Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data

Xing-Xing Shen,* Xiaofan Zhou,* Jacek Kominek,† Cletus P. Kurtzman,‡,1 Chris Todd Hittinger,†,1 and Antonis Rokas*,1

*Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee 37235, †Laboratory of Genetics, Genome Center of Wisconsin, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, Wisconsin 53706, and ‡Mycotoxin Prevention and Applied Microbiology Research Unit, National Center for Agricultural Utilization Research, Agricultural Research Service, U.S. Department of Agriculture, Peoria, Illinois 61604

ORCID IDs: 0000-0002-2879-6317 (X.Z.); 0000-0002-1916-0122 (J.K.); 0000-0001-5088-7461 (C.T.H.); 0000-0002-7248-6551 (A.R.)

ABSTRACT Understanding the phylogenetic relationships among the yeasts of the subphylum Saccharomycotina is a prerequisite for understanding the evolution of their metabolisms and ecological lifestyles. In the last two decades, the use of rDNA and multilocus data sets has greatly advanced our understanding of the yeast phylogeny, but many deep relationships remain unsupported. In contrast, phylogenomic analyses have involved relatively few taxa and lineages that were often selected with limited considerations for covering the breadth of yeast biodiversity. Here we used genome sequence data from 86 publicly available yeast genomes representing nine of the 11 known major lineages and 10 nonyeast fungal outgroups to generate a 1233-gene, 96-taxon data matrix. Species phylogenies reconstructed using two different methods (concatenation and coalescence) and two data matrices (amino acids or the first two codon positions) yielded identical and highly supported relationships between the nine major lineages. Aside from the lineage comprised by the family Pichiaceae, all other lineages were monophyletic. Most interrelationships among yeast species were robust across the two methods and data matrices. However, eight of the 93 internodes conflicted between analyses or data sets, including the placements of: the clade defined by species that have reassigned the CUG codon to encode serine, instead of leucine; the clade defined by a whole genome duplication; and the species *Ascoidea rubescens*. These phylogenomic analyses provide a robust roadmap for future comparative work across the yeast subphylum in the disciplines of taxonomy, molecular genetics, evolutionary biology, ecology, and biotechnology. To further this end, we have also provided a BLAST server to query the 86 Saccharomycotina genomes, which can be found at http://y1000plus.org/blast.

KEYWORDS
phylogenomics
maximum likelihood
incongruence
genome completeness
nuclear markers

Molecular phylogenetic analyses show that the fungal phylum Ascomycota is comprised of three monophyletic subphyla that share a common ancestor from ~500 MYA (Kurtzman and Robnett 1994; Sugiyama *et al.* 2006; Taylor and Berbee 2006; James *et al.* 2006; Liu *et al.* 2009): the Saccharomycotina (syn. Hemiascomycota; *e.g.*, *Saccharomyces*, *Pichia*, *Candida*), the Pezizomycotina (syn. Euascomycota; *e.g.*, *Aspergillus*, *Neurospora*), and the Taphrinomycotina (syn. Archaeascomycota; *e.g.*, *Schizosaccharomyces*, *Pneumocystis*).

Yeasts of the fungal subphylum Saccharomycotina exhibit remarkably diverse heterotrophic metabolisms, which have enabled them to successfully partition nutrients and ecosystems and inhabit every continent and every major aquatic and terrestrial biome (Hittinger *et al.* 2015). While yeast species were historically identified by metabolic differences, recent studies have shown that many of these classic characters are subject to rampant homoplasy, convergence, and parallelism (Hittinger *et al.* 2004; Hall and Dietrich 2007; Wenger *et al.* 2010; Slot and Rokas 2010; Lin and Li 2011; Wolfe *et al.* 2015). Despite the considerable progress in classifying yeasts using multilocus DNA sequence data, critical gaps remain (Kurtzman and Robnett 1998, 2003, 2007, 2013; Nguyen *et al.* 2006; Kurtzman *et al.* 2008, 2011; Kurtzman and Suzuki 2010); many genera are paraphyletic or polyphyletic, while circumscriptions at or above the family level are often poorly supported (Hittinger *et al.* 2015).

In recent years, phylogenomic analyses based on data matrices comprised of hundreds to thousands of genes from dozens of taxa have provided unprecedented resolution to several, diverse branches of the tree of life (Song *et al.* 2012; Salichos and Rokas 2013; Liang *et al.* 2013; Xi *et al.* 2014; Wickett *et al.* 2014; Whelan *et al.* 2015). Although the genomes of several dozen yeast species are currently available (Hittinger *et al.* 2015), published phylogenomic studies contain at most 25 yeast genomes (Rokas *et al.* 2003; Fitzpatrick *et al.* 2006; Liu *et al.* 2009; Medina *et al.* 2011; Salichos and Rokas 2013; Marcet-Houben and Gabaldón 2015; Shen *et al.* 2016; Riley *et al.* 2016).

A robustly resolved backbone yeast phylogeny will be of great benefit, not only to the study of yeast biodiversity, but also to diagnosticians seeking to identify and treat yeast infections, to biotechnologists harnessing yeast metabolism to develop advanced biofuels, and to biologists designing computational and functional experiments. Toward that end, here we have used genome sequence data from 86 publicly available yeast genomes representing 9 of the 11 major lineages and 10 nonyeast fungal outgroups to reconstruct the backbone of the Saccharomycotina yeast phylogeny.

MATERIALS AND METHODS

Data acquisition

The workflow used to assemble the data sets for the inference of the backbone phylogeny of Saccharomycotina yeasts is described in Figure 1. To assemble a data set with the greatest possible taxonomic sampling as of January 11, 2016, we first collected all Saccharomycotