X-axis). Each algorithm was run using the parameter value yielding the highest ACCURACY. All values depicted in the graphs are shown in Table S2.1.

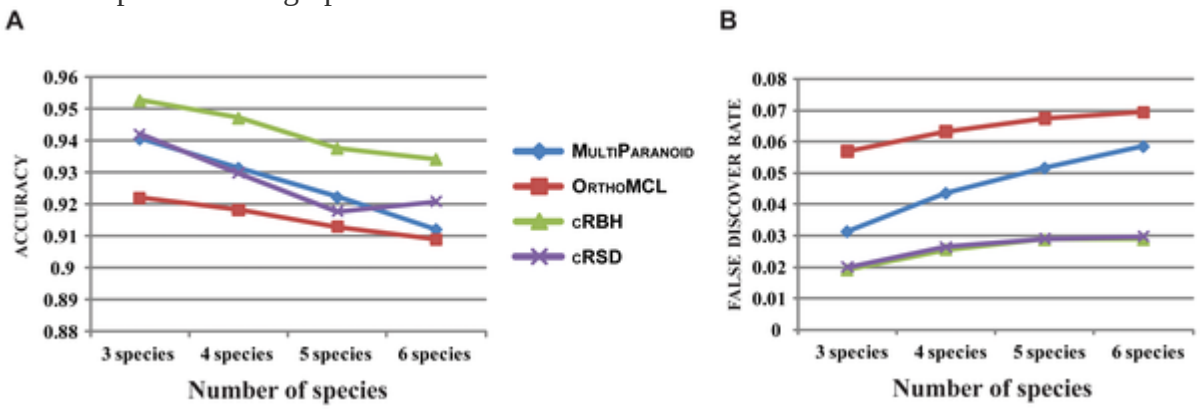


Figure 2.5. The ACCURACY **and** FDR **of ortholog prediction algorithms across five orthogroup classes with different gene retention patterns.**
The five classes are described in Figure 2.1. (**A**) The accuracy of ortholog prediction algorithms (shown on the Y-axis) is plotted against the five classes (shown on the X-axis). (**B**) The FDR of ortholog prediction algorithms (shown on the Y-axis) is plotted against the five classes (shown on the X-axis). Each algorithm was run using the parameter value yielding the highest ACCURACY. All values depicted in the graphs are shown in Table S2.1.

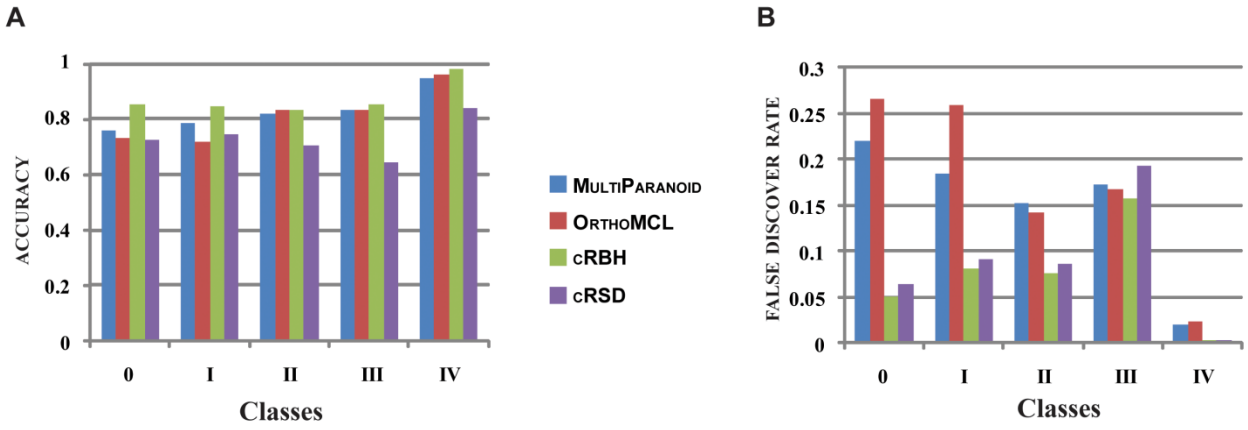
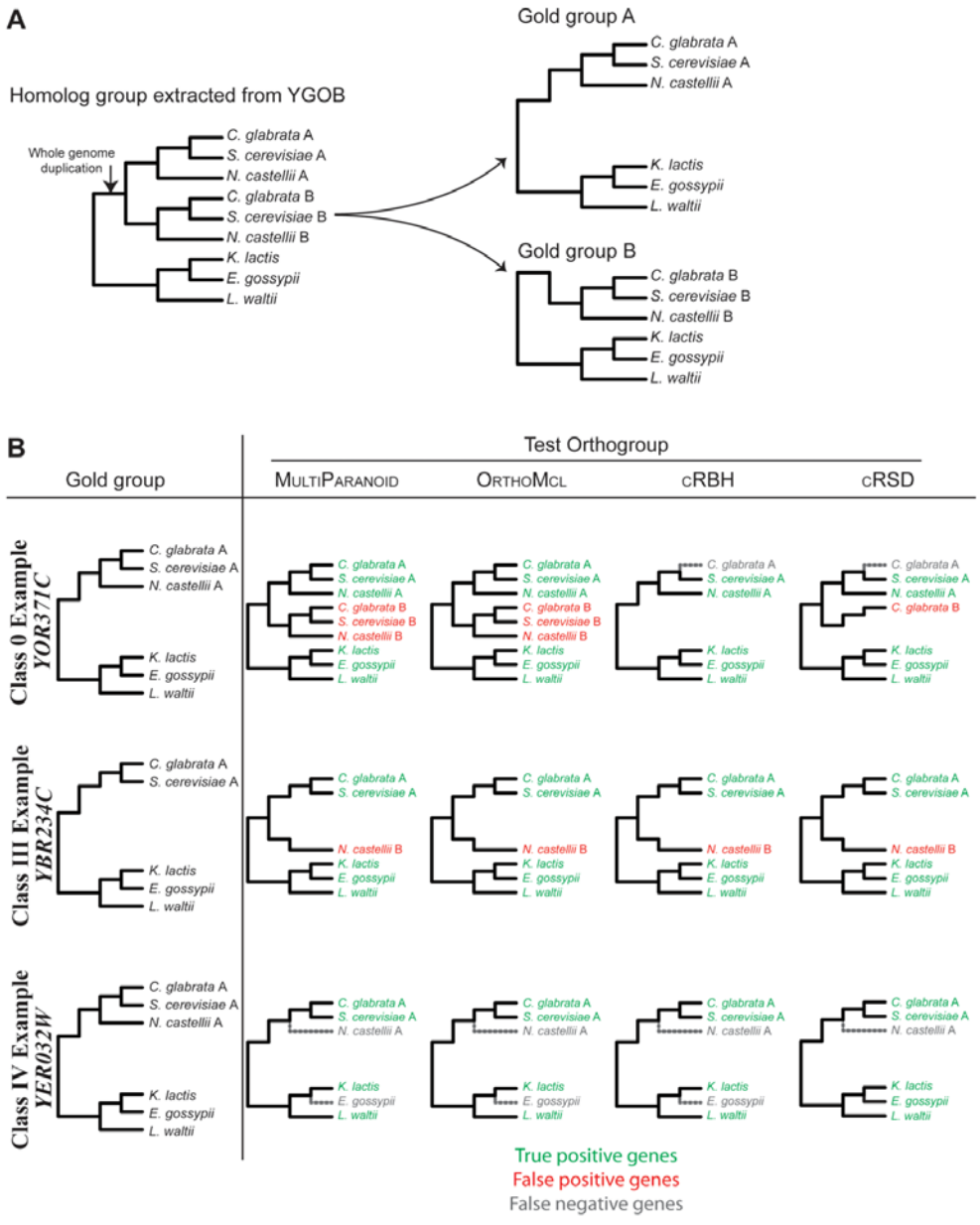


Figure 2.6. Examples of the behavior of the four algorithms in predicting orthogroups from gold groups belonging to three different classes.
(**A**) Construction of gold groups (gold groups A and B) from the set of homologous gene groups from the YGOB. Each test group is evaluated against only against the gold group that had the best match. (**B**) The orthogroups for three different gold groups belonging to classes 0, III and IV predicted by the four different algorithms. The gold group is shown on the left-most column. The *S. cerevisiae* gene name for each of the three gold groups is shown on the left. Genes correctly predicted as belonging to each orthogroup (true positives) are shown in green, genes incorrectly predicted as belonging to each orthogroup (false positives) are shown in red, whereas genes present in a gold group that were not predicted to belong to this or any other test group (false negatives) are shown in grey.



Comparing Algorithm Performance across Different Parameter Values

Ranging the cut-off parameter value of the MULTIPARANOID algorithm had minor effects on its performance. All analyses with cut-off values >0 yielded identical results with higherSENSITIVITY and ACCURACY, but lower SPECIFICITY relative to the default cut-off value of zero. The ORTHOMCL algorithm did not exhibit any clear trade-off between SENSITIVITY andSPECIFICITY with increasing inflation rate values. Specifically, predictions using inflation rate values ≥ 3.5 had both lower SENSITIVITY and SPECIFICITY. The algorithm had almost equalSENSITIVITY for values < 3, with the best SPECIFICITY and ACCURACY obtained when the inflation rate was 1.5. The CRBH algorithm had the highest SENSITIVITY and ACCURACY when *r* was 0.3, although similar values were obtained when *r* was not set (default) or when *r* was 0.4. In general, *r* values greater than 0.4 decreased the SENSITIVITY of the algorithm by excluding increasing numbers of putative orthologs, but increased its SPECIFICITY.

For CRSD,SENSITIVITY and ACCURACY remain largely stable and optimal for *a* values ≥ 0.4. SENSITIVITYwas highest at $a = 0.4$, whereas ACCURACY and SPECIFICITY were both highest at $a = 1.5$. In general, the algorithm produced a limited number of false positives, which resulted in both highACCURACY and low FDR.

The performance of all ortholog algorithms across different parameter values is summarized in Figure 2.3. Our results suggest that CRBH is the most accurate algorithm. Specifically, CRBH had the highest ACCURACY (0.934, for $r = 0.3$), followed by CRSD (0.921, for $a = 1.5$), MULTIPARANOID (0.912, for any cut-off >0) and ORTHOMCL (0.909, for inflation rate = 1.5) (Figure 2.3). Higher SENSITIVITY is typically associated with either higher numbers of true positives or lower number of false negatives. Across the range of all parameters for all

algorithms, ORTHOMCL showed the highest SENSITIVITY (inflation rate = 1), followed by CRBH (r = 0.3), MULTIPARANOID (for cut-off >0) and CRSD (for a = 0.4) (Figure 2.3). In contrast, higherSPECIFICITY is typically associated with lower numbers of false positives. Across the range of all parameters for all algorithms, CRBH has the highest SPECIFICITY (for r = 0.9), followed byCRSD (for a = 0.1), MULTIPARANOID (for cut-off = 0) and ORTHOMCL (for inflation rate = 1.5) (Figure 2.3).

Comparing Algorithm Performance Using a Varying Number of Species and Across Different Gene Loss Classes

To evaluate the performance of each algorithm under a varying number of species, we ran the algorithms for all possible combinations of three, four and five species (Figure 2.4). Once again,CRBH had the highest ACCURACY (Figure 2.4A) and the lowest FDR across all taxon numbers (Figure 2.4B), followed by CRSD.

To investigate how the existence of ‘trap’ gold groups affected the performance of the four ortholog prediction algorithms, we compared their ACCURACY and FDR across the five different gold group classes (Figure 2.1C). Overall, all four algorithms had higher FDR values in paralog-containing classes (classes 0 through III) than in paralog-lacking classes (class IV) (Figure 2.5).CRBH had the highest ACCURACY and the lowest FDR values across all classes. However, not all algorithms exhibited the same behavior across the five classes. For example, whereas CRBH and CRSD had their highest FDR values in class III, ORTHOMCL and MULTIPARANOID had their highest FDR values in class 0, due to the larger number of paralogs (Figures 2.5, 2.6). Finally, note that in class IV, where all paralogs from the same track

(track A or B) have been lost, all algorithms perform well, but CRBH still showed the highest ACCURACY and the lowest FDR.

DISCUSSION

More than twenty orthology prediction algorithms and databases have been developed, which can be divided into three main groups: graph-based (orthology is inferred from sequence similarity), tree-based (orthology is inferred from phylogeny), and hybrid-based (orthology is inferred from both phylogeny and sequence similarity) [8]. In this study, we compared the performance of four popular graph-based clustering algorithms (MULTIPARANOID, ORTHOMCL,CRBH and CRSD) that predict orthogroups for use in molecular phylogenetics. We did not include tree-based and hybrid algorithms because ortholog prediction on large datasets typically requires faster algorithms, and because the reliance of these algorithms on knowledge of the gene family (e.g., [18]) or species phylogeny (e.g., [19]) can render them inappropriate for downstream phylogenetic studies (but see [49]). Furthermore, the use of YGOB as our reference dataset required the availability of standalone algorithms that could make predictions on user-provided datasets.

For the majority of orthogroup predictions, all methods showed high ACCURACY and low FDR(Figures 2.3, 2.4, 2.5), a finding consistent with their similarity in algorithm construction and popularity in the literature. However, our results also suggested that CRBH outperformed all other three algorithms in almost all of our comparisons (Figures 2.3, 2.4, 2.5). These results directly pertain to on-going debates about the choice of ortholog prediction algorithms for downstream evolutionary, genomic and functional analyses [8], [10], [24], [25], [26]. However, the selection of the optimal ortholog prediction

algorithm for inferring orthologous genes and groups across such a remarkably wide range of fields and applications is a complex problem that is likely to be influenced by many parameters.

Curated Ortholog Databases as Gold Standards for Algorithm Evaluation

Several different benchmarks have been used to assess the ACCURACY of ortholog prediction algorithms [8]. However, the lack of 'gold' standard reference datasets has made interpretations of relative performance challenging. For example, several recent comparative studies have yielded contradictory results [10], [24], [25], [26], but the degree to which this lack of common high-quality reference sets contributes to these conflicts is largely unknown. To circumvent these issues, we employed a highly accurate genomic database of homologs to evaluate directly ortholog prediction algorithms (see also [19], [32]). We think that our gold group set has strong potential to become one such 'gold' standard for the evaluation of ortholog prediction algorithms. Of course, our dataset stems from species inhabiting a single small twig of the tree of life. Thus, it remains an open question whether these results hold across branches of the tree of life, or whether ACCURACY in ortholog prediction in different branches will require several different approaches. As more genomes from several clades of the tree of life are sequenced[50] we anticipate that highly accurate homolog databases, like the YGOB [31], will become commonplace and more densely populated with orthologs from several additional species (e.g.,[51]), thus greatly facilitating algorithm evaluation and testing the generality (or not) of findings such as those reported in this study.

One potential limitation of such reference databases is that their construction might be possible only from genomes of close relatives. This is so, because accurate annotation of orthologs between distantly related species is much more challenging; at greater evolutionary distances

protein homology is frequently reduced to homology between domains [52], domain shuffling is commonplace [53], and independent data, such as synteny conservation, that are highly informative for accurate annotation of orthologs between closely related species, become less useful [54]. Nevertheless, our findings (see also [19], [32]) suggest that evaluation approaches against high-quality 'gold standard' databases [31], [51] are likely to be a very useful addition to existing benchmarks [8], [24], [25] in the quest to accurately infer orthologs on a genome-wide scale.

Simpler Algorithms Can Sometimes Be Better

The usefulness of ortholog identification in several downstream genomic, molecular and evolutionary analyses, coupled with the abundance of genomic data from diverse organisms, has spurred the development of several ortholog prediction algorithms [8]. Thus, we were surprised to find that CRBH, a conservative clustering version of the simplest and earliest-developed of the four algorithms tested that drops instead of resolving inconsistencies [4], [6],[12], [13], [55], was consistently (e.g., across several parameter values and varying numbers of species) the best ortholog predictor. In agreement with our results, a recent phylogenetic and functional assessment of ortholog prediction algorithms and databases also found that RBH performed well and its predictions were, in several instances, better than those of more complex algorithms [25]. The superior performance of CRBH and CRSD may be partially explained by the fact that ORTHOMCL and MULTIPARANOID are designed to also include inparalogs in their orthogroup predictions (Figure 2.6). Using our evaluation pipeline, this design can raise significantly the number of false positives, thus decreasing the algorithms' ACCURACY and SPECIFICITY, but increasing the algorithms' FDR

and SENSITIVITY. However, when the algorithms were tested on class IV orthogroups, which comprise the majority of gold groups (1,957 orthogroups or ~70%) and have lost all paralogs from the same track (Figure 2.1C), CRBH still performed better by showing a very low FDR, high ACCURACY and SPECIFICITY and almost equal SENSITIVITY as ORTHOMCL, the most sensitive algorithm (Figure 2.3). Although this difference in performance could be due to the inclusion of other paralogs that did not originate through the WGD, the existence of other paralogs is unlikely to account fully for it. For example, analysis of a dataset that contained only genes belonging to class IV gold groups, an inparalogs-free dataset, also showed that CRBH and CRSD have the highest ACCURACY and lowest FDR. Finally, the set of single-copy orthogroups obtained from ORTHOMCL and MULTIPARANOID is much smaller than the total number of predicted orthogroups and shows much lower SENSITIVITY and ACCURACY. This suggests that the popular approach of using these algorithms for orthogroup prediction in molecular phylogenetic studies is less accurate than the use of algorithms designed to predict orthogroups that contain a single gene from each species, like CRSD and CRBH.

When tested on the class III groups (Figure 2.1), in which the pattern of gene loss forced all algorithms to place single-copy paralogs in the same orthogroup, all algorithms showed very high FDR values (Figures 2.1, 2.5). CRBH was again the best performing algorithm, partly due to the effect of the filtering parameter *r* in dropping putative orthogroups composed of distantly related paralogs. Note that the lack of a 'gold' reference dataset or the adoption of an evaluation strategy based on majority-rule predictions would have not permitted us to identify the failing of these algorithms for class III orthogroups, and would have instead considered most of them as likely true.

Choosing the Right Algorithm for Orthologous Gene Group Prediction

Our results suggest that simpler algorithms, like CRBH and CRSD, might be better choices for many downstream evolutionary analyses than more complex ones in cases where the objective is to identify orthogroups and that the trend of several studies toward using more complex ortholog prediction strategies is not always justified. One of the criteria used in our selection of algorithms was for ones whose orthogroup predictions would be appropriate for use in phylogenetic analyses. Thus, we did not evaluate tree-based or hybrid-based algorithms. However, such algorithms could be much more appropriate for orthogroup prediction in several other contexts, e.g., for functional annotation. For example, the SYNERGY algorithm [19], [56], which integrates information from similarity searches, gene trees, and synteny in its orthogroup predictions has been shown to be more accurate than RBH [19], and likely to be a much better choice for evolutionary genomics and functional studies. Similarly, because RBH, RSD and their clustering extensions are limited to finding orthogroups that contain a single gene from each species, they will fail to detect the presence of inparalogs, and in contrast to algorithms such as SYNERGY [19], [56], MULTIPARANOID [10] and ORTHOMCL [3], are probably of no use for studying gene family evolution.

ACKNOWLEDGMENTS

We thank members of the Rokas lab for valuable comments on this work. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University.

Author Contributions

Conceived and designed the experiments: LS AR. Performed the experiments: LS. Analyzed the data: LS AR. Contributed reagents/materials/analysis tools: LS AR. Wrote the paper: LS AR.

REFERENCES

1. Fitch WM (1970) distinguishing homologous from analogous proteins. Syst Zool 19: 99–113.

2. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39: 309–338.

3. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.

4. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, et al. (1998) Predicting function: From genes to genomes and back. J Mol Biol 283: 707–725.

5. Mirny LA, Gelfand MS (2002) Using orthologous and paralogous proteins to identify specificity determining residues. Genome Biol 3: preprint0002.0001–0002.0020.

6. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.

7. Mindell DP, Meyer A (2001) Homology evolving. Trends Ecol Evol 16: 434–440.

8. Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM (2008) The quest for orthologs: finding the corresponding gene across genomes. Trends Genet 24: 539–551.

9. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

10. Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22: e9–15.

11. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314: 1041–1052.

12. Bork P, Ouzounis C, Casari G, Schneider R, Sander C, et al. (1995) Exploring theMycoplasma capricolum genome: a minimal cell reveals its physiology. Mol Microbiol 16: 955–967.

13. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, et al. (1996) Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. Curr Biol 6: 279–291.

14. Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. Bioinformatics 19: 1710–1711.

15. DeLuca TF, Wu IH, Pu J, Monaghan T, Peshkin L, et al. (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. Bioinformatics 22: 2044–2046.

16. Dessimoz C, Boeckmann B, Roth ACJ, Gonnet GH (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. Nucleic Acids Res 34: 3309–3316.

17. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96: 2896–2901.

18. Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, et al. (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics 22: 699–707.

19. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics 23: i549–558.

20. Storm CEV, Sonnhammer ELL (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18: 92–99.

21. Storm CEV, Sonnhammer ELL (2003) Comprehensive analysis of orthologous protein domains using the HOPS database. Genome Res 13: 2353–2362.

22. Zmasek CM, Eddy SR (2002) RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 3: 14.

23. van Noort V, Snel B, Huynen MA (2003) Predicting gene function by conserved co-expression. Trends Genet 19: 238–242.

24. Hulsen T, Huynen MA, de Vlieg J, Groenen PM (2006) Benchmarking ortholog identification methods using functional genomics data. Genome Biol 7: R31.

25. Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5: e1000262.

26. Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE 2: e383.

27. Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34: D363–D368.

28. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2010) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 38: D5–D16.

29. Cummings MP, Otto SP, Wakeley J (1995) Sampling properties of DNA sequence data in phylogenetic analysis. Mol Biol Evol 12: 814–822.

30. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798–804.

31. Byrne KP, Wolfe KH (2005) The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15: 1456–1461.

32. Akerborg O, Sennblad B, Arvestad L, Lagergren J (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci U S A 106: 5714–5719.

33. Gordon JL, Byrne KP, Wolfe KH (2009) Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. Plos Genetics 5: e1000485.

34. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.

35. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440: 341–345.

36. Kurtzman CP (2003) Phylogenetic circumscription of Saccharomyces, Kluyveromycesand other members of the Saccharomycetaceae, and the proposal of the new generaLachancea, Nakaseomyces, Naumovia, Vanderwaltozyma and Zygotorulaspora. FEMS Yeast Res 4: 233–245.

37. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. (1996) Life with 6000 genes. Science 274: 546, 563–567.

38. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, et al. (2004) The Ashbya gossypiigenome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science 304: 304–307.

39. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, et al. (2004) Genome evolution in yeasts. Nature 430: 35–44.

40. Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428: 617–624.

41. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, et al. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301: 71–76.

42. O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33: D476–480.

43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

44. Salichos L, Rokas A (2010) The diversity and evolution of circadian clock proteins in fungi. Mycologia 102: 269–278.

45. Grossetete S, Labedan B, Lespinet O (2010) FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. BMC Genomics 11: 81.

46. Kent BN, Salichos L, Gibbons JG, Rokas A, Newton IL, et al. (2011) Complete bacteriophage transfer in a bacterial endosymbiont (Wolbachia) determined by targeted genome capture. Genome Biol Evol 3: 209–218.

47. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, et al. (2005) Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A 102: 5483–5488.

48. Sokal RR, Rohlf FJ (1995) Biometry: the principles and practice of statistics in biological research. New York: Freeman. xix.

49. Lemoine F, Lespinet O, Labedan B (2007) Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. BMC Evol Biol 7: 237.

50. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 34: D332–334.

51. Fitzpatrick DA, O'Gaora P, Byrne KP, Butler G (2010) Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser. BMC Genomics 11: 290.

52. Koonin EV (2001) Computational genomics. Curr Biol 11: R155–158.

53. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA (2004) Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 14: 208–216.

54. Ehrlich J, Sankoff D, Nadeau JH (1997) Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics 147: 289–296.

55. Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52: 540–542.

56. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. Nature 449: 54–61.

SUPPLEMENTARY FIGURES, TABLES & TEXT

Table S2.1. The ACCURACY, SENSITIVITY, SPECIFICITY and FDR values of ortholog prediction algorithms across a range of parameter values (S2.1A), using varying numbers of species (S2.1B), and across five orthogroup classes with different gene retention patterns (S2.1C).

Table_S2.1A. Figure 2.3 Raw Data

ORTHOMCL													
'inflation rate' parameter value	**0.10**	**0.50**	**1.00**	**1.50**	**2.00**	**2.50**	**3.00**	**3.50**	**5.00**	**7.50**	**10.00**	**100.00**	
ACCURACY	0.90	0.90	0.90	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	**Data for Figure 3A**
SENSITIVITY	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.95	0.96	0.94	0.94	**Data for Figure 3B**
SPECIFICITY	0.42	0.42	0.42	0.43	0.42	0.42	0.42	0.39	0.37	0.39	0.35	0.35	

MULTIPARANOID							
'cut-off' parameter value	**0.00**	**0.01**	**0.05**	**0.10**	**0.20**	**>0.3**	
ACCURACY	0.90	0.91	0.91	0.91	0.91	0.91	**Data for Figure 3A**
SENSITIVITY	0.95	0.96	0.96	0.96	0.96	0.96	**Data for Figure 3B**
SPECIFICITY	0.59	0.58	0.58	0.58	0.58	0.58	

cRBH											
'r' parameter value	**0.00**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**	
ACCURACY	0.93	0.93	0.93	0.93	0.93	0.92	0.92	0.90	0.89	0.83	**Data for Figure 3A**
SENSITIVITY	0.95	0.95	0.94	0.95	0.94	0.94	0.93	0.90	0.89	0.81	**Data for Figure 3B**
SPECIFICITY	0.82	0.82	0.82	0.83	0.84	0.83	0.85	0.86	0.87	0.89	

cRSD											
'α' parameter value	**0.10**	**0.40**	**0.50**	**0.60**	**0.70**	**1.00**	**1.50**	**2.00**	**2.50**	**5.00**	
ACCURACY	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	**Data for Figure 3A**
SENSITIVITY	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	**Data for Figure 3B**
SPECIFICITY	0.81	0.83	0.84	0.84	0.83	0.84	0.85	0.84	0.84	0.83	

Table S2.1B. Figure 2.4 Raw Data

	Number of Species				
ACCURACY	**3 species**	**4 species**	**5 species**	**6 species**	
MULTIPARANOID	0.94	0.93	0.92	0.91	**Data for Figure 4A**
ORTHOMCL	0.92	0.92	0.91	0.91	
CRBH	0.95	0.95	0.94	0.93	

CRSD	0.94	0.93	0.92	0.92	

FALSE DISCOVERY RATE	**3 species**	**4 species**	**5 species**	**6 species**	
MULTIPARANOID	0.03	0.04	0.05	0.06	
ORTHOMCL	0.06	0.06	0.07	0.07	**Data for Figure 4B**
CRBH	0.02	0.03	0.03	0.03	
CRSD	0.02	0.03	0.03	0.03	

Table S2.1C. Figure 2.5 Raw Data

	Classes					
ACCURACY	**0**	**I**	**II**	**III**	**IV**	
MULTIPARANOID	0.86	0.85	0.84	0.85	0.98	
ORTHOMCL	0.73	0.72	0.83	0.83	0.96	**Data for Figure 5A**
CRBH	0.76	0.79	0.82	0.83	0.95	
CRSD	0.72	0.74	0.70	0.64	0.84	

FALSE DISCOVERY RATE	**0.00**	**I**	**II**	**III**	**IV**	
MULTIPARANOID	0.22	0.18	0.15	0.17	0.02	
ORTHOMCL	0.27	0.26	0.14	0.17	0.02	**Data for Figure 5B**
CRBH	0.05	0.08	0.07	0.16	0.00	
CRSD	0.06	0.09	0.09	0.19	0.00	

Text S2.1.Analytical description of the evaluation algorithm. For the 'defined' predicted orthogroups ('defined' test groups), a gene that was present in both the test group and its corresponding gold group was considered as true positive (TP), whereas a gene that was only present in the test group, but not in the corresponding gold group, was considered as false positive (FP). In general:

(All genes used in the comparison) = FP + TP + FN + TN (1)

We distinguished FP genes into those that are found in the set of corresponding gold groups (FP_{in}) and those that are not found in the set of corresponding gold groups (FP_{out}):

$$FP = FP_{in} + FP_{out} \quad (2)$$

In addition, we distinguished FN genes to those genes belonging to gold groups that are absent from their corresponding test groups (FN_m) and to those genes belonging to gold groups that were not matched by any test group (FN_{nm}). Thus:

$$FN = FN_m + FN_{nm} \quad (3)$$

We calculated TP, FP_{in}, FP_{out}, and FN_m values by comparing test groups with their corresponding gold groups. Furthermore, in cases where an algorithm predicted fewer test groups than expected based on the number of gold groups (2,723 for all classes, 210 for Class 0, 149 for Class I, 188 for Class II, 219 for Class III, 1,957 for Class IV), we estimated the FN_{nm} value using the equation:

FN_{nm} = (‘number of gold groups’ – ‘number of defined groups’) x ‘average number of genes per gold group’ (4)

We then used the TP, FP and FN values for the ‘defined’ test genes to estimate true positive (TP*), false positive (FP*), and false negative (FN*) values for the ‘undefined’ test genes according to:

TP* = TP x (number of ‘undefined’ test genes / number of ‘defined’ test genes) (5)

FP* = FP x (number of ‘undefined’ test genes / number of ‘defined’ test genes) (6)

FN* = FN x (number of ‘undefined’ test genes / number of ‘defined’ test genes) (7)

Finally, we estimated the numbers of total true positive (tTP), total false positive (tFP), total false negative (tFN) and total true negative (tTN) genes according to:

$$tTP = TP + TP^* \quad (8)$$

$$tFP = FP + FP^* \quad (9)$$

$$tFN = FN + FN^* \quad (10)$$

$$tTN = \text{‘number of genes in proteome set’} - tTP - tFP - tFN \quad (11)$$

CHAPTER III

NOVEL INFORMATION THEORY-BASED MEASURES FOR QUANTIFYING INCONGRUENCE AMONG PHYLOGENETIC TREES

Leonidas Salichos[1], Alexandros Stamatakis[2,3] and Antonis Rokas[1,4]

[1]*Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA*
[2]*The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg D-69118, Germany*
[3]*Institute for Theoretical Informatics, Karlsruhe Institute of Technology, D-76131, Karlsruhe, Germany*
[4]*Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA*

This chapter is published in *Mol. Biol. Evol, , February 7, 2014.* 31(5):1261–1271

ABSTRACT

Phylogenies inferred from different data matrices often conflict with each other necessitating the development of measures that quantify this incongruence. Here, we introduce novel measures that use information theory to quantify the degree of conflict or incongruence among all non-trivial bipartitions present in a set of trees. The first measure, Internode Certainty (IC), calculates the degree of certainty for a given internode by considering the frequency of the bipartition defined by the internode (internal branch) in a given set of trees jointly with that of the most prevalent conflicting bipartition in the same tree set. The second measure, IC All (ICA), calculates the degree of certainty for a given internode by considering the frequency of the bipartition defined by the internode in a given set of trees in conjunction with that of all conflicting bipartitions in the same underlying tree set. Finally, the Tree Certainty (TC) and Tree Certainty All (TCA) measures are the sum of IC and ICA values across all internodes of a phylogeny, respectively. IC, ICA, TC, and TCA can be calculated from different types of data that contain non-trivial bipartitions, including from bootstrap replicate trees, gene trees or individual characters. Given a set of phylogenetic trees, the IC and ICA values of a given internode reflect its specific degree of incongruence, and the TC and TCA values describe the global degree of incongruence between trees in the set. All four measures are implemented and freely available in version 8.0.0 and subsequent versions of the widely-used program RAxML.

INTRODUCTION

Phylogenetic trees constructed from different genes frequently contradict each other, giving rise to incongruence[1,2]. For example, several recent studies examining hundreds of genes in fungi[3,4], plants[5] and mammals[6] found that the vast majority of gene trees are not topologically congruent

either with each other or with the species phylogeny. This incongruence can be due to analytical factors stemming from either inadequate sample sizes[1,7,8] or the misfit between data and evolutionary models[9,10] or due to biological factors such as horizontal gene transfer, lineage sorting, introgression, and hybridization[11–13].

Although the challenge of detecting and appropriately handling incongruence has vexed systematists for decades[7,14,15], the recent realization that a large number of gene trees will typically disagree with the species phylogeny has highlighted the importance and value of measures that capture and quantify incongruence[3]. Incongruence tests can be broadly classified[16] into tests that assess incongruence between characters[17–24] and tests that assess incongruence between trees[25–27]. Note that both character-based and tree-based incongruence tests rely on phylogenetic trees; however, in character-based tests, the assessment of incongruence is focused on the differences between how the distinct data sets fit the trees, whereas in tree-based tests, the assessment of incongruence focuses on the difference between the trees[16]. For example, the character-based measure developed by Shimodaira and Hasegawa (1999) relies on bootstrap resampling of characters to identify whether any one or more of a set of trees best explains the data, whereas Rodrigo's topology-based measure relies on the distribution of tree distances among bootstrap replicate trees to examine the degree of incongruence between sets of characters[25]. Although several of these measures are extremely useful in practice, they frequently lack generality because they depend on a particular optimality criterion[19,22,28] or clade support measure[23,25].

A particularly interesting group of tree-based methods for handling incongruence and summarizing conflict are consensus methods[29]. Because each internode (or internal branch) in a phylogenetic tree represents a bipartition that separates two sets of taxa (e.g., Fig. 3.1 shows a

bipartition *a, b, c, d, e | f, g, h, i, j* that divides the internode between nodes 1 and 5 into taxon sets {*a, b, c, d, e*} and {*f, g, h, i, j*}), a set of trees can be effectively summarized into a consensus tree that depicts only those bipartitions that are 'representative' of the set. For example, the majority-rule consensus (MRC) approach[29] calculates the shared bipartitions across all trees in a set and displays only those shared by the majority of the trees. Consequently, each internode in the MRC tree has a value that corresponds to either the number or the percentage of individual phylogenetic trees that contain the bipartitions created by splitting up the tree at this internode. Although consensus methods have been extremely useful and very popular in summarizing agreement and incongruence, they do not provide information on the next most prevalent conflicting bipartition, or more generally, on the distribution of conflicting bipartitions. For example, when a MRC tree reports that 51 out of 100 phylogenetic trees contain a specific bipartition, whether the second most prevalent, yet conflicting bipartition, is supported by the remaining 49 phylogenetic trees or by only 5 of these is not known. Information about the distribution of conflicting bipartitions, however, can be informative because the first type of conflict in the previous example (51% *versus* 49%) shows that both bipartitions receive almost identical support, whereas the second type (51% *versus* 5%) suggests that the first bipartition represents the sole strongly supported bipartition. Although phylogenetic inference programs typically report the distribution of bipartitions from a set of trees, including those That do not appear in the MRC tree[30,31], and several methods have been developed to visualize the phylogenetic conflict on each internode[32–34], measures that also incorporate conflicting bipartitions to quantify incongruence have so far been lacking.

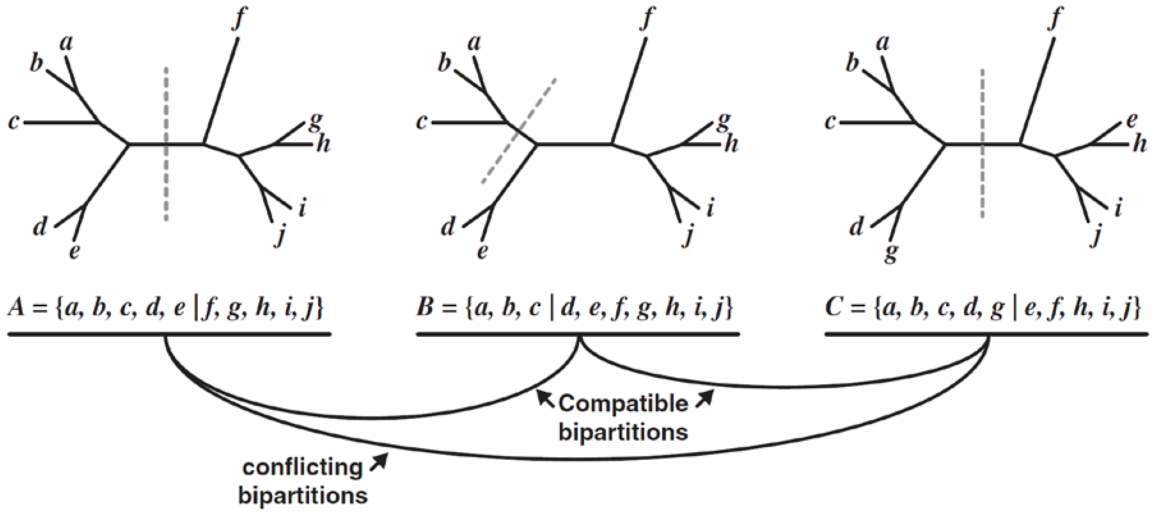
We introduce four related measures that, given a set of trees or characters defining bipartitions, can be used to quantify the degree of incongruence for a given internode, or for an entire tree. The quantification of incongruence or conflict in all four measures is based on Shannon's entropy, a common uncertainty measure for a random variable[35]. The first two measures, Internode Certainty (IC) and Internode Certainty All (ICA), quantify the degree of certainty for each individual internode by considering the two most prevalent conflicting bipartitions (IC) or all most prevalent conflicting bipartitions (ICA), by providing the log magnitude of their difference. The other two measures, Tree Certainty (TC) and Tree Certainty All (TCA), are the sums of IC and ICA values, respectively over all internodes in a phylogeny. In this study, we present the theory of the four measures and illustrate by example how they can be applied to different types of data and biological questions. Finally, we describe how they have been implemented in the widely-used program RAxML.

Four Novel Measures that Use Information Theory to Quantify Incongruence

Phylogenetic trees that represent evolutionary relationships among different genes or taxa are acyclic connected graphs that consist of nodes connected by edges or branches. Each internal branch (or internode) in a phylogenetic tree can also be represented as a bipartition or split that divides the taxa into two disjoint partitions (Fig. 3.1). Therefore, any measure that quantifies internode support will also represent the support for the given bipartition. By considering each internode as a bipartition, any unrooted fully bifurcating phylogenetic tree with k taxa will contain k–3 non-trivial bipartitions (i.e., k–3 bipartitions, each of which divides the $k = m + n$ taxa in the tree into two partitions of m and n taxa, respectively where $m \geq 2$ and $n \geq 2$). If two phylogenetic trees with the same number of taxa k are topologically identical, then the total

number of unique non-trivial bipartitions is still only *k*–3 because the union of the set of bipartitions induced by this second tree with the set of bipartitions induced by the first shows that there are no unique non-trivial bipartitions that are only present in one tree but absent from the other. In contrast, if two phylogenetic trees are incongruent, then the set of phylogenetic trees will contain more than *k*–3 bipartitions, where each of the additional bipartitions represent bipartitions that conflict with one or more of the *k*–3 bipartitions.

Figure 3.1. Compatible and conflicting bipartitions. Bipartition $A=\{a, b, c, d, e \mid f, g, h, i, j\}$ is composed of the partitions $A_1=\{a, b, c, d, e\}$ and $A_2=\{f, g, h, i, j\}$, where *a*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*, and *j* are taxa. Bipartition $B=\{a, b, c \mid d, e, f, g, h, i, j\}$ is composed of the partitions $B_1=\{a, b, c\}$ and $B_2=\{$d, e, f, g, h, i, j$\}$, and bipartition $C=\{a, b, c, d, g \mid e, f, h, i, j\}$ is composed of the partitions $C_1=\{a, b, c, d, g\}$ and $C_2=\{e, f, h, i, j\}$. Bipartitions *A* and *B* are compatible because one of the intersections of their bipartition pairs ($A2 \cap B1$) is empty. Bipartitions *B* and *C* are compatible for the same reason ($B_1 \cap C_2$ is empty). In contrast, bipartition *C* conflicts or is incompatible with bipartition *A* because none of the four intersections ($A_1 \cap C_1$, $A_1 \cap C_2$, $A_2 \cap C_1$, $A_2 \cap C_2$) is empty.



Compatible and Conflicting Bipartitions

Two bipartitions $A = X_1 \mid X_2$ and $B = Y_1 \mid Y_2$ from the same taxon set are compatible if and only if at leastone of the intersections of the four bipartition pairs ($X_1 \cap Y_1$, $X_1 \cap Y_2$, $X_2 \cap Y_1$, $X_2 \cap Y_2$) is

empty[29,34]. If this condition is not met, then the bipartitions are said to be incompatible or incongruent or to conflict with one another.

Example. Let us consider the bipartition $A = \{a, b, c, d, e \mid f, g, h, i, j\}$, comprised by the partitions $A_1 = \{a, b, c, d, e\}$ and $A_2 = \{f, g, h, i, j\}$, where *a*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i* and *j* are taxon names. Let us also consider a second bipartition from the same set of taxa $B = \{a, b, c \mid d, e, f, g, h, i, j\}$, comprised by the partitions $B_1 = \{a, b, c\}$ and $B_2 = \{d, e, f, g, h, i, j\}$ (Fig. 3.1). Bipartition *B* does not conflict with bipartition *A* because $A_2 \cap B_1$ is empty. In contrast, bipartition $C = \{a, b, c, d, g \mid e, f, h, i, j\}$, comprised by the partitions $C_1 = \{a, b, c, d, g\}$ and $C_2 = \{e, f, h, i, j\}$, conflicts or is incompatible with bipartition *A* because none of the four intersections ($A_1 \cap C_1$, $A_1 \cap C_2$, $A_2 \cap C_1$, $A_2 \cap C_2$) is empty (Fig. 3.1).

Shannon's Entropy and Internode Certainty

Shannon's entropy measures the amount of uncertainty in random variables[35]. For two equally probable events, for example "head or tails" in a fair coin toss, the amount of uncertainty is equal to 1. However, if the coin is not fair the uncertainty of the outcome decreases proportionally to the coin's 'unfairness'. In general, for a random variable X with a set of *n* possible values $\{X_1, X_2 \ldots X_n\}$ Shannon's entropy $H(X)$ is defined as

$$H(X) = -\sum_{n=1}^{n} P(X_n) \log\left(P(X_n)\right)$$

where $P(X_n)$ is the probability of outcome X_n . In its simplest form, if variable *X* consists of only two possible outcomes X_1 and X_2, Shannon's entropy is equal to

$$H(X) = -\sum_{n=1}^{2} P(X_n) \log_2 \left(P(X_n) \right)$$

In phylogenetics, let us consider variable $H(X)$ as the entropy that measures the amount of uncertainty of support for a given internode with the set of possible values being the values of the two most prevalent conflicting bipartitions ($n = 2$) for that internode (i.e., $X = \{X_1, X_2\}$), with X_1 being the frequency of support for the bipartition that defines the internode. For these two bipartitions X_1 and X_2 we define $H(X)$ as the Internode Uncertainty:

$$\textit{Internode Uncertainty} = -\sum_{n=1}^{2} P(X_n) \log_2 \left(P(X_n) \right) =$$

$$= \quad P(X_1) \log_2 \left(P(X_1) \right) + P(X_2) \log_2 \left(P(X_2) \right)$$

where $P(X_1) = X_1 / (X_1 + X_2)$, $P(X_2) = X_2 / (X_1 + X_2)$, and $P(X_1) + P(X_2) = 1$.

Because internode support measures typically quantify the degree of support for a given internode, rather than the lack thereof, we reverse the sign of the equation and add log2 (n) to it so that the measure corresponds to *certainty* rather than *uncertainty*. Thus, we define Internode Certainty (IC) as

$$IC = \log_2(n) + \sum_{n=1}^{2} P(X_n)\log_2\left(P(X_n)\right) \quad =$$

$$= \quad 1 + P(X_1)\log_2\left(P(X_1)\right) + P(X_2)\log_2\left(P(X_2)\right)$$

where $P(X_1) = X_1 / (X_1 + X_2)$, $P(X_2) = X_2 / (X_1 + X_2)$, and $P(X_1) + P(X_2) = 1$.

For a given internode, IC values correspond to the magnitude of conflict between the bipartition that defines the internode and the most prevalent conflicting bipartition in the given tree set. For example, IC values at or close to 1 indicate the absence of conflict for the bipartition defined by a given internode, whereas IC values at or close to 0 indicate equal support for both bipartitions and hence maximum conflict.

So far, we have assumed that the frequency of the bipartition that defines the internode is equal or higher than the frequency of the most prevalent bipartition, that is, $P(X_1) \geq P(X_2)$. However, in some cases it may happen that we need to calculate the IC of an internode that was included in the consensus tree (depending on the type of consensus tree constructed, see below) whose bipartition frequency is actually smaller than the frequency of a conflicting bipartition, that is $P(X_1) \leq P(X_2)$. To distinguish between cases where $P(X_1) \geq P(X_2)$ from cases where $P(X_1) \leq P(X_2)$, we reverse the sign of the IC value for all cases where $P(X_1) \leq P(X_2)$. Thus, negative IC values indicate that the internode of interest conflicts with a bipartition that has higher frequency, and IC values at or close to –1 indicate an almost complete absence of support for the bipartition defined by the given internode and an almost absolute support for the conflicting bipartition. The behavior of the IC measure for a range of different values of X_1 and X_2 is shown in Fig. 3.2.

Examples. Let us consider a set of 100 gene trees, from which 62 gene trees support bipartition X_1, which appears on the MRC tree, and 6 gene trees support the conflicting bipartition X2 (which does not appear on the MRC tree). In this case,

$P(X_1) = X_1 / (X_1 + X_2) = 62 / (62 + 6) = 0.91$, and $P(X_2) = X_2 / (X_1 + X_2) = 6 / (62 + 6) = 0.09$. Thus,

$IC = 1 + P(X_1) \log_2 ((P(X_1)) + P(X_2) \log_{2(} P(X_2)) = 1 + 0.91 * \log_2(0.91) + 0.09 * \log_2(0.09)$ **=0.57**

If $X_1 = 52$ gene trees and the conflicting bipartition $X_2 = 29$ gene trees, then

$P(X_1) = X_1 / (X_1 + X2) = 52 / (52 + 29) = 0.64$, and $P(X_2) = X_2 / (X_1 + X_2) = 29 / (52 + 29) = 0.36$. Thus,

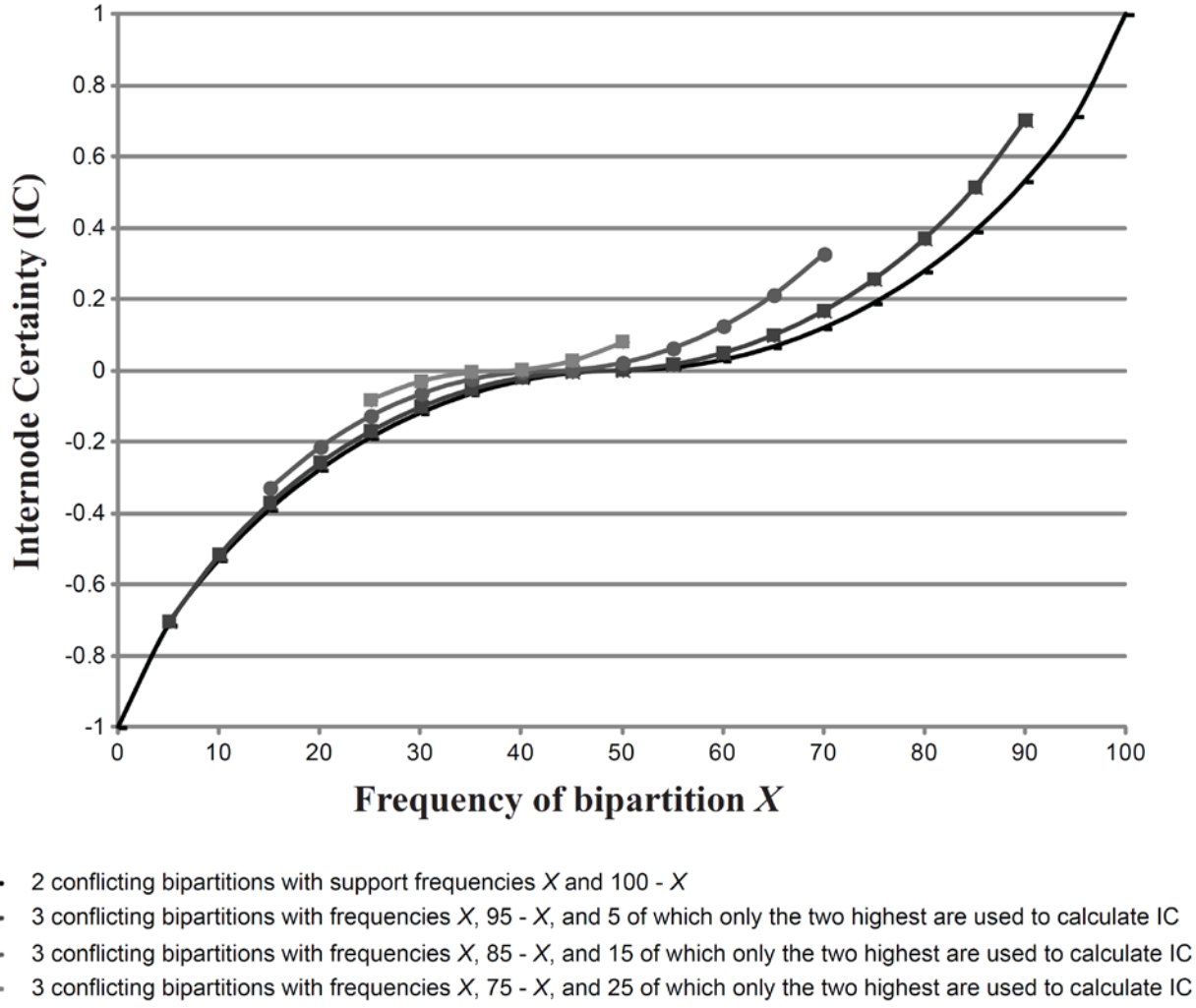
$IC = 1 + P(X_1) \log_2 (P(X_1)) + P(X_2) \log_2 (P(X_2)) = 1 + 0.64 * \log_2(0.64) + 0.36 * \log_2(0.36)$ **=0.06**

Finally, if an internode is defined by a bipartition X_1 supported by 5 gene trees and the conflicting bipartition X_2 is support by 55 gene trees, then

$P(X_1) = X_1 / (X_1 + X_2) = 5 / (5 + 55) = 0.08$, and $P(X_2) = X_2 / (X_1 + X_2) = 55 / (5 + 55) = 0.92$. Thus,

$IC = 1 + P(X_1) \log_2(P(X_1)) + P(X_2) \log_2(P(X_2)) = 1 + 0.08 * \log_2(0.08) + 0.92 * \log_2(0.92) =$ **–0.59**

Fig. 3.2. Visualizing IC for the two most prevalent conflicting bipartitions of a given internode. The default curve represents the case of only two conflicting bipartitions for one internode (only two partitions: {X, 100 – X}). Out of 100 total trees, when 60 trees recover the first bipartition, the remaining 40 will support the second and conflicting bipartition. In the presence of three conflicting bipartitions for a given internode (e.g., {65, 30, 5}), when the two most prevalent bipartitions are considered, the percentage of trees supporting the first bipartition is equal to 65 / (65 + 30), whereas the percentage of trees supporting the second conflicting bipartition is equal to 30 / (65 + 30). The reason that we do not include the number of trees containing the third bipartition is that we want IC to measure the magnitude of certainty conveyed by the two most prevalent bipartitions. This way, IC will be zero when the two most prevalent conflicting bipartitions have equal frequencies.



Extending IC to Include All Prevalent Conflicting Bipartitions

The IC can be extended to consider all *n* prevalent conflicting bipartitions for a given internode, that is $X = \{X_1, X_2 \ldots X_n\}$. This measure, which we name Internode Certainty All (ICA), can be calculated using

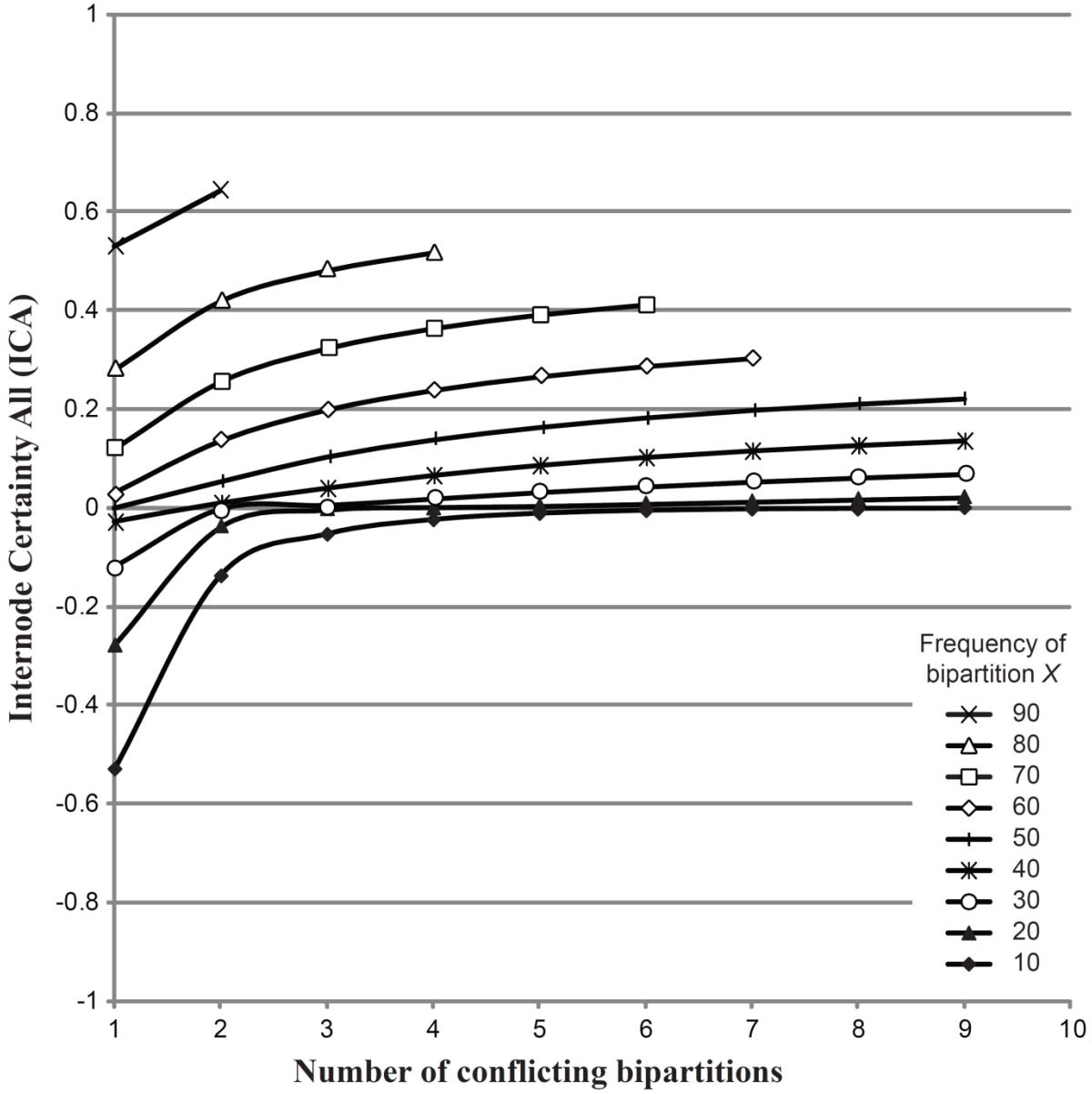
$$\text{ICA}=\log_n(n)+ P(X_1) \log_n (P(X_1))+ P(X_1) \log_n (P(X_1))+\ldots+ P(X_n) \log_n(P(X_n))$$

where $P(X_1) = X_1 / (X_1 + X_2 + \ldots + X_n)$, $P(X_2)= X_2 / (X_1 + X_2 + \ldots + X_n)$, … ,

$P(X_n) = X_n / (X_1 + X_2 + \ldots + X_n)$, and $P(X_1) + P(X_2)+ \ldots + P(X_n)= 1$.

Because the number of bipartitions that conflict with a given internode in large phylogenetic tree sets can be high, as well as because conflicting bipartitions whose frequency is very low have little impact on the certainty value of a given internode, we restrict the ICA to consider only bipartitions whose frequency is ≥ 5% because this represents a reasonable trade-off between speed and accuracy. To distinguish between cases where $P(X_1)$ is greater than or equal to each single one of the frequencies for all conflicting bipartitions from cases where $P(X_1)$ is lower than one or more conflicting bipartitions, we reverse the sign of the ICA for all cases where $P(X_1)$ is lower. Thus, ICA values at or near 1 indicate the absence of any conflict for the bipartition defined by a given internode, whereas ICA values at or near 0 indicate that one or more conflicting bipartitions have almost equal support. Negative ICA values indicate that the internode of interest conflicts with one or more bipartitions that exhibit a higher frequency and ICA values at or near 1 indicate the absence of support for the bipartition defined by a given internode. The behavior of the ICA measure for a range of different values of $X_1, X_2, \ldots, X_n$ is shown in Fig. 3.3.

Fig. 3.3. Visualizing ICA for all the most prevalent conflicting bipartitions of a given internode. For simplicity, calculations were performed using a 2-variable system (X : Y … Y) with the number of conflicting bipartitions increasing. For example, the open triangle line on the graph illustrates the behavior of ICA when the frequency of the most strongly supported bipartition for a given internode is 80, with the remaining 20% equally divided among all conflicting bipartitions (e.g., if there is one conflicting bipartition it will have a frequency of 20%, if there are two conflicting bipartitions each one will have a frequency of 10%, etc.).



Examples. Let us consider a set of 100 gene trees, from which 80 gene trees support bipartition X_1, 6 gene trees support the conflicting bipartition X_2, and 5 gene trees support the conflicting bipartition X_3. In this case,

$P(X_1) = X_1 / (X_1 + X_2 + X_3) = 80 / (80 + 6 + 5) = 0.88$,

$P(X_2) = X_2 / (X_1 + X_2 + X_3) = 6 / (80 + 6 + 5) = 0.07$, and

$P(X_3) = X_3 / (X_1 + X_2 + X_3) = 5 / (80 + 6 + 5) = 0.05$. Thus,

ICA=1+ $P(X_1) \log_3 (P(X_1)) + P(X_2) \log_3 (P(X_2)) + P(X_3) \log 3 (P(X_3))$ =

$=1+0.88* \log_3 (0.88) +0.07* \log_3 (0.07) +0.05*\log_3 (0.05) =$ **0.59**

If $X_1 = 52$ gene trees and the conflicting bipartitions $X_2 = 29$ gene trees and $X_3 = 19$ gene trees, then

$P(X_1) = X_1 / (X_1 + X_2 + X_3) = 52 / (52 + 29 + 19) = 0.52$,

$P(X_2) = X_2 / (X_1 + X_2 + X_3) = 29 / (52 + 29 + 19) = 0.29$, and

$P(X_3) = X_3 / (X_1 + X_2 + X_3) = 19 / (52 + 29 + 19) = 0.19$. Thus,

ICA=1+ $P(X_1) \log 3 (P(X_1)) + P(X_2) \log_3 (P(X_2)) + P(X_3) \log_3 (P(X_3))$ =

$=1+0.52 *\log_3 (0.52) +0.29* \log_3 (0.29) + 0.19* \log_3 (0.19) =$ **0.08**

Finally, if $X_1 = 5$ gene trees and the conflicting bipartitions $X_2 = 15$ gene trees and $X_3 = 11$ gene trees, then

$P(X_1) = X_1 / (X_1 + X_2 + X_3) = 5 / (5 + 15 + 11) = 0.16$,

$P(X_2) = X_2 / (X_1 + X_2 + X_3) = 15 / (5 + 15 + 11) = 0.48$, and

$P(X_3) = X_3 / (X_1 + X_2 + X_3) = 11 / (5 + 15 + 11) = 0.36$. Thus,

ICA=1+ $P(X_1) \log_3 (P(X_1)) + P(X_2) \log_3 (P(X_2)) + P(X_3) \log_3 (P(X_3))$ =

$=1+0.16 *\log_3 (0.16) + 0.48* \log_3 (0.48) + 0.36* \log_3 (0.36) =$ **0.08**

However, because $P(X_1) \leq P(X_2)$ and $P(X_1) \leq P(X_3)$, the sign of the ICA value is reversed to -0.08.

Tree Certainty

Given that empirical examinations of the support frequencies of internodes in a phylogeny suggest that they are generally independent from each other[3], it is reasonable to assume that the mutual information or dependence between internodes in a phylogenetic tree is very small. Thus, the sum of all IC or ICA values across a phylogeny can be used to quantify changes in the degree of incongruence produced by the phylogenetic analysis of a given data set when analyzed with a variety of protocols or methods. Thus, for the complete set of k – 3 internodes (internal branches) in a phylogeny, where k is the number of taxa, we define the Tree Certainty (TC) as

$$TC = \sum_{i=1}^{i=k-3} IC_i$$

and Tree Certainty All (TCA) as

$$TCA = \sum_{i=1}^{i=k-3} ICA_i$$

The maximum TC or TCA value is equal to $k - 3$ and indicates a comprehensive absence of conflict in the phylogeny. When comparing phylogenies with different taxon numbers, a normalized value of TC or TCA can also be obtained by dividing the TC value by $k - 3$, the number of internodes in the phylogeny.

Applications of IC, ICA, TC, and TCA

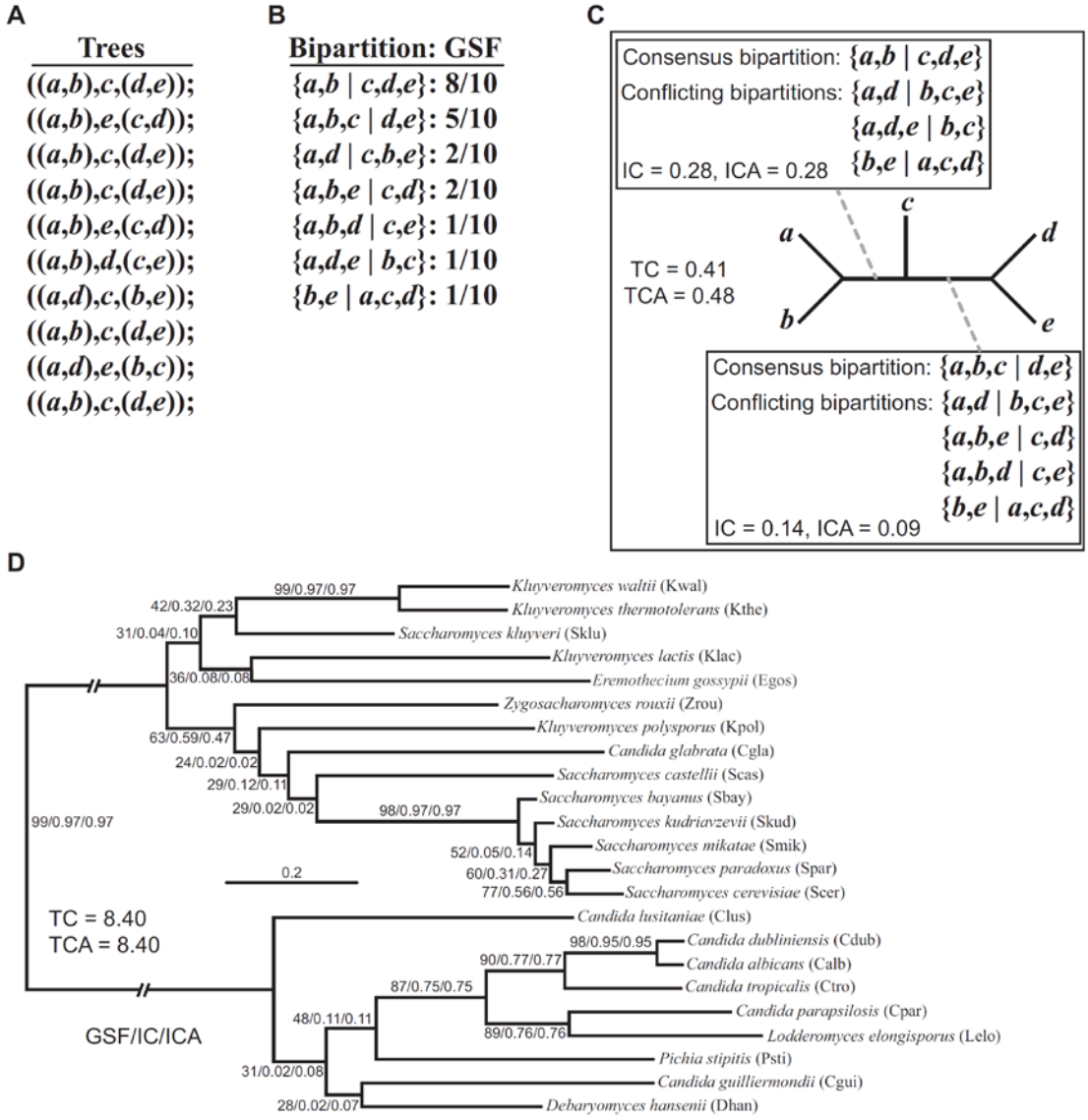
All four measures can be used to quantify incongruence on any dataset that contains bipartitions, including from bootstrap replicate trees, gene trees or individual characters (e.g., from morphology, from large-scale and rare genomic changes, or from individual sites in a sequence alignment). To demonstrate the utility of the four measures, we discuss three commonly used data types here where one can deploy IC, ICA, TC, and TCA to quantify incongruence.

IC, ICA, TC, and TCA Can Quantify Incongruence in Sets of Trees

The most straightforward use of the four measures is for quantifying incongruence on a set of trees (Fig. 3.4); often, this set is comprised of the gene trees obtained from analysis of several different genes collected from the same set of taxa. In this case, calculation of the four measures will be based on the frequency values of the bipartitions present in the entire set of gene trees; note that, the frequency value of a bipartition is also known as gene support frequency or GSF and reflects the percentage of gene trees that contain the bipartition[36]. When quantifying incongruence in a set of gene trees, the IC and ICA values of a given internode will reflect the degree of incongruence for that internode in the set of gene trees, and the TC and TCA values will reflect the degree of incongruence between the individual gene trees across the entire phylogeny. When applied to a dataset of 1,070 gene trees from 23 taxa, the IC and ICA values revealed high levels of incongruence in several internodes of the extended majority-rule consensus phylogeny and enabled us to distinguish between internodes that have similar GSF values but very different degrees of conflict (Fig. 3.4D). Specifically, the placement of *Saccharomyces bayanus* and of *Zygosaccharomyces rouxii* received 52% and 62% GSF, whereas their IC values were 0.05 and 0.59 and their ICA values were 0.14 and 0.47, respectively (Fig.

3.4D). This marked difference between the GSF and the IC / ICA values of the two internodes is a result of the absence of well-supported bipartitions that conflict with the placement of *Z. rouxii* and the presence of well-supported bipartitions that conflict with the placement of *S. bayanus*[3,37].

Fig. 3.4. IC, ICA, TC, and TCA can quantify incongruence in any set of trees or bipartitions. Given a set of trees (panel A) that defines a set of bipartitions (panel B), one can use the four measures to quantify incongruence (panel C). For example, examination of 1,070 gene trees revealed the presence of extensive incongruence in a phylogeny of 23 yeast taxa (panel D) (values near internodes correspond to GSF / IC / ICA values).



When analyzing phylogenetic trees from a single gene or set of genes (multiple genes in supermatrix), it is standard practice to calculate the robustness of support for each internode of the gene tree via bootstrapping[38]. One can thus use the set of bootstrap replicate trees for a given gene to calculate IC, ICA, TC, and TCA. In this case, calculation of the measures will be based on the frequency values of the bipartitions present in the entire set of bootstrap replicate trees, which are better known as bootstrap support values. When quantifying incongruence in a set of bootstrap replicate trees from a single gene, the IC and ICA values of a given internode will reflect the degree of incongruence for that internode in the set of bootstrap replicate trees, and the TC and TCA values will reflect the degree of incongruence between the individual bootstrap replicate trees across the entire gene phylogeny. For example, in our recent study[3] we ranked 1,070 genes from 23 yeast species based on their TC value as calculated from each gene's bootstrap trees. Interestingly, concatenation analysis of the 131 genes with the highest TC placed *C. glabrata* in a position that is also supported by several distinct rare genomic changes[39], a result that contradicts both the analysis of all 1,070 genes as well as previously published phylogenomic analyses[4,40–42].

IC, ICA, TC, and TCA Can Quantify Incongruence in Sets of Bipartitions

The four measures can also be calculated from a set of partially resolved trees or even directly from bipartitions (Fig. 3.4B, C). For example, the bipartitions present in each gene tree rarely receive equal support; the bootstrap consensus tree of virtually every gene shows that certain internodes receive higher bootstrap support or IC / ICA values, indicating that the degree of congruence of phylogenetic signals as well as the degree of "noise" from a given gene differs widely across internodes. Thus, it may frequently be desirable to use only a genes' highly

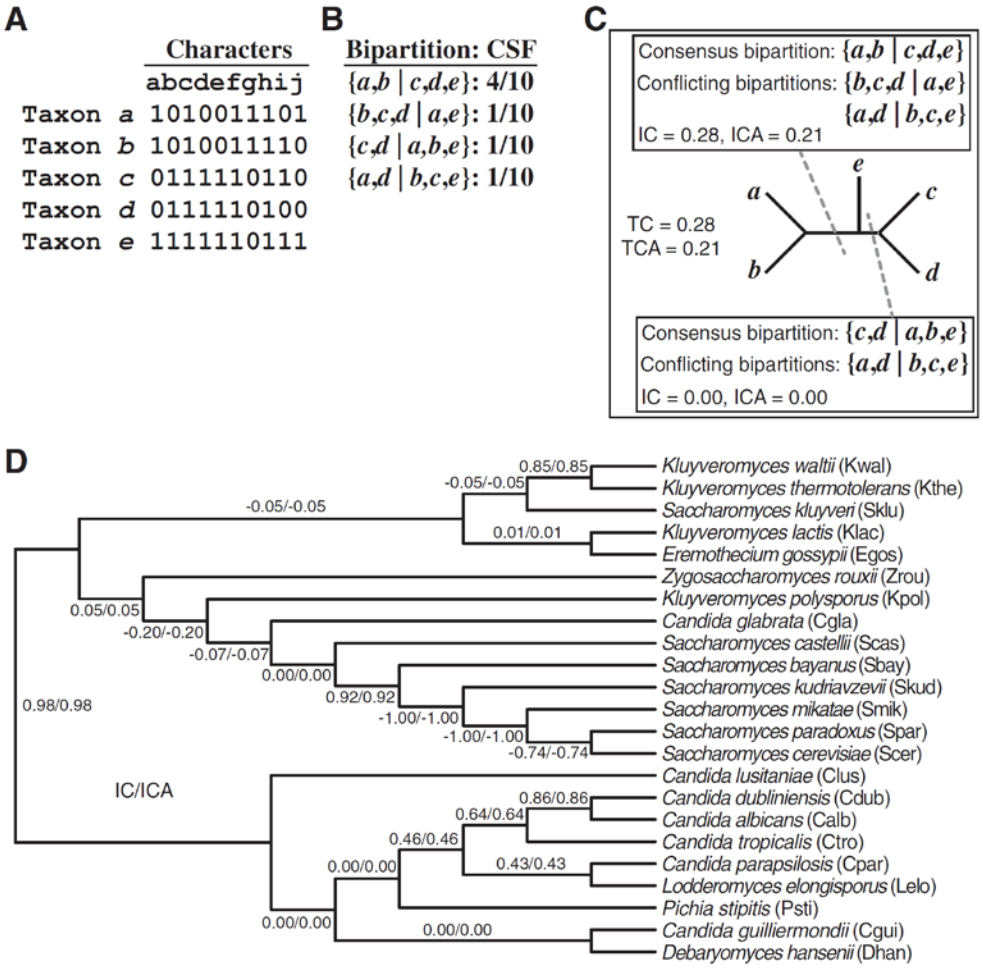
supported bipartitions in the inference of consensus phylogenies (one can easily select the highly supported bipartitions in the bootstrap consensus tree of a given gene by "collapsing" all internodes with bootstrap support values below a certain threshold using software such as the CONSENSE program in the PHYLIP package[30]). In this case, calculation of the four measures will be exclusively based on the frequency values of those bipartitions that received high support (e.g., high bootstrap support) or present low conflict in the entire set of gene bootstrap consensus trees. Thus, the IC and ICA values of a given internode in the consensus tree will reflect the degree of incongruence for that internode among only the group of highly supported bipartitions present in the set of gene trees, whereas the TC and TCA values will reflect the degree of incongruence between highly supported bipartitions across the entire phylogeny. Note that, the use of IC or ICA overcomes potential issues when only a small number of highly supported bipartitions are associated with a given internode by measuring the degree of incongruence independently of the number of bipartitions taken into consideration. For example, both the IC and the ICA value for the sister group *Saccharomyces cerevisiae* and *S. paradoxus* calculated from an analysis of 1,070 gene trees from 23 yeast taxa is 0.56 (Fig. 3.4D). In contrast, both the IC and ICA values calculated using only those bipartitions that received $\geq$ 80% bootstrap support in individual gene analyses of the same 1,070 genes are 0.85, suggesting that most of the observed incongruence in the resolution of this internode stems from conflict among weakly supported bipartitions.

IC, ICA, TC, and TCA Can Quantify Incongruence in Sets of Individual Characters

Because the four measures can be applied to any dataset that contains taxon bipartitions one can extend their use to quantifying the level of phylogenetic conflict on any character in which the

distribution of character states is such that it splits the taxon set into two non-trivial bipartitions (Fig. 3.5). Assuming a character with two states 0 and 1 from a set of $k = m + n$ taxa, where $m \geq 2$ and $n \geq 2$, any site with a character state distribution of $(0_1 \ldots 0_m, 1_1 \ldots 1_n)$ corresponds to the bipartition {*m* taxa} / {*n* taxa}. Thus, one can use IC or ICA to quantify the degree of incongruence for a given bipartition defined by a character across a set of characters by considering the number of characters supporting that bipartition jointly with the number of characters supporting the most prevalent bipartition that conflicts with it (IC) or jointly with the numbers of characters supporting all most prevalent bipartitions that conflict with it (ICA). Note that, much like GSF reflects the frequency of bipartitions in a set of trees, the frequency value of a bipartition defined by a character reflects the percentage of characters that support the bipartition, which we denote as character support frequency (CSF). Examples of characters that can be used to define bipartitions include rare genomic changes[43], indels, sites that contain a single substitution between amino acids that differ radically in their physicochemical properties[44], binary morphological characters, as well as any other binary characters. For example, analysis of 20,289 sites that contain single radical substitutions (defined as substitutions with a blosum62 matrix score ≤ -3), from the dataset of 1,070 genes from 23 yeast taxa, also known as RGC_CAMs[44], showed that the bipartitions defined by such sites were more incongruent than the bipartitions present in the 1,070 gene trees.

Fig. 3.5. IC, ICA, TC, and TCA can quantify incongruence in any set of characters that define bipartitions. Given a set of characters (panel A) that defines a set of bipartitions (panel B), one can use the four measures to quantify incongruence (panel C). For example, examination of 20,289 sites that contain single radical substitutions (defined as substitutions with a blosum62 matrix score ≤ –3), from the dataset of 1,070 genes from 23 yeast taxa showed that the bipartitions defined by such sites not only lacked information about several internodes of the yeast phylogeny but also displayed considerable levels of incongruence.



Using TC and TCA to Evaluate the Impact of Different Practices in Data Analysis

Summing the IC or ICA values across all internodes of a phylogeny amounts to the phylogeny's TC or TCA, respectively. One useful application of the TC and TCA measures is for comparing the relative impact of different analytical practices on incongruence. For example, one could calculate the TC and TCA values of the extended MRC phylogeny constructed from the gene

trees estimated from analysis of 100 genes with only those sites that do not contain missing data and compare it with the TC / TCA measured from the eMRC phylogeny constructed from analysis of the same 100 genes in which only sites with more than 50% data missing are excluded. In this case, the practice with the highest TC / TCA value will be that one that displays the lowest degree of incongruence among the 100 gene trees. In contrast, a high decrease in TC / TCA may indicate that a particular data filtering approach increases incongruence across the phylogeny. For example, examination of the TC of the trees from the 100 slowest-evolving genes in a data matrix comprised of 1,070 genes from 23 yeast taxa showed that they had a substantially lower TC than the TC calculated by considering all 1,070 gene trees[3].

Calculating IC, ICA, TC, and TCA using the RAxML software

We implemented the score calculations of the four measures in RAxML[45] (version 7.7.8, available via https://github.com/stamatak/standard-RAxML), taking advantage of already available efficient data structures for performing calculations on bipartitions[46]. For a full description of the commands for calculation of the four measures and an example, please see the manual (Supplementary Text File) and test dataset (Supplementary Data file). Given a set of gene trees, RAxML can directly calculate a MRC as well as an eMRC tree on this set that is annotated by the respective IC and ICA values. The particularly compute-intensive inference of eMRC trees (finding the optimal eMRC tree is, in fact, NP-hard[47]) relies on the fast parallel implementation presented in Aberer et al.[46]. It can also compute stricter MRC trees with arbitrary threshold settings that range between 51 and 99%. Furthermore, we have implemented an option that allows for drawing IC scores onto a given, strictly bifurcating reference tree (e.g., the best-known ML tree).

Note that, the IC and ICA values are represented as branch labels, since, as is the case for bootstrap support values, information associated to bipartitions of a tree always refers to its internodes (internal branches) and *not* its nodes. Each tree viewer (e.g., Dendroscope[34]) that can properly parse the Newick tree format is able to display these branch labels. The rationale for not providing IC values as node labels is that some tree viewers may not properly rotate the node labels when the user reroots the tree, leading to an erroneous internal branch-to-IC-value association.

When calculating the IC and ICA values on extended MRC trees or onto a given reference tree it may occur that, the bipartition that has been included in the tree has lower support than one or more conflicting bipartitions (see also above). In this case, RAxML will display a warning to the user and annotate the internode with a negative IC value. Note that, this is not only a theoretical possibility when using extended MRC trees, but a frequent observation for bipartitions that have low frequency in a gene tree set or that have low bootstrap support in a set of bootstrap replicate trees.

RAxML also calculates the TC and TCA values as well as their relative values that are normalized by the maximum possible TC / TCA values for a given phylogeny. Finally, we have implemented a verbose output option that allows users to further scrutinize particularly interesting conflicting bipartitions. In verbose mode RAxML will generate two types of output files: one set of files containing the bipartition included in the MRC tree and its corresponding conflicting bipartitions in Newick format and an output file listing all bipartitions (included and conflicting) with their IC and ICA values in a PHYLIP-like format.

DISCUSSION

To tackle gene incongruence, phylogeneticists often resort to creating concatenated data matrices comprised of tens or hundreds of genes[1,48–51]. Because the vast majority of concatenation studies assesses robustness in inference using bootstrapping, an extremely useful measure of robustness of inference when data are limited[38] but one that in the presence of large amounts of data will nearly always result in 100% support[3,10,41] numerous studies purport to have resolved long-standing phylogenetic problems. However, different phylogenomic studies focused on the same internodes sometimes provide contradicting, but equally robustly supported, answers[49,50,52,53], suggesting that incongruence is not ameliorated, but rather masked, by these practices. Consequently, accurate phylogenetic inference requires not only large amounts of data and absolute bootstrap support, but also demonstration that the data do not contain substantial amounts of conflicting phylogenetic signal[3]. Thus, accurate inference requires methods that identify and quantify conflicts in phylogenetic signal.

To quantify the degree of incongruence present in phylogenomic data matrices, we developed two novel measures, IC and ICA, which quantify the degree of conflict on each specific internode of a phylogeny and two novel measures, TC and TCA, which quantify the degree of conflict for the whole tree. All four measures can be used for a wide variety of different phylogenetic markers, from individual characters to gene trees to genomic characters (Figs. 4 and 5) and are meant to provide simple, fast and intuitive measurements that identify the presence of incongruence in a phylogenomic data matrix rather than to elucidate the root cause(s) of the observed incongruence. Even though the absolute values of our measures are not aimed to provide statistical significance, the degree of certainty calculated derives from the amount of information on each internode. For example, in the case of IC the degree of certainty corresponds

to the ratio between the most prevalent and the next most prevalent, but conflicting, bipartition (Fig. 2). If the most prevalent bipartition is supported by 95% of the data and the next most prevalent conflicting bipartition is supported by the remaining 5%, then the value of the IC measure will be approximately 0.71, whereas if the two most prevalent conflicting bipartitions have the same frequency of support, then IC will equal zero.

Compared to the very popular incongruence length difference test[28], our measures can easily be applied to the study of a single internode or the whole tree, to study one or many data partitions, and are not dependent on a particular optimality criterion. Compared to topology constraint tests, such as the Kishino-Hasegawa (KH) test[20], the Shimodaira-Hasegawa (SH) test[23], and the Approximately Unbiased (AU) test[54], there is no need for a priori tree selection and multiple internodes can be examined simultaneously very quickly. The price of this speed and flexibility, however, is that our tests are not designed to test specific phylogenetic hypotheses or provide estimates of statistical significance; in many ways, our measures are designed to quickly identify incongruence in phylogenomic data matrices, enabling users to further explore its causes using more custom methods.

Our IC, ICA, TC, and TCA measures do not distinguish whether a low degree of certainty is the result of strong conflicts in phylogenetic signal, or random noise due to be absence of any signal. In other words, incongruence between trees does not necessarily indicate conflicting support, because incongruent trees are also the null expectation when a data matrix contains no phylogenetic signal (although, differences between IC and ICA values may alert for the presence of more than two signals). In such cases, users are advised to examine whether the tree distance distribution of observed trees deviates significantly from randomness by using a tree distance method[3,4], such as the Robinson-Foulds tree distance[55], prior to inferring that the low degree of

certainty in a data matrix is the result of strong conflicts in phylogenetic signal. Other alternatives include employing the more computationally-intensive topology constraint KH, SH, or AU tests[20,23,54].

One potential drawback when applying the IC, ICA, TC, and TCA measures is their values may not be representative when small numbers of characters or gene trees are used. Although this is a general problem that influences all measures, including bootstrap support (BS) and gene support frequency (GSF), our measures are likely to be most informative when applied to large amounts of data (e.g., hundreds of characters or dozens of genes or hundreds of bootstrap replicates). Our TC and TCA measures also assume that the support frequencies of internodes in a phylogeny are independent from each other. Even though this is an approximation, previous results suggest that the application of a variety of standard practices aimed at reducing incongruence, such as removal of unstable or fast-evolving taxa, do not affect IC and ICA values across the entire phylogeny; rather, their effects are largely localized on one particular internode[3]. It should be noted that such a focus on a single internode or a small, local neighborhood of an internode represents a common approximation in phylogenetics and is frequently used to design search heuristics or statistical tests such as the aLRT test[56].

Finally, IC, ICA, TC, and TCA measures, as currently implemented in RAxML, cannot be applied on datasets with missing data (for example when some genes are missing from certain taxa), because dealing with trees that only contain subset of the overall taxon set is computationally substantially more challenging and requires the appropriate adaptation and/or extension of supertree methods. Hence, the solution to this problem is not straightforward, but we hope to address this challenging issue in the near future.

ACKNOWLEDGMENTS

We thank Christoph Hahn for testing early RAxML implementations of these measures and for constructive feedback. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. This work was supported by the National Science Foundation (DEB-0844968) and by institutional funding from the Heidelberg Institute for Theoretical Studies.

REFERENCES

1. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*. 2003;425(6960):798-804. doi:10.1038/nature02053.

2. Rokas A, Chatzimanolis S. From gene-scale to genome-scale phylogenetics: the data flood in, but the challenges remain. *Methods Mol Biol*. 2008;422:1-12. doi:10.1007/978-1-59745-581-7_1.

3. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*. 2013;497(7449):327-31. doi:10.1038/nature12130.

4. Hess J, Goldman N. Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: Yeasts revisited. *PLoS One*. 2011;6(8).

5. Zhong B, Liu L, Yan Z, Penny D. Origin of land plants using the multispecies coalescent model. *Trends Plant Sci*. 2013;18(9):492-495.

6. Song S, Liu L, Edwards S V., Wu S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci*. 2012;109(37):14942-14947.

7. Bull J. J., J. P. Huelsenbeck, Clifford W. Cunningham DLS and PJW. Partitioning and Combining Data in Phylogenetic Analysis. *Syst Biol*. 1993;42(3):384-397.

8. Cummings MP, Otto SP, Wakeley J. Sampling properties of DNA sequence data in phylogenetic analysis. *Mol Biol Evol*. 1995;12(5):814-822.

9. Swofford, D.L. et al. Phylogenetic inference. In: Hillis, David M., Craig Moritz BKM, ed. *Molecular Systematics*. 2nd ed. Sinauer Associates, Inc; 1996:407-514.

10. Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. Statistics and truth in phylogenomics. *Mol Biol Evol*. 2012;29(2):457-472.

11. Pamilo P, Nei M. Relationships between gene trees and species trees. *Mol Biol Evol*. 1988;5(5):568-583.

12. Slowinski J, Page R. How should species phylogenies be inferred from sequence data? *Syst Biol*. 1999;48:814-825.

13. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*. 2009;24(6):332-340.

14. Cunningham CW. Can three incongruence tests predict when data should be combined? *Mol Biol Evol*. 1997;14(7):733-740.

15. Huelsenbeck JP, Bull JJ, Cunningham CW. Combining data in phylogenetic analysis. *Trends Ecol Evol*. 1996;11(4):152-158. doi:10.1016/0169-5347(96)10006-9.

16. Planet PJ. Tree disagreement: Measuring and testing incongruence in phylogenies. *J Biomed Inform*. 2006;39(1 SPEC. ISS.):86-102.

17. Wilson E. A consistency test for phylogenies based on contemporaneous species. *Syst Zool*. 1965;14:214–220.

18. Le Quesne W. A method of selection of characters in numerical taxonomy. *Syst Zool*. 1969;18:201-205.

19. Templeton AR. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution (N Y)*. 1983;37(2):221-244. doi:10.2307/2408332.

20. Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol*. 1989;29(2):170-179.

21. Faith D. Cladistic permutation tests for monophyly and nonmonophyly. *Syst Zool*. 1991;40:366–375.

22. Baker RH, DeSalle R. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Syst Biol*. 1997;46(4):654-673. doi:10.1093/sysbio/46.4.654.

23. Shimodaira H, Hasegawa M. Letter to the Editor Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol*. 1999;16(8):1114-1116. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21178118.

24. Goldman N, Anderson JP, Rodrigo AG. Likelihood-based tests of topologies in phylogenetics. *Syst Biol*. 2000;49(4):652-670.

25. Rodrigo AG, Kelly-Borges M, Bergquist PR, Bergquist PL. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *New Zeal J Bot*. 1993;31(3):257-268. doi:10.1080/0028825X.1993.10419503.

26. Thorley JL WM. Testing the phylogenetic stability of early tetrapods. *J Theor Biol*. 1999;200:343-344.

27. Thorley JL, Page RD. RadCon: phylogenetic tree comparison and consensus. *Bioinformatics*. 2000;16(5):486-487. doi:10.1093/bioinformatics/16.5.486.

28. Farris JS, Kallersjo M, Kluge AG BC. Testing significance of incongruence. *Cladistics*. 1994;10:315-319.

29. Bryant D. A classification of consensus methods for phylogenetics. *DIMACS Ser Discret Math Theor Comput Sci*. 2003;61:163-184.

30. Felsenstein J. Phylip: phylogeny inference package (version 3.2). *Cladistics*. 1989;5:164-166. doi:10.1111/j.1096-0031.1989.tb00562.x.

31. Swofford DL. Phylogenetic Analysis Using Parsimony * (and other methods). Version 4. *Options*. 2002:1-142. doi:10.1159/000170955.

32. Lento GM, Hickson RE, Chambers GK, Penny D. Use of spectral analysis to test hypotheses on the origin of pinnipeds. *Mol Biol Evol*. 1995;12(1):28-52. doi:7877495.

33. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 2006;23(2):254-267.

34. Huson DH, Scornavacca C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst Biol*. 2012;61(6):1061-1067.

35. Shannon CE. A Mathematical Theory of Communication. *Bell Syst Tech J*. 1948;27(3):379-423. doi:10.1002/j.1538-7305.1948.tb01338.x.

36. Gadagkar SR, Rosenberg MS, Kumar S. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *J Exp Zool Part B Mol Dev Evol*. 2005;304(1):64-74.

37. Yu Y, Degnan JH, Nakhleh L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet*. 2012;8(4).

38. Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution (N Y)*. 1985;39:783-791. Available at: http://www.jstor.org/stable/2408678.

39. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*. 2006;440(7082):341-345. doi:10.1038/nature04562.

40. Hittinger CT, Rokas A, Carroll SB. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci U S A*. 2004;101(39):14144-14149.

41. Rokas A, Carroll SB. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol*. 2005;22(5):1337-1344.

42. Fitzpatrick DA, Logue ME, Stajich JE, Butler G. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol*. 2006;6:99.

43. Rokas A, Holland PWH. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol*. 2000;15(11):454-459.

44. Rogozin IB, Wolf YI, Carmel L, Koonin E V. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. *Mol Biol Evol*. 2007;24(4):1080-1090.

45. Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22(21):2688-2690.

46. Aberer AJ, Pattengale ND, Stamatakis A. Parallelized phylogenetic post-analysis on multi-core architectures. *J Comput Sci*. 2010;1(2):107-114.

47. Phillips C, Warnow TJ. The asymmetric median tree — A new model for building consensus trees. *Discret Appl Math*. 1996;71(1-3):311-335.

48. Rokas A, Krüger D, Carroll SB. Animal evolution and the molecular signature of radiations compressed in time. *Science*. 2005;310(5756):1933-1938.

49. Dunn CW, Hejnol A, Matus DQ, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. 2008;452(7188):745-749.

50. Philippe H, Derelle R, Lopez P, et al. Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Curr Biol*. 2009;19(8):706-712.

51. Regier JC, Shultz JW, Zwick A, et al. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*. 2010;463(7284):1079-1083.

52. Smith SA, Wilson NG, Goetz FE, et al. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*. 2011;480(7377):364-367.

53. Kocot KM, Cannon JT, Todt C, et al. Phylogenomics reveals deep molluscan relationships. *Nature*. 2011;477(7365):452-456. doi:10.1038/nature10382.

54. Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*. 2001;17(12):1246-1247.

55. Robinson DF, Foulds LR. *Comparison of Phylogenetic Trees*.; 1981:131-147.

56. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol*. 2006;55(4):539-552. doi:10.1080/10635150600755453.

SUPPLEMENTARY MATERIAL

Manual for calculating Internode Certainty (IC), Internode Certainty All (ICA), Tree Certainty (TC), and Tree Certainty All (TCA) in RAxML [Provided as Supplementary Text File to: Salichos, L., A. Stamatakis, and A. Rokas (2013). Novel Information Theory-Based Metrics for Quantifying Incongruence among Phylogenetic Trees. *Manuscript under review*]

Disclaimers

Score calculations of the IC, ICA, TC, and TCA metrics have been implemented in the widely-used program RAxML (version 7.7.8, available via https://github.com/stamatak/standard-RAxML) (Stamatakis 2006). RAxML users are strongly encouraged to always check for and use the latest RAxML version on GITHUB. User support is provided via the following Google group: https://groups.google.com/forum/?hl=de#!forum/raxml. Users should avoid contacting the authors directly with inquiries about the code, but to post their question on the RAxML Google group. Users are encouraged to examine past answers to questions, which can be easily searched via keywords.

Users of the IC, ICA, TC, and TCA metrics are kindly requested to cite the following papers when using them:

Salichos, L., and A. Rokas (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497: 327-331

Stamatakis, A. (2006) RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics 22: 2688-2690

Salichos, L., A. Stamatakis, and A. Rokas (2013). Novel Information Theory-Based Metrics for Quantifying Incongruence among Phylogenetic Trees. ***Manuscript under review***

Manual

The implementation of the IC, ICA, TC, and TCA metrics relies on the efficient data structures that are already available in RAxML for performing calculations on tree bipartitions/splits [2].

Given a set of gene trees, RAxML can directly calculate a majority rule consensus (MRC; MR in RAxML terminology) as well as an extended MRC tree (MRE in RAxML terminology) on this set that has every internode (that is, internal branch) annotated by their respective IC and ICA scores. For instance, to compute the IC, ICA, TC, and TCA scores for a given set of gene trees on a MRC tree you would type:

```
./raxmlHPC -L MR -z 1070_yeast_genetrees.tre -m GTRCAT -n T1
```

where **-L MR** specifies that the scores will be displayed on the MRC tree computed by RAxML, **-z 1070_yeast_genetrees.tre** specifies the filename that contains the set of gene trees (which are the maximum likelihood trees from the 1,070 yeast genes analyzed by Salichos, and Rokas 2013, and which are provided as supplementary data to this manuscript), **-m GTRCAT** is an arbitrary substitution model (this will have no effect whatsoever, but is required as input to RAxML), and **-n T1** is the run ID

that is appended to output files. RAxML will automatically build the MRC tree, annotate it with the IC and ICA scores, and report both in an output file named **RAxML_MajorityRuleConsensusTree_IC.T1**, which will look like this:

(Scer,Spar,(Smik,(Skud,(Sbay,(Scas,(Cgla,(Kpol,(Zrou,((Clus,((Psti,((Ctro,(Calb,Cdub):1.0[0.95,0.95]):1.0[0.77,0.77],(Cpar,Lelo):1.0[0.76,0.76]):1.0[0.75,0.75]):1.0[0.11,0.11],(Cgui,Dhan):1.0[0.02,0.07]):1.0[0.02,0.08]):1.0[0.97,0.97],((Sklu,(Kwal,Kthe):1.0[0.97,0.97]):1.0[0.32,0.23],(Agos,Klac):1.0[0.08,0.08]):1.0[0.04,0.10]):1.0[0.59,0.47]):1.0[0.02,0.02]):1.0[0.11,0.11]):1.0[0.02,0.02]):1.0[0.97,0.97]):1.0[0.05,0.14]):1.0[0.30,0.27]):1.0[0.54,0.54]);

For each internode or internal branch of the constructed MRC tree, RAxML will assign an **length[x,y]**branch label, where **length** corresponds to the branch's length (because this is a MRC tree, all internal branch lengths have been arbitrarily set to 1.0 by default), **x** corresponds to the IC score and **y** to the ICA score.

RAxML will also calculate the TC and TCA scores for the MRC tree, as well as the relative TC and TCA scores that are normalized by the maximum possible TC and TCA scores for a fully bifurcating tree from the same number of taxa. The scores are displayed in the terminal output and in the **RAxML_info.runID** standard output file associated with the run (in this case **RAxML_info.T1**) and will look like this:

Tree certainty for this tree: 7.642240

Relative tree certainty for this tree: 0.382112

Tree certainty including all conflicting bipartitions (TCA) for this tree: 7.580023

Relative tree certainty including all conflicting bipartitions (TCA) for this tree: 0.379001

Given a set of gene trees, RAxML can also directly calculate an extended MRC tree on this set that has every internode (that is, internal branch) annotated by their respective IC and ICA scores. The particularly compute-intensive inference of extended MRC trees (finding the optimal extended MRC tree is, in fact, NP-hard; Phillips, and Warnow 1996) relies on RAxML's fast parallel implementation (presented in Aberer, Pattengale, and Stamatakis 2010). Thus if you use the PThreads version of RAxML, this part will run in parallel. To compute IC, ICA, TC and TCA scores on an extended MRC tree you would type:

```
./raxmlHPC -L MRE -z 1070_yeast_genetrees.tre -m GTRCAT -n T2
```

RAxML can compute MRC and extended MRC trees, using both fully bifurcating and partially resolved / multifurcating trees as an input. RAxML can also compute stricter MRC trees with arbitrary threshold settings that range between 51 and 100%. For instance, by typing

```
./raxmlHPC -L T_75 -z 1070_yeast_genetrees.tre -m GTRCAT -n T3
```

RAxML will display IC, ICA, TC and TCA scores on a MRC tree that only includes those bipartitions that have ≥ 75% support.

We have also implemented an option (**-f i**) that allows the user to calculate and display IC, ICA, TC and TCA scores onto a given, strictly bifurcating reference tree (for example, the best-known ML tree). This is analogous to the standard **-f b** option in RAxML that draws bootstrap support values from a set of bootstrap trees onto a reference phylogeny. The option can be invoked by typing

```
./raxmlHPC -f i -t yeast_concatenationtree.tre -z
1070_yeast_genetrees.tre -m GTRCAT -n T4
```

Note that, the tree contained in file **yeast_concatenationtree.tre** needs to be strictly bifurcating and contain branch lengths. In this example, the **yeast_concatenationtree.tre** file is the best-known maximum likelihood tree recovered by concatenation analysis of the 1,070 yeast genes (Salichos, and Rokas 2013). Using this command, RAxML will annotate the tree in **yeast_concatenationtree.tre** with the IC and ICA scores, and report both in an output file named **RAxML_IC_Score_BranchLabels.T4**, which will look like this:

```
(((((((Clus:0.47168135428609103688,((((Lelo:0.30356174702769450624,Cpa
r:0.25490874239480920682):0.13023178275857649755[0.76,0.76],(Ctro:0.18
383414558272206940,(Calb:0.04124660275465741321,Cdub:0.042908015883968
32289):0.14526604486383792869[0.95,0.95]):0.123558250286546558 73[0.77,
0.77]):0.17335821030783615804[0.75,0.75],Psti:0.42255112174261910685):
0.07862882822310976461[0.11,0.11],(Cgui:0.45961028886034632768,Dhan:0.
28259245937168109286):0.05586015476156453580[0.02,0.07]):0.08116340505
230199009[0.02,0.08]):1.03598510402913923656[0.97,0.97],((Agos:0.53332
956655591512440,Klac:0.47072785596320687596):0.081320063577044271 46[0.
08,0.08],((Kthe:0.17123899487739652203,Kwal:0.17320923240031221857):0.
```

2562011749511056701 9[0.97,0.97],Sklu:0.24833228915799765435):0.0564699 2617871094550[0.32,0.23]):0.052363061872351221 45[0.04,0.10]):0.1068651 7691208799463[0.59,0.47],Zrou:0.41307833685563782877):0.03792570537296 727218[0.02,0.02],Kpol:0.43287284049576529865):0.04560341693136910068[0.11,0.11],Cgla:0.49584136365135367264):0.04363310339731014259[0.02,0. 02],Scas:0.37212829744050218705):0.293621339962805150 14[0.97,0.97],(Sk ud:0.06926467973344750673,(Smik:0.06535810850036427588,(Scer:0.0428584 8856634000975,Spar:0.03030513540244994877):0.025067190660568425 96[0.54 ,0.54]):0.024593232915558628 50[0.30,0.27]):0.025242238670262769 07[0.05 ,0.14],Sbay:0.06506923220637816918);

For each internode or internal branch of this output tree RAxML will assign a **length[x,y]** branch label, where **length** corresponds to the branch's length, **x** corresponds to the IC score and **y** to the ICA score. RAxML will also display the TC and TCA scores of this tree both in the terminal output and in the **RAxML_info.T4** output file associated with the run.

It should further be noted that the IC and ICA scores are represented as branch labels, since, as is the case for bootstrap support values, information associated to splits/bipartitions of a tree always refers to branches and not nodes. Each tree viewer (e.g., Dendroscope; Huson, and Scornavacca 2012) that can properly parse the Newick tree format is able to display these branch labels. The rationale for not providing IC and ICA scores as node labels is that, some viewers may not properly rotate the node labels when the tree is re-rooted by the user, which will lead to an erroneous branch-IC/ICA-score association. When calculating IC and ICA scores on extended MRC trees or when drawing IC and ICA scores onto a given reference tree it may occur that the bipartition that has been included in the tree has lower support than one or more conflicting bipartitions. In this case, RAxML will report IC and ICA scores on the inferred tree with negative signs.

Finally, we have implemented a verbose output option that allows users to further scrutinize particularly interesting conflicting bipartitions. Verbose mode is activated by adding the **-C** command line switch to any of the above examples. In verbose mode RAxML will generate two types of output files: One set of files containing one included bipartition and the corresponding conflicting bipartitions in Newick format (called **RAxML_verboseIC.runID.0 … RAxML_verboseIC.runID.N-1**, where **N** is the number of bipartitions in the tree) and an output file that lists all bipartitions (included and conflicting) in a PHYLIP-like format (called **RAxML_verboseSplits.runID**).

For example, by adding **-C** to the previous command

```
./raxmlHPC -f i -t yeast_concatenationtree.tre -z
1070_yeast_genetrees.tre -m GTRCAT -n T5 -C
```

will produce 20 files (one for each of the 20 bipartitions present in the **yeast_concatenationtree.tre**) named **RAxML_verboseIC.T5.0**, **RAxML_verboseIC.T5.1**, …, **RAxML_verboseIC.T5.19**

For example, the **RAxML_verboseIC.T5.0** file will look like this:

```
((Cpar, Lelo),(Scer, Smik, Skud, Cgla, Kpol, Zrou, Kwal, Kthe, Agos,
Klac, Clus, Cgui, Psti, Ctro, Calb, Cdub, Dhan, Sklu, Scas, Sbay,
Spar));
((Cpar, Ctro, Calb, Cdub),(Scer, Smik, Skud, Cgla, Kpol, Zrou, Kwal,
Kthe, Agos, Klac, Clus, Cgui, Psti, Lelo, Dhan, Sklu, Scas, Sbay,
Spar));
```

where the first Newick string represents the bipartition that was included in the **yeast_concatenationtree.tre** and all following Newick strings represent the corresponding conflicting bipartitions in descending order of their frequency of occurrence. In the case of the **RAxML_verboseIC.T5.0** file the first bipartition, which is included in the **yeast_concatenationtree.tre** conflicts with only one other bipartition, which is listed as the second bipartition.

Analogously, the output file that lists all bipartitions (included and conflicting) in a PHYLIP-like format (**RAxML_verboseSplits.T5**), looks like this:

1. Scer

2. Smik

3. Skud

4. Cgla

5. Kpol

6. Zrou

7. Kwal

8. Kthe

9. Agos

10. Klac

11. Clus

12. Cgui

13. Psti

14. Cpar

15. Lelo

16. Ctro

17. Calb

18. Cdub

19. Dhan

20. Sklu

21. Scas

22. Sbay

23. Spar

partition:

----- ----- --- ----- --- 956/89.345794/0.761406**

----- ----- ---*- *-- --- 39/3.644860/0.761406**

partition:

----- ----- ----- --- --- 1051/98.224299/0.949483**

----- ----- ----- **--- --- 6/0.560748/0.949483

.

.

.

partition:

--* ***** ***** ***** **- 641/59.906542/0.303620**

--- ----- ----- ----- -*- 148/13.831776/0.303620**

-*- ***** ***** ***** *-- 114/10.654206/0.303620**

partition:

-** ***** ***** ***** **- 825/77.102804/0.545775**

```
-**-- ----- ----- ----- -**     87/8.130841/0.545775
```

Here each block that starts with the **`partition`** keyword contains a specific bipartition and all corresponding conflicting bipartitions in descending order. The **`x/y/z`** scores correspond to the frequency of the bipartition (**`x`**), the support percentage (also known as gene support frequency; **`y`**), and the IC score (**`z`**).

References

Aberer, AA, ND Pattengale, A Stamatakis. 2010. Parallelized phylogenetic post-analysis on multi-core architectures. Journal of Computational Science 1:107-114.

Huson, DH, C Scornavacca. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Systematic Biology 61:1061-1067.

Phillips, C, TJ Warnow. 1996. The asymmetric median tree - a new model for building consensus trees. Discrete Applied Mathematics 71:311-335.

Salichos, L, A Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327-331.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

CHAPTER IV

INFERRING ANCIENT DIVERGENCES REQUIRES GENES WITH STRONG PHYLOGENETIC SIGNALS

Leonidas Salichos and Antonis Rokas

Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA

This chapter is published in *NATURE, 497,* 327–331, 16 May 2013

ABSTRACT

To tackle incongruence, the topological conflict between different gene trees, phylogenomic studies couple concatenation with practices such as rogue taxon removal or the use of slowly evolving genes. Phylogenomic analysis of 1,070 orthologues from 23 yeast genomes identified 1,070 distinct gene trees, which were all incongruent with the phylogeny inferred from concatenation. Incongruence severity increased for shorter internodes located deeper in the phylogeny. Notably, whereas most practices had little or negative impact on the yeast phylogeny, the use of genes or internodes with high average internode support significantly improved the robustness of inference. We obtained similar results in analyses of vertebrate and metazoan phylogenomic data sets. These results question the exclusive reliance on concatenation and associated practices, and argue that selecting genes with strong phylogenetic signals and demonstrating the absence of significant incongruence are essential for accurately reconstructing ancient divergences.

INTRODUCTION

Concatenation, the compilation and analysis of hundreds of genes as a single dataset, has become the standard approach for inferring deep branches of the tree of life[1-5]. However, incongruence stemming from either analytical errors in gene history reconstruction[6,7] or the action of biological processes[8], evidenced by disagreements between phylogenomic studies[9-14], argues that the histories of some lineages are better depicted by or more closely resemble networks of highly related trees[15] and that concatenation might not be as robust as confidence indices indicate. To tackle incongruence, studies have adopted several practices, such as removing unstable taxa[1-3], which although useful are not always effective[16-18].

The *Saccharomyces* and *Candida* yeasts are excellent for examining phylogenomic practices in the presence of incongruence, due to the presence of conflicting gene trees[7,19], and the availability of two synteny databases[20,21] for genome-wide identification of high-quality orthologs, minimizing the risk of incongruence from hidden paralogy[22,23] and horizontal gene transfer[24]. Importantly, levels of sequence divergence between yeasts are intermediate to those observed between vertebrates and animals, making them an appropriate model for the study of ancient divergences.

Analyses on 1,070 genes from 23 yeast genomes showed that although concatenation resolved the species phylogeny, several internodes of the extended majority-rule consensus (eMRC) phylogeny of the 1,070 underlying gene trees (GTs) were weakly supported. None of the 1,070 GTs agreed with each other, with the concatenation phylogeny or with the eMRC phylogeny. By developing a novel measure to quantify the observed incongruence and evaluate standard practices aimed at reducing it, we found that such practices had little impact. In agreement with theory[9,16,25,26], incongruence was more severe for shorter internodes deeper on the phylogeny. Remarkably, the selection of genes whose bootstrap consensus trees had high average clade support, or of highly supported internodes, significantly reduced incongruence, arguing that inference in deep time critically depends on identifying molecular markers with strong phylogenetic signal.

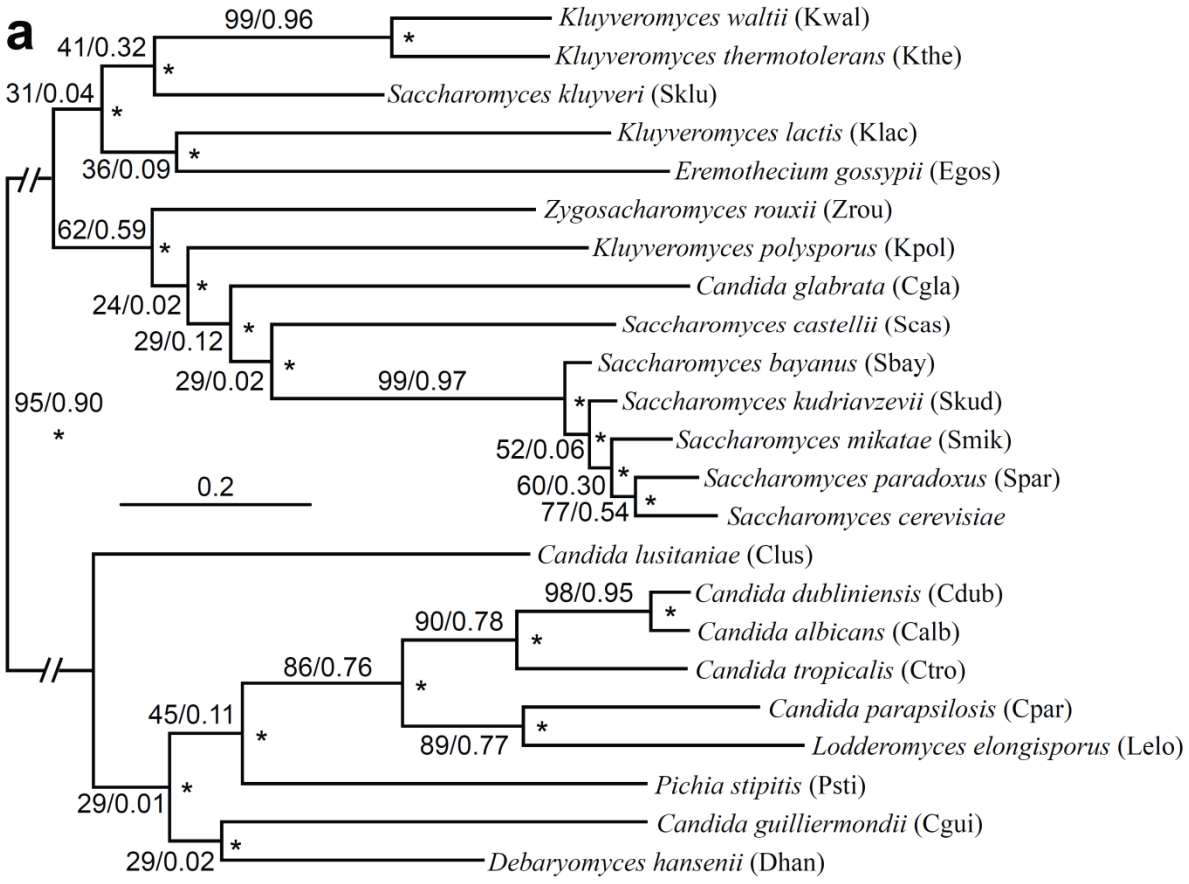
All Gene Trees Differ From Species Phylogeny

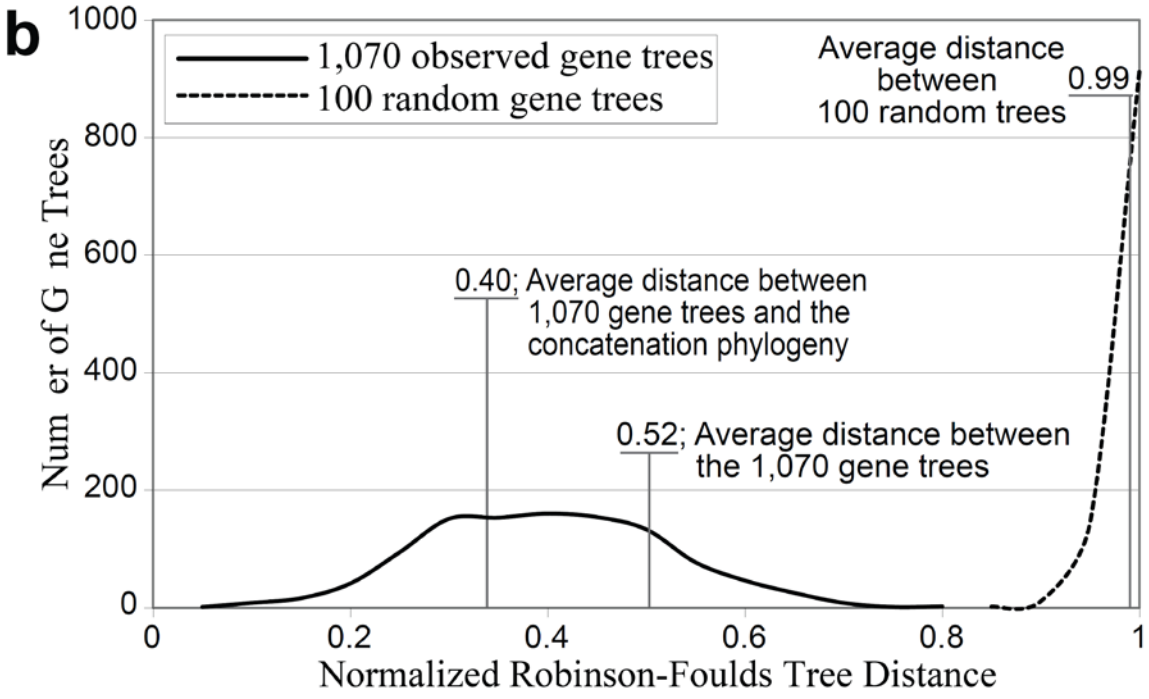
We assembled a dataset of 1,070 groups of orthologous genes (orthogroups) from 23 yeast genomes[20,21,27] (Methods and Supplementary Table 1). Maximum likelihood analysis of the concatenation of all 1,070 orthogroups yielded a species phylogeny where all 20 internodes

exhibited 100% bootstrap support (BS) (Fig. 4.1a); we obtained identical results using one other maximum likelihood and one other Bayesian inference software (Supplementary Fig. 4.1). Remarkably, all 1,070 GTs were topologically distinct and none matched the topology inferred by concatenation analysis (Fig. 4.1b). However, the average tree distance between the 1,070 GTs was much lower (normalized Robinson-Foulds tree distance[28] = 0.52, i.e., two GTs differed on average in 10.4 out of their 20 bipartitions) than that between randomly generated trees of the same taxon number (0.99, i.e., two trees differed on average on 19.8/20 bipartitions), indicating that the yeast GTs have similar evolutionary histories.

Summarizing the 1,070 GTs into an eMRC phylogeny yielded a topology identical with the concatenation phylogeny (Fig. 4.1a). However, although 11/20 internodes in the eMRC phylogeny had >50% gene support frequency (GSF), 5 of the remaining 9 internodes had GSF <30% (Fig. 4.1a). Furthermore, the most prevalent conflicts to most of these weakly supported internodes had substantial GSF values (Supplementary Table 2). Take, for example, the relative placement of *C. glabrata*, *S. castellii*, and the *Saccharomyces* sensu stricto clade where 5 uniquely shared chromosomal rearrangements and a substantially higher number of uniquely shared gene losses between *C. glabrata* and *S. cerevisiae* indicate that *S. castellii* divergence preceded that of *C. glabrata* from the *Saccharomyces* sensu stricto clade[22]. Even though concatenation provided 100% BS for the apparently incorrect grouping of *S. castellii* with the sensu stricto species (Fig. 4.1a), only 311/1,070 GTs (29%) favored it, whereas 214 (20%) inferred the *C. glabrata* – *Saccharomyces* sensu stricto one.

Figure 4.1 | The yeast species phylogeny recovered from the concatenation analysis of 1,070 genes disagrees with every single gene tree, despite absolute bootstrap support. **a,** The yeast species phylogeny recovered from concatenation analysis of 1,070 genes using maximum likelihood. Asterisks (*) denote internodes that received 100% bootstrap support by the concatenation analysis. Values near internodes correspond to gene support frequency and internode certainty, respectively. **b,** The distribution of the agreement between the bipartitions present in the 1,070 individual gene trees and the concatenation phylogeny, as well as the distribution of the agreement between the bipartitions present in 100 randomly generated trees of equal taxon number and the concatenation phylogeny, measured using the normalized Robinson-Foulds tree distance. Average distances between the 1,070 gene trees and the concatenation phylogeny, between the 1,070 gene trees with each other, and between 100 randomly generated gene trees of equal taxon number with each other, are also shown. The phylogeny of the 23 yeast species analyzed in this study is unrooted and contains 20 non-trivial bipartitions; because the divergence of Saccharomyces and Candida lineages is well established, the mid-point rooting of the phylogeny is shown for easier visualization.





A Novel Measure That Considers Incongruence

To quantify incongruence we developed Internode Certainty (IC), which evaluates support for a given internode by considering its frequency in a given set of trees jointly with that of the most prevalent conflicting bipartition in the same set of trees. Like phylogenetic network methods developed for visualizing phylogenetic conflicts[15], IC relies on the bipartitions present in trees, each of which is a split of the taxa into two mutually exclusive non-empty groups. Compared to other incongruence measures[29-32], IC is not character-based[29-31], it does not depend on an optimality criterion[29-31] or clade support metric[32], and can be applied to any set of trees. For example, if the entire set of GTs is used, the IC of a given internode will reflect the amount of information available for that internode in the set of GTs by considering the internode's GSF jointly with the GSF of the most prevalent bipartition that conflicts with the internode. If the set of bootstrap replicate trees for a given gene is used, then IC will be calculated based on BS

values. IC values near 0 indicate the presence of an almost equally supported bipartition that conflicts with the inferred internode, whereas values near 1 indicate the absence of conflict. Examination of the eMRC phylogeny showed 9/20 internodes with IC <0.3, which corresponds to a <4:1 ratio between the support for the inferred internode to that of its most prevalent conflicting bipartition, and 7/20 with IC <0.1 (<7:3 ratio) (Fig. 4.1a and Supplementary Fig. 4.2). Because IC measures the degree of conflict for every internode, it is more informative than GSF. For example, whereas the placement of *S. bayanus* and the placement of *Z. rouxii* received 52% and 62% GSF, their ICs were 0.06 and 0.59, respectively (Fig. 4.1a). This marked difference in IC values of the two internodes despite similar GSF values is because there was strong secondary signal only in the case of *S. bayanus*[33] (29% GSF for grouping *S. bayanus* with *S. kudriavzevii*), but not in the case of *Z. rouxii* (Supplementary Table 2). Furthermore, comparison of the sums of IC values across trees of a given taxon number (Tree Certainty; TC) can be used to quantify changes in the degree of incongruence between trees inferred using different datasets or methods.

Standard Practices Do Not Reduce Incongruence

To test whether we could decrease incongruence, we evaluated the effect of several standard phylogenomic practices purported to do so on the inference of the yeast phylogeny (Fig. 4.2). Specifically, we tested the effect of:

(1) removing sites containing gaps as well as of "rogue" genes producing alignments of bad quality (Supplementary Fig. 4.3),

(2) removing unstable and fast-evolving species (Supplementary Figs 4.5, 4.6, and 4.7),

(3) using only genes that recover a particular internode widely regarded as known or well established from prior data (Supplementary Figs 4.6 and 4.7),

(4) using only slowly evolving genes (Supplementary Fig. 4.8), and

(5) using conserved amino acid substitutions or indels (Supplementary Fig. 4.9).

Whereas the first three practices did not have a substantial effect on the inference and support of the yeast phylogeny, the use of slowly evolving genes and conserved sites increased incongruence across many internodes of the yeast phylogeny (Fig. 4.2). Furthermore, the removal of unstable or fast-evolving species from the *Saccharomyces* lineage had no effect on, often highly ambiguous, internodes in the *Candida* lineage and *vice versa* (Supplementary Figs 4.5 and 4.6), arguing that the impact of removing "rogue" taxa was not only minimal but also highly localized.

Figure 4.2 | The effect of phylogenomic practices on the inference of the yeast phylogeny. The first column (Treatment) indicates the specific phylogenomic practice tested, the second (avGSF) the average gene support frequency of the internodes of the yeast phylogeny, the third (TC) the tree certainty of the yeast phylogeny, the fourth (#↑ | ↓ GSF) the numbers of internodes of the yeast phylogeny where GSF increases or decreases by more than 3%, and the fifth (# ↑ | ↓ IC) the numbers of internodes of the yeast phylogeny where IC increases or decreases by more than 0.03. Because the maximum value of IC for a given internode is 1, the maximum value of TC for a given phylogeny is the number of internodes, which will equal K-3, where K is the number of taxa used. In the analyses concerned with the removal of poorly aligned genes, only genes whose alignment length after gap removal is ≥x% of original one were used. In the analyses concerned with the use of bipartitions, only those bipartitions that displayed BS ≥60%, ≥70%, or ≥80% in the bootstrap consensus trees of the 1,070 genes were used to construct eMRC phylogenies, which were then compared with the default analysis.

