ARTICLE

doi:10.1038/nature12130

Inferring ancient divergences requires genes with strong phylogenetic signals

Leonidas Salichos[1] & Antonis Rokas[1]

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.

Concatenation, the compilation and analysis of hundreds of genes as a single data set, has become the standard approach for determining major 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], that are useful but not always effective[16–18].

The *Saccharomyces* and *Candida* yeasts are excellent for examining phylogenomic practices in the presence of incongruence, owing 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 orthologues, 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 of 1,070 groups of orthologues (below we refer to groups of orthologues simply as 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 were weakly supported. None of the 1,070 gene trees agreed with each other, with the concatenation phylogeny or with the eMRC phylogeny. The novel measure we developed to quantify the observed incongruence showed that standard practices aimed at reducing incongruence had little impact. In agreement with current theoretical models[9,16,25,26], incongruence was more severe for shorter internodes that are deeper on the phylogeny. Notably, the selection of genes whose bootstrap consensus trees had high average clade support, or the selection of highly supported internodes, significantly reduced incongruence, arguing that inference in deep time depends critically on the identification of molecular markers with strong phylogenetic signals.

All gene trees differ from species phylogeny

We assembled a data set of 1,070 genes from 23 yeast genomes[20,21,27] (Methods and Supplementary Table 1). Maximum-likelihood analysis of the concatenation of all 1,070 genes produced a species phylogeny in which all 20 internodes exhibited 100% bootstrap support (Fig. 1a); we obtained identical results using Bayesian inference and one other type of maximum-likelihood software (Supplementary Fig. 1). Notably, all 1,070 gene trees were topologically distinct and none matched the topology inferred from concatenation analysis (Fig. 1b). However, the average tree distance between the 1,070 gene trees was much lower (normalized Robinson–Foulds[28] tree distance = 0.52; that is, two gene trees differed, on average, in 10.4 out of their 20 bipartitions) than that between randomly generated trees of the same taxon number (0.99; that is, two trees differed on average in 19.8 out of 20 bipartitions), indicating that the yeast gene trees have similar evolutionary histories.

Summarizing the 1,070 gene trees into an eMRC phylogeny produced a topology that was identical to the concatenation phylogeny (Fig. 1a). However, although 11 out of 20 internodes in the eMRC phylogeny had a gene-support frequency (GSF) of greater than 50%, 5 of the remaining 9 internodes had a GSF of less than 30% (Fig. 1a). Furthermore, the most prevalent conflicts to most of these weakly supported internodes had substantial GSF values (Supplementary Table 2). For example, the relative positions of *Candida glabrata*, *Saccharomyces castellii* and the *Saccharomyces* 'sensu stricto' clade suggest that there are five uniquely shared chromosomal rearrangements and a substantially higher number of uniquely shared gene losses between *C. glabrata* and *S. cerevisiae*, which indicates that divergence of *S. castellii* preceded that of *C. glabrata* from the *Saccharomyces* sensu stricto clade[22]. Although concatenation provided 100% bootstrap support for the apparently incorrect grouping of *S. castellii* with the *Saccharomyces* sensu stricto clade (Fig. 1a), only 311 out of 1,070 gene trees (29%) favoured it, whereas 214 (20%) favoured the grouping of *C. glabrata* with the *Saccharomyces* sensu stricto clade.

A novel measure that considers incongruence

To quantify incongruence, we developed 'internode certainty', 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], internode certainty relies on the bipartitions present in trees, each of which is a split of

[1]Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee 37235, USA.