Treatment	Treatment details	Average GSF	Tree certainty	GSF increases	GSF decreases	IC increases	IC decreases
Removal of sites containing gaps	Default analysis	60.02	8.35	-	-	-	-
	All sites with gaps are excluded	58.17	7.91	0	5	0	7
	All sites with ≥50% gaps are excluded	60.04	8.23	0	0	1	2
Removal of poorly aligned genes	Default analysis (x = 50%; 1,070 genes)	60.00	8.35	-	-	-	-
	Poor alignments removed (x = 70%; 374 genes)	60.24	8.42	2	1	4	3
Removal of quickly evolving or unstable species	*C. lusitaniae* (unstable)	62.22	8.15	1	0	2	2
	S. castellii (unstable)	62.08	8.20	1	0	1	1
	K. polysporus (fast and unstable)	63.30	8.33	3	0	1	1
	E. gossypii (fast and unstable)	61.93	7.98	2	0	0	4
	C. glabrata (fast and unstable)	63.10	8.30	3	0	1	2
	K. lactis (fast and unstable)	61.86	7.99	2	1	0	3
	E. gossypii, K. lactis	63.91	7.88	1	1	0	3
	E. gossypii, C. glabrata, K. lactis	67.32	7.88	3	0	1	3
Selection of genes that recover specific bipartitions	(*C. glabrata, S. bayanus, S. kudriavzevii, S. mikatae, S. cerevisiae, S. paradoxus*)	65.88	9.47	4	1	6	3
	(*Z. rouxii, K. polysporus, C. glabrata, S. bayanus, S. castellii, S. kudriavzevii, S. mikatae, S. cerevisiae, S. paradoxus*)	63.34	8.62	3	0	0	4
	(*C. tropicalis, C. dubliniensis, C. albicans*)	61.20	8.62	1	0	0	0
Selection of the most slowly evolving genes	The 100 slowest evolving genes	52.20	6.76	1	10	2	9
Selection of genes whose bootstrap consensus trees have high average BS	Genes with average BS ≥ 60% (904 genes)	62.17	8.59	4	0	2	0
	Genes with average BS ≥ 70% (545 genes)	65.68	9.18	14	0	12	0
	Genes with average BS ≥ 80% (131 genes)	70.56	9.92	15	0	14	0
Selection of genes whose bootstrap consensus trees have high tree certainty	Using only the 904 genes with the highest TC	62.26	8.72	6	0	2	0
	Using only the 545 genes with the highest TC	66.06	9.37	13	0	12	0
	Using only the 131 genes with the highest TC	71.20	10.28	16	0	12	1
Selection of bipartitions with high BS in the bootstrap consensus trees of genes	Using only bipartitions that have ≥60% BS	NA	10.11	-	-	14	0
	Using only bipartitions that have ≥70% BS	NA	10.70	-	-	16	0
	Using only bipartitions that have ≥80% BS	NA	11.32	-	-	15	0

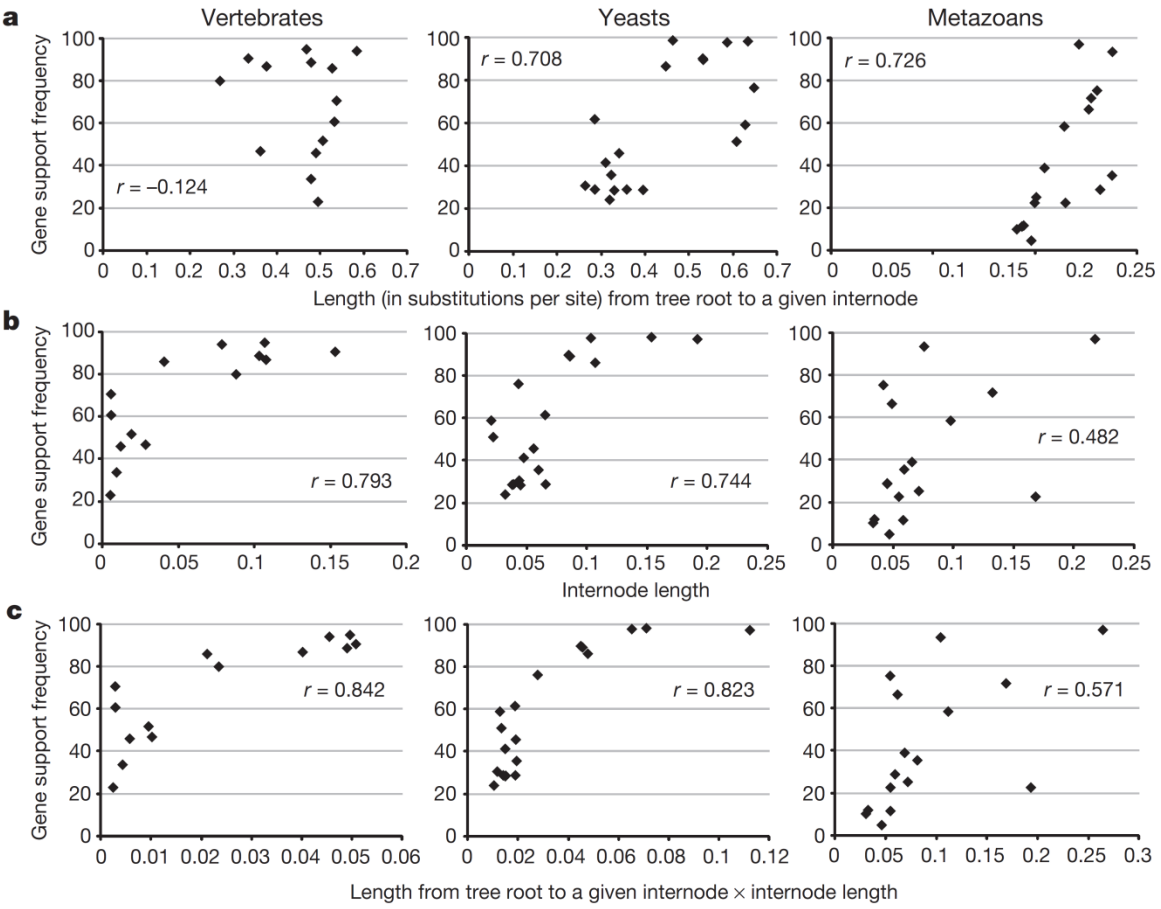
Average GSF	Tree certainty
48	5.18
50	5.69
52	6.20
54	6.71
56	7.22
58	7.73
60	8.24
62	8.75
64	9.26
66	9.77
68	10.28
70	10.79
72	11.30

Support depends on internode length and depth

Examination of whether the degree of incongruence, as measured by low GSF, correlated with internode length and depth, as measured by branch lengths, showed that incongruence was stronger in early divergent and short internodes (Fig. 4.3), in agreement with theoretical

expectations[9,16,25,26]. To test if this relationship holds in other lineages, we generated a dataset of 1,086 orthogroups from 18 vertebrate species, which has higher sequence similarity than the yeast one (61% vs 44% average pairwise aa similarity, respectively), and a dataset of 225 orthogroups from 21 metazoan species, which has lower sequence similarity (29% average pairwise aa similarity). The vertebrate genes yielded 299 distinct GTs (average normalized Robinson-Foulds tree distance = 0.42). Concatenation analysis inferred an absolutely supported species phylogeny; however, this phylogeny was topologically identical to 15 GTs and eMRC analysis showed that 4/15 internodes had GSF <50% and IC <0.3 (Supplementary Fig. 4.10a-c). Similarly, the 225 metazoan genes yielded 224 distinct GTs (average normalized Robinson-Foulds tree distance = 0.72). Concatenation analysis inferred 14/18 internodes with 100% BS despite that it was not topologically identical to any of the 225 GTs and that 10/18 internodes had <50% GSF and <0.1 IC (Supplementary Fig. 4.10d-f). Interestingly, incongruence was significantly correlated only with short internodes in the (less divergent) vertebrates, nearly equally significantly with both internode length and internode depth in yeasts, and more significantly with internode depth than with internode length in the (more divergent) metazoans (Fig. 4.3).

Figure 4.3 | Incongruence is more prevalent in shorter internodes located deeper on the phylogeny. The correlation (Pearson's r) between a measure of internode support (gene support frequency or GSF) with internode length and depth was measured for each internode present in three datasets that show lower (vertebrates, 1,086 genes), intermediate (yeasts, 1,070 genes) and higher (metazoans, 225 genes) levels of sequence divergence. a, GSF is positively correlated with internode length in yeasts and metazoans. b, GSF is positively correlated with the root to internode length in all three lineages, indicating that internodes placed deeper in the phylogeny typically have lower GSF. c, GSF is positively correlated with the product of internode length and root to internode length in all three lineages.



Strong Signal Reduces Incongruence

To test whether the selection of genes with stronger phylogenetic signal reduced incongruence, we analyzed three datasets comprised of genes whose bootstrap consensus trees showed average BS across all internodes ≥ 60% (904 genes), ≥ 70% (545 genes), or ≥ 80% (131 genes),and three datasets comprised of the 904, 545, or 131 genes whose bootstrap consensus trees had the

highest TC. Selecting genes with high average BS or high TC significantly reduced incongruence across many, but not all, internodes (Fig. 4.2, and Supplementary Figs 4.11 and 4.12). Concatenation analysis of the sets of genes with average BS ≥ 60%, and ≥ 70% (and of the 904 and the 545 genes with the highest TC) yielded the same species phylogeny as when all genes were analyzed. Remarkably, analysis of genes with average BS ≥ 80%, as well as of the 131 genes with the highest TC, yielded the correct placement of *C. glabrata* (Supplementary Fig. S4.11c,f), a result that, to our knowledge, has not been observed in any concatenation-based yeast phylogenomic analysis[7,34-37], suggesting that high BS is a good indicator of a gene's phylogenetic usefulness, but also that concatenating genes with high BS reduces incongruence and improves resolution.

We also tested whether selecting internodes with high BS decreased incongruence by extracting only those bipartitions that displayed BS values ≥ 60%, ≥ 70%, and ≥ 80% from every one of the 1,070 genes' bootstrap consensus trees and then using them to construct new eMRC phylogenies (Supplementary Figs 4.12 and 4.13). One advantage of working with taxon bipartitions, rather than genes, is that we can quantify a given internode's IC from only the subset of bipartitions that highly support or conflict with that internode. This practice significantly increased IC values for ≥ 14 internodes relative to the phylogeny of Figure 4.1a and showed the highest TC of all our analyses (Fig. 4.2). Interestingly, while IC for most internodes increased when we increased the BS threshold, this was not the case for several of the most difficult to resolve internodes (Supplementary Fig. 4.13d), suggesting that those few genes that show high BS for short internodes deep in the phylogeny strongly conflict with each other. We obtained similar results when we performed the same analyses on the vertebrate and metazoan datasets (Supplementary Fig. 4.14).

Standard Practices Can Mislead

Aiming to infer the yeast phylogeny, we constructed and analyzed 1,070 yeast genes. Had we relied solely on concatenation and standard phylogenomic practices we would have recovered an absolutely supported phylogeny similar to those obtained by major phylogenomic studies[1,3-5,11,12,16,19]. However, examination of the signal in GTs showed that concatenation masked the considerable incongruence present in several internodes. Thus, while analyses of ~20% of the genes typically present in a yeast genome definitively support many internodes of the yeast phylogeny, the topology of a considerable number of others remains uncertain (Supplementary Figs 4.15 and 4.16).

Our finding that incongruence correlates with early divergent and short internodes indicates that analytical factors are major contributors; however, it is likely that biological factors have also contributed. "Species tree" methods use coalescent theory to estimate the species phylogeny from the individual GTs allowing for lineage sorting, a common biological explanation for GTs incongruent with the species phylogeny[8]. Unfortunately, many such methods assume that analytical errors in inference are minimal, a valid assumption for most shallow clades but one that is untenable for the deeply divergent clades of the yeast phylogeny. For example, analysis of our dataset with the average unit-ranking method[38] yielded a species phylogeny where all the internodes with very low GSF and IC values were extremely short, largely because all incongruence was considered to be due to variation in coalescent depth across GTs (Supplementary Fig. 4.17a). Not surprisingly, these coalescent unit-based branch lengths were highly correlated with internodes' GSF and IC values (Supplementary Fig. 4.17b). Furthermore, bootstrapping of this dataset inferred a highly supported species phylogeny (Supplementary Fig. 4.17a), again contradicting our findings of extensive conflict in certain internodes.

PERSPECTIVE

These results argue that elimination of the observed incongruence between phylogenomic studies[1,3,4,11,12] will require three fundamental revisions to current practices. First, we should abandon using BS on concatenation analyses of large datasets. Developed at a time when high-throughput sequencing was unimaginable, the bootstrap is an extremely useful measure of sampling error, that is the robustness in inference when data are limited[39], such as when a single gene is analyzed. Given the availability and ease of generating genome-scale data[40], relying on bootstrap to analyze phylogenomic datasets is misleading, not only because sampling error is minimal but also because its application will, even in the presence of significant conflict[9] or systematic error[6,16], almost always result in 100% values[9,19,41].

The second critical revision necessary is that we carefully examine the signal present in individual genes[16,29-32,42] and their trees[15]. Our results indicate that the subset of genes with strong phylogenetic signal is more informative than the whole, arguing for a conditional combination approach than a total evidence one[42]. Preferably, such analyses should be combined with internode-specific approaches[31] because the latter can uncover internodes that harbor multiple conflicting phylogenetic signals. As the IC measure shows (Supplementary Fig. 4.2), the amount of information for a given internode supported by 50% of GTs with the other 50% being uninformative is far greater from that when the other 50% of the GTs harbors significant support for two or three alternative conflicting topologies. Whereas in the first case the gene trees strongly suggest that the internode is resolved, in the second there is reason to be cautious. Finally, we need to begin explicitly identifying internodes that, despite the use of genome-scale datasets, robust study designs, and powerful algorithms, are poorly supported. We argue that the on-going debate around phylogenies inferred in different phylogenomic studies[10] concerns

internodes that are poorly supported by individual GTs. Identifying these internodes and distinguishing them from ones supported by a significant fraction of genes and lack conflicts will go way beyond helping pinpoint challenging internodes, allowing us to identify the broad contours of the network of highly related gene histories that is the tree of life. Perhaps most importantly, it will focus the attention of researchers to develop novel phylogenomic approaches and markers to more accurately decipher the most challenging ancient branches of life's genealogy from the DNA record.

METHODS SUMMARY

Using synteny and orthology information present in the YGOB[20] and CGOB[21] databases from 23 yeast genomes[20,21,27], we constructed an initial dataset of *2,651 orthogroups, which following quality control (see Methods), was reduced to the final 1,070. We also used the complete gene sets from 18 vertebrate and 21 metazoan species and used the cRBH algorithm[23] to identify 1,086 vertebrate and 225 metazoan orthologous groups of genes. Orthogroups were aligned using MAFFT[43], the best fit evolutionary model was inferred using ProtTest[44], and the maximum likelihood tree was estimated using RAxML[45]. Extended majority rule consensus trees were inferred using PHYLIP[46] and custom perl scripts. A series of different datasets were constructed using custom perl scripts.* Internode Certainty (IC), was *calculated according to:*

$$IC = log_2(2) + p\left(\frac{x_1}{x_1 + x_2}\right) log_2\left(p\left(\frac{x_1}{x_1 + x_2}\right)\right) + p\left(\frac{x_2}{x_1 + x_2}\right) log_2\left(p\left(\frac{x_2}{x_1 + x_2}\right)\right)$$

where x_1 and x_2 are the frequencies of the first and second most prevalent conflicting bipartitions for a given internode.

METHODS

Data Matrix Construction

We used the complete sets of annotated genes from 23 yeast genomes[20,21,27,47] (Supplementary Table S1) and, using the synteny and orthology information present in the YGOB[20] and CGOB[21] databases, we constructed an initial dataset of 2,651 orthologous groups of genes that had representatives in all 23 genomes. This reliance on two highly accurate and manually curated synteny databases and the requirement for a given ortholog to be present in all 23 species greatly minimized errors in orthology inference due to hidden paralogy[23,48]. It also avoided the inclusion of any horizontally transferred genes present in some, but not all, species as well as any horizontally transferred genes present in regions that lack synteny conservation. For any potentially horizontally transferred gene to be included in our data matrix, it would have had to have been gained in some, but not all, yeast species used in our study and it would have had to replace the native gene and take up its position on the chromosome, which has never been observed in yeasts[24,49-51] and is likely very rare.

The nucleotide sequences of all genes were translated to amino acids (aa) taking into account that in certain species in the *Candida* lineage the CUG codon encodes for the amino acid Serine rather than Leucine. Using alignment quality and individual gene length filtering criteria described below, we then reduced the number of orthogroups to the final 1,070. Examination of the functional annotation–as defined by the Gene Ontology consortium[52]–of the 1,070 *S. cerevisiae* orthologs using the GOstat software[53] showed that this gene set is statistically overrepresented for several different functional categories, such as cellular metabolic process, cellular component organization and biogenesis, and ribosome assembly and biogenesis, in other words, for categories associated with standard cell housekeeping functions. Analysis of different

ortholog subsets (e.g., of the 131 genes whose bootstrap consensus trees show the highest average bootstrap support (BS)) show that they are too statistically overrepresented for many fewer, but the same, functions.

We also created two additional datasets from the complete sets of annotated genes from 18 vertebrate and 21 metazoan species (Supplementary Table S1). The two datasets were constructed using the cRBH algorithm[23], and comprised of 1,086 vertebrate and 225 metazoan orthologous groups of genes. To avoid constructing orthogroups that contained very distant homologs we set the filtering parameter of the cRBH algorithm[23], which considers the degree by which the two proteins differed in sequence length or BLAST alignment, *to r* = 0.3.

For each species, for reasons of space and convenience, we constructed a corresponding acronym using the first letter from the genus name and the three first letters from the species name (e.g., the acronym for *Saccharomyces cerevisiae* is "Scer"). All data matrices are available from the authors upon request.

Gene Alignment and Filtering Criteria

To minimize the use of orthogroups that contained sequences whose annotation was problematic or which resulted in alignments of low quality, we applied various filtering criteria. We first excluded, prior to alignment, all orthogroups with an average sequence length ≤ 150 amino acids. Second, we aligned all orthogroups using the MAFFT software[43], *with the default settings*, and excluded orthogroups whose alignment after removing all positions that contained gaps was $\leq 50\%$ of the original alignment length.

Gene Tree Inference

For each orthogroup, the best fit evolutionary model, which typically consisted of an empirically-determined aa substitution matrix (e.g., WAG[54]), empirically-measured aa state frequencies, and accounted for heterogeneity in evolutionary rates among sites by using the gamma distribution as well as by allowing for a given proportion of sites to be invariable, was selected using ProtTest[44]. The unrooted phylogenetic tree of each and every orthogroup, also called gene tree (GT), was then inferred using *RAxML*[45].

Species Phylogeny Inference Using Concatenation and Extended Majority-Rule Consensus Approaches

For the concatenation analysis, orthogroup alignments were analyzed as a single supermatrix. An unrooted concatenation species phylogeny was then inferred under the "PROTGAMMAIWAGF" model of aa substitution in *RAxML*[45]*, and confirmed with GARLI*[55] *as well as with MrBayes*[56]*. The unrooted extended majority rule consensus (eMRC) phylogeny that consisted of those bipartitions* that appear in more than half of the *maximum likelihood estimated GTs, as well as of additional compatible bi*partitions that appear in less than half of the GTs[57,58]*, was inferred using the CONSENSE program in PHYLIP*[46]*. The eMRC phylogeny of bipartitions with high BS was constructed using custom perl scripts.* Because the divergence of *Saccharomyces* and *Candida* lineages is well established, all phylogenies shown in figures have been mid-point rooted at the internode that separates these two lineages for easier visualization.

Species Phylogeny Inference Using a Consensus Phylogenetic Network Approach

A consensus phylogenetic network was constructed based on the 1,070 GTs estimated by maximum likelihood using the median network construction algorithm in the SplitsTree4 software[15] with a threshold of 0.1.

Tree Distance Estimation

Distances between trees were estimated using the normalized Robinson-Foulds tree distance[28], as calculated by RAxML[45]. Sets of 100 random trees for 23 taxa (yeasts), 18 taxa (vertebrates), and 21 taxa (metazoans), were generated using the random tree generator in the T-REX webserver[59], using the random tree generation procedure described by Kuhner and Felsenstein[60].

Internode Certainty (IC)

A phylogenetic tree is an acyclic connected graph that represents evolutionary relationships among different genes or taxa and consists of nodes that are connected by edges or internodes. Phylogenetic trees can also be represented in a variety of other ways. One useful depiction is as sets of bipartitions (or splits). In this representation, each internode in a phylogenetic tree is viewed as a bipartition between two sets of taxa. For example, given a set of five species (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus*), one example of a bipartition is the one that separates the set of *S. cerevisiae*, *S. paradoxus*, and *S. mikatae* from the set of *S. kudriavzevii* and *S. bayanus*.

Information from multiple phylogenetic trees from the same set of taxa is typically summarized using consensus trees. For example, the majority-rule consensus approach[57] calculates the shared bipartitions across all phylogenetic trees and displays only those shared by their majority. Consequently, each internode in the majority-rule consensus tree typically contains a value that

corresponds to the percentage of individual trees that contain a given bipartition, but does not provide any information about the next most prevalent conflicting bipartition, or more generally, about the distribution of bipartitions that conflict with the internode. For example, if a consensus tree reports that 51 out of 100 phylogenetic trees contain a specific bipartition, we are not informed whether the second most prevalent conflicting bipartition is present in the remaining 49 trees or in 5 of the remaining trees. However, the first case (51% vs 49%) would indicate that both bipartitions have nearly equal support, whereas the second case (51% vs 5%) would indicate that the first bipartition is the only strongly supported bipartition for this internode. Consensus phylogenetic networks[15,61], which are potentially hyperdimensional graphs inferred from all bipartitions present above a certain frequency in a given set of trees, are very useful in visualizing such conflicting bipartitions. To quantify the degree of incongruence, as well as examine whether incongruence is reduced when standard phylogenomic practices are applied, we developed internode certainty (IC), a measure that provides robust quantitative measures of the information conveyed by conflicting bipartitions for each internode.

Description of IC. Shannon's entropy measures the amount of certainty found in a random variable[62]. For example, when tossing a fair coin, heads or tails are equally probable and so the amount of certainty we have about the outcome is 0, whereas if the coin is not fair, our certainty about the toss outcome will be high. Similarly, we can quantify the certainty that we have in the inference of a given internode in a phylogenetic tree, by introducing a function that is maximized in the absence of any conflicting bipartitions, but is minimized in the presence of equally prevalent conflicting bipartitions. IC quantifies the certainty of a bipartition that appears on a phylogenetic tree (i.e., of a given internode) by considering its frequency of occurrence against

that of the second most prevalent conflicting bipartition. Specifically, for the two most prevalent conflicting bipartitions:

$$IC = log_2(2) + p\left(\frac{x_1}{x_1 + x_2}\right) log_2\left(p\left(\frac{x_1}{x_1 + x_2}\right)\right) + p\left(\frac{x_2}{x_1 + x_2}\right) log_2\left(p\left(\frac{x_2}{x_1 + x_2}\right)\right)$$

where x_1 and x_2 are the frequencies of the first and second most prevalent conflicting bipartitions for a given internode.

IC, as well as the related measure Tree Certainty (see below), can be measured on any given set of trees. For example, if the entire set of GTs is used, the IC value of a given internode will reflect the amount of information available for that internode in the set of GTs by considering the internode's gene support frequency (GSF) jointly with the GSF of the most prevalent bipartition that conflicts with the internode. If the set of bootstrap replicate trees for a given gene is used, then IC will be calculated based on BS values (instead of GSF values). IC can also be measured on any given set of bipartitions. For example, any two-state character that is variable across *x* species can be thought of as a bipartition, as it splits the set of taxa into two distinct groups. Thus, one can use IC to measure the amount of information available for a given bipartition, and quantify the extent of incongruence, by considering the number of characters supporting that bipartition jointly with the number of characters supporting the most prevalent bipartition that conflicts with the internode.

Example #1. Let us assume that there are four prevalent conflicting bipartitions with frequencies of 40%, 10%, 10% and 10%, respectively for a given internode. In this case,

$$\mathrm{IC} = 1 + \frac{40}{40+10} log_2\left(\frac{40}{40+10}\right) + \frac{10}{40+10} log_2\left(\frac{10}{40+10}\right) \approx 0.28$$

Example #2. Let us assume that there are four prevalent conflicting bipartitions with frequencies of 40%, 40%, 10% and 10%, respectively for a given internode. In this case,

$$\mathrm{IC} = 1 + \frac{40}{40+40} log_2\left(\frac{40}{40+40}\right) + \frac{40}{40+40} log_2\left(\frac{40}{40+40}\right) = 0.00$$

Tree Certainty (TC). We define Tree Certainty (TC) as the sum of all IC values across all internodes of a phylogenetic tree.

Evaluation of Phylogenomic Practices

Removing positions or genes with gaps. We used custom perl scripts to modify our default alignments by removing sites that contained either ≥50% gaps or any gap. We also tested whether the removal of genes producing alignments of bad quality by filtering genes whose alignment length after removal of all gap-containing sites was ≤70% of the original alignment length (instead of the ≤50% threshold used in the default analysis).

*Removing species from dataset***.** We removed several different unstable and fast-evolving species from the default dataset, singly and in combination. After each removal, the new orthogroups were re-aligned, a new best-fit evolutionary model was identified, and the phylogenetic analysis was performed again with the new alignment and model.

*Selection of genes that recover specific bipartitions***.** For the 100 hundred bootstrap replicate trees constructed from each gene, we used the CONSENSE program in the PHYLIP package to generate the bootstrap consensus tree as well as its bipartitions. Using custom perl scripts, we then extracted all genes that supported the three following bipartitions: (1) [*C. albicans, C. dubliniensis, C. tropicalis*], (2) [*C. glabrata, K. polysporus, S. bayanus, S. castellii, S. cerevisiae, S. kudriavzevii, S. mikatae, S. paradoxus, Z. rouxii*], and (3) [*C. glabrata, S. bayanus, S. cerevisiae, S. kudriavzevii, S. mikatae, S. paradoxus*]. We then used the selected genes and their GTs to infer a species phylogeny using concatenation and eMRC analysis.

*Selecting slow-evolving genes***.** The 100 slowest-evolving genes were identified by calculating the 100 genes whose GTs had the smallest sum of branch lengths.

Selecting single rare but conserved aa substitutions or indels. To reduce the effect of homoplasy for early divergent internodes, many studies have suggested the use of rare substitution types[63] as well as insertions or deletions (indels)[64]. We constructed three datasets by extracting all sites from our 1,070 gene alignments that contained (1) a single radical aa substitution (defined as a substitution with a blosum62 matrix score ≤ -3) (20,289 sites), (2) a single substitution between aa that differ radically in their physicochemical properties[63] (4,075 sites), or (3) a single indel that spans 7 or more aa (2,474 sites). The presence of any of these three types of sites instantly parts a set of x species into two groups of taxa or, equivalently, into two bipartitions ($0_1 \ldots 0_m$ and $1_1 \ldots 1_n$), where $m \geq 2$ species contain the "0" character state, $n \geq 2$ species contain the "1" character state, and $m + n = x$. To quantify the extent of incongruence of each type of site on a given internode we used IC to measure the amount of information available for that internode by

considering the number of characters supporting that internode jointly with the number of characters supporting the most prevalent bipartition that conflicts with the internode.

Selecting genes with high average BS or high TC. For every gene from the default dataset, we estimated the average BS value of all 20 internodes of its bootstrap consensus tree. We also used the set of bootstrap replicate trees for every gene to calculate the IC value of every internode in its bootstrap consensus tree. Thus calculated, the IC value reflects the amount of information available for that internode in the set of bootstrap replicate trees because it considers the internode's BS jointly with the BS of the most prevalent bipartition that conflicts with the internode. We then calculated the TC value for each gene by summing the IC values of all internodes in its bootstrap consensus tree. Finally, we used these average BS and TC values to construct six subsets of orthogroups: three with genes having average BS ≥ 60% (904 genes), ≥ 70% (545 genes) and ≥ 80% (131 genes), as well as three datasets of the 904, 545, and 131 genes with the highest TC.

Selecting bipartitions with high BS. For every gene from the default dataset, we extracted all bipartitions from its bootstrap consensus tree that had BS ≥60%, ≥70% and ≥80%. We then used each one of these three sets of highly supported bipartitions to construct eMRC species phylogenies with custom perl scripts.

Estimating root–to-node and internode length. We calculated the *root-to-node* length as the sum of all branch lengths from the midpoint of the rooted concatenation species phylogeny to the

focal node. As *internode* length, we considered the branch length of the internode leading to the focal node.

ACKNOWLEDGMENTS

We thank Ken Polzin for providing a script that identified alignment sites that contained single substitutions between amino acids that differ in their physicochemical properties. We thank members of the Rokas lab and Brian O'Meara for valuable comments on this work. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. This work was supported by the National Science Foundation (DEB-0844968).

Author Contribution Statement

Author Contributions L.S. and A.R. conceived and designed experiments; L.S. performed experiments; L.S. and A.R. analyzed data and wrote the paper.

REFERENCES

1. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745-749 (2008).
2. Rokas, A., Kruger, D. & Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. *Science* **310**, 1933-1938 (2005).
3. Philippe, H. *et al.* Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **19**, 706-712 (2009).
4. Schierwater, B. *et al.* Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. *PLoS Biol.* **7**, e20 (2009).
5. Regier, J. C. *et al.* Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**, 1079-1083 (2010).
6. Phillips, M. J., Delsuc, F. D. & Penny, D. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **21**, 1455-1458 (2004).
7. Hess, J. & Goldman, N. Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: yeasts revisited. *PLoS One* **6**, e22783 (2011).

8. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332-340 (2009).
9. Rokas, A. & Carroll, S. B. Bushes in the tree of life. *PLoS Biol.* **4**, e352 (2006).
10. Philippe, H. *et al.* Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* **9**, e1000602 (2011).
11. Kocot, K. M. *et al.* Phylogenomics reveals deep molluscan relationships. *Nature* **477**, 452-456 (2011).
12. Smith, S. A. *et al.* Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* **480**, 364-367 (2011).
13. Bourlat, S. J. *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444**, 85-88 (2006).
14. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965-968 (2006).
15. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
16. Regier, J. C. *et al.* Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst Biol* **57**, 920-938 (2008).
17. Regier, J. C. & Zwick, A. Sources of signal in 62 protein-coding nuclear genes for higher-level phylogenetics of arthropods. *PLoS One* **6**, e23408 (2011).
18. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-577 (2007).
19. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798-804 (2003).
20. Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15**, 1456-1461 (2005).
21. Fitzpatrick, D. A., O'Gaora, P., Byrne, K. P. & Butler, G. Analysis of gene evolution and metabolic pathways using the *Candida* Gene Order Browser. *BMC Genomics* **11**, 290 (2010).
22. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341-345 (2006).
23. Salichos, L. & Rokas, A. Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One* **6**, e18755 (2011).
24. Slot, J. C. & Rokas, A. Multiple *GAL* pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc. Natl. Acad. Sci. USA* **107**, 10136-10141 (2010).
25. Mossel, E. & Steel, M. A phase transition for a random cluster model on phylogenetic trees. *Math Biosci* **187**, 189-203 (2004).
26. Townsend, J. P., Su, Z. & Tekle, Y. I. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. *Systematic Biol* **61**, 835-849 (2012).
27. Scannell, D. R. *et al.* The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces sensu stricto* genus. *G3* **1**, 11-25 (2011).

28. Robinson, D. R. & Foulds, L. R. Comparison of phylogenetic trees. *Math Biosci* **53**, 131-147 (1981).
29. Farris, J. S., Kallersjo, M., Kluge, A. G. & Bult, C. Testing significance of incongruence. *Cladistics* **10**, 315-319 (1994).
30. Templeton, A. R. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. *Evolution* **37**, 221-244 (1983).
31. Baker, R. H. & DeSalle, R. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Systematic Biol* **46**, 654-673 (1997).
32. Rodrigo, A. G., Kelly-Borges, M., Bergquist, P. G. & Bergquist, P. L. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *New Zeal J Bot* **31**, 257-268 (1993).
33. Yu, Y., Degnan, J. H. & Nakhleh, L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.* **8**, e1002660 (2012).
34. Hittinger, C. T., Rokas, A. & Carroll, S. B. Parallel inactivation of multiple *GAL* pathway genes and ecological diversification in yeasts. *Proc. Natl. Acad. Sci. USA* **101**, 14144-14149 (2004).
35. Rokas, A. & Carroll, S. B. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol. Biol. Evol.* **22**, 1337-1344 (2005).
36. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**, 225-231 (2006).
37. Fitzpatrick, D. A., Logue, M. E., Stajich, J. E. & Butler, G. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.* **6**, 99 (2006).
38. Liu, L., Yu, L., Pearl, D. K. & Edwards, S. V. Estimating species phylogenies using coalescence times among sequences. *Systematic Biol* **58**, 468-477 (2009).
39. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791 (1985).
40. Hittinger, C. T., Johnston, M., Tossberg, J. T. & Rokas, A. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. *Proc. Natl. Acad. Sci. USA* **107**, 1476-1481 (2010).
41. Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L. & Tamura, K. Statistics and truth in phylogenomics. *Mol. Biol. Evol.* **29**, 457-472 (2012).
42. Cunningham, C. W. Can three incongruence tests predict when data should be combined? *Mol. Biol. Evol.* **14**, 733-740 (1997).
43. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286-298 (2008).
44. Abascal, F., Zardoya, R. & Posada, D. Prottest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105 (2005).
45. Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).
46. PHYLIP (Phylogeny Inference Package) (Distributed by the Author, Department of Genetics, University of Washington, Seattle, 1993).
47. Dujon, B. Yeast evolutionary genomics. *Nat. Rev. Genet.* **11**, 512-524 (2010).

48. Scannell, D. R., Butler, G. & Wolfe, K. H. Yeast genome evolution-the origin of the species. *Yeast* **24**, 929-942 (2007).
49. Hall, C., Brachat, S. & Dietrich, F. S. Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot. Cell* **4**, 1102-1115 (2005).
50. League, G. P., Slot, J. C. & Rokas, A. The *ASP3* locus in *Saccharomyces cerevisiae* originated by horizontal gene transfer from *Wickerhamomyces*. *FEMS Yeast Res.* **12**, 859-863 (2012).
51. Novo, M. *et al.* Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proc Natl Acad Sci U S A* **106**, 16333-16338 (2009).
52. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25-29 (2000).
53. Beissbarth, T. & Speed, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464-1465 (2004).
54. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691-699 (2001).
55. Zwickl, D. J. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion* Ph.D. thesis, The University of Texas at Austin, (2006).
56. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).
57. Bryant, D.A classification of consensus methods for phylogenetics. in *Bioconsensus* (eds M. Janowitz *et al.*) 163-184 (2003).
58. Felsenstein, J. *Inferring Phylogenies*. (Sinauer, 2003).
59. Alix, B., Boubacar, D. A. & Vladimir, M. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res.* **40**, W573-579 (2012).
60. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459-468 (1994).
61. Holland, B. R., Huber, K. T., Moulton, V. & Lockhart, P. J. Using consensus networks to visualize contradictory evidence for species phylogeny. *Mol. Biol. Evol.* **21**, 1459-1461 (2004).
62. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379-423 (1948).
63. Rogozin, I. B., Wolf, Y. I., Carmel, L. & Koonin, E. V. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. *Mol. Biol. Evol.* **24**, 1080-1090 (2007).
64. Belinky, F., Cohen, O. & Huchon, D. Large-scale parsimony analysis of metazoan indels in protein-coding genes. *Mol. Biol. Evol.* **27**, 441-451 (2010).

SUPPLEMENTARY FIGURES & TABLES

Supplementary Table 1. The Taxonomy of the Organisms Used in this Study

Organism (acronym)	Taxonomy
Yeasts	
Kluyveromyces waltii (Kwal)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Kluyveromyces thermotolerans (Kthe)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Saccharomyces kluyveri (Sklu)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Kluyveromyces lactis (Klac)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Eremothecium gossypii (Egos)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Zygosacharomyces rouxii (Zrou)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Kluyveromyces polysporus (Kpol)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Candida glabrata (Cgla)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Saccharomyces castellii (Scas)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Saccharomyces bayanus (Sbay)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Saccharomyces kudriavzevii (Skud)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Saccharomyces mikatae (Smik)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Saccharomyces paradoxus (Spar)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Saccharomyces cerevisiae (Scer)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Candida lusitaniae (Clus)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Candida dubliniensis (Cdub)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Candida albicans (Calb)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Candida tropicalis (Ctro)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Candida parapsilosis (Cpar)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Lodderomyces elongisporus (Lelo)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae;
Pichia stipitis (Psti)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Candida guilliermondii (Cgui)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Debaryomyces hansenii (Dhan)	Fungi;Ascomycota;Saccharomycetes;Saccharomycetaceae
Verterbrates	
Xenopus tropicalis (Xtro)	Animalia;Chordata;Aphibia;Anura;Pipidae
Gallus gallus (Ggal)	Animalia;Chordata;Aves;Galliformes;Phasianidae
Monodelphis domestica (Mdom)	Animalia;Chordata;Monodelphis;Mammalia;Didelphimorphia
Bos taurus (Btau)	Animalia;Chordata;Mammalia;Artiodactyla;Bovidae
Equus caballus (Ecab)	Animalia;Chordata;Mammalia;Perissodactyla;Equidae
Canis familiaris (Cfam)	Animalia;Chordata;Mammalia;Carnivora;Canidae
Macaca mulatta (Mmul)	Animalia;Chordata;Mammalia;Primates;Cercopithecidae
Pongo pygmaeus (Ppyg)	Animalia;Chordata;Mammalia;Primates;Hominidae
Homo sapiens (Hsap)	Animalia;Chordata;Mammalia;Primates;Hominidae
*Pan troglodytes (*Ptro)	Animalia;Chordata;Mammalia;Primates;Hominidae
Rattus norvegicus (Rnor)	Animalia;Chordata;Mammalia;Rodentia;Muridae
Mus musculus (Mmus)	Animalia;Chordata;Mammalia;Rodentia;Muridae

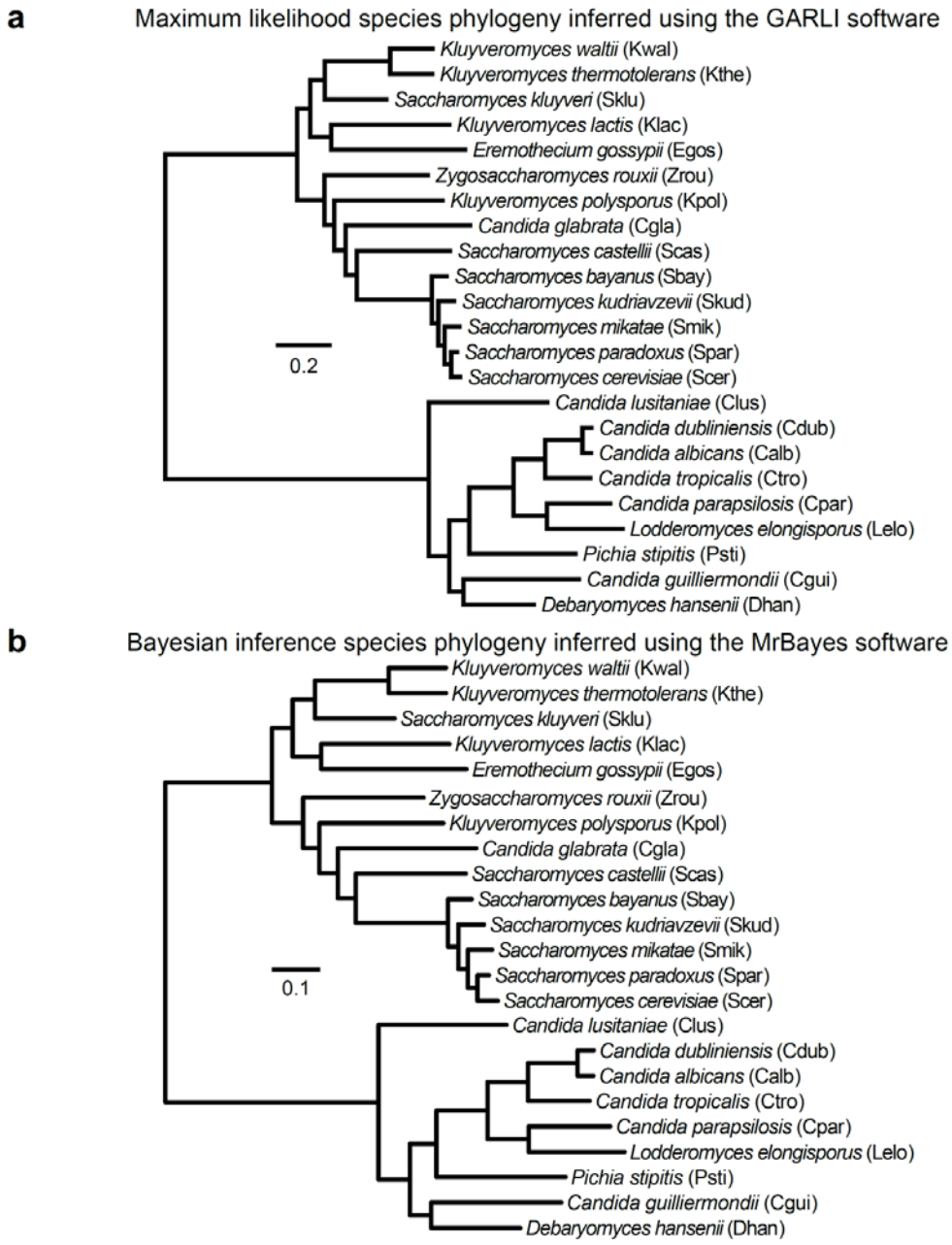
Cavia porcellus (Cpor)	Animalia;Chordata;Mammalia;Rodentia;Caviidae
Danio rerio (Drer)	Animalia;Chordata;Actinopterygii;Cypriniformes;Cyprinidae
Oryzias latipes (Olat)	Animalia;Chordata;Actinopterygii;Beloniformes;Adrianichthyidae
Tetraodon nigroviridis (Tnig)	Animalia;Chordata;Actinopterygii;Tetraodontiformes;Tetraodontidae
Takifugu rubripes (Trub)	Animalia;Chordata;Actinopterygii;Tetraodontiformes;Tetraodontidae
Gasterosteus aculeatus (Gacu)	Animalia;Chordata;Actinopterygii;Gasterosteiformes;Gasterosteidae
Metazoa	
Strongylocentrotus purpuratus (Spur)	Animalia;Echinodermata;Echinoidea;Echinoida;Strongylocentrotidae
Branchiostoma floridae (Bflo)	Animalia;Chordata;Leptocardii;Amphioxiformes;Branchiostomidae
Ciona intestinalis (Cint)	Animalia;Chordata;Ascidiaceae;Enterogona;Cionidae
Mus musculus (Mmus)	Animalia;Chordata;Mammalia;Rodentia;Muridae
Gallus gallus (Ggal)	Animalia;Chordata;Aves;Galliformes;Phasianidae
Homo sapiens (Hsap)	Animalia;Chordata;Mammalia;Primates;Hominidae
Xenopus tropicalis (Xtro)	Animalia;Chordata;Aphibia;Anura;Pipidae
Danio rerio (Drer)	Animalia;Chordata;Actinopterygii;Cypriniformes;Cyprinidae
Helobdella robusta (Hrob)	Animalia;Annelida;Clitellata;Rhynchobdellida;Glossiphoniidae
Lottia gigantea (Lgig)	Animalia;Mollusca;Gastropoda;Patellogastropoda;Lottiidae
Caenorhabditis elegans (Cele)	Animalia;Nematoda;Secernentea;Rhabditida;Rhabditidae
Schistosoma mansoni (Sman)	Animalia;Platyhelminthes;Digenea;Strigeidida
Ixodes scapularis (Isca)	Animalia;Arthropoda;Arachnida;Ixodida;Ixodidae
Daphnia pulex (Dpul)	Animalia;Arthropoda;Branchiopoda;Cladocera;Daphniidae
Apis mellifera (Amel)	Animalia;Arthropoda;Insecta;Hymenoptera;Apidae
Tribolium castaneum (Tcas)	Animalia;Arthropoda;Insecta;Coleoptera;Tenebrionidae
Drosophila melanogaster (Dmel)	Animalia;Arthropoda;Insecta;Diptera;Drosophilidae
Bombyx mori (Bmor)	Animalia;Arthropoda;Insecta;Lepidopteroa;Bombycidae
Monosiga brevicollis (Mbre)	hoanoflagellida;Codonosigidae
Nematostella vectensis (Nvec)	Animalia;Cnidaria;Anthozoa;Actiniaria;Edwardsiidae
Trichoplax adhaerens (Tadh)	Animalia;Placozoa;Tricoplacia;Tricoplaciformes;Trichoplacidae

Supplementary Table 2.Bipartitions that Significantly Conflict with the Bipartitions Recovered in the Concatenation Phylogeny of 23 Yeast Genomes

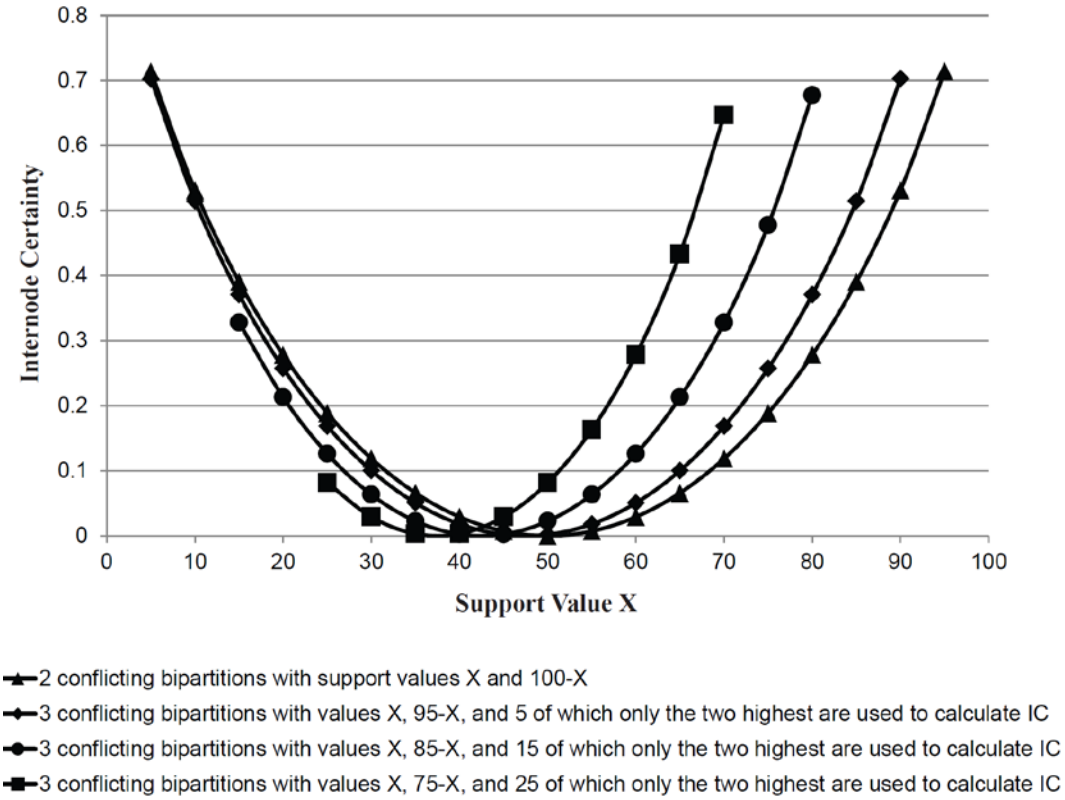
Primary Tree Bipartition	GSF	IC	[Conflicting Bipartition]:GSF value
Kthe, Kwal	99	0.96	None
Calb, Cdub	98	0.95	None
Sbay, Scer, Skud, Smik, Spar	99	0.97	None
Calb, Cdub, Cgui, Clus, Cpar, Ctro, Dhan, Lelo, Psti	95	0.90	None
Calb, Cdub, Ctro	90	0.78	None
Cpar, Lelo	89	0.77	None
Calb, Cdub, Cpar, Ctro, Lelo	86	0.76	None
Scer, Spar	77	0.54	[Sbay,Skud,Smik,Spar]:8; [Smik,Spar]:5
Cgla, Kpol, Sbay, Scas, Scer, Skud, Smik, Spar, Zrou	62	0.59	[Calb,Cdub,Cgla,Cgui,Clus,Cpar,Ctro,Dhan,Lelo,Psti]:6; [Calb,Cdub,Cgui,Clus,Cpar,Ctro,Dhan,Lelo,Psti,Zrou]:5
Scer, Smik, Spar	60	0.30	[Sbay,Skud,Smik]:14; [Sbay,Scer,Skud,Spar]:11; [Sbay,Skud,Smik,Spar]:8; [Skud,Smik]:7 ; [Scer,Skud,Spar]:6
Scer, Skud, Smik ,Spar	52	0.06	[Sbay,Skud]:29; [Sbay,Skud,Smik]:14; [Sbay,Scer,Smik,Spar]:11; [Sbay,Scer,Skud,Spar]:11; [Sbay,Skud,Smik,Spar]:8
Calb, Cdub, Cpar, Ctro, Lelo, Psti	45	0.11	[Cgui,Clus,Dhan,Psti]:20; [Dhan,Psti]:11; [Cgui,Dhan,Psti]:10; [Calb,Cdub,Cgui,Clus,Cpar,Ctro,Dhan,Lelo]:8; [Calb,Cdub,Clus,Cpar,Ctro,Lelo]:5; [Clus,Dhan,Psti]:5
Kthe, Kwal, Sklu	41	0.32	[Agos,Kthe,Kwal]:9; [Calb,Cdub,Cgui,Clus,Cpar,Ctro,Dhan,Kthe, Kwal,Lelo,Psti]:9; [Klac,Sklu]:8; [Agos,Klac,Kthe,Kwal]:7; [Agos, Klac,Sklu]:7; [Agos,Sklu]:7; [Cgla,Kpol,Kthe,Kwal,Sbay,Scas, Scer,Skud,Smik,Spar,Zrou]:6; [Klac,Kthe,Kwal]:5; [Cgla,Kpol, Sbay,Scas,Scer,Sklu,Skud,Smik,Spar,Zrou]:5
Agos, Klac	36	0.09	[Agos,Cgla,Kpol,Kthe,Kwal,Sbay,Scas,Scer,Sklu,Skud,Smik,Spar,Zrou]:17; [Cgla,Klac,Kpol,Kthe,Kwal,Sbay,Scas,Scer,Sklu,Skud, Smik,Spar,Zrou]:13; [Agos,Kthe,Kwal,Sklu]:13; [Klac,Kthe,Kwal, Sklu]:10; [Agos,Kthe,Kwal]:9; [Klac,Sklu]:8; [Agos,Sklu]:7; [Cgla, Klac,Kpol,Sbay,Scas,Scer,Skud,Smik,Spar,Zrou]:7; [Klac,Kthe, Kwal]:5
Agos, Klac, Kthe, Kwal, Sklu	31	0.04	[Agos,Calb,Cdub,Cgui,Clus,Cpar,Ctro,Dhan,Klac,Lelo,Psti]:19; [Calb,Cdub,Cgui,Clus,Cpar,Ctro,Dhan,Klac,Lelo,Psti]:17; [Agos, Calb,Cdub,Cgui,Clus,Cpar,Ctro,Dhan,Lelo,Psti]:13; [Calb,Cdub, Cgui,Clus,Cpar,Ctro,Dhan,Kthe,Kwal,Lelo,Psti]:9; [Agos,Cgla, Klac,Kpol,Sbay,Scas,Scer,Skud,Smik,Spar,Zrou]:7; [Cgla,Klac, Kpol,Sbay,Scas,Scer,Skud,Smik,Spar,Zrou]:7; [Cgla,Kpol,Kthe, Kwal,Sbay,Scas,Scer,Skud,Smik,Spar,Zrou]:6; [Cgla,Kpol,Sbay, Scas,Scer,Sklu,Skud,Smik,Spar,Zrou]:5

Cgla, Sbay, Scas, Scer, Skud, Smik, Spar	29	0.12	[Cgla,Kpol]:12; [Kpol,Scas]:10; [Kpol,Sbay,Scas,Scer,Skud,Smik, Spar,Zrou]:9; [Kpol,Sbay,Scas,Scer,Skud,Smik,Spar]:8; [Cgla, Zrou]:8; [Kpol,Sbay,Scer,Skud,Smik,Spar]:8; [Sbay,Scer,Skud, Smik,Spar,Zrou]:7; [Sbay,Scas,Scer,Skud,Smik,Spar,Zrou]:7; [Cgla,Kpol,Scas]:6; [Agos,Klac,Kpol,Kthe,Kwal,Sbay,Scas,Scer, Sklu,Skud,Smik,Spar,Zrou]:6; [Scas,Zrou]:5
Sbay, Scas, Scer, Skud, Smik, Spar	29	0.02	[Cgla,Sbay,Scer,Skud,Smik,Spar]:20; [Cgla,Scas]:17; [Kpol, Scas]:10; [Kpol,Sbay,Scer,Skud,Smik,Spar]:8; [Sbay,Scer,Skud, Smik,Spar,Zrou]:7; [Cgla,Kpol,Scas]:6; [Scas,Zrou]:5
Calb, Cdub, Cgui, Cpar, Ctro, Dhan, Lelo, Psti	29	0.01	[Cgui,Clus,Dhan]:24; [Cgui,Clus,Dhan,Psti]:20; [Cgui,Clus]:20; [Calb,Cdub,Clus,Cpar,Ctro,Dhan,Lelo,Psti]:16; [Clus,Dhan]:12; [Calb,Cdub,Clus,Cpar,Ctro,Lelo,Psti]:9; [Calb,Cdub,Cgui,Clus, Cpar,Ctro,Dhan,Lelo]:8; [Calb,Cdub,Cgui,Clus,Cpar,Ctro,Lelo, Psti]:6; [Clus,Dhan,Psti]:5; [Calb,Cdub,Clus,Cpar,Ctro,Lelo]:5
Cgui, Dhan	29	0.02	[Cgui,Clus]:20; [Calb,Cdub,Cpar,Ctro,Dhan,Lelo,Psti]:18; [Calb,Cdub,Clus,Cpar,Ctro,Dhan,Lelo,Psti]:16; [Clus,Dhan]:12; [Dhan,Psti]:11; [Calb,Cdub,Cgui,Cpar,Ctro,Lelo,Psti]:6; [Calb, Cdub,Cgui,Clus,Cpar,Ctro,Lelo,Psti]:6; [Clus,Dhan,Psti]:5
Cgla, Kpol, Sbay, Scas, Scer, Skud, Smik, Spar	24	0.02	[Kpol,Zrou]:17; [Cgla,Sbay,Scas,Scer,Skud,Smik,Spar,Zrou]:15; [Kpol,Sbay,Scas,Scer,Skud,Smik,Spar,Zrou]:9; [Cgla,Zrou]:8; [Sbay,Scer,Skud,Smik,Spar,Zrou]:7; [Sbay,Scas,Scer,Skud,Smik, Spar,Zrou]:7; [Calb,Cdub,Cgla,Cgui,Clus,Cpar,Ctro,Dhan,Lelo, Psti]:6; [Scas,Zrou]:5

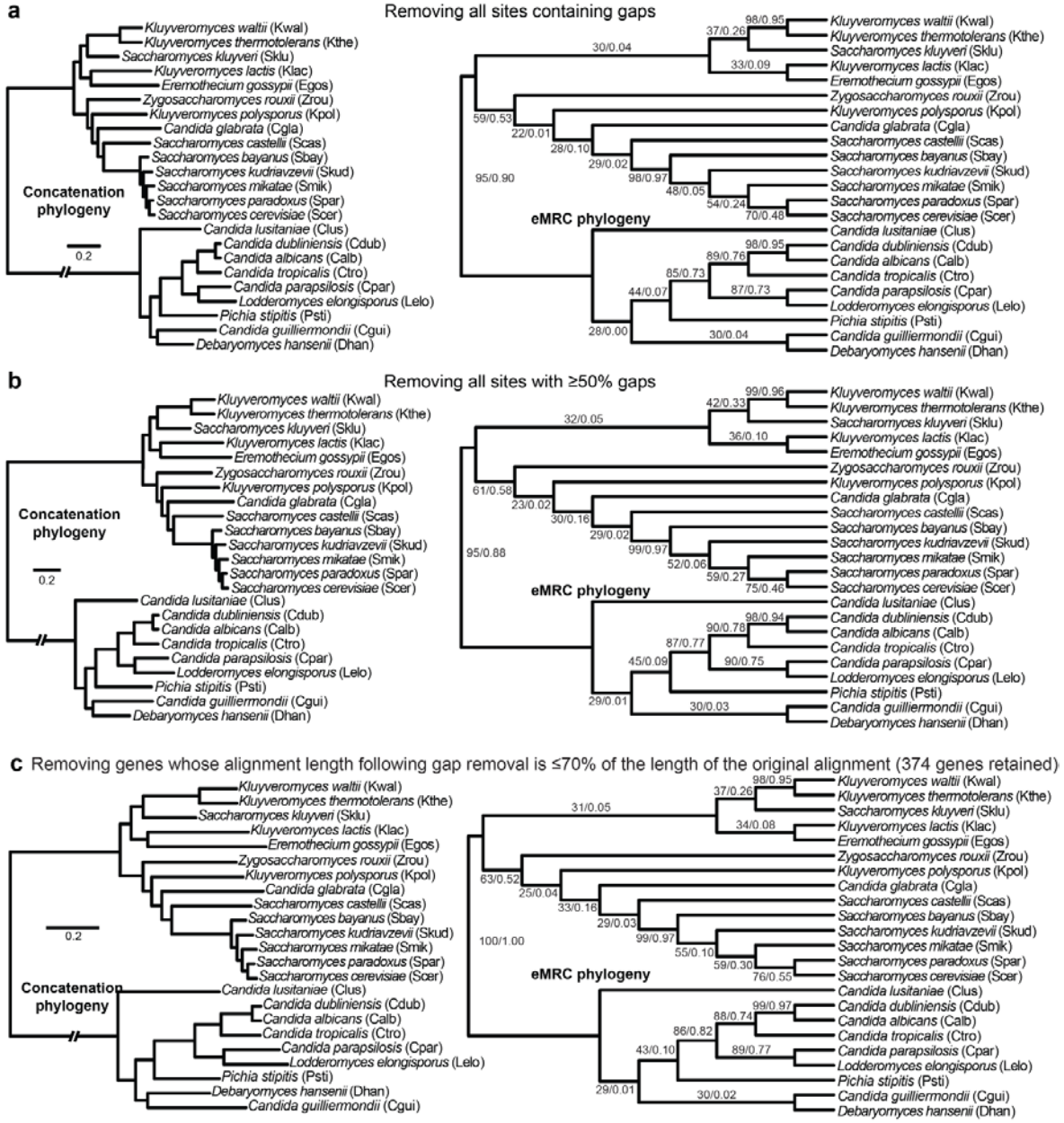
Supplementary Figures 4.1 | The topology of the yeast phylogeny recovered from concatenation analyses using one other maximum likelihood software (GARLI) and one other Bayesian inference (MrBayes) software was identical to the topology recovered by maximum likelihood analysis using the RAxML software. a, The yeast species phylogeny recovered from concatenation analysis of 1,070 genes using maximum likelihood as implemented in the GARLI software. All internodes received 100% bootstrap support. **b**, The yeast species phylogeny recovered from concatenation analysis of 1,070 genes using Bayesian inference as implemented in the MrBayes software. All internodes had 100% posterior probability.



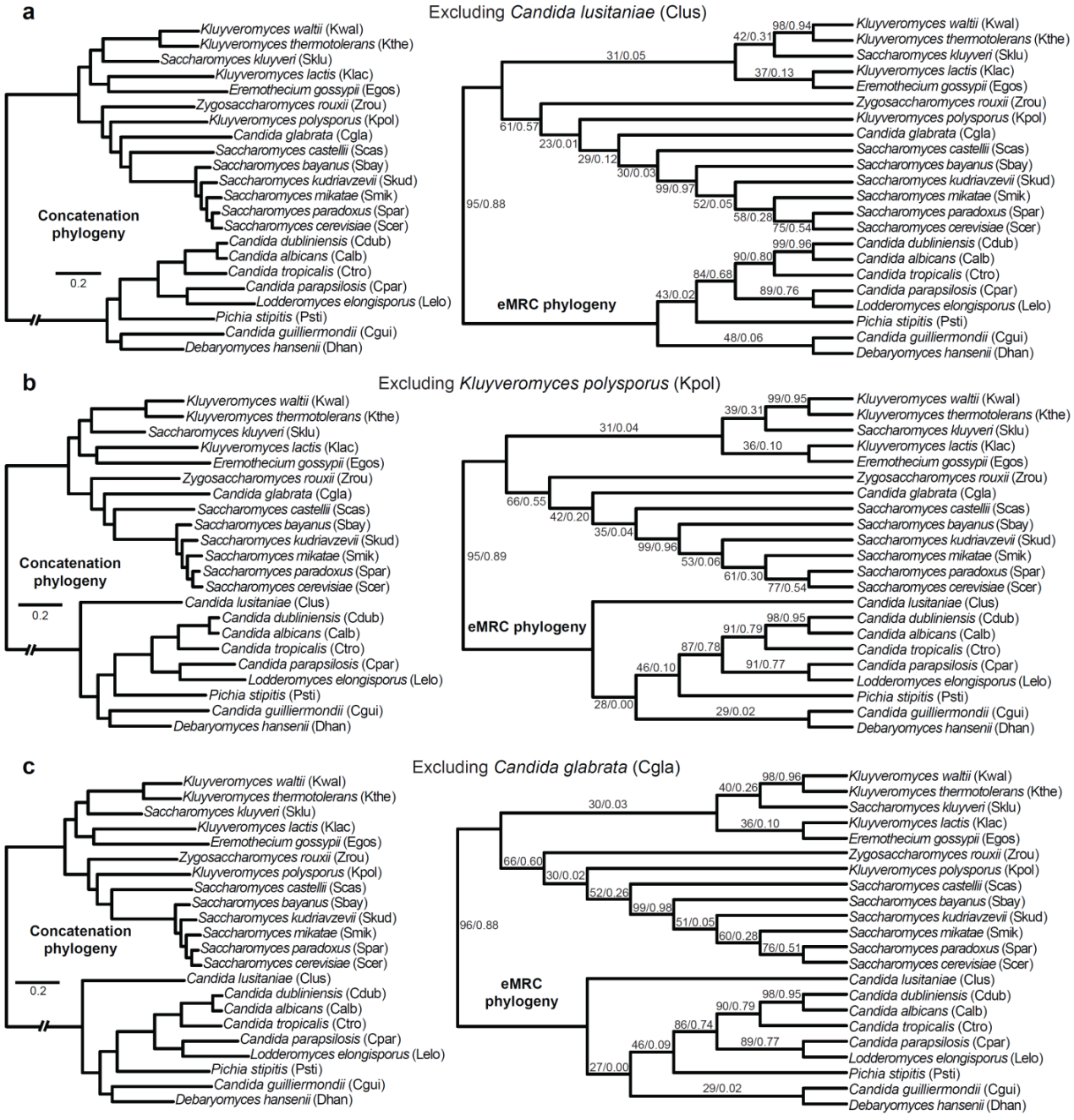
Supplementary Figures 4.2 | Representative values of the new measure Internode Certainty (IC) for a range of representative support values of two most prevalent and conflicting bipartitions for a given internode. Each plot on the graph depicts how IC (Y-axis) varies in response to the relative support of conflicting bipartitions on a given internode. IC can be measured on any given set of trees. For example, if the entire set of gene trees (GTs) is used, the IC value of a given internode will reflect the amount of information available for that internode in the set GTs by considering the internode's gene support frequency jointly with the frequency of the most prevalent bipartition that conflicts with the internode. If the set of bootstrap replicate trees for a given gene is used, then IC will be calculated based on bootstrap support values. From the right to the left of the graph, the first of the four plots shown with triangle symbols corresponds to the case of only two conflicting bipartitions for one internode with support values X and 100–X. For example, given 100 total GTs, if 60 of them support bipartition 1, the remaining 40 will support the conflicting bipartition. The second, third and fourth of the four plots (shown with diamond, circle, and square symbols, respectively) correspond to case where there are three conflicting bipartitions for one internode, but only the two most prevalent ones are considered. For example, in the plot with the diamond symbols, given 100 total GTs, if 60 of them support bipartition 1, 35 will support the conflicting bipartition 2, because conflicting bipartition 3 has been set to be supported by 5 GTs. Thus, when the two most prevalent ones are considered, the percentage of GTs supporting the first bipartition will be equal to 60/(60+35), whereas the percentage of GTs supporting the second bipartition will be 35/(60+35). The reason that the number of GTs that support the third conflicting bipartition is not included is because we want IC to measure the magnitude of certainty conveyed by the two most prevalent bipartitions. This way, IC will equal zero when the two most prevalent bipartitions are equally prevalent (in this example that would be the case if bipartitions 1 and 2 were each supported by 42.5 GTs each).

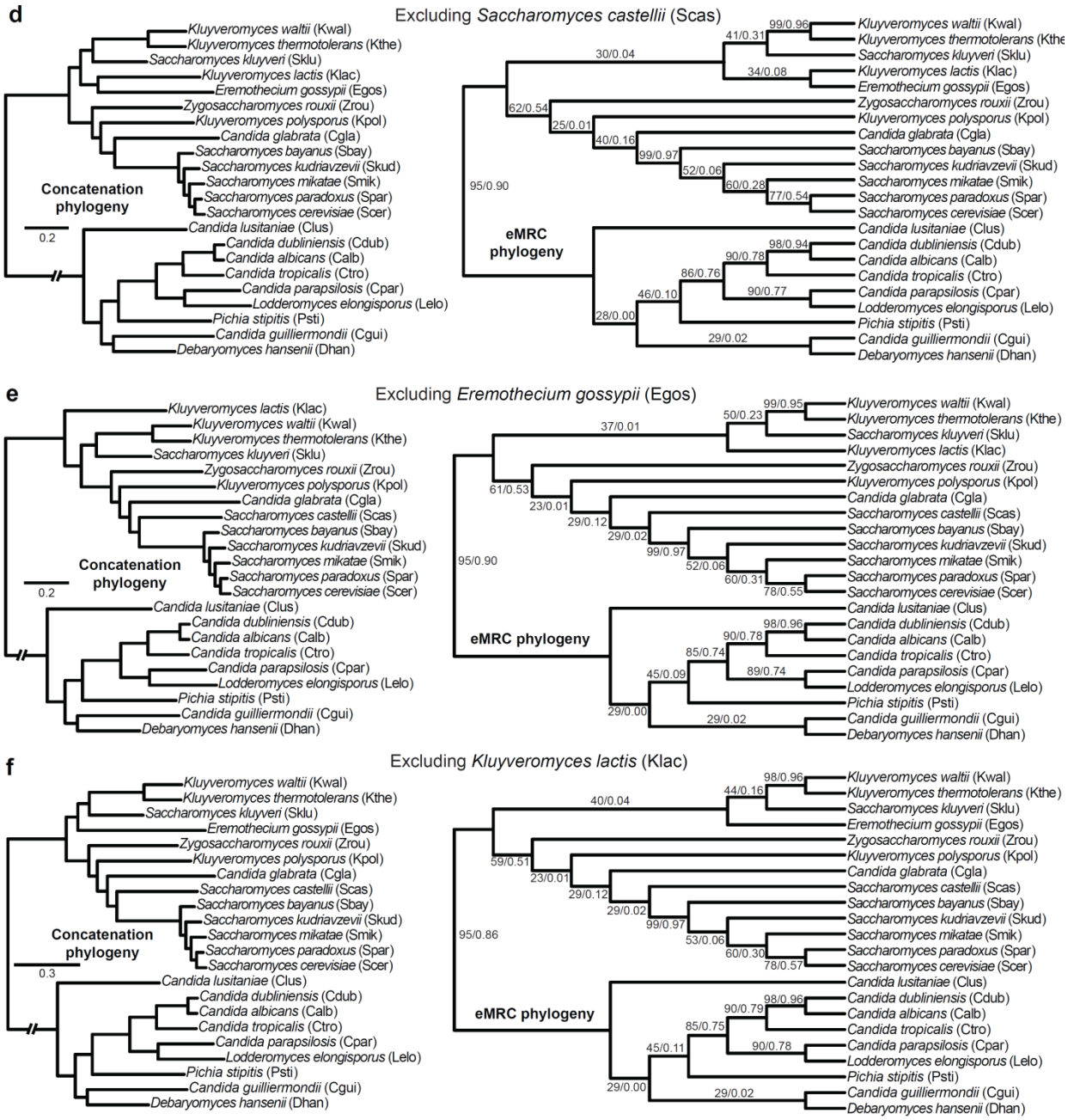


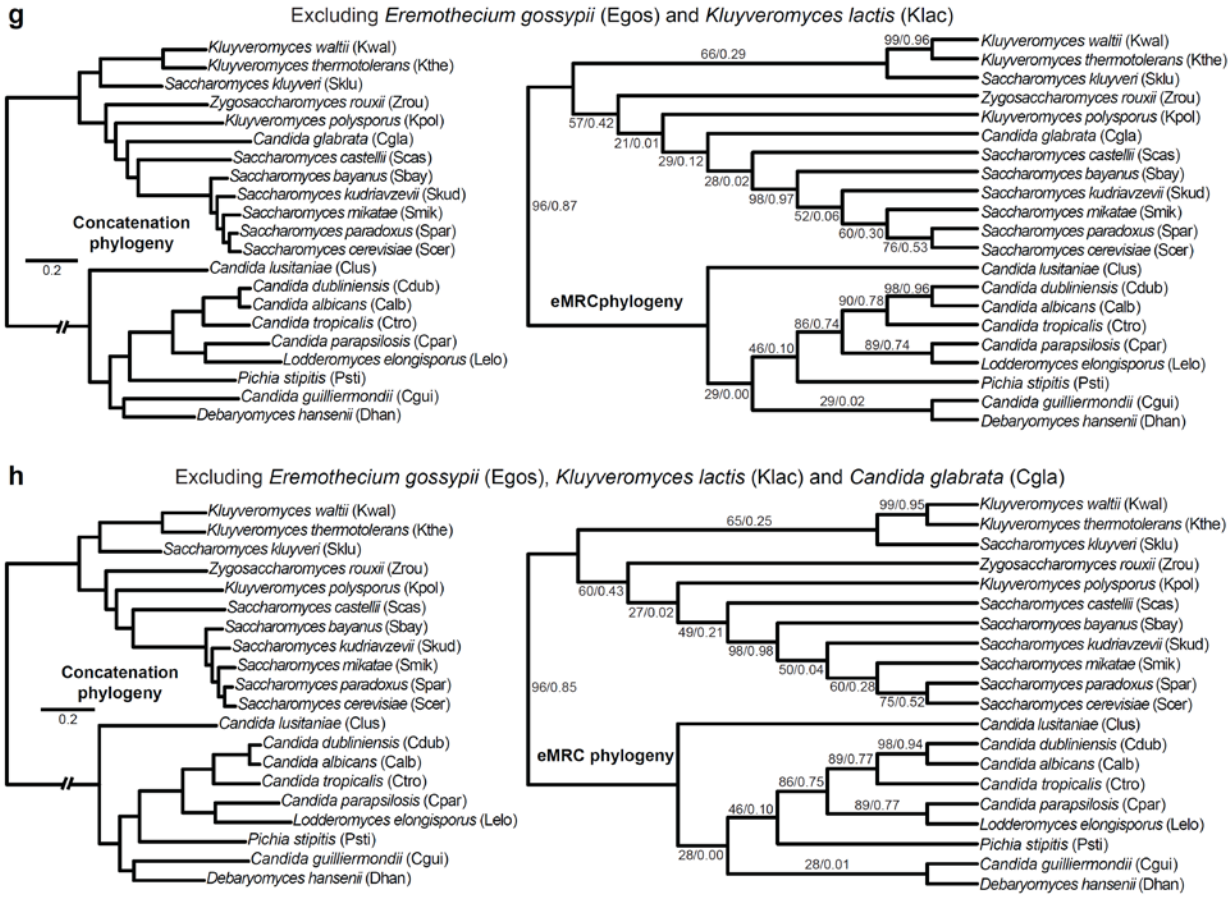
Supplementary Figures 4.3 | Removal of sites containing gaps or of poorly aligned genes does not significantly improve the yeast phylogeny inferred by concatenation and eMRC approaches. Each panel shows the yeast species phylogeny inferred from concatenation analysis (left panel) and from extended majority rule consensus (eMRC) analysis (right panel). All internodes of phylogenies inferred by concatenation received 100% bootstrap support unless otherwise indicated. Values near internodes of phylogenies inferred by eMRC analysis correspond to gene support frequency and internode certainty, respectively. **a**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following removal of all sites containing gaps. **b**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following removal of all sites where ≥50% of the character states are gaps. **c**, Concatenation (left) and eMRC (right) phylogenies of the 374 genes whose alignment length following removal of all gaps is ≥70% of the length of the original alignment.



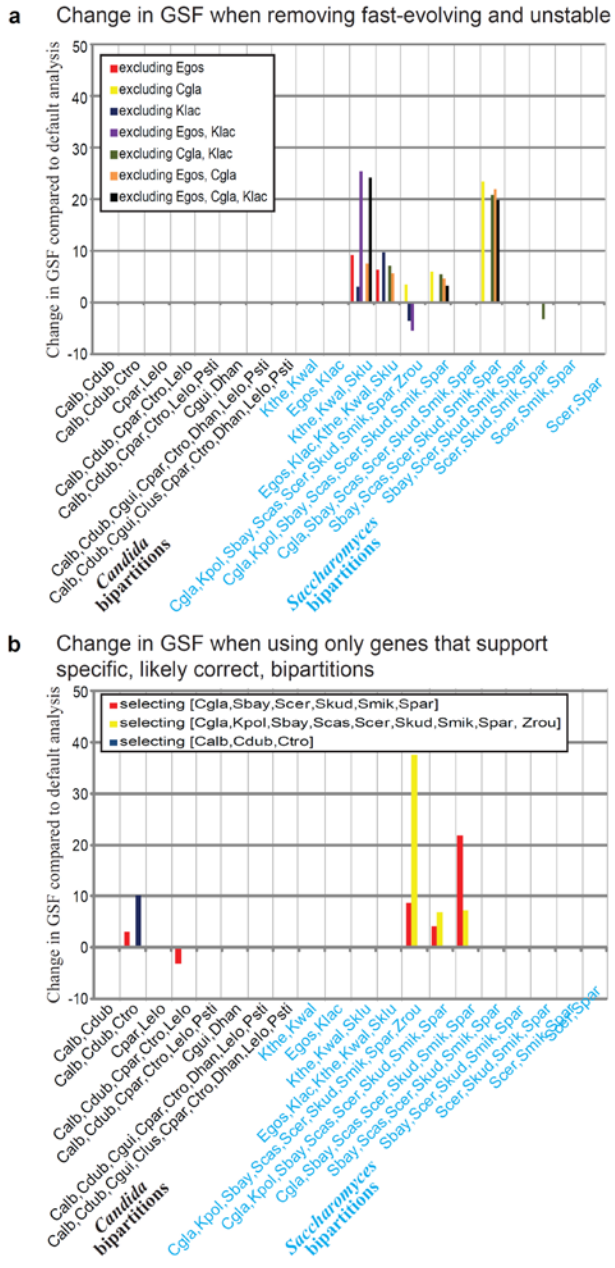
Supplementary Figures 4.4 | Removal of one or more unstable or fast-evolving species has, if any, a minor and local effect on the yeast phylogeny inferred by concatenation and eMRC approaches. Each panel shows the yeast species phylogeny inferred from concatenation analysis (left panel) and from extended majority rule consensus (eMRC) analysis (right panel) following removal of one or more unstable or fast-evolving species from the analysis. All internodes of phylogenies inferred by concatenation received 100% bootstrap support unless otherwise indicated. Values near internodes of phylogenies inferred by eMRC analysis correspond to gene support frequency and internode certainty, respectively. **a**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following the removal of the unstable taxon *Candida lusitaniae*. **b**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following the removal of the fast-evolving and unstable taxon *Kluyveromyces polysporus*. **c**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following the removal of the fast-evolving and unstable taxon *Candida glabrata*. **d**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following the removal of the unstable taxon *Saccharomyces castellii*. **e**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following the removal of the fast-evolving and unstable taxon *Eremothecium gossypii*. **f**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following the removal of the fast-evolving and unstable taxon *Kluyveromyces lactis*. **g**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following the removal of both *Eremothecium gossypii* and *Kluyveromyces lactis*. **h**, Concatenation (left) and eMRC (right) phylogenies of all 1,070 genes following the removal of *Eremothecium gossypii, Kluyveromyces lactis* and *Candida glabrata*.


a
Excluding Candida lusitaniae (Clus)
Concatenation phylogeny
eMRC phylogeny
0.2
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
98/0.94
42/0.31
31/0.05
37/0.13
61/0.57
23/0.01
29/0.12
30/0.03
99/0.97
52/0.05
95/0.88
58/0.28
75/0.54
99/0.96
90/0.80
84/0.68
89/0.76
43/0.02
48/0.06
b
Excluding Kluyveromyces polysporus (Kpol)
Concatenation phylogeny
eMRC phylogeny
0.2
99/0.95
39/0.31
31/0.04
36/0.10
66/0.55
42/0.20
35/0.04
99/0.96
53/0.06
95/0.89
61/0.30
77/0.54
98/0.95
91/0.79
87/0.78
91/0.77
46/0.10
28/0.00
29/0.02
c
Excluding Candida glabrata (Cgla)
Concatenation phylogeny
eMRC phylogeny
0.2
98/0.96
40/0.26
30/0.03
36/0.10
66/0.60
30/0.02
52/0.26
99/0.98
51/0.05
96/0.88
60/0.28
76/0.51
98/0.95
90/0.79
86/0.74
89/0.77
46/0.09
27/0.00
29/0.02

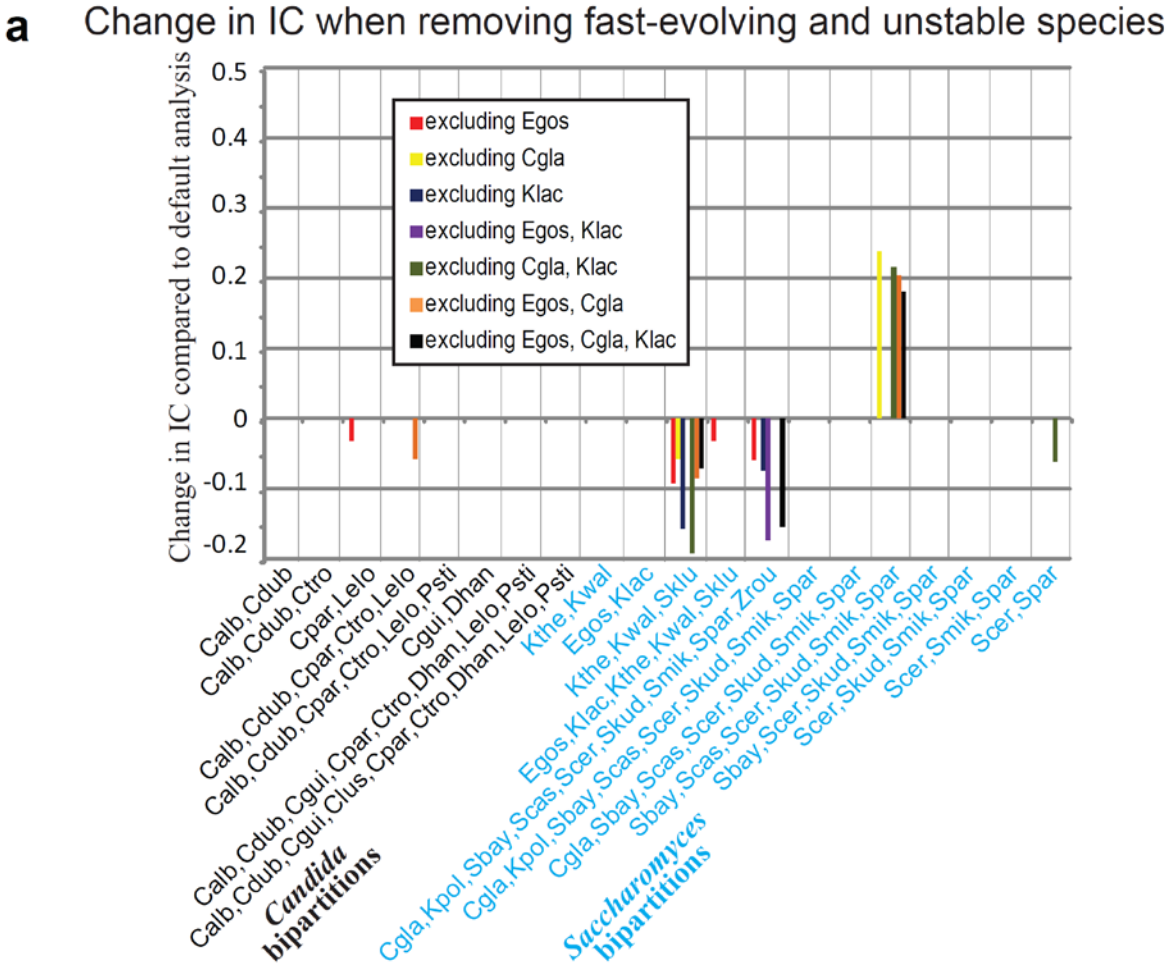

d
Excluding Saccharomyces castellii (Scas)
Concatenation phylogeny
0.2
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Candida lusitaniae (Clus)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
eMRC phylogeny
99/0.96
41/0.31
30/0.04
34/0.08
62/0.54
25/0.01
40/0.16
99/0.97
52/0.06
60/0.28
77/0.54
95/0.90
98/0.94
90/0.78
86/0.76
90/0.77
46/0.10
28/0.00
29/0.02
e
Excluding Eremothecium gossypii (Egos)
Concatenation phylogeny
0.2
Kluyveromyces lactis (Klac)
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Candida lusitaniae (Clus)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
eMRC phylogeny
99/0.95
50/0.23
37/0.01
61/0.53
23/0.01
29/0.12
29/0.02
99/0.97
52/0.06
60/0.31
78/0.55
95/0.90
98/0.96
90/0.78
85/0.74
89/0.74
45/0.09
29/0.00
29/0.02
f
Excluding Kluyveromyces lactis (Klac)
Concatenation phylogeny
0.3
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Eremothecium gossypii (Egos)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Candida lusitaniae (Clus)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
eMRC phylogeny
98/0.96
44/0.16
40/0.04
59/0.51
23/0.01
29/0.12
29/0.02
99/0.97
53/0.06
60/0.30
78/0.57
95/0.86
98/0.96
90/0.79
85/0.75
90/0.78
45/0.11
29/0.00
29/0.02


g
Excluding Eremothecium gossypii (Egos) and Kluyveromyces lactis (Klac)
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Concatenation phylogeny
0.2
Candida lusitaniae (Clus)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
99/0.96
66/0.29
57/0.42
21/0.01
29/0.12
28/0.02
98/0.97
52/0.06
60/0.30
76/0.53
96/0.87
eMRCphylogeny
98/0.96
90/0.78
86/0.74
89/0.74
46/0.10
29/0.00
29/0.02
h
Excluding Eremothecium gossypii (Egos), Kluyveromyces lactis (Klac) and Candida glabrata (Cgla)
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Concatenation phylogeny
0.2
Candida lusitaniae (Clus)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
99/0.95
65/0.25
60/0.43
27/0.02
49/0.21
98/0.98
50/0.04
60/0.28
75/0.52
96/0.85
eMRC phylogeny
98/0.94
89/0.77
86/0.75
89/0.77
46/0.10
28/0.00
28/0.01