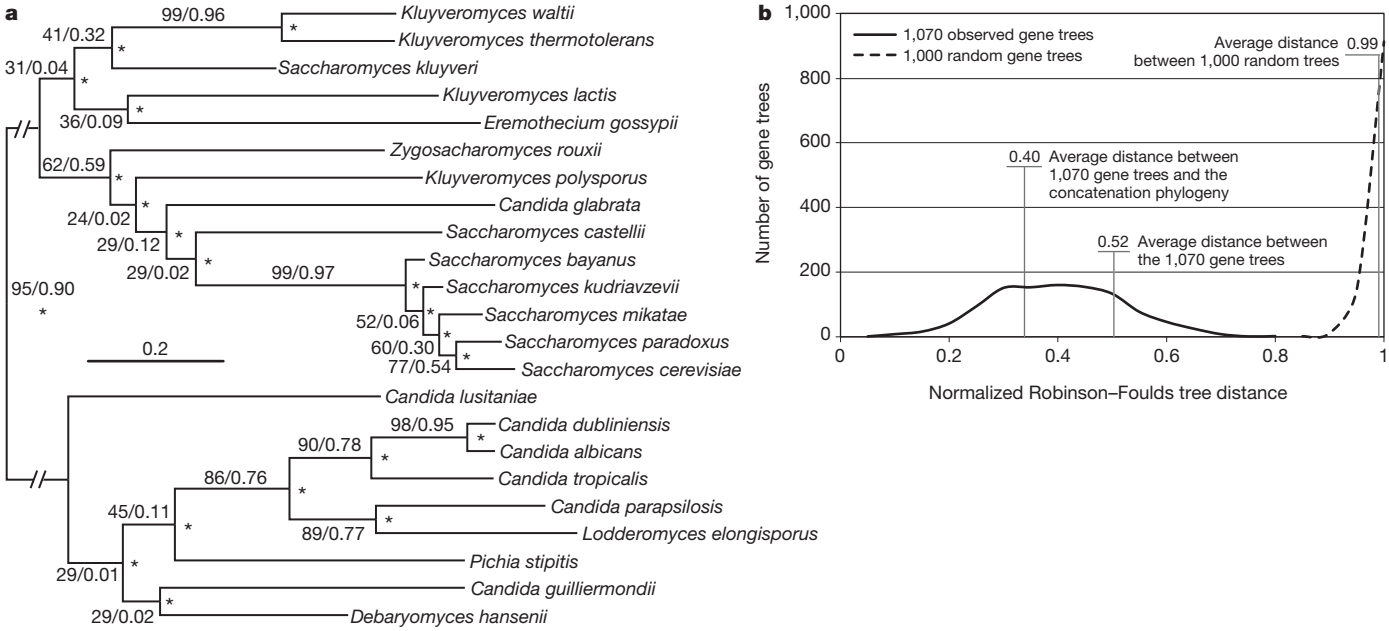

Figure 1 | The yeast species phylogeny recovered from the concatenation analysis of 1,070 genes disagrees with every gene tree, despite absolute bootstrap support. **a**, The yeast species phylogeny recovered from concatenation analysis of 1,070 genes using maximum likelihood. Asterisks denote internodes that received 100% bootstrap support by the concatenation analysis. Values near internodes correspond to gene-support frequency and internode certainty, respectively. The scale bar is in units of amino-acid substitutions per site. **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 1,000 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 themselves, and between 1,000 randomly generated gene trees that have equal taxon numbers, are also shown. The phylogeny of the 23 yeast species analysed 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.

the taxa into two mutually exclusive non-empty groups. Compared to other incongruence measures[29–32], internode certainty is not character-based[29–31], 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 gene trees is used, the internode certainty of a given internode will reflect the amount of information available for that internode in the set of gene trees 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 internode certainty will be calculated based on bootstrap-support values. Internode-certainty values near zero indicate the presence of an almost equally supported bipartition that conflicts with the inferred internode, whereas values close to one indicate the absence of conflict. Examination of the eMRC phylogeny showed that 9 out of 20 internodes had an internode certainty of less than 0.3, which corresponds to a less than 4:1 ratio between the support for the inferred internode and its most prevalent conflicting bipartition, and that 7 out of 20 internodes had an internode certainty of less than 0.1 (a ratio of less than 7:3) (Fig. 1a and Supplementary Fig. 2).