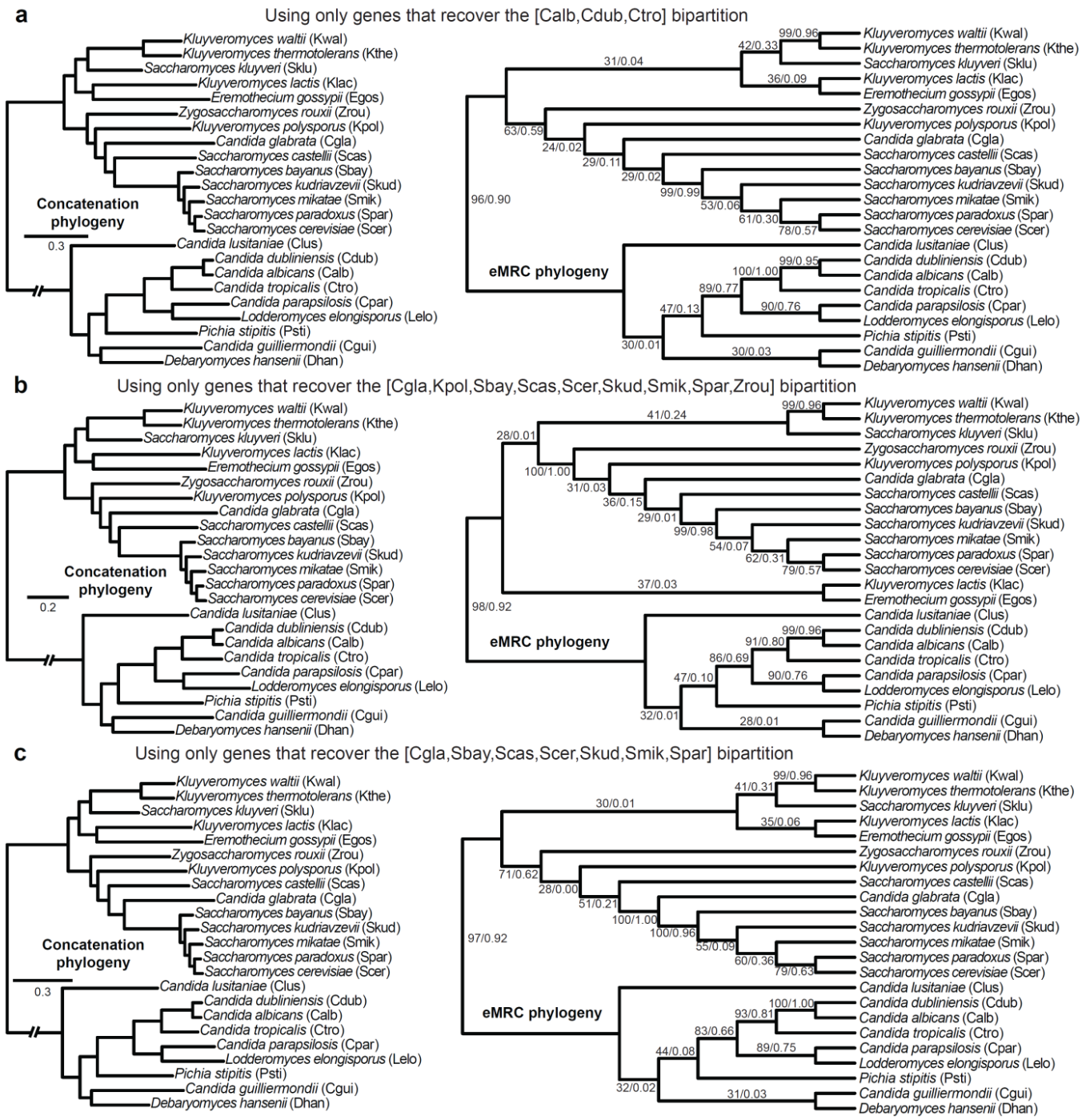
Supplementary Figures 4.5 | Removal of fast-evolving and unstable species has, if any, a minor and effect on GSF and IC values of internodes of the yeast phylogeny. The X-axis shows the 20 bipartitions present in the yeast phylogeny suggested by concatenation analysis and the Y-axis the percent change in gene support frequency (GSF) or Internode Certainty (IC) observed for each bipartition between the treatment (removal of fast-evolving and unstable species) and the default analysis (all species included). Only GSF changes ≥3% are shown. **a,** Change in the GSF values of the 20 bipartitions present in the yeast phylogeny when *C. glabrata*, *C. lusitaniae*, *K. polysporus*, and *S. castellii* are removed individually. **b**, Change in the IC values of the 20 bipartitions present in the yeast phylogeny when *C. glabrata*, *C. lusitaniae*, *K. polysporus*, and *S. castellii* are removed individually.



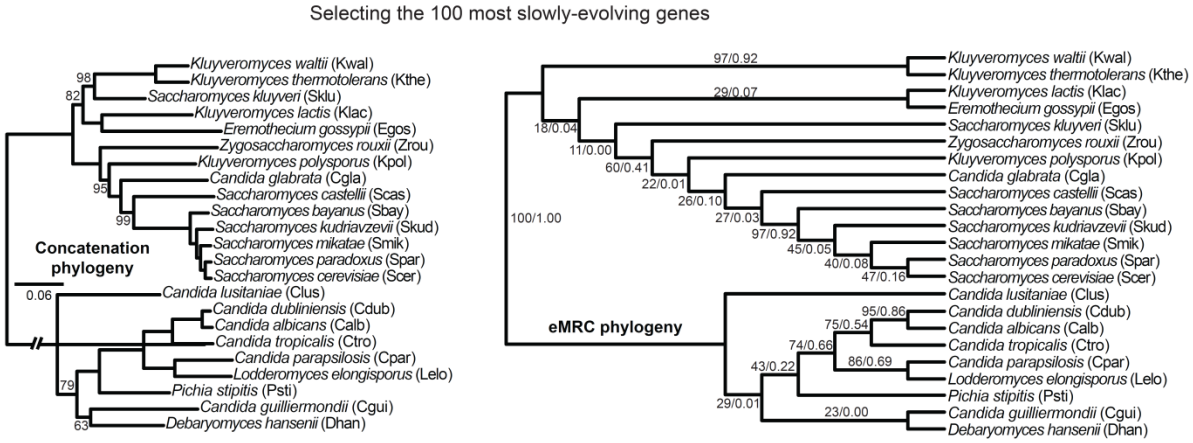
Supplementary Figures 4.6 | Removal of fast-evolving and unstable species or the exclusive use of genes that recover specific bipartitions has a minor and typically local effect on IC values of internodes of the yeast phylogeny. The X-axis shows the 20 bipartitions present in the yeast phylogeny suggested by concatenation analysis and the Y-axis the percent change in Internode Certainty (IC) observed for each bipartition between the treatment (removal of fast-evolving and unstable species or of genes that fail to recover specific clades) and the default analysis (all species and genes included). Only GSF changes ≥3% are shown. **a**, The individual or combined removal of *E. gossypii* (Egos), *K. lactis* (Klac), and *C. glabrata* (Cgla), three of the fastest evolving species as well as of those whose phylogenetic position is most unstable from the dataset has a minor and local effect on the IC of neighboring internodes. **b**, The selection of genes whose individual topologies recover well-established bipartitions of the yeast phylogeny has a minor effect on the IC of internodes of the yeast phylogeny. Note that the [*C. albicans*, *C. dubliniensis*, *C. tropicalis*] (abbreviated [Calb, Cdub, Ctro]) bipartition has 90% GSF in the extended majority rule consensus (eMRC) phylogeny reconstructed from the 1,070 individual gene trees, the [*C. glabrata*, *K. polysporus*, *S. bayanus*, *S. castellii*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*, *Z. rouxii*] (abbreviated [Zrou, Kpol, Cgla, Sbay, Skud, Smik, Scer, Spar]) bipartition has 62% GSF, and the [*C. glabrata*, *S. bayanus*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*] (abbreviated [Cgla, Sbay, Skud, Smik, Scer, Spar]) bipartition has 20% GSF. This last bipartition does not appear in the eMRC phylogeny but, as discussed in the main text, several independent rare genomic changes strongly suggest that it is the correct one.



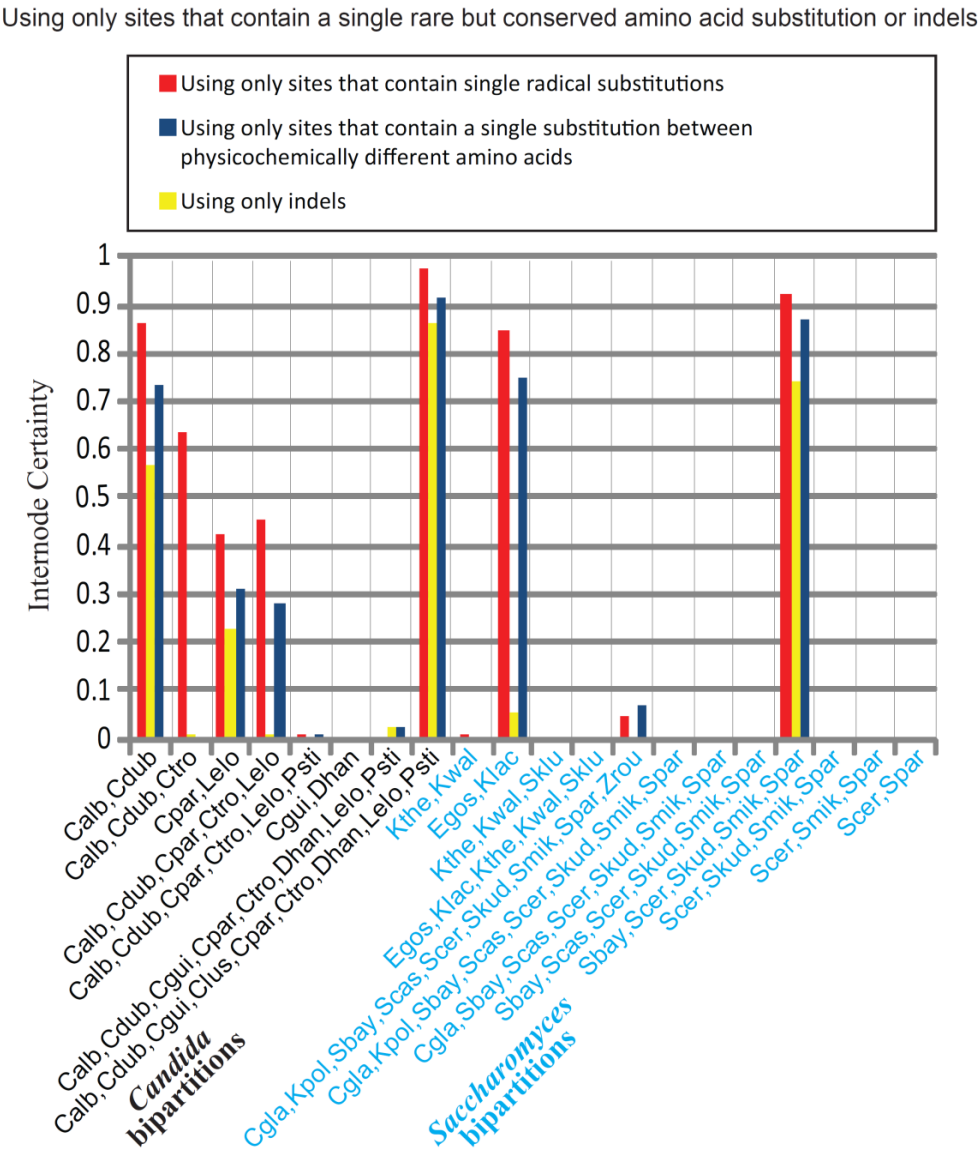
Supplementary Figures 4.7 | Selection of genes that support specific bipartitions has, if any, a minor and local effect on the yeast phylogeny inferred by concatenation and eMRC approaches. Each panel shows the yeast species phylogeny inferred from concatenation analysis (left panel) and from extended majority rule consensus (eMRC) analysis (right panel) following the selection and use of genes that recover specific bipartitions. All internodes of phylogenies inferred by concatenation received 100% bootstrap support unless otherwise indicated. Values near internodes of phylogenies inferred by eMRC analysis correspond to gene support frequency and internode certainty, respectively. **a**, Concatenation (left) and eMRC (right) phylogenies using only the genes that recover the [*C. albicans*, *C. dubliniensis*, *C. tropicalis*] (abbreviated [Calb, Cdub, Ctro]) bipartition. **b**, Concatenation (left) and eMRC (right) phylogenies using only the genes that recover the [*C. glabrata*, *K. polysporus*, *S. bayanus*, *S. castellii*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*, *Z. rouxii*] (abbreviated [Zrou, Kpol, Cgla, Sbay, Skud, Smik, Scer, Spar]) bipartition. **c**, Concatenation (left) and eMRC (right) phylogenies using only the genes that recover the [*C. glabrata*, *S. bayanus*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*] (abbreviated [Cgla, Sbay, Skud, Smik, Scer, Spar]) bipartition. This last bipartition does not appear in the eMRC phylogeny but, as discussed in the main text, several independent rare genomic changes strongly suggest that it is the correct one.


a
Using only genes that recover the [Calb,Cdub,Ctro] bipartition
Concatenation phylogeny
0.3
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Candida lusitaniae (Clus)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
eMRC phylogeny
99/0.96
42/0.33
31/0.04
36/0.09
63/0.59
24/0.02
29/0.11
29/0.02
99/0.99
96/0.90
53/0.06
61/0.30
78/0.57
99/0.95
100/1.00
89/0.77
90/0.76
47/0.13
30/0.01
30/0.03
b
Using only genes that recover the [Cgla,Kpol,Sbay,Scas,Scer,Skud,Smik,Spar,Zrou] bipartition
Concatenation phylogeny
0.2
eMRC phylogeny
41/0.24
28/0.01
100/1.00
31/0.03
36/0.15
29/0.01
99/0.98
54/0.07
62/0.31
79/0.57
37/0.03
98/0.92
99/0.96
91/0.80
86/0.69
90/0.76
47/0.10
32/0.01
28/0.01
c
Using only genes that recover the [Cgla,Sbay,Scas,Scer,Skud,Smik,Spar] bipartition
Concatenation phylogeny
0.3
eMRC phylogeny
99/0.96
41/0.31
30/0.01
35/0.06
71/0.62
28/0.00
51/0.21
100/1.00
100/0.96
55/0.09
97/0.92
60/0.36
79/0.63
100/1.00
93/0.81
83/0.66
89/0.75
44/0.08
32/0.02
31/0.03

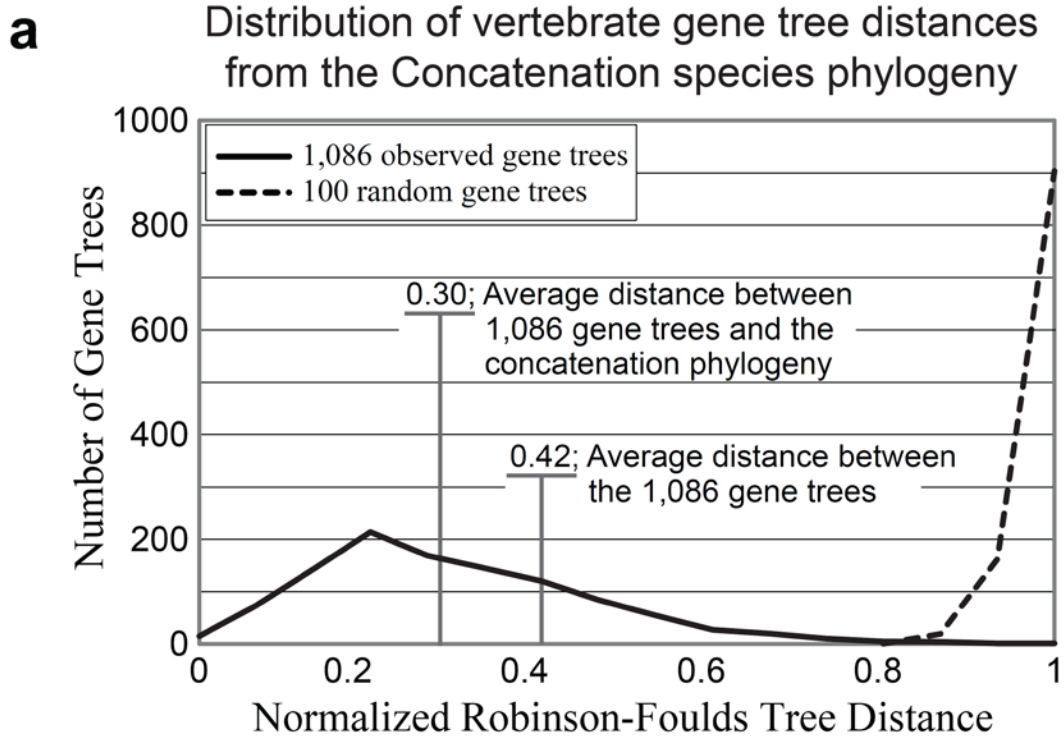
Supplementary Figures 4.8 | Selection of the 100 slowest-evolving genes has a large, negative effect on GSF and IC values of internodes of the yeast phylogeny inferred by concatenation and eMRC approaches. Each panel shows the yeast species phylogeny inferred from concatenation analysis (left panel) and from extended majority rule consensus (eMRC) analysis (right panel) following the selection and use of the 100 slowest-evolving genes. All internodes of phylogenies inferred by concatenation received 100% bootstrap support unless otherwise indicated. Values near internodes of phylogenies inferred by eMRC analysis correspond to gene support frequency and internode certainty, respectively.

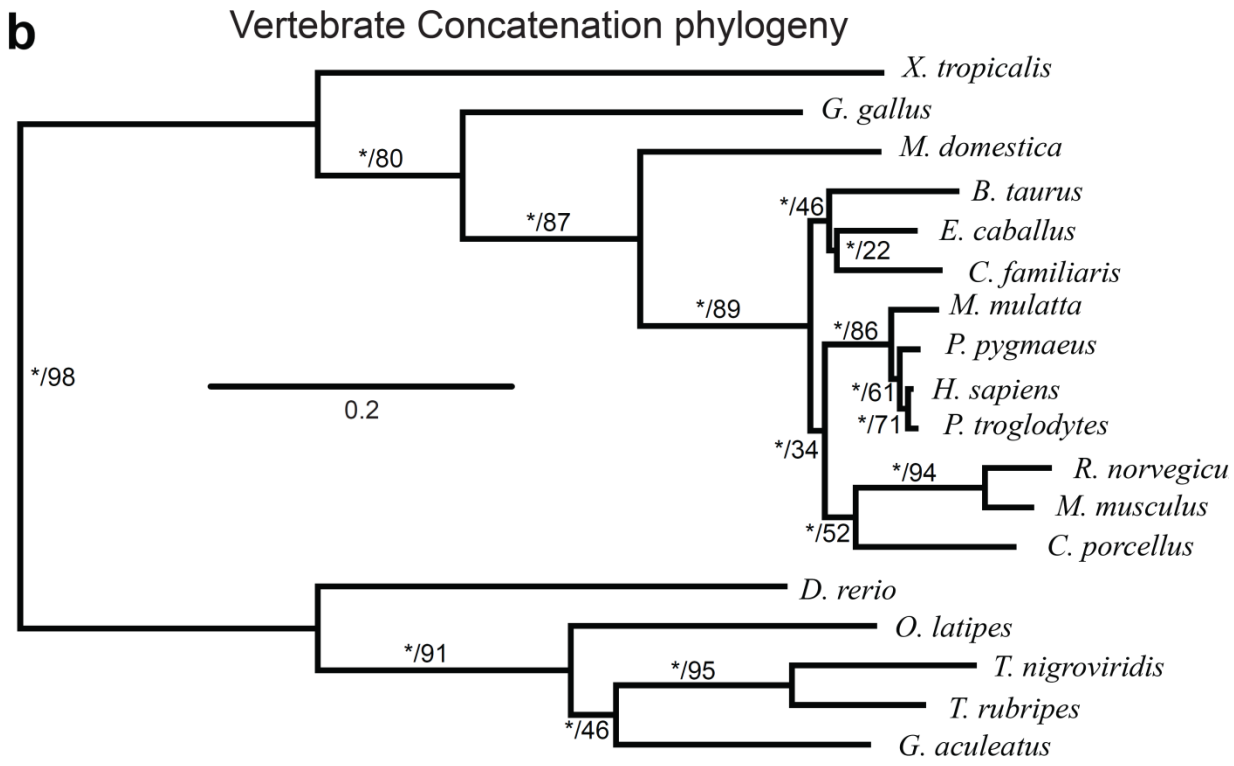


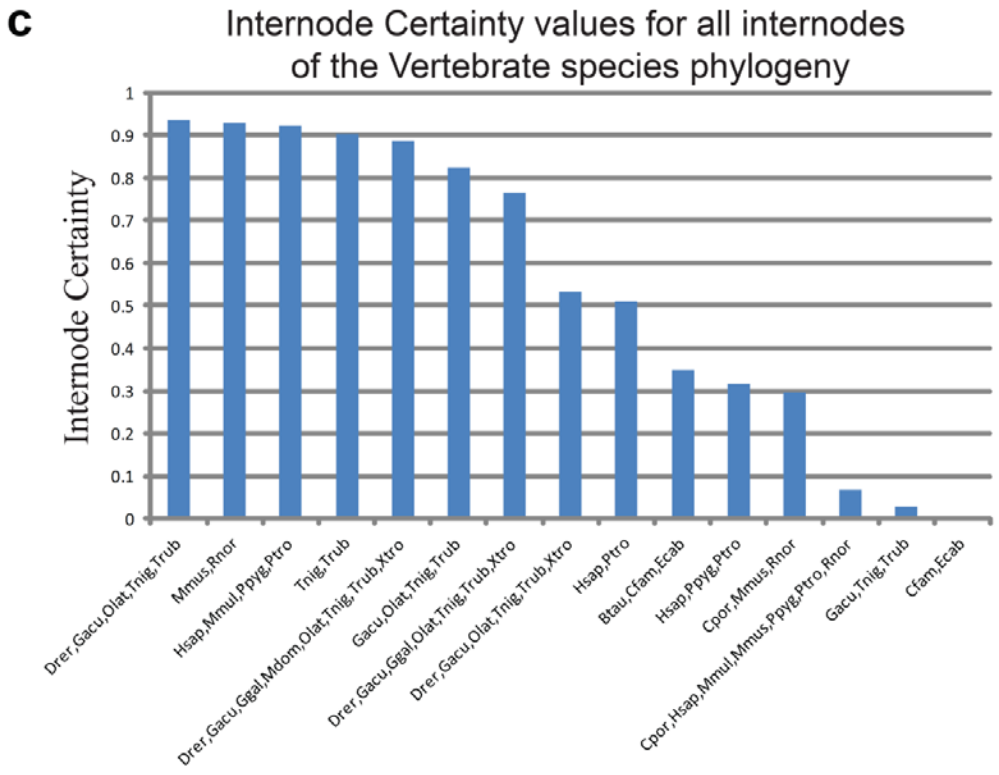
Supplementary Figures 4.9 | Selection of sites that contain a single rare but conserved amino acid substitution or indels has a large, negative effect on GSF and IC values of internodes of the yeast phylogeny. The X-axis shows the 20 bipartitions present in the yeast phylogeny suggested by concatenation analysis and the Y-axis the percent change in Internode Certainty (IC) observed for each bipartition between the treatment (selection of specific sites or indels) and the default analysis (all sites included). Only GSF changes ≥3% are shown. The red bars correspond to changes in IC when using only the 20,289 sites that contain single radical substitutions (defined as a substitution with a blosum62 matrix score ≤–3), the blue bars correspond to changes in IC when using only the 4,075 sites that contain a single substitution between amino acids that differ radically in their physicochemical properties, and the yellow bars correspond to changes in IC when using only the 2,474 characters which mark the presence / absence of a single indel that spans 7 or more aa among the 23 yeast species.



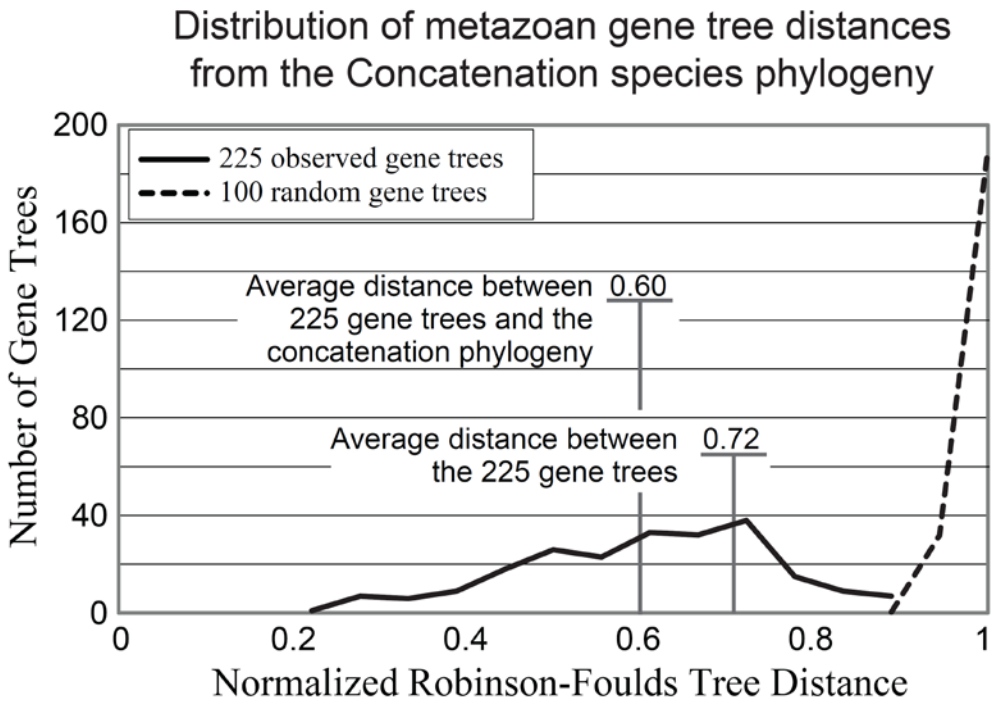
Supplementary Figures 4.10 | High levels of incongruence in Vertebrate and Metazoan phylogenomic datasets despite the inference of highly supported phylogenies by concatenation analysis. a, The distribution of the agreement between the bipartitions present in the 1,086 individual gene trees (GTs) and the vertebrate concatenation phylogeny, as well as the distribution of the agreement between the bipartitions present in 100 randomly generated trees of equal taxon number and the concatenation phylogeny, measured using the normalized Robinson-Foulds tree distance. The average tree distances between the 1,086 GTs and the concatenation phylogeny as well as between the 1,086 GTs with each other are also shown. **b**, The vertebrate species phylogeny recovered from concatenation analysis of 1,086 genes using maximum likelihood. The extended majority rule consensus (eMRC) phylogeny is topologically identical to the concatenation phylogeny. Values near internodes correspond to bootstrap support and gene support frequency (GSF), respectively. Asterisks (*) denote internodes that received 100% bootstrap support by the concatenation analysis. **c**, The distribution of Internode Certainty (IC) values for all internodes of the vertebrate species phylogeny. **d**, The distribution of the agreement between the bipartitions present in the 225 individual GTs and the metazoan concatenation phylogeny, as well as the distribution of the agreement between the bipartitions present in 100 randomly generated trees of equal taxon number and the concatenation phylogeny, measured using the normalized Robinson-Foulds tree distance. The average tree distances between the 225 GTs and the concatenation phylogeny as well as between the 225 GTs with each other are also shown. **e**, The metazoan species phylogeny recovered from concatenation analysis of 225 genes using maximum likelihood. The eMRC phylogeny is topologically identical to the concatenation phylogeny. Values near internodes correspond to bootstrap support and gene support frequency, respectively. Asterisks (*) denote internodes that received 100% bootstrap support by the concatenation analysis. **f**, The distribution of IC values for all internodes of the metazoan species phylogeny. Note that GSF and IC values indicate the existence of numerous internodes in the vertebrate and especially in the metazoan phylogeny that are supported by a small percentage of gene trees and have very small or zero IC values.



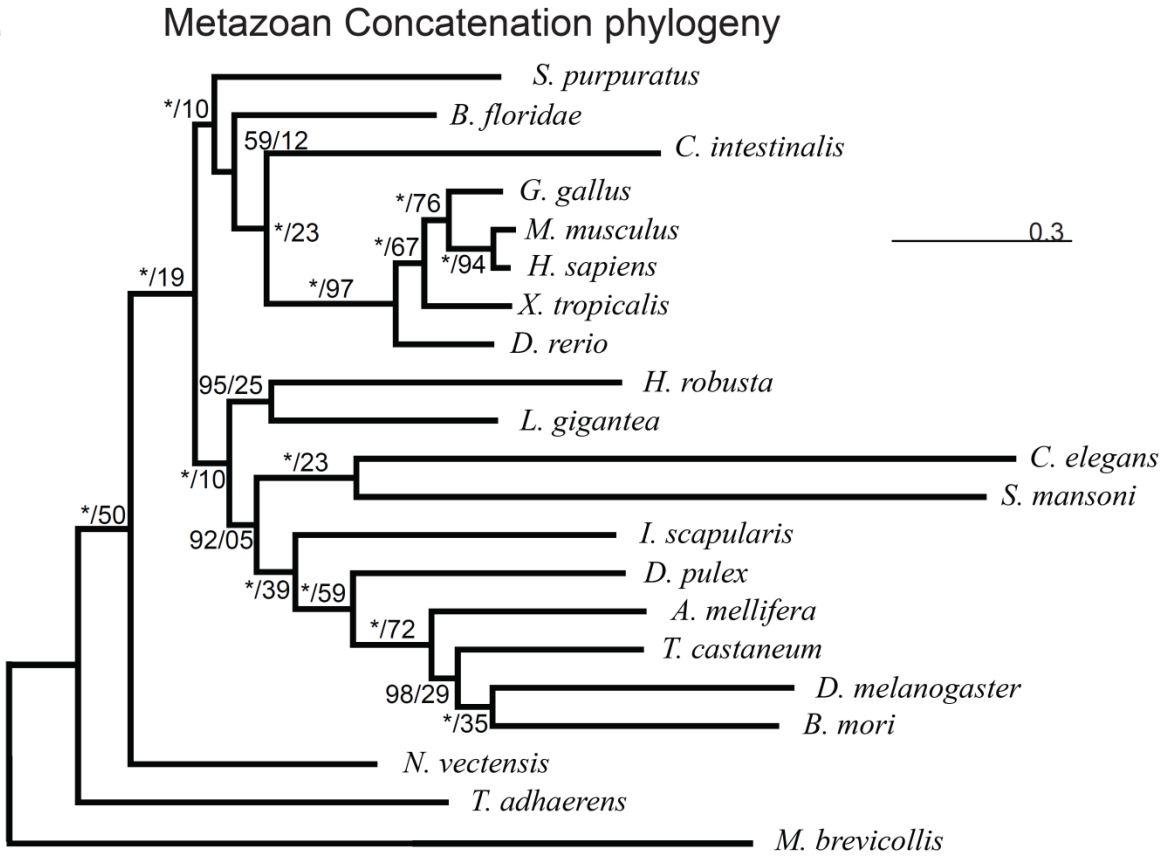

b
Vertebrate Concatenation phylogeny
X. tropicalis
G. gallus
M. domestica
B. taurus
E. caballus
C. familiaris
M. mulatta
P. pygmaeus
H. sapiens
P. troglodytes
R. norvegicu
M. musculus
C. porcellus
D. rerio
O. latipes
T. nigroviridis
T. rubripes
G. aculeatus
*/80
*/87
*/46
*/22
*/89
*/86
*/61
*/71
*/34
*/94
*/52
*/98
0.2
*/91
*/95
*/46


c
Internode Certainty values for all internodes
of the Vertebrate species phylogeny
Internode Certainty
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Drer,Gacu,Olat,Tnig,Trub
Mmus,Rnor
Hsap,Mmul,Ppyg,Ptro
Tnig,Trub
Drer,Gacu,Ggal,Mdom,Olat,Tnig,Trub,Xtro
Gacu,Olat,Tnig,Trub
Drer,Gacu,Ggal,Olat,Tnig,Trub,Xtro
Drer,Gacu,Olat,Tnig,Trub,Xtro
Hsap,Ptro
Btau,Cfam,Ecab
Hsap,Ppyg,Ptro
Cpor,Mmus,Rnor
Cpor,Hsap,Mmul,Mmus,Ppyg,Ptro,Rnor
Gacu,Tnig,Trub
Cfam,Ecab

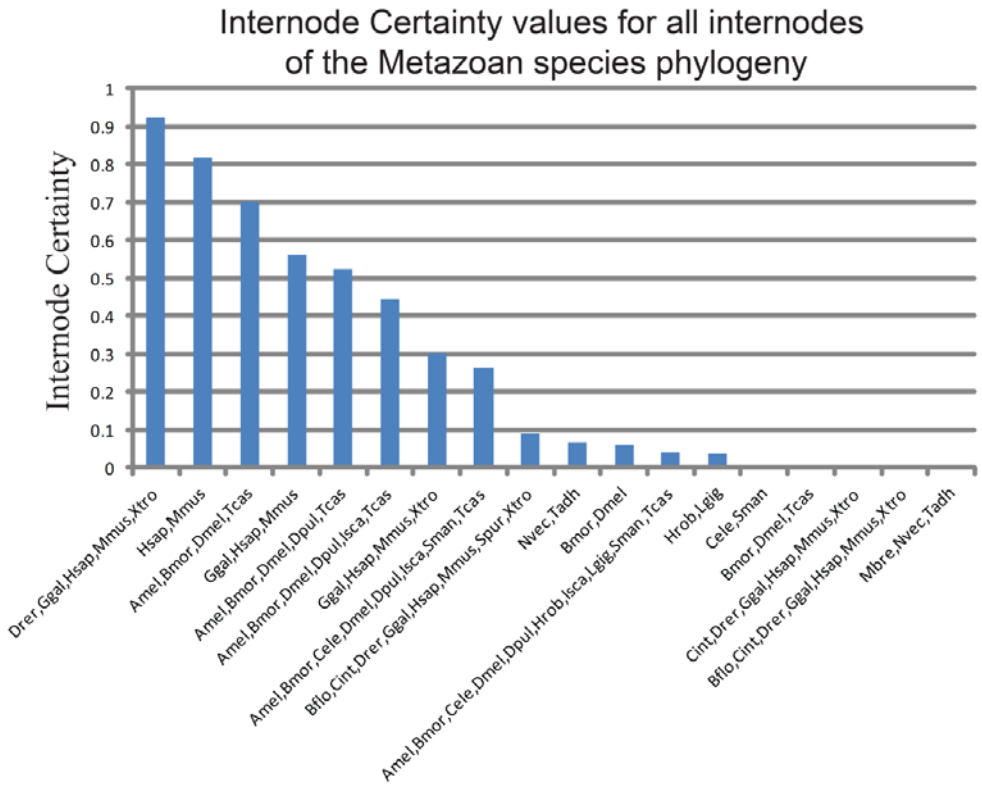
d


Distribution of metazoan gene tree distances
from the Concatenation species phylogeny
225 observed gene trees
100 random gene trees
Average distance between 225 gene trees and the concatenation phylogeny 0.60
Average distance between the 225 gene trees 0.72
Number of Gene Trees
Normalized Robinson-Foulds Tree Distance
0
40
80
120
160
200
0
0.2
0.4
0.6
0.8
1