Given that internode certainty measures the degree of conflict for every internode, it is more informative than GSF. For example, the placement of *Saccharomyces bayanus* and the placement of *Zygosaccharomyces rouxii* received 52% and 62% GSF, whereas their internode certainties were 0.06 and 0.59, respectively (Fig. 1a). This marked difference in the internode-certainty values of the two internodes (despite similar GSF values) is a result of the strong secondary conflicting signal in *S. bayanus* only[33] (29% GSF for grouping *S. bayanus* with *Saccharomyces kudriavzevii*), and not in *Z. rouxii* (Supplementary Table 2). Furthermore, comparison of the sums of internode-certainty values across trees of a given taxon number ('tree certainty') can be used to quantify changes in the degree of incongruence between trees inferred from different data sets 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. 2). Specifically, we tested the effect of: removing sites containing gaps as well as 'rogue' genes that produced alignments of bad quality (Supplementary Fig. 3); removing unstable and quickly evolving species (Supplementary Figs 4–6); using only genes that recover a particular internode widely regarded as certain, or well established, from prior data (Supplementary Figs 6 and 7); using only slowly evolving genes (Supplementary Fig. 8); and using conserved amino acid substitutions or indels (Supplementary Fig. 9).

The first three practices did not have a substantial effect on the inference and support of the yeast phylogeny, whereas the use of slowly evolving genes and conserved sites increased incongruence across many internodes of the yeast phylogeny (Fig. 2). Furthermore, the removal of unstable or quickly evolving species from the *Saccharomyces* lineage had no effect on, often highly ambiguous, internodes in the *Candida* lineage and vice versa (Supplementary Figs 5 and 6), arguing that the impact of removing rogue taxa was not only minimal but also highly localized.

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. 3). This is consistent with theoretical expectations[9,16,25,26]. To test whether this relationship is the same in other lineages, we generated a data set of 1,086 genes from 18 vertebrate species, which has higher sequence similarity than the yeast data set (61% versus 44% average pairwise amino acid similarity, respectively), and a data set of 225 genes from metazoan species,

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

Figure 2 | Differences in yeast phylogenies inferred from different phylogenomic practices. The specific phylogenomic practice tested (Treatment), the average GSF of the internodes of the yeast phylogeny, the tree certainty (TC) of the yeast phylogeny, the numbers of internodes of the yeast phylogeny in which GSF increases or decreases by more than 3% (GSF increases and GSF decreases), and the numbers of internodes of the yeast phylogeny in which internode certainty increases or decreases by more than 0.03 (internode certainty (IC) increases and IC decreases). As the maximum value of internode certainty for a given internode is 1, the maximum value of tree certainty 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 greater than or equal to a certain percentage, x, of the original alignment were used. In the analyses concerned with the use of bipartitions, only those bipartitions that displayed bootstrap support greater or equal to 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. NA, not applicable.

which has lower sequence similarity (29% average pairwise amino acid similarity). The vertebrate genes produced 299 distinct gene trees (average normalized Robinson–Foulds tree distance = 0.42). Concatenation analysis suggested a completely supported species phylogeny; however, this phylogeny was topologically identical to 15 gene trees and eMRC analysis showed that 4 out of 15 internodes had a GSF of less than 50% and internode certainty of less than 0.3 (Supplementary Fig. 10a–c). Similarly, the 225 metazoan genes produced 224 distinct gene trees (average normalized Robinson–Foulds tree distance = 0.72). Concatenation analysis suggested 14 out of 18 internodes with 100% bootstrap support, despite the fact that it was not topologically identical to any of the 225 gene trees and that 10 out of 18 internodes had less than 50% GSF and less than 0.1 internode certainty (Supplementary Fig. 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. 3).

Strong signal reduces incongruence

To test whether the selection of genes with stronger phylogenetic signal reduced incongruence, we analysed three data sets comprising genes whose bootstrap consensus trees showed average bootstrap support across all internodes that was greater than or equal to 60% (904 genes), 70% (545 genes), or 80% (131 genes), and three data sets comprising the 904, 545 or 131 genes whose bootstrap consensus trees had the highest tree certainty. Selecting genes with high average bootstrap support or high tree certainty significantly reduced incongruence across many, but not all, internodes (Fig. 2, and Supplementary Figs 11 and 12). Concatenation analysis of the sets of genes with average bootstrap support that was greater than or equal to 60% and 70% (and of the 904 and the 545 genes with the highest tree certainty) produced the same species phylogeny as when all genes were analysed. Notably, analysis of genes with average bootstrap support that was greater than or equal to 80%, as well as of the 131 genes with the highest tree certainty, produced the correct placement of *C. glabrata* (Supplementary Fig. 11c, f). This result, to our knowledge, has not been observed in any concatenation-based yeast phylogenomic analysis[7,34–37], and suggests that high bootstrap support is a good indicator of a gene's phylogenetic usefulness, but also that concatenating genes with high bootstrap support reduces incongruence and improves resolution.

We also tested whether selecting internodes with high bootstrap support decreased incongruence by extracting only those bipartitions that displayed bootstrap support 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 12 and 13). One advantage of working with taxon bipartitions, rather than genes, is that we can quantify a given internode's internode certainty from only the subset of bipartitions that highly support or conflict with that internode. This practice significantly increased internode certainty values for ≥ 14 internodes relative to the phylogeny of Fig. 1a and showed the highest tree certainty of all our analyses (Fig. 2). Interestingly, while internode certainty for most internodes increased when we increased the bootstrap support threshold, this was not the case for several of the most difficult to resolve internodes (Supplementary Fig. 13d), suggesting that those few genes that show high bootstrap support for short internodes deep in the phylogeny strongly conflict with each other. We obtained similar results when we

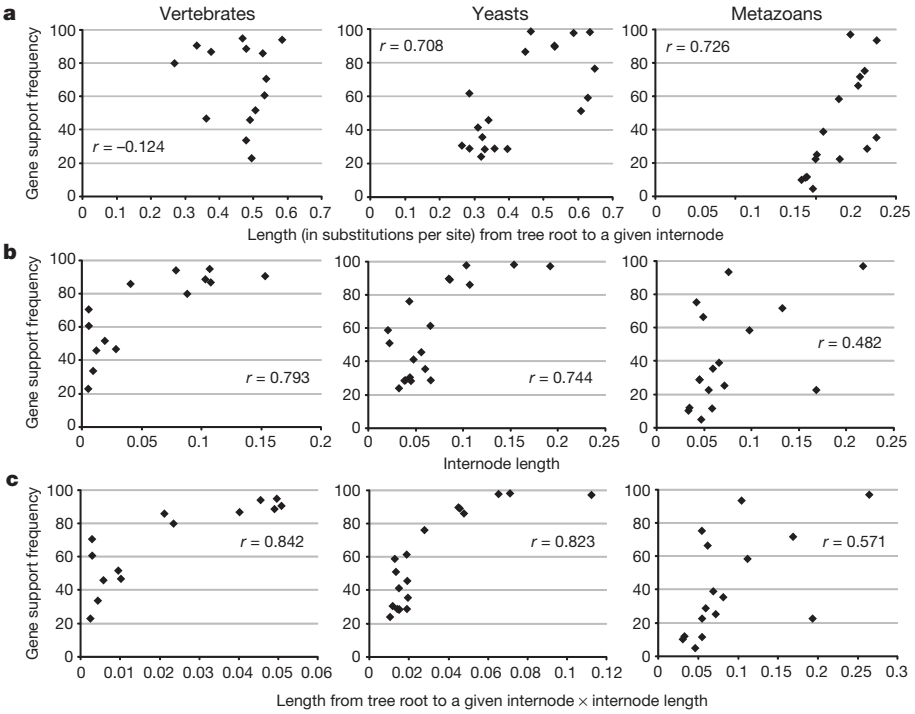

Figure 3 | Incongruence is more prevalent in shorter internodes located deeper on the phylogeny. The correlation (Pearson's *r*) between a measure of internode support (GSF) with internode length and depth was measured for each internode present in three data sets that show lower (vertebrates, 1,086 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.

performed the same analyses on the vertebrate and metazoan data sets (Supplementary Fig. 14).