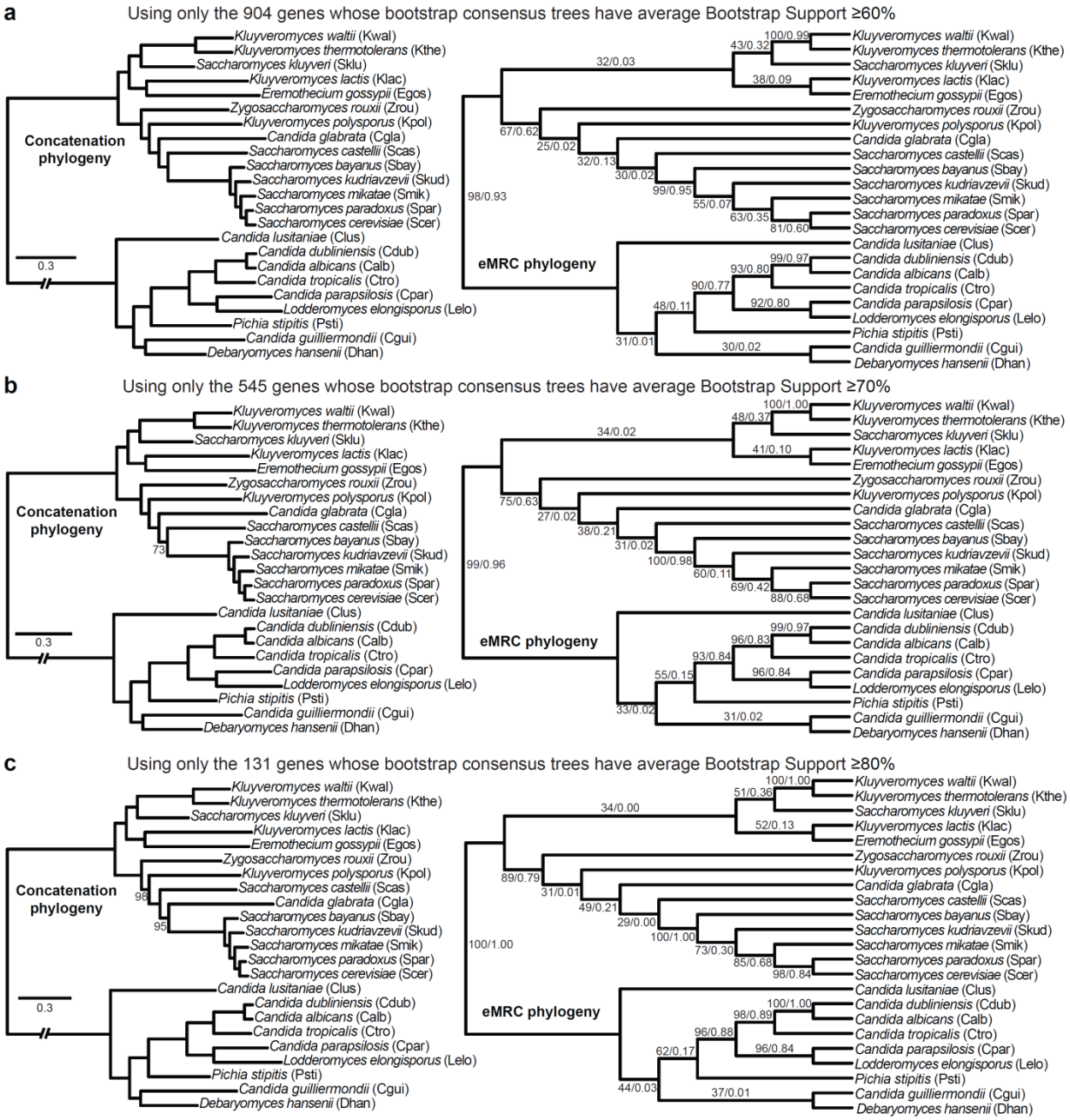
e

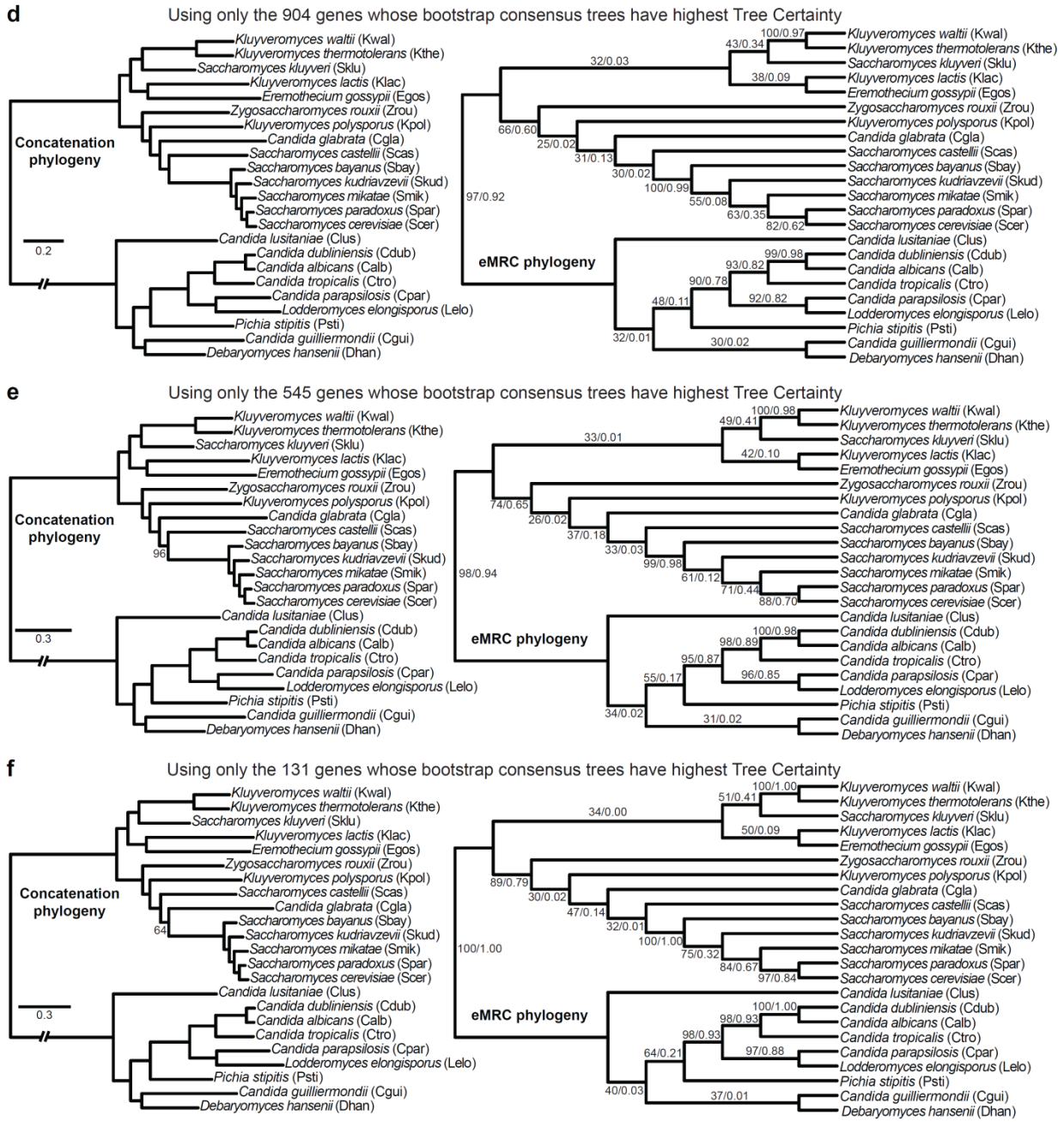

Metazoan Concatenation phylogeny
S. purpuratus
B. floridae
C. intestinalis
G. gallus
M. musculus
H. sapiens
X. tropicalis
D. rerio
H. robusta
L. gigantea
C. elegans
S. mansoni
I. scapularis
D. pulex
A. mellifera
T. castaneum
D. melanogaster
B. mori
N. vectensis
T. adhaerens
M. brevicollis
0.3
*/10
59/12
*/23
*/76
*/67
*/94
*/97
*/19
95/25
*/10
*/23
*/50
92/05
*/39
*/59
*/72
98/29
*/35

f

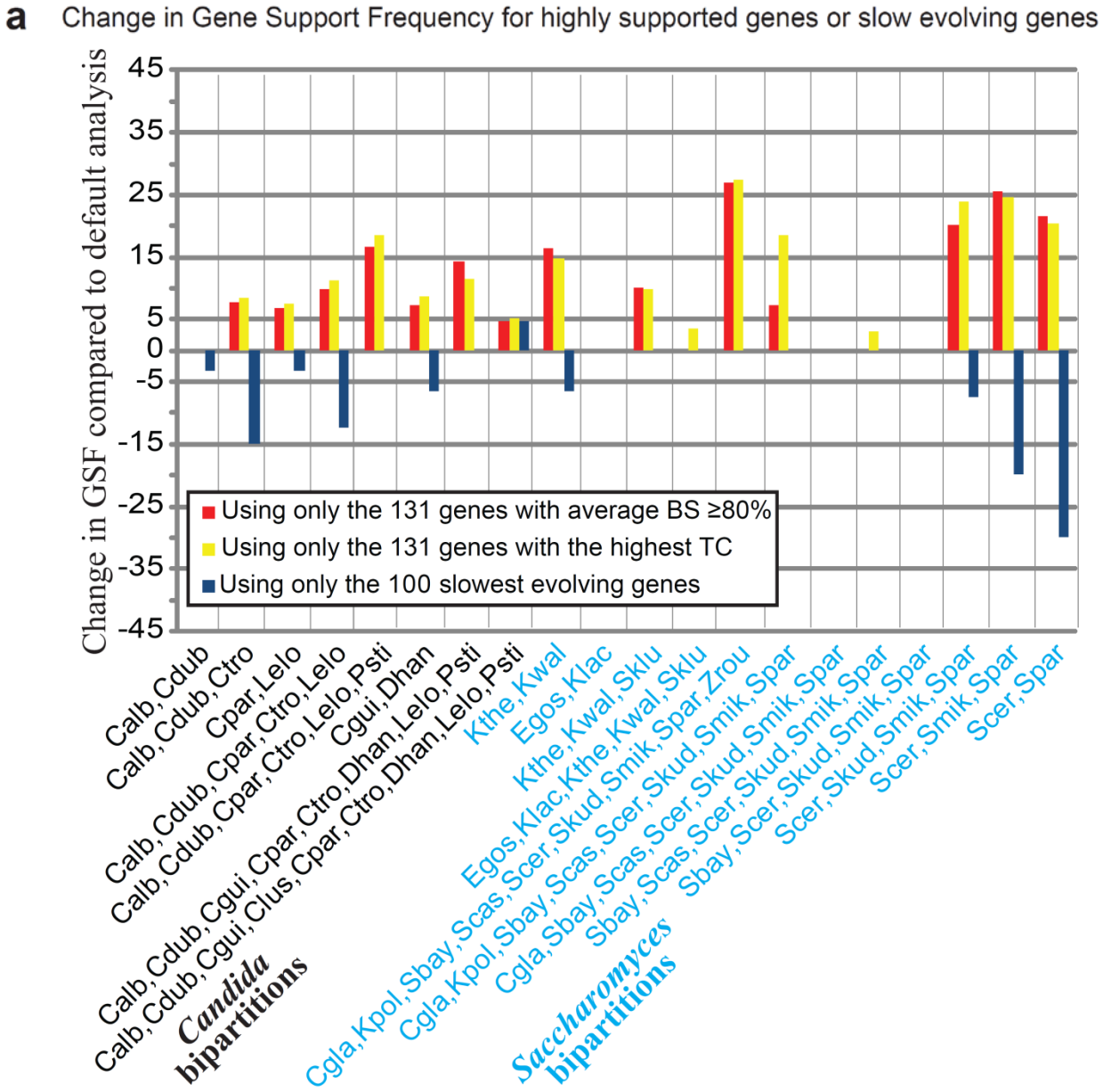

Internode Certainty values for all internodes of the Metazoan species phylogeny
Internode Certainty
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Drer,Ggal,Hsap,Mmus,Xtro
Hsap,Mmus
Amel,Bmor,Dmel,Tcas
Ggal,Hsap,Mmus
Amel,Bmor,Dmel,Dpul,Tcas
Amel,Bmor,Dmel,Dpul,Isca,Tcas
Ggal,Hsap,Mmus,Xtro
Amel,Bmor,Cele,Dmel,Dpul,Isca,Sman,Tcas
Bflo,Cint,Drer,Ggal,Hsap,Mmus,Spur,Xtro
Nvec,Tadh
Bmor,Dmel
Amel,Bmor,Cele,Dmel,Dpul,Hrob,Isca,Lgig,Sman,Tcas
Hrob,Lgig
Cele,Sman
Bmor,Dmel,Tcas
Cint,Drer,Ggal,Hsap,Mmus,Xtro
Bflo,Cint,Drer,Ggal,Hsap,Mmus,Xtro
Mbre,Nvec,Tadh

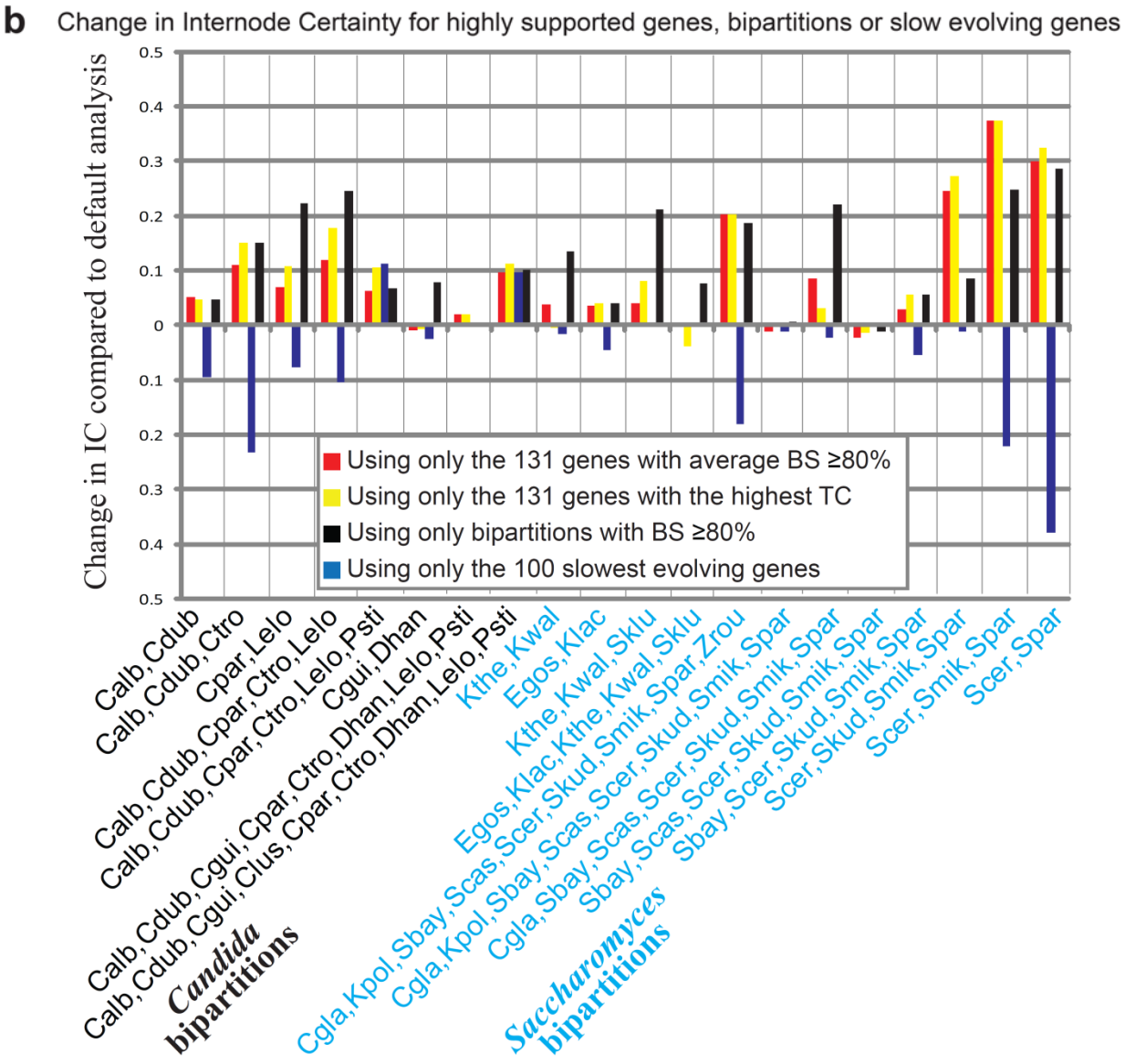
Supplementary Figures 4.11 | Selection of genes whose bootstrap consensus trees have high average Bootstrap Support (avBS) or Tree Certainty (TC) has a large, positive effect on GSF and IC values of internodes of the yeast phylogeny inferred by concatenation and eMRC approaches. Each panel shows the yeast species phylogeny inferred from concatenation analysis (left panel) and from extended majority rule consensus (eMRC) analysis (right panel) following the selection of genes whose trees have high average bootstrap support (BS) or Tree Certainty (TC). All internodes of phylogenies inferred by concatenation received 100% bootstrap support unless otherwise indicated. Values near internodes of phylogenies inferred by eMRC analysis correspond to gene support frequency and internode certainty, respectively. **a**, Concatenation (left) and eMRC (right) phylogenies of the 904 genes whose gene trees have average BS ≥60%. **b**, Concatenation (left) and eMRC (right) phylogenies of the 545 genes whose gene trees have average BS ≥70%. **c**, Concatenation (left) and eMRC (right) phylogenies of the 131 genes whose gene trees have average BS ≥80%. **d**, Concatenation (left) and eMRC (right) phylogenies of the 904 genes with the highest TC. **e**, Concatenation (left) and eMRC (right) phylogenies of the 545 genes with the highest TC. **f**, Concatenation (left) and eMRC (right) phylogenies of the 131 genes with the highest TC.


a
Using only the 904 genes whose bootstrap consensus trees have average Bootstrap Support ≥60%
Concatenation phylogeny
eMRC phylogeny
b
Using only the 545 genes whose bootstrap consensus trees have average Bootstrap Support ≥70%
Concatenation phylogeny
eMRC phylogeny
c
Using only the 131 genes whose bootstrap consensus trees have average Bootstrap Support ≥80%
Concatenation phylogeny
eMRC phylogeny
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Candida lusitaniae (Clus)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
0.3

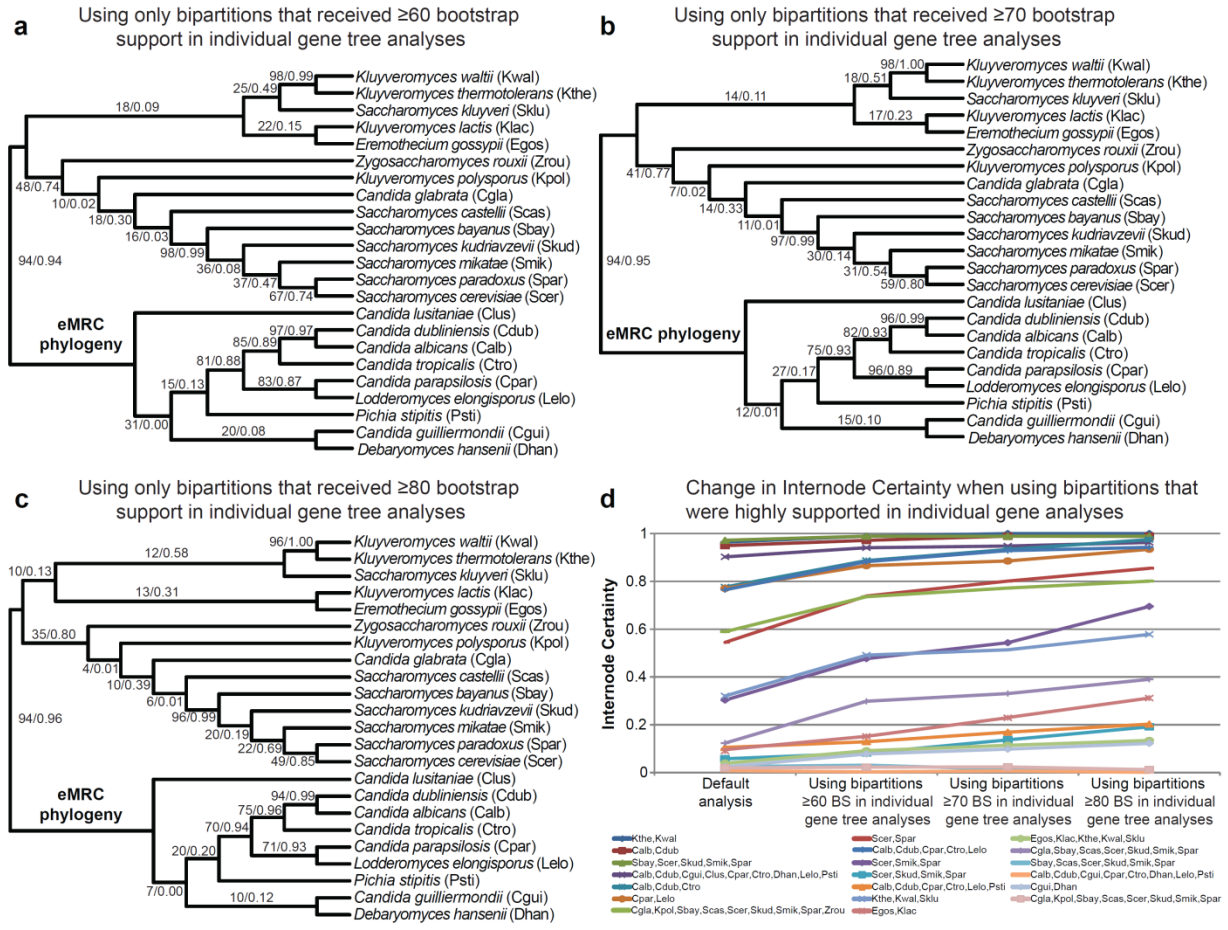

d
Using only the 904 genes whose bootstrap consensus trees have highest Tree Certainty
Concatenation phylogeny
0.2
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Candida lusitaniae (Clus)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
eMRC phylogeny
97/0.92
66/0.60
25/0.02
31/0.13
32/0.03
30/0.02
100/0.99
55/0.08
63/0.35
82/0.62
43/0.34
100/0.97
38/0.09
99/0.98
93/0.82
90/0.78
48/0.11
92/0.82
32/0.01
30/0.02
e
Using only the 545 genes whose bootstrap consensus trees have highest Tree Certainty
Concatenation phylogeny
0.3
96
eMRC phylogeny
98/0.94
74/0.65
26/0.02
37/0.18
33/0.03
99/0.98
61/0.12
71/0.44
88/0.70
33/0.01
49/0.41
100/0.98
42/0.10
98/0.89
100/0.98
95/0.87
96/0.85
55/0.17
34/0.02
31/0.02
f
Using only the 131 genes whose bootstrap consensus trees have highest Tree Certainty
Concatenation phylogeny
0.3
64
eMRC phylogeny
100/1.00
89/0.79
30/0.02
47/0.14
32/0.01
100/1.00
75/0.32
84/0.67
97/0.84
34/0.00
51/0.41
100/1.00
50/0.09
98/0.93
100/1.00
98/0.93
97/0.88
64/0.21
40/0.03
37/0.01

Supplementary Figures 4.12 | Selecting highly supported genes or bipartitions has a large, positive effect on GSF and IC values of internodes of the yeast phylogeny. The X-axis shows the 20 bipartitions present in the yeast phylogeny suggested by concatenation analysis and the Y-axis the percent change in Gene Support Frequency (GSF) and Internode Certainty (IC) observed for each bipartition between the treatment (selection of highly supported genes or internodes) and the default analysis. Only GSF changes ≥3% and IC changes ≥0.03 are shown. The red bars correspond to changes in IC when using only the 131 genes with average bootstrap support ≥80%, the yellow bars correspond to changes in IC when using only the 131 genes with the highest Tree Certainty, the black bars correspond to changes in IC when using only those bipartitions found in the bootstrap consensus trees of individual genes that had bootstrap support ≥80%, and the blue bars correspond to changes in IC when using only the 100 slowest-evolving genes. **a**, Change in GSF for highly supported genes or slow evolving genes. **b**, Change in IC for highly supported genes, bipartitions or slow evolving genes.

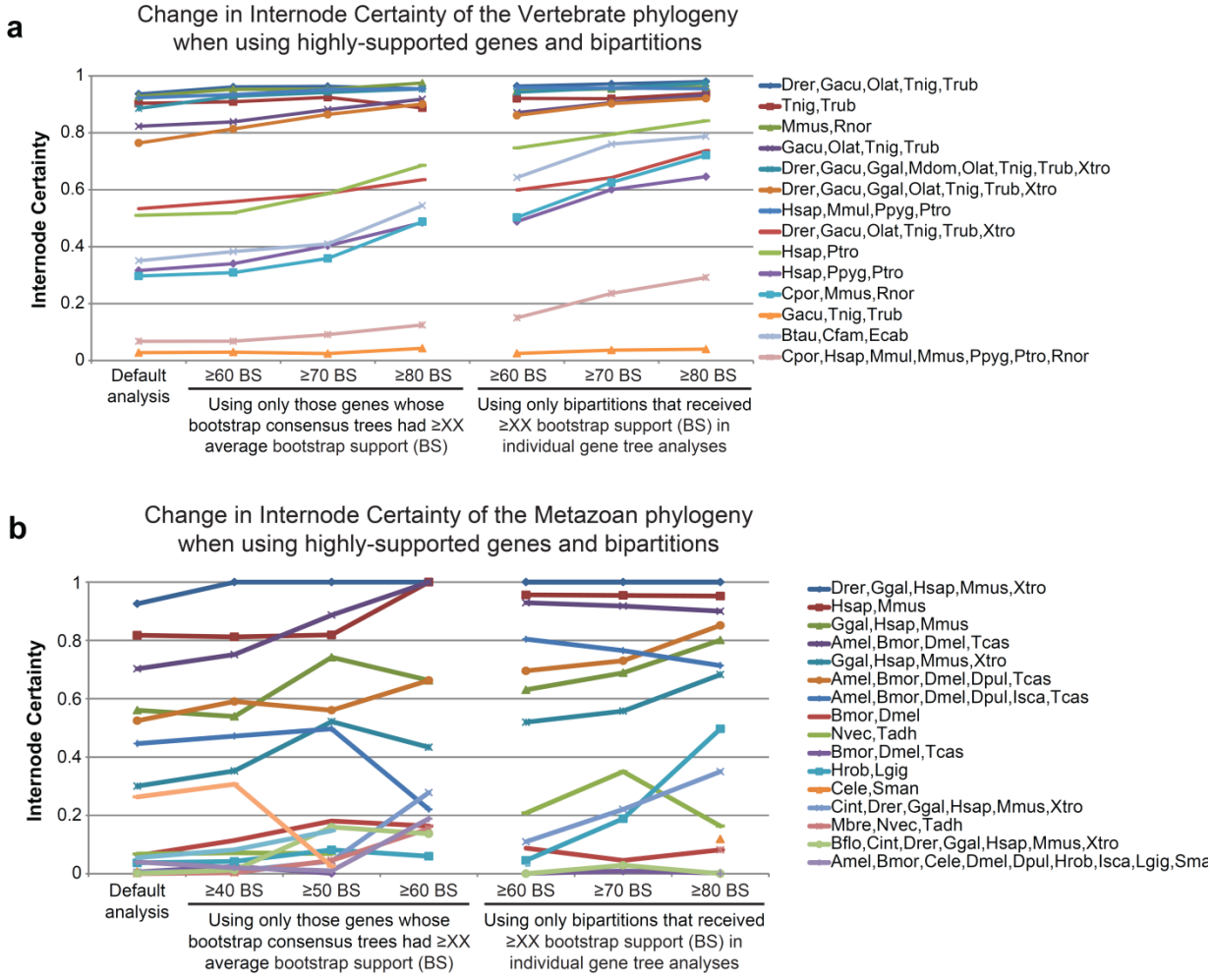



b
Change in Internode Certainty for highly supported genes, bipartitions or slow evolving genes
Change in IC compared to default analysis
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
Using only the 131 genes with average BS ≥80%
Using only the 131 genes with the highest TC
Using only bipartitions with BS ≥80%
Using only the 100 slowest evolving genes
Calb,Cdub
Calb,Cdub,Ctro
Cpar,Lelo
Calb,Cdub,Cpar,Ctro,Lelo
Calb,Cdub,Cpar,Ctro,Lelo,Psti
Cgui,Dhan
Calb,Cdub,Cgui,Cpar,Ctro,Dhan,Lelo,Psti
Calb,Cdub,Cgui,Clus,Cpar,Ctro,Dhan,Lelo,Psti
Candida bipartitions
Kthe,Kwal
Egos,Klac
Kthe,Kwal,Sklu
Egos,Klac,Kthe,Kwal,Sklu,Spar,Zrou
Cgla,Kpol,Sbay,Scas,Scer,Skud,Smik,Spar
Cgla,Kpol,Sbay,Scas,Scer,Skud,Smik,Spar
Cgla,Sbay,Scas,Scer,Skud,Smik,Spar
Sbay,Scas,Scer,Skud,Smik,Spar
Sbay,Scer,Skud,Smik,Spar
Scer,Skud,Smik,Spar
Scer,Smik,Spar
Scer,Spar
Saccharomyces bipartitions

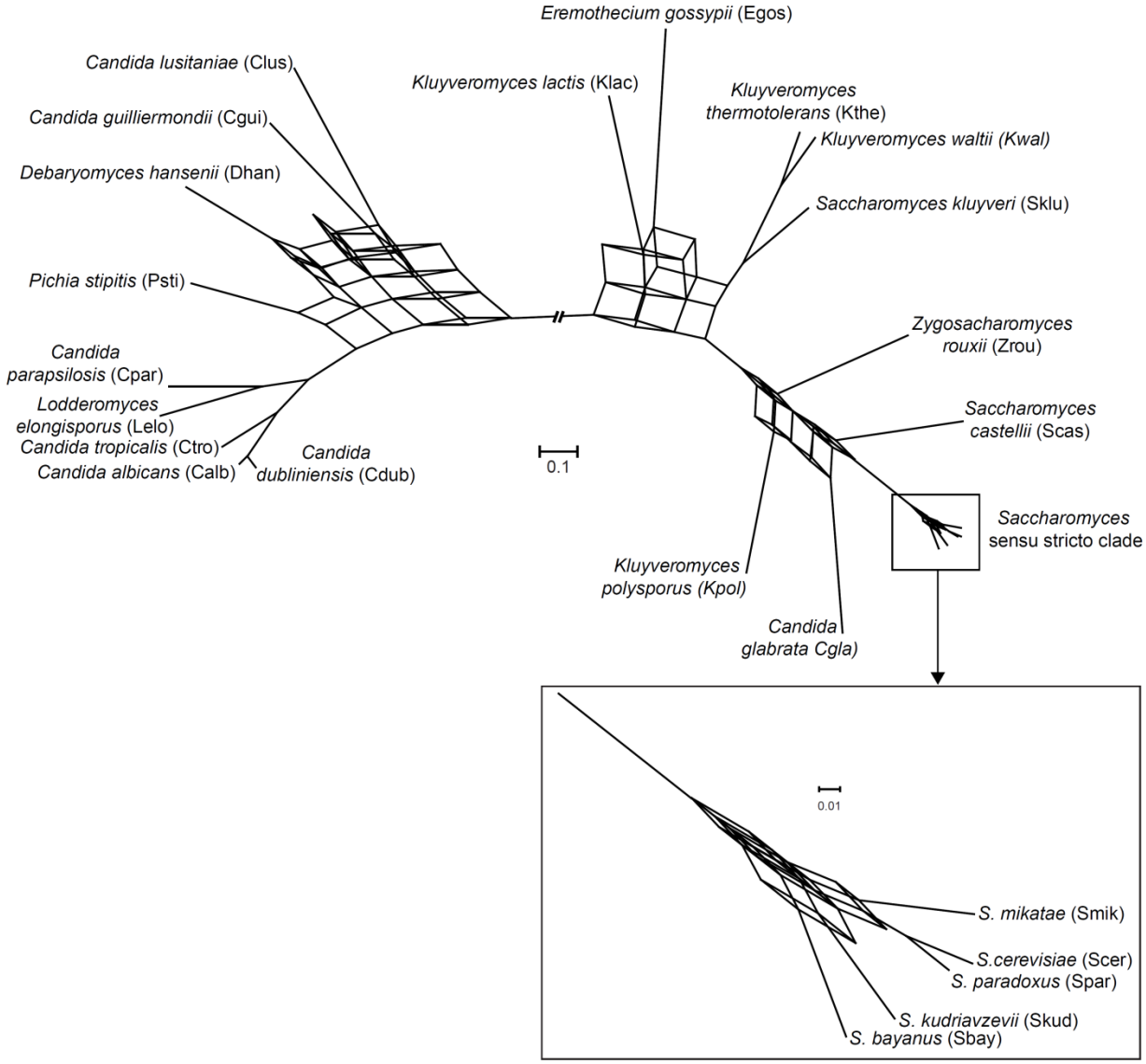
Supplementary Figures 4.13 | Selection of highly supported bipartitions from the bootstrap consensus trees of individual genes has a large, positive effect on the IC values of internodes of the yeast phylogeny inferred by the eMRC approach. The first three panels show the yeast species phylogeny inferred from extended majority rule consensus (eMRC) analysis following the selection of bipartitions that had high bootstrap support (BS) in the bootstrap consensus trees of individual genes. Values near internodes correspond to the percentage of bootstrap consensus trees of individual genes in which this specific bipartition received high BS and to internode certainty (IC), respectively. **a**, The eMRC phylogeny inferred from selecting bipartitions that had BS ≥60% in individual gene analyses. **b**, The eMRC phylogeny inferred from selecting bipartitions that had BS ≥70% in individual gene analyses. **c**, The eMRC phylogeny inferred from selecting bipartitions that had BS ≥80% in individual gene analyses. **d**, Plot that illustrates the change in IC of internodes relative to the values obtained in the default analysis associated with the use of bipartitions that had high bootstrap support (BS) in the bootstrap consensus trees of individual genes. Each line of different color depicts the IC value obtained for a given internode in the default analysis (Fig. 1a), when using only bipartitions that had BS ≥60%, BS ≥70%, and BS ≥80%.



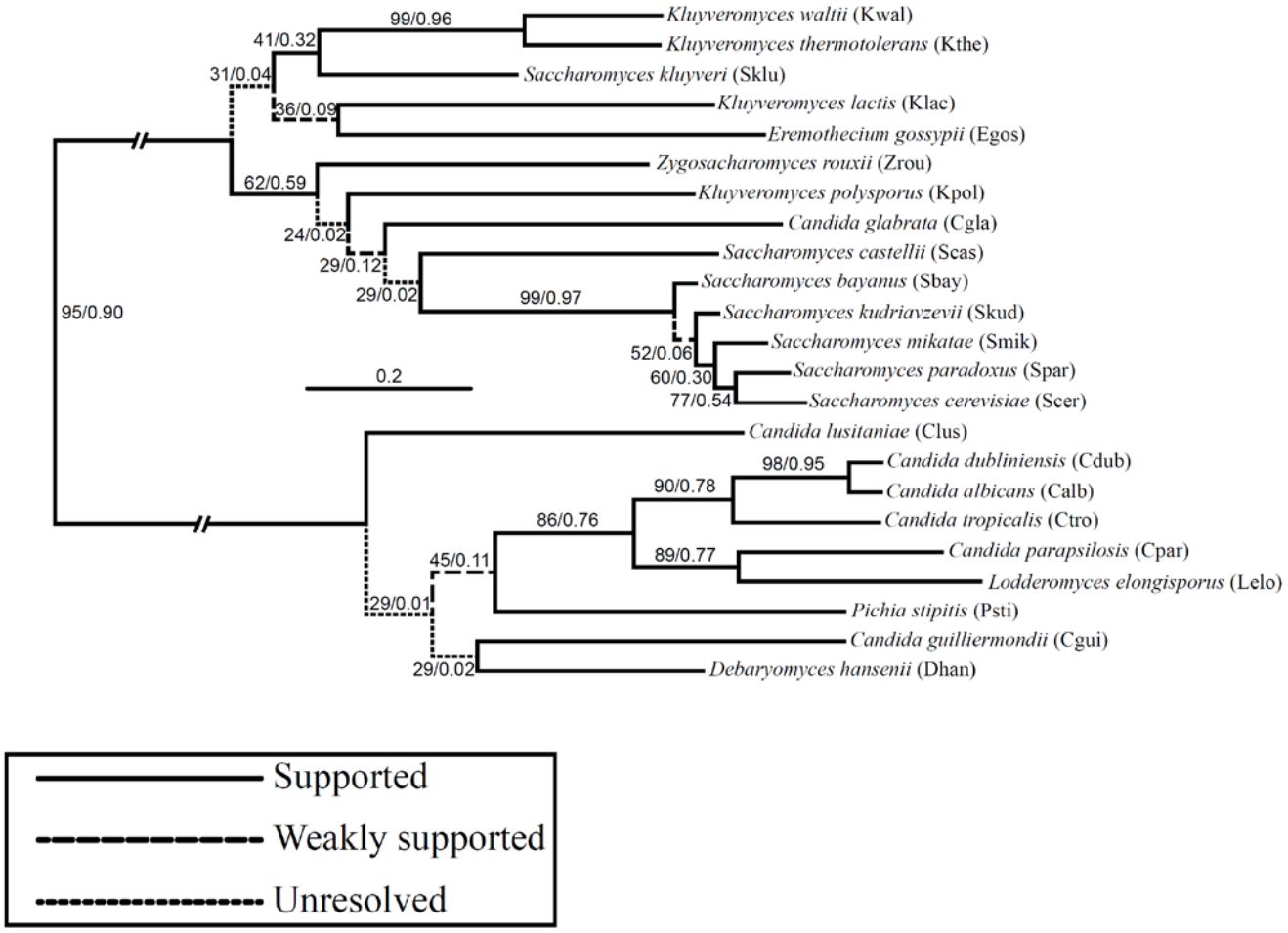
Supplementary Figures 4.14 | Selection of highly supported genes and bipartitions has a large, positive effect on IC values of internodes of the vertebrate and metazoan phylogenies. a, Plot that illustrates the change in IC of internodes of the vertebrate phylogeny relative to the values obtained in the default analysis associated with the use of genes whose bootstrap consensus trees have high average bootstrap support (BS) or with the use of bipartitions that had high BS in the bootstrap consensus trees of individual genes. Each line of different color depicts the IC value obtained for a given internode in the default analysis (Supplementary Fig. S10b), when using only genes with average BS ≥60%, BS ≥70%, and BS ≥80%, as well as when using only bipartitions that had BS ≥60%, BS ≥70%, and BS ≥80%. **b**, Plot that illustrates the change in IC of internodes of the metazoan phylogeny relative to the values obtained in the default analysis associated with the use of genes whose bootstrap consensus trees have high average bootstrap support (BS) or with the use of bipartitions that had high BS in the bootstrap consensus trees of individual genes. Each line of different color depicts the IC value obtained for a given internode in the default analysis (Supplementary Fig. S10e), when using only genes with average BS ≥40%, BS ≥50%, and BS ≥60%, as well as when using only bipartitions that had BS ≥60%, BS ≥70%, and BS ≥80%.



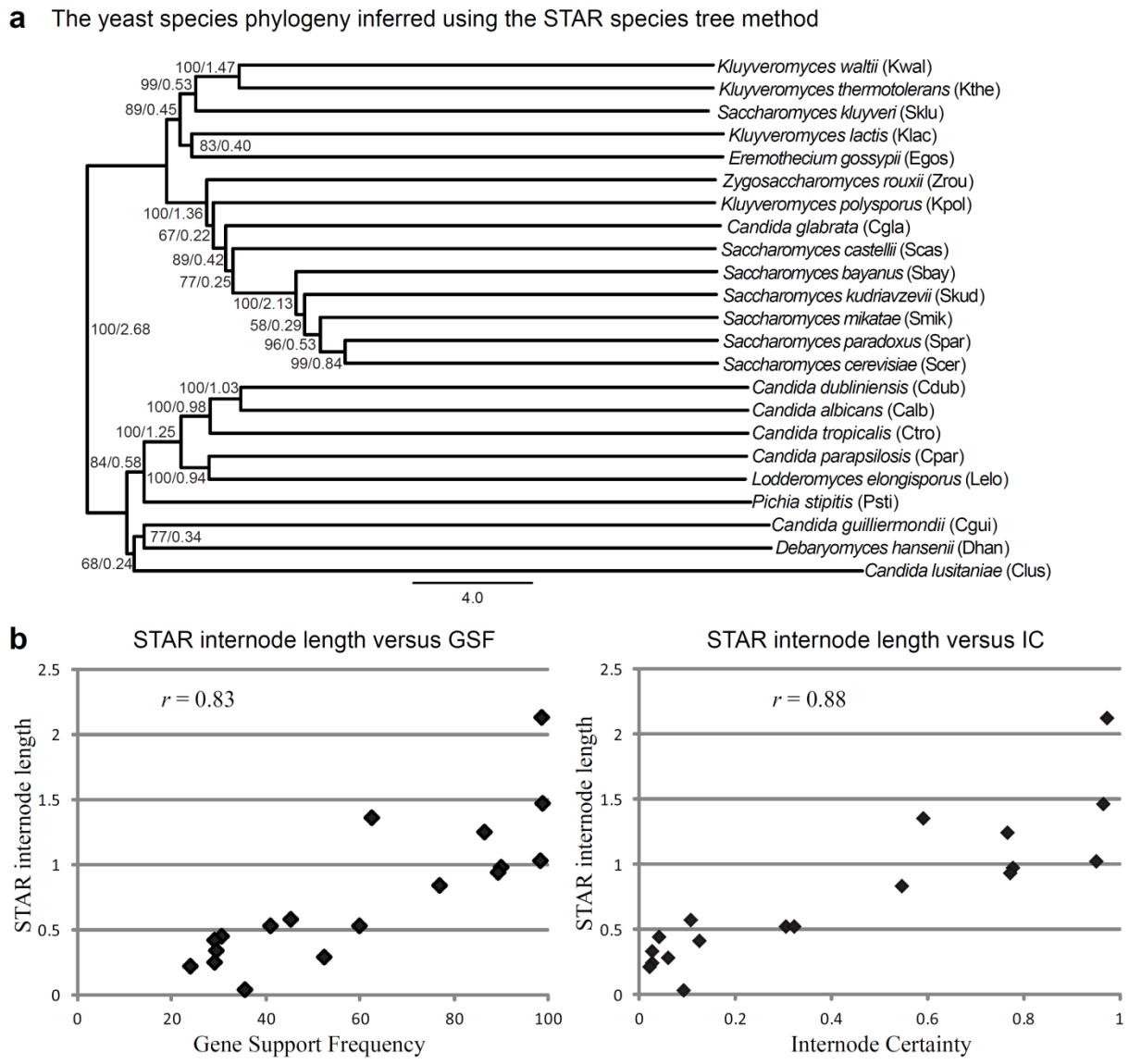
Supplementary Figures 4.15 | The phylogenetic consensus network that describes the 1,070 yeast gene histories. The consensus network inferred using the 1,070 maximum likelihood gene trees under the median network construction algorithm in the SplitsTree4 software. Boxes in the network denote internodes that harbor significant conflict, with the length of each branch in each box being proportional to the number of GTs that support it. Only branches that are present in at least 10% of the GTs are shown in the network.



Supplementary Figures 4.16 | Supported, weakly supported, and unresolved internodes in the yeast phylogeny. Values near internodes correspond to gene support frequency and internode certainty, respectively calculated from the 1,070 yeast gene histories. Note that the validity of certain internodes marked as "unresolved" is supported by independent data (e.g., rare genomic changes).



Supplementary Figures 4.17 | The yeast phylogeny inferred using a "species tree" method that accounts for variation between the 1,070 gene histories is highly supported and has extremely short internodes whose coalescent unit lengths are highly correlated with gene support frequency and internode certainty values. Using the 1,070 gene dataset, we inferred a yeast species phylogeny under the coalescent model and average ranks of gene coalescence times, as implemented in the STAR species tree method. **a**, The yeast species phylogeny under the coalescent. Values near internodes correspond to bootstrap support and internode length in coalescence units, respectively. The inferred topology is identical to the phylogeny shown in Figures 4.1a, except with respect to the placement of *Candida lusitaniae*. **b**, The lengths of internodes in the phylogeny inferred using the STAR species tree method, measured in average coalescent units, is highly correlated with internodes' Gene Support Frequency (left panel) and Internode Certainty (right panel) values. The strength of each correlation is indicated by *r*, Pearson's correlation coefficient.



CHAPTER V

EXAMINATION OF FACTORS THAT INFLUENCE PHYLOGENETIC INCONGRUENCE IN A YEAST MODEL CLADE

Leonidas Salichos[1] and Antonis Rokas[1,2]

[1]*Department of Biological Sciences, Vanderbilt University, VU Station B #35-1634, Nashville, TN, 37235, United States of America*
[2]*Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA*

ABSTRACT

The reconstruction of ancient divergences continues to confound molecular phylogeneticists due to the presence of substantial amounts of incongruence between gene trees. Apart from biological events that can cause gene histories to differ from the species one, several previous studies have suggested a link between the functional characteristics of genes (e.g., GC content) and their degree of incongruence. Identifying the influence of different factors on phylogenetic inference is critical because they render phylogenetic studies vulnerable to systematic error and model misspecification but also because they can help identify genes that are more informative markers of phylogeny. In this study, we examined the degree to which 10 diverse functional and evolutionary characteristics -extracted from 1,070 groups of orthologous genes (orthogroups) from 23 yeast species- are correlated with 4 phylogenetic gene measures of incongruence, two referring to the orthogroup's conflicting phylogenetic signal and the other two to the orthogroup's tree when compared against all other gene trees. Overall, we found that GC content, the percentage of variable sites, codon bias and codon adaptation, provided the highest correlation with the levels of gene incongruence both within and across gene trees. Genes with low GC content or low GC variance across taxa, with higher percentage of variable sites or relative long branches, as well as genes with lower codon bias, seem to be less incongruent. On the other end of the spectrum, genes that exhibit few or many physical interactions, much conserved genes, and genes with high or low codon adaptation appear to increase gene incongruence. A principal component regression analysis showed that variance based on different functional factors can explain approximately 15-20% of the total variance in gene tree incongruence. Thus, even though several functional and evolutionary properties of genes contribute significantly to the incongruence between a given gene tree and the species

phylogeny, our results indicate that a large amount of the observed incongruence remains unexplained. Selecting genes based on their phylogenetic properties remains the safest way to reduce incongruence.

INTRODUCTION

Phylogenomic data matrices from diverse clades of the tree of life typically exhibit extensive phylogenetic incongruence between the trees of the individual genes that comprise them[1–12]. In general, the reasons for observing phylogenetic incongruence may be characterized as either analytical or biological. Biological reasons involve cases where the history of genes is genuinely different from the species phylogeny[13]. In contrast, analytical reasons involve cases where the data are not representative of the whole population[5,14,15]or the models of evolution are misspecified[16,17], and are typically distinguished into two types, sampling error and systematic error[16].

Several factors contribute to sampling and systematic error. Factors that may increase sampling error are taxon sampling[6,18,19] and the availability of data[15,16,20,21]. Importantly, incongruence stemming from sampling errors can be detected and overcome by including more data; the same is typically true of biological reasons. In contrast, incongruence stemming from systematic error cannot be overcome by increases in the amount of data[5,22]. Example of factors that may contribute to systematic error are base composition and compositional heterogeneity, sequence length, mutation rate,the branching pattern of a phylogeny, branch length and others[18,23–31,32]. In their 2003 study, using a matrix of 106 orthologous groups, Rokas et al. identified that bootstrap support obtained from individual gene tree analyses was significantly correlated with gene properties such as gene size, long branches, GC content, or the percentage of variable sites[5].

Subsequent analyses of the same dataset showed an additional dependence on gene stationarity (genes exhibiting similar base frequencies among taxa)[33] and that, occasionally, the length of the branches resulted in the misplacement for some of the eight taxa[19,21].

In Salichos and Rokas 2013, using a dataset of 1,070 groups of orthologous genes or orthogroups constructed based on syntenic information, sequence similarity and manual curation from 23 yeast species, we showed that gene tree incongruence was highly correlated with short internodes at the base of the phylogeny. By comparing 1,070 yeast genes against the species phylogeny, we discovered great differences among the gene trees, as well as between the gene trees and the species phylogeny. Moreover, with the use of two novel phylogenetic measures that quantify incongruence, namely internode certainty (IC) and tree certainty (TC), we showed that by selecting genes or gene tree bipartitions that exhibit high TC (genes) or IC (bipartitions) values, we were able to decrease the levels of incongruence much more than when we applied standard practices such as the removal of rogue taxa, genes or sites. However, the factors that contribute to these high levels of incongruence and may render a gene more informative still remain largely unexplored.

In this study, we first estimate the correlation between a set of 10 functional gene factors (% GC content, variance in GC content across sequences of the orthogroup, % of variable sites, sum of gene tree branch lengths, codon bias and codon adaptation[34], number of physical or genetic interactions per gene - retrieved from the Saccharomyces Genome Database[35]-, gene expression - as estimated from Busby et al 2011[36]-, number of paralogs per gene) and a set of 4 phylogenetic measures (including gene Tree Certainty (TC)[37], orthogroup tree's average bootstrap support (AvBS)[38], gene tree's Robinson-Foulds[39] mean distance (mRF) and Robinson-Foulds variance of distances per gene). My results indicate a significant correlation between phylogenetic

incongruence and many functional gene properties, including % GC content, % variable sites, codon bias and codon adaptation. Second, using a sliding-window approach, we test the behavior of these functional factors in terms of contributing or not to incongruence, across a range of the gene values in ascending order, by constructing majority consensus trees for each sliding window of 100 genes. These analysis show that genes with low GC content or low GC variance across taxa, with higher percentage of variable sites or relative long branches and genes with lower codon bias, seem to provide MRC trees with higher TC. On the contrary, genes linked with few or many physical interactions, as well as much conserved genes provide MRC trees with very low TC.T hird, using a principal component regression, we examine the linearity of functional factors against mRF, a measure that demonstrates the topological distance of one gene against all others. Based on this analysis, we find that approximately 18% of total topological variance can be directly explained by functional gene factors. However, even though these factors may play a significant role in driving gene incongruence, they still cannot be utilized as effective markers for selecting informative genes

RESULTS

Using a dataset of 1,070 orthogroups, we assigned for each orthogroup a set of functional measures like % GC content, % GC variance across the orthogroup, % of variable sites, branch length of its gene tree, codon bias, codon adaptation, number of genetic or physical interactions, gene expression, and number of gene paralogs. Values for the last six factors were obtained based on the *Saccharomyces cerevisiae* gene ortholog. For each orthogroup, we also calculated 4 measures of incongruence including gene TC, AvBS, mRF and RF variance. Gene TC and AvBS refer to the incongruence observed based on the orthogroup's conflicting phylogenetic

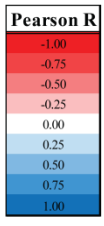
signals, while mRF and RF distance variance refer to how incongruent is the orthogroup's gene tree when compared against all other genes trees.

Significant Link Between Gene Factors and Phylogenetic Gene Incongruence

Initially, we estimated the correlation among the set of 10 factors and 4 phylogenetic measures. In table 5.1, we summarize the results for every correlation analysis between all gene factors and phylogenetic measures. All phylogenetic measures showed a very high correlation (positive or negative) with each other. Significant correlation was also observed between functional factors like % GC content, % of variable sites, codon bias, codon adaptation, sum of branch lengths, the number of physical interactions and gene expression. Finally, % GC content, % of variable sites, codon bias and codon adaptation provided a significant correlation with values of mRF having a Pearson's coefficient of > 0.2, suggesting involvement in gene's incongruence.

Table 5.1. A correlation analysis. For every orthogroup, we estimated the percentage of GC content, the variance of GC content, the percentage of variable sites, the sum of all branch lengths from the gene tree, the number of physical interactions (PI), the logPI, the number of genetic interactions (GI), the logGI, the codon adaptation index (CAI), the codon bias index (CBI), the expression levels of the *S.cerevisiae* (Scer) ortholog, the number of close paralogs to the Scer ortholog, the mean Robinson –Foulds distance against all other genes, the variance of all 1069 pairwise RF distances, the average bootstrap support for its gene tree (AvBS) and the tree's Tree Certainty. Then we calculated the correlation across different gene factors and phylogenetic measures. Values represent the Pearson's coefficient R.

	% GC	% GC var	%vs	BrLen	PI	logPI	GI	logGI	CAI	CBI	SGE	# SH	mRF	Rfvar	AvBS	gTC
% GC																
% GC variance	0.23															
% variable sites (%vs)	-0.36	0.16														
Gene Tree Branch Length	-0.35	0.21	0.66													
# of Physical Interactions (PI)	0.01	-0.10	-0.19	-0.22												
logphys	0.01	-0.14	-0.24	-0.29	0.71											
# of Genetic Interactions (GI)	-0.06	-0.01	-0.04	0.01	0.08	0.13										
logGi	-0.03	-0.03	-0.06	-0.01	0.09	0.14	0.77									
Codon Adaptation Index (CAI)	0.28	-0.15	-0.42	-0.46	0.23	0.31	-0.03	-0.02								
Codon Bias Index (CBI)	0.42	-0.08	-0.47	-0.54	0.23	0.31	-0.03	-0.01	0.93							
Scer gene expression (SGE)	0.20	-0.06	-0.17	-0.18	0.06	0.10	0.03	0.00	0.41	0.36						
# of Scer homologs (#SH)	0.05	-0.03	-0.09	-0.03	0.01	0.01	0.08	0.08	0.05	0.06	-0.01					
RF mean distance (mRF)	0.34	0.20	-0.30	-0.19	-0.01	-0.03	-0.07	-0.06	0.24	0.28	0.08	0.05				
RF variance	-0.33	-0.18	0.28	0.21	0.02	0.05	0.07	0.06	-0.22	-0.27	-0.08	-0.02	-0.87			
Average Bootstrap (AvBS)	-0.31	-0.25	0.27	0.15	0.00	0.03	0.02	0.03	-0.18	-0.21	-0.07	0.00	-0.66	-0.21		
gene Tree Certainty (gTC)	-0.35	-0.26	0.31	0.20	-0.02	0.01	0.02	0.02	-0.22	-0.25	-0.09	-0.01	-0.66	-0.25	0.93	



A Sliding Window Approach

Using a sliding-window analysis with a step of 20 and a window of 100 genes, we ordered all 1,070 genes in an ascending order based on their value for each different functional factor. Then, for every window of 100 genes and their respective gene-trees, each time we inferred the majority-rule consensus tree (MRC), while also calculating the MRC's tree certainty (TC). We repeated this process for each phylogenetic measure. Our results indicated that selecting genes with high GC content or high GC variance increases incongruence (measured by the lower TC values). Genes whose trees have short branch lengths also increase incongruence, but the longest branches may provide lower TC too. The same trend in a greater magnitude seems to apply for the number of physical interactions per gene, whereby selecting genes with small or large numbers of physical interactions results in lower TC values. Genes with higher values of Codon Adaptation Index or Codon Bias Index appear to provide higher values of TC with very high correlation, but the effect on TC does not deviate much. Finally, genes with low expression seem to provide higher TC values, but overall correlation is not that high (figure 5.1). As expected, phylogenetic properties showed an extremely high correlation with TC, while the first 100 genes with the smallest mRF presented the highest TC across all datasets (figure 5.2).

Figure 5.1. A sliding window approach for functional factors. Using a sliding window approach, with a window of 100 genes and a step of 10, we plot the majority consensus tree's Tree Certainty (y axis) vs the average (x axis) value of 100 gene's GC content, variance of GC content, percentage of variable sites, tree's total branch length, number of physical interactions, genetic interactions, Codon adaptation index, Codon Bias Index and gene expression. By plotting them in an ascending order, overall, none factor achieves particular high TC values, although many factors appear to behave differently across the spectrum of their values. Genes with lower GC content appears to show the least amounts of incongruence (close to 0.5). On the contrary, much conserved genes, appear to provide very high levels of incongruence (close to 0.3).

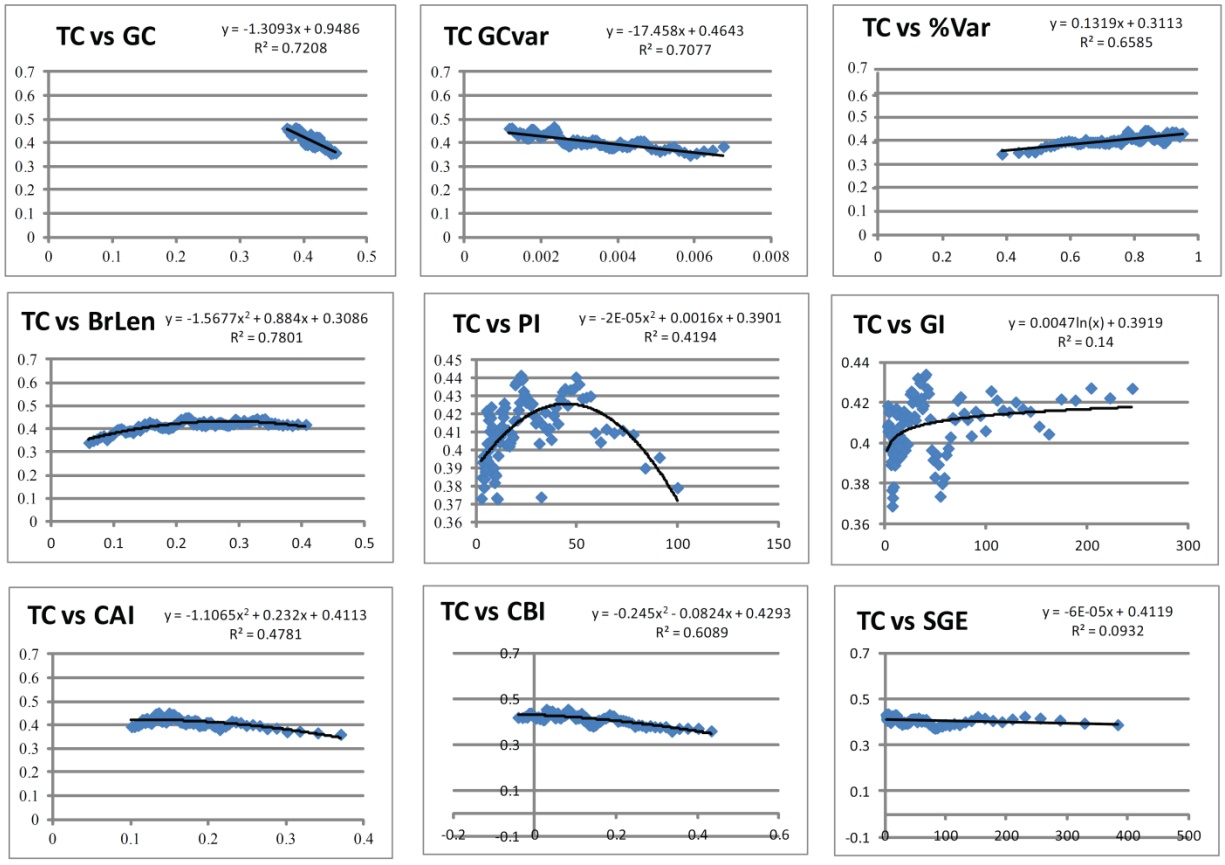
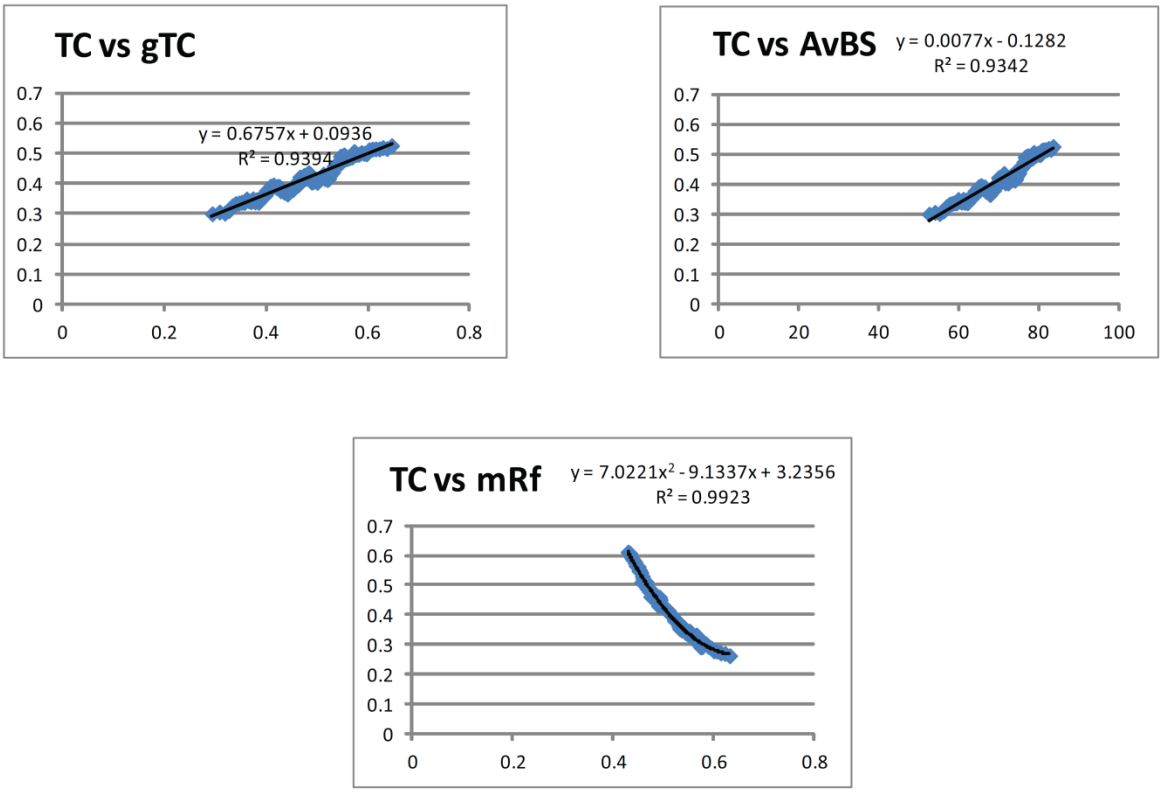


Figure 5.2. A sliding window approach phylogenetic measures. Using a sliding window approach, with a window of 100 genes and a step of 10, we plot the majority consensus tree's Tree Certainty (y axis) vs the average (x axis) value of 100 gene tree's Tree Certainty (gTC), average bootstrap support (AvBS) and mean Robinson-Foulds distance. By plotting them in an ascending order, overall, all phylogenetic measures achieve high TC values (demonstrating low incongruence). mRF, not only provides the highest degree of correlation, but by selecting genes with low mRF gives TC > 0.6.



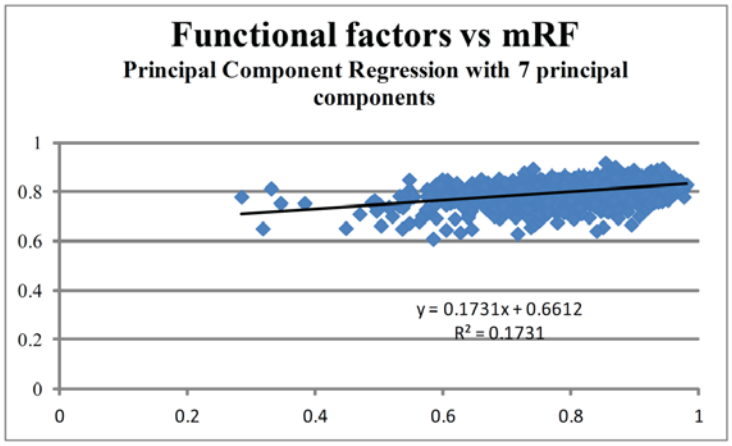
Regression analysis

To decrease the number of dimensions in our data set, we performed a Principal Component Analysis (PCA) on the set of functional factors. Then, given the extremely high correlation between mRF and majority rule consensus TC, we used mRF to perform a regression analysis against those components, as well against the entire set of functional factors. Overall, the highest degree of variance was explained using 7 principal components accounting for ~18% of the total

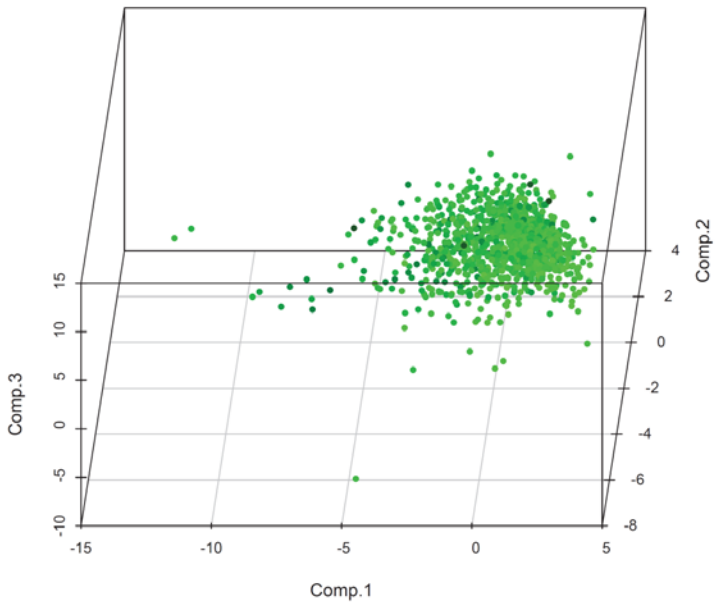
variance (figure 5.3a). In figure 5.3b, we show a 3d representation of the three first Principal components, colored by mRF.

Figure 5.3. Principal Component Regression analysis a) By performing a principal component regression analysis of all functional factors against each gene's mean Robinson-Foulds distance we show that about 17% of the total variance of gene incongruence in the dataset can be explained by functional factors In b) we present a 3-d scatterplot of the first 3 principal components, where each gene's values has been colored based on their mRF value. Darker colors signify higher mean distance and therefore a more evolutionary diverged gene.

a



b



Testing the predictability of functional factors to predict gene incongruence

In the last step of our analysis, we randomly divided our dataset into two groups of 700 (training dataset) and 370 genes (testing). Using the training dataset, we trained a simple neural network of different layers and asked to predict each orthologous group's mRF value based on its functional properties. Then, we performed a simple correlation analysis against the true mRF values for the test dataset. The best neural network consisted of 20 hidden layers, and provided an R^2 of ~0.15.

DISCUSSION

The existence of topological conflict in recent phylogenomic analyses on ancient divergences[40–44]together with an increasing number of studies that demonstrate extremely high levels of gene incongruence[5,10,12,38,45]continue to confound phylogeneticists. In general, phylogenetic incongruence can be attributed into two main reasons: biological or analytical[16]. As biological reasons, we consider events when some genes have a different history than their respective species such as incomplete lineage sorting, hybridization or horizontal gene transfer[45–47]. As, analytical, we consider the type of error that stems from either small sample sizes[5,14,15] or the misspecification of the evolutionary model[16,17].

By performing a principal component regression analysis, we found that at least 17% of the total variance of gene-tree incongruence can be directly attributed to analytical reasons and gene factors like the percentage of GC content, codon bias, codon adaptation and percentage of variable sites. However, a large amount of this variance remains unexplained. Given that our analysis consisted of 1070 orthogroups based on syntenic information and without any missing data, the effect of sampling error[5,22](small sample size), horizontal gene transfer and paralogy[48]

should be insignificant. Thus, we consider that the remaining variance of gene tree incongruence could be potentially explained by the existence of additional analytical reasons, the loss of phylogenetic signal in short internodes deep in time, together with biological reasons such as incomplete lineage sorting or hybridization. However, distinguishing between the three later reasons that may drive gene incongruence is an extremely difficult puzzle. One typical example is the high phylogenetic conflict observed for the topologies of *S. bayanus* and *S. kudriavzevii* [38]. Our current models have a great difficulty in ascertaining whether this conflict is the result of hybridization or incomplete lineage sorting (but see [49,50]).

To tackle incongruence, several phylogenomic studies have adopted various approaches (see Salichos and Rokas 2013) including the selection of only a subset of genes based on specific gene properties. Such properties may refer to various gene factors, for example retaining only slowly involving genes[40,41,51–54], the use of gene markers[55], 'good' genes that support known topologies[55,56], as well as stationary genes[33]. In our analysis, we explored the behavior of 10 such functional gene factors (and their combination) but we were not able to identify any factor that stands out and could serve as reliable marker for selecting informative genes. Furthermore, we found that, in some cases, selecting for highly conserved genes could be detrimental in resolving ancient divergences. In contrast, by selecting genes based on phylogenetic measures such as gene TC, AvBS or mRF we were able to observe a dramatic decrease in gene incongruence.

METHODS

The dataset

To perform our analysis, we used the gene dataset from Salichos and Rokas 2013. This dataset consists of 1070 orthologous groups, without any missing data, constructed using synteny and orthology information present in the YGOB[57] and CGOB[58] databases from 23 yeast genomes.

Analysis and calculation of phylogenetic properties

Genes were aligned using MAFFT[59], the best-fit evolutionary model for each gene-tree was determined using ProtTest[60], and the maximum likelihood tree was estimated using RAxML. Moreover, using RAxML[61], we also calculated Internode Certainty (IC)[37] and Tree Certainty (TC)[37] for each gene tree, TC for the sliding window approach and Robinson-Foulds(RF)[39] gene tree distance. For each gene, to estimate the mean RF distance, we averaged over all 1069 pairwise distances against every other gene. Size of homolog gene family was calculated using OrthoMCL[62] with an inflation parameter of 1.5. The percentage of variable sites per genes, average bootstrap support per gene tree and branch lengths were calculated using custom perl scripts. The sliding window approach was also performed using custom perl scripts and RAxML. CBI and CAI[34] for orthogroups were calculated using codonw[63].

Functional gene factors

The number of physical and genetic interactions per gene were retrieved from the Saccharomyces Genome Database[35]. Information concerning *Saccharomyces cerevisiae* gene expression was retrieved from Busby et al., 2011[35]. Raw counts were averaged per gene and expression values were normalized using RPKM.

Statistical analysis

All statistical and data mining analyses including data normalization, correlation, regression with Principal Component Analysis and the construction neural networks were performed using the R project[64].

ACKNOWLEDGMENTS

We thank members of the Rokas laboratory for valuable comments on this work. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. This work was supported by the National Science Foundation (DEB-0844968).

Author Contributions

L.S. and A.R. conceived and designed experiments; L.S. carried out experiments; L.S. and A.R. analysed data and wrote the paper.

The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to A.R. (antonis.rokas@vanderbilt.edu)

REFERENCES

1. Satta Y, Klein J, Takahata N. DNA archives and our nearest relative: the trichotomy problem revisited. *Mol Phylogenet Evol*. 2000;14(2):259-275.

2. Giribet G, Edgecombe GD, Wheeler WC. Arthropod phylogeny based on eight molecular loci and morphology. *Nature*. 2001;413(6852):157-161. doi:10.1038/35093097.

3. Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W. Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature*. 2001;413(6852):154-157.

4. Kopp A, True JR. Phylogeny of the Oriental Drosophila melanogaster species group: a multilocus reconstruction. *Syst Biol*. 2002;51(5):786-805. doi:10.1080/10635150290102410.

5. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*. 2003;425(6960):798-804. doi:10.1038/nature02053.

6. Rokas A, Carroll SB. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol*. 2005;22(5):1337-1344.

7. Delsuc F, Brinkmann H, Chourrout D, Philippe H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*. 2006;439(7079):965-968.

8. Marlétaz Ferdinaz, Elise Martin, Yvan Perez, Daniel Papillon, Xavier Caubit, Christopher J. Lowe, Bob Freeman, Laurent Fasano, Carole Dossat, Patrick Wincker, Jean Weissenbach YLP. Chaetognath phylogenomics: a protostome with deuterostome-like development. *Curr Biol*. 2006;16(15).

9. Matus DQ, Copley RR, Dunn CW, et al. Broad taxon and gene sampling indicate that chaetognaths are protostomes. *Curr Biol*. 2006;16(15).

10. Pollard DA, Iyer VN, Moses AM, Eisen MB. Widespread discordance of gene trees with species tree in drosophila: Evidence for incomplete lineage sorting. *PLoS Genet*. 2006;2(10):1634-1647.

11. Rokas A, Chatzimanolis S. From gene-scale to genome-scale phylogenetics: the data flood in, but the challenges remain. *Methods Mol Biol*. 2008;422:1-12. doi:10.1007/978-1-59745-581-7_1.

12. Hess J, Goldman N. Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: Yeasts revisited. *PLoS One*. 2011;6(8).

13. Baum DA. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon*. 2006;56(2):417-426. Available at: papers2://publication/uuid/51000653-3ADB-4080-8AEA-E54898ADE024.

14. Bull J. J., J. P. Huelsenbeck, Clifford W. Cunningham DLS and PJW. Partitioning and Combining Data in Phylogenetic Analysis. *Syst Biol*. 1993;42(3):384-397.

15. Cummings MP, Otto SP, Wakeley J. Sampling properties of DNA sequence data in phylogenetic analysis. *Mol Biol Evol*. 1995;12(5):814-822.

16. Swofford, D.L. et al. Phylogenetic inference. In: Hillis, David M., Craig Moritz BKM, ed. *Molecular Systematics*. 2nd ed. Sinauer Associates, Inc; 1996:407-514.

17. Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. Statistics and truth in phylogenomics. *Mol Biol Evol*. 2012;29(2):457-472.

18. Graybeal A. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol*. 1998;47(1):9-17. doi:10.1080/106351598260996.

19. Hedtke SM, Townsend TM, Hillis DM. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol*. 2006;55(3):522-529. doi:10.1080/10635150600697358.

20. Philippe H, Chenuil A, Adoutte A. Can the Cambrian Explosion Be Inferred through Molecular Phylogeny. *Dev Suppl 1994*. 1994:Suppl.: 15-25. Available at: <Go to ISI>://A1994QK89700003.

21. Gatesy J DR and WN. How Many Genes Should a Systematist Sample? Conflicting Insights from a Phylogenomic Matrix Characterized by Replicated Incongruence. *Syst Biol*. 2006;56(2):355-363.

22. Phillips MJ, Delsuc F, Penny D. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol*. 2004;21(7):1455-1458.

23. Fiala KL, Soakal RR. Factors Determining the Accuracy of Cladogram Estimation: Evaluation Using Computer Simulation. *Evolution (N Y)*. 1985;39(3):609-622. doi:10.2307/2408656.

24. Foster PG, Hickey DA. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol*. 1999;48(3):284-290. doi:10.1007/PL00006471.

25. Gowri-Shankar V, Rattray M. On the correlation between composition and site-specific evolutionary rate: Implications for phylogenetic inference. *Mol Biol Evol*. 2006;23(2):352-364.

26. Jermiin L, Ho SY, Ababneh F, Robinson J, Larkum AW. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol*. 2004;53(4):638-643. doi:10.1080/10635150490468648.

27. Nesnidal MP, Helmkampf M, Bruchhaus I, Hausdorf B. Compositional heterogeneity and phylogenomic inference of metazoan relationships. *Mol Biol Evol*. 2010;27(9):2095-2104.

28. Yang Z. On the best evolutionary rate for phylogenetic analysis. *Syst Biol*. 1998;47(1):125-133.

29. Huelsenbeck JP, Rannala B. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science*. 1997;276(5310):227-232. doi:10.1126/science.276.5310.227.

30. Charleston MA, Hendy MD, Penny D. The effects of sequence length, tree topology, and number of taxa on the performance of phylogenetic methods. *J Comput Biol*. 1994;1(2):133-151.

31. Huelsenbeck JP. Performance of Phylogenetic Methods in Simulation. *Syst Biol*. 1995;44(1):17-48. doi:10.1093/sysbio/44.1.17.

32. Felsenstein J. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Syst Biol*. 1978;27(4):401-410. doi:10.1093/sysbio/27.4.401.

33. Collins TM, Fedrigo O, Naylor GJP. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. *Syst Biol*. 2005;54(3):493-500. doi:10.1080/10635150590947339.

34. Sharp PM, Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 1987;15(3):1281-1295.

35. Cherry JM, Hong EL, Amundsen C, et al. Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res*. 2012;40(D1).

36. Busby MA, Gray JM, Costa AM, et al. Expression divergence measured by transcriptome sequencing of four yeast species. *BMC Genomics*. 2011;12(1):635. doi:10.1186/1471-2164-12-635.

37. Salichos L, Stamatakis A, Rokas A. Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Mol Biol Evol*. 2014;31(5):1261-71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24509691.

38. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*. 2013;497(7449):327-31. doi:10.1038/nature12130.

39. Robinson DF, Foulds LR. *Comparison of Phylogenetic Trees*.; 1981:131-147.

40. Dunn CW, Hejnol A, Matus DQ, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. 2008;452(7188):745-749.

41. Philippe H, Derelle R, Lopez P, et al. Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Curr Biol*. 2009;19(8):706-712.

42. Smith SA, Wilson NG, Goetz FE, et al. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*. 2011;480(7377):364-367.

43. Schierwater B, Eitel M, Jakob W, et al. Concatenated analysis sheds light on early metazoan evolution and fuels a modern “urmetazoon” hypothesis. *PLoS Biol*. 2009;7(1).

44. Kocot KM, Cannon JT, Todt C, et al. Phylogenomics reveals deep molluscan relationships. *Nature*. 2011;477(7365):452-456. doi:10.1038/nature10382.

45. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*. 2009;24(6):332-340.

46. Pamilo P, Nei M. Relationships between gene trees and species trees. *Mol Biol Evol*. 1988;5(5):568-583.

47. Slowinski J, Page R. How should species phylogenies be inferred from sequence data? *Syst Biol*. 1999;48:814-825.

48. Koonin E V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 2005;39:309-338. doi:10.1146/annurev.genet.39.073003.114725.

49. Yu Y, Degnan JH, Nakhleh L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet*. 2012;8(4).

50. DeBiasse MB, Nelson BJ, Hellberg ME. Evaluating summary statistics used to test for incomplete lineage sorting: Mito-nuclear discordance in the reef sponge Callyspongia vaginalis. *Mol Ecol*. 2014;23(1):225-238.

51. Rodriguez-Ezpeleta N, Brinkmann H, Burger G, et al. Toward Resolving the Eukaryotic Tree: The Phylogenetic Positions of Jakobids and Cercozoans. *Curr Biol*. 2007;17(16):1420-1425.

52. Zhang N, Zeng L, Shan H, Ma H. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytol*. 2012;195(4):923-937.

53. Regier JC, Shultz JW, Ganley ARD, et al. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst Biol*. 2008;57(6):920-938.

54. Lang JM, Darling AE, Eisen JA. Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices. *PLoS One*. 2013;8(4).

55. Capella-Gutierrez S, Kauff F, Gabaldón T. A phylogenomics approach for selecting robust sets of phylogenetic markers. *Nucleic Acids Res*. 2014:1-11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24476915.

56. Regier JC, Shultz JW, Zwick A, et al. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*. 2010;463(7284):1079-1083.

57. Byrne KP, Wolfe KH. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*. 2005;15(10):1456-1461.

58. Fitzpatrick DA, O'Gaora P, Byrne KP, Butler G. Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser. *BMC Genomics*. 2010;11:290.

59. Katoh K, Kuma KI, Toh H, Miyata T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 2005;33(2):511-518.

60. Abascal F, Zardoya R, Posada D. ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*. 2005;21(9):2104-2105.

61. Stamatakis A. The RAxML 7.0.4 Manual. *Bioinformatics*. 2006;22(21):2688–2690.

62. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13(9):2178-2189.

63. Peden JF. Analysis of Codon Usage. 1999. Available at: http://codonw.sourceforge.net/.

64. R Development Core Team. The R Project for Statistical Computing. Team RDC, ed. *Text*. 2002;2009(July 2003):1-9. Available at: http://www.r-project.org/.

CHAPTER VI

CONCLUSIONS

In this last chapter, I would like to address and summarize epigrammatically, the basic conclusions of my dissertation thesis, as they were thoroughly presented throughout the previous chapters of my dissertation thesis.

Chapter II: Evaluating ortholog prediction algorithms in a yeast model clade

Having a quality set of orthogroups is the keystone for every phylogenic analysis. By evaluating 4 graphed-based ortholog prediction algorithms in a yeast model, I found that they all perform very well in datasets deprived of paralogy, but their accuracy decreases dramatically when paralogy is rampant. Moreover, my evaluation of these algorithms showed that sometimes simpler is better, as cRBH, a simple clustering algorithm for reciprocal best hit outperformed all other three algorithms in almost every category.

Chapter III: Novel information theory-based measures for quantifying incongruence among phylogenetic trees

With the advent of phylogenomics, most recent phylogenetic studies do not depend on single few genes. However, despite the abundance of data and the high bootstrap support the use of hundreds of genes brought, many phylogenomic studies continue to present conflicting topologies, all supported with high confidence. For this reason, I developed 4 novel measures - IC, ICA, TC, TCA- based on information theory and Shannon's entropy, aiming at quantifying

incongruence and the uncertainty existing in many conflicting and ambiguous clades. These measures are independent from the species tree, optimality criteria, they can be used for many different types of data-characters, including molecular data, indels or other genomic characters, they can be used as optimality criterion, they are relatively straightforward and easy to use and they have been integrated in the latest version of RAxML, a very popular open-source software for constructing phylogenies. In the near future, I 'll be working with collaborators to extend these measures for datasets that contain missing data.

Chapter IV: Inferring ancient divergences requires genes with strong phylogenetic signals

As mentioned previously, the use of concatenation in several studies, has presented conflicting topologies with high support. By concatenating 1070 high quality orthogroups from a yeast model clade, I inferred the yeast species tree, which was at least partially wrong based on syntenic information. By examining the individual gene trees, I discovered that all gene trees differed from the inferred species tree, as well as with each other. Using IC, I was able to unmask excessive levels of gene tree incongruence and show that clades with high bootstrap support, were extremely ambiguous. Moreover, using TC, I demonstrated that several high-profile and widely used methods that intend to decrease incongruence, have little, no or negative effect. Moreover, I introduced two new methods which were able to dramatically decrease phylogenetic incongruence. However, even with the use of these methods, I was not able to resolve at least four short basal internodes on the tree of yeast, possibly due to reasons of incomplete lineage sorting, hybridization, or simply the loss of any phylogenetic signal, after more than 200 million years of evolution. However, the development of methods that can distinguish between these three reasons of incongruence, still remains as a strong puzzle and a

future aim. It should be mentioned, that I also obtained similar results using two more datasets; Vertebrates and Metazoa.

Chapter V: Examination of factors that influence phylogenetic incongruence in a yeast model clade

In this chapter, I examined several functional gene factors to find whether they play a role in driving this excessive gene incongruence that I previously described.and whether genes that show some of these properties may be selected for phylogenetic markers. Overall, I found that some of these factors show a significant correlation with gene incongruence. Moreover, by implementing a principal component regression analysis on these functional gene factors, I was able to explain more than 17% of the total variance of gene incongruence. However, my results also showed that they cannot be selected for and used as reliable phylogenetic markers, despite their often use as such by many researchers.