Standard practices can mislead

We constructed and analysed 1,070 yeast genes with the aim of determining the yeast phylogeny. If only concatenation and standard phylogenomic practices had been used, this analysis would have produced 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 gene trees showed that concatenation masked the considerable incongruence present in several internodes. Thus, while analyses of approximately 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 15 and 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 gene trees, allowing for lineage sorting, a common biological explanation for gene trees that are 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 data set with the average unit-ranking method[38] produced a species phylogeny in which all the internodes with very low GSF and internode-certainty values were extremely short, mainly because all incongruence was considered to be due to variation in coalescent depth across gene trees (Supplementary Fig. 17a). Not surprisingly, these coalescent unit-based branch lengths were highly correlated with internodes' GSF and internode-certainty values (Supplementary Fig. 17b). Furthermore, bootstrapping of this data set produced a highly supported species phylogeny (Supplementary Fig. 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 bootstrap support on concatenation analyses of large data sets. Bootstrapping was developed long before the discovery of high-throughput sequencing, and it is an extremely useful measure of sampling error—that is, the robustness of inference when data are limited[39]—such as when a single gene is analysed. Given the availability and ease of generating genome-scale data[40], relying on bootstrap to analyse phylogenomic data sets is misleading, not only because sampling error is minimal but also because its application will, even in the presence of notable conflict[9] or systematic error[6,16], almost always result in 100% values[9,19,41].

The second critical revision 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 full set of genes, suggesting that phylogenomic analyses using conditional combination approaches, rather than approaches based on total evidence, may be more powerful[42]. Preferably, such analyses would be combined with internode-specific approaches[31] because the latter can uncover internodes that harbour several conflicting phylogenetic signals. As the internode certainty measure shows (Supplementary Fig. 2), the amount of information for a given internode that is supported by 50% of gene trees, with the remaining 50% being uninformative, is far greater than in cases in which the remaining 50% of the gene trees support two or three alternative conflicting topologies. In the first case the gene trees strongly suggest that the internode is resolved, whereas in the second there is reason to be cautious.

Finally, it is necessary to identify explicitly internodes that, despite the use of genome-scale data sets, robust study designs and powerful algorithms, are poorly supported. We argue that the ongoing debate regarding phylogenies inferred from different phylogenomic studies[10] concerns internodes that are poorly supported by individual gene trees. Identifying these internodes and distinguishing them from those that are supported by a considerable fraction of genes and that lack conflicts will be far more beneficial than simply helping to pinpoint challenging internodes. It should enable 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 data set of 2,651 genes that, 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 genes. Genes were aligned using MAFFT[43], the best-fit evolutionary model was determined using ProtTest[44], and the maximum-likelihood tree was estimated using RAxML[45]. eMRC trees were determined using the Phylogeny Inference Package (PHYLIP; J. Felsenstein, University of Washington, Seattle; http://evolution.genetics.washington.edu/phylip.html) and custom Perl scripts. A series of different data sets was constructed using custom Perl scripts. Internode certainty, our novel measure that evaluates support for a given internode by considering its GSF (or bootstrap support) jointly with that of the most prevalent conflicting bipartition in the entire set of gene trees (or bootstrap replicate trees), was calculated according to the equation:

$$\text{Internode certainty} = \log_2(2) + p(x_1/(x_1+x_2))\log_2(p(x_1/(x_1+x_2))) + p(x_2/(x_1+x_2))\log_2(p(x_2/(x_1+x_2)))$$

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

Full Methods and any associated references are available in the online version of the paper.

Received 6 December 2012; accepted 28 March 2013.
Published online 8 May 2013.

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. *Syst. 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 (1995).
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. *Syst. 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. *N. Z. 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. *Syst. 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).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank K. 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 laboratory and B. 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 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.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.R. (antonis.rokas@vanderbilt.edu).

METHODS

Data matrix construction. We used the complete sets of annotated genes from 23 yeast genomes[20,21,27,46] (Supplementary Table 1) and, using the synteny and orthology information present in the YGOB[20] and CGOB[21] databases, we constructed an initial data set of 2,651 groups of orthologues (referred to below simply as 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 orthologue to be present in all 23 species greatly minimized errors in orthology inference due to hidden paralogy[23,47]. 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[24,48–50] and is probably very rare.

The nucleotide sequences of all genes were translated to amino acids, 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 genes to 1,070. Examination of the functional annotation—as defined by the Gene Ontology consortium[51]—of the 1,070 *S. cerevisiae* orthologues using the GOstat software[52] 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; that is, for categories associated with standard cell housekeeping functions. Analysis of different orthologue subsets (for example, of the 131 genes whose bootstrap consensus trees show the highest average bootstrap support) shows that for many of these subsets, some of the same functions were statistically overrepresented.

We also created two additional data sets from the complete sets of annotated genes from 18 vertebrate and 21 metazoan species (Supplementary Table 1). The two data sets were constructed using the cRBH algorithm[23], and were comprised of 1,086 vertebrate and 225 metazoan genes. To avoid constructing groups of orthologues that contained very distant homologues, 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 (for example, the acronym for *S. cerevisiae* is 'Scer'; see Supplementary Information). All data matrices are available from the authors on request.

Gene alignment and filtering criteria. To minimize the use of genes that contained sequences whose annotation was problematic, or which resulted in alignments of low quality, we applied various filtering criteria. We first excluded, before alignment, all genes that had an average sequence length of less than or equal to 150 amino acids. Second, we aligned all genes using the MAFFT software[43], using the default settings, and excluded genes whose alignments after removing all positions that contained gaps were less than or equal to 50% of the original alignment length.

Gene-tree inference. For each gene, the best-fit evolutionary model was selected using ProtTest[44]. The models typically consisted of an empirically determined amino acid substitution matrix (for example, WAG[53]), empirically measured amino acid 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. The unrooted phylogenetic tree of each and every gene, also called the gene tree, was then inferred using RAxML[45].

Species phylogeny inferred from concatenation and eMRC approaches. For the concatenation analysis, gene alignments were analysed as a single supermatrix. An unrooted concatenation species phylogeny was then inferred under the 'PROTGAMMAIWAGF' model of amino acid substitution in RAxML[45], and confirmed with GARLI[54] as well as with MrBayes[55]. The unrooted eMRC phylogeny that consisted of those bipartitions that appear in more than half of the maximum-likelihood estimated gene trees, as well as of additional compatible bipartitions that appear in less than half of the gene trees[56,57], was inferred from the CONSENSE program in PHYLIP. The eMRC phylogeny of bipartitions with high bootstrap support was constructed using custom Perl scripts. As 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 gene trees 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 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[58], using the random tree generation procedure described by Kuhner and Felsenstein[59].

Internode certainty. 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*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *S. kudriavzevii* and *S. bayanus*), one example of a bipartition is 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 MRC approach[56] calculates the shared bipartitions across all phylogenetic trees and displays only those shared by their majority. Consequently, each internode in the MRC 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 about 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% versus 49%) would indicate that both bipartitions have nearly equal support, whereas the second case (51% versus 5%) would indicate that the first bipartition is the only strongly supported bipartition for this internode. Consensus phylogenetic networks[15,60], 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, a measure that provides robust quantitative measures of the information conveyed by conflicting bipartitions for each internode.

Shannon's entropy measures the amount of certainty found in a random variable[61]. For example, when tossing a fair coin, heads or tails are equally probable and so the amount of certainty that 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 deduction 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. Internode certainty quantifies the certainty of a bipartition that appears on a phylogenetic tree (that is, 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:

$$\text{Internode certainty} = \log_2(2) + p(x_1/(x_1 + x_2))\log_2(p(x_1/(x_1 + x_2))) + p(x_2/(x_1 + x_2))\log_2(p(x_2/(x_1 + x_2)))$$

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

Internode certainty, as well as the related measure called tree certainty (see below), can be measured on any given set of trees. For example, if the entire set of gene trees is used, the internode certainty value of a given internode will reflect the amount of information available for that internode in the set of gene trees 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 internode certainty will be calculated based on bootstrap support values (instead of GSF values). Internode certainty 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 internode certainty 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. For example, if we assume that there

are four prevalent conflicting bipartitions with frequencies of 40%, 10%, 10% and 10%, respectively, for a given internode, then:

$$\text{Internode certainty} = 1 + (40/(40+10))\log_2(40/(40+10)) + (10/(40+10))\log_2(10/(40+10)) \approx 0.28$$

As another example, if we assume that there are four prevalent conflicting bipartitions with frequencies of 40%, 40%, 10% and 10%, respectively, for a given internode, then:

$$\text{Internode certainty} = 1 + (40/(40+40))\log_2(40/(40+40)) + (40/(40+40))\log_2(40/(40+40)) = 0.00$$

We define tree certainty as the sum of all internode-certainty values across all internodes of a phylogenetic tree.

Evaluation of phylogenomic practices. To remove positions or genes with gaps, we used custom Perl scripts to modify our default alignments by removing sites that contained greater than or equal to 50% gaps, or sites that contained any gap. We also tested whether the removal of genes producing alignments of bad quality improved inference of the species phylogeny by filtering genes whose alignment length after removal of all gap-containing sites was less than or equal to 70% of the original alignment length (instead of the less than or equal to 50% threshold used in the default analysis).

We removed several different unstable and quickly evolving species from the default data set, singly and in combination. After each removal, the gene sequences were re-aligned, a new best-fit evolutionary model was identified, and the phylogenetic analysis was carried out again with the new alignment and model.

To select genes that recover specific bipartitions, for the 100 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: first, *Candida albicans*, *Candida dubliniensis* and *Candida tropicalis*; second, *C. glabrata*, *Kluyveromyces polysporus*, *S. bayanus*, *S. castellii*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus* and *Z. rouxii*; and third, *C. glabrata*, *S. bayanus*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae* and *S. paradoxus*. We then used the selected genes and their gene trees to infer a species phylogeny using concatenation and eMRC analysis.

The 100 most slowly evolving genes were identified by calculating the 100 genes whose gene trees had the smallest sum of branch lengths.

To reduce the effect of homoplasy for early divergent internodes, many studies have suggested the use of rare substitution types[62] as well as insertions or deletions (indels)[63]. We constructed three data sets by extracting all sites from our 1,070 gene alignments that contained first, a single radical amino acid substitution (defined as a substitution with a blosum62 matrix score of less than or equal to –3) (20,289 sites); second, a single substitution between amino acids that differ radically in their physicochemical properties[62] (4,075 sites); or third, a single indel that spans seven or more amino acids (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 internode certainty 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.

For every gene from the default data set, we estimated the average bootstrap support 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 internode-certainty value of every internode in its bootstrap consensus tree. When the internode certainty is calculated in this way, its value reflects the amount of information available for that internode in the set of bootstrap replicate trees because it considers the internode's bootstrap support jointly with the bootstrap support of the most prevalent bipartition that conflicts with the internode. We then calculated the tree-certainty value for each gene by summing the internode-certainty values of all internodes in its bootstrap consensus tree. Finally, we used these average bootstrap support and tree-certainty values to construct six subsets of orthogroups: three with genes that have average bootstrap support values that are greater than or equal to 60% (904 genes), 70% (545 genes) and 80% (131 genes), as well as three data sets of the 904, 545 and 131 genes with the highest tree certainty.

For every gene from the default data set, we extracted all bipartitions from its bootstrap consensus tree that had bootstrap support that was greater than or equal to 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.

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

46. Dujon, B. Yeast evolutionary genomics. *Nature Rev. Genet.* **11,** 512–524 (2010).
47. Scannell, D. R., Butler, G. & Wolfe, K. H. Yeast genome evolution-the origin of the species. *Yeast* **24,** 929–942 (2007).
48. Hall, C., Brachat, S. & Dietrich, F. S. Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot. Cell* **4,** 1102–1115 (2005).
49. 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).
50. 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. USA* **106,** 16333–16338 (2009).
51. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
52. Beissbarth, T. & Speed, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20,** 1464–1465 (2004).
53. 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).
54. Zwickl, D. J. *Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion.* Ph.D. thesis, Univ. Texas at Austin (2006).
55. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
56. Bryant, D. in *Bioconsensus* (eds M. Janowitz *et al.*) 163–184 (American Mathematical Society and DIMACS, 2003).
57. Felsenstein, J. *Inferring Phylogenies*. (Sinauer, 2003).
58. 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–W579 (2012).
59. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11,** 459–468 (1994).
60. 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).
61. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27,** 379–423 (1948).
62. 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).
63. Belinky, F., Cohen, O. & Huchon, D. Large-scale parsimony analysis of metazoan indels in protein-coding genes. *Mol. Biol. Evol.* **27,** 441–451 (2010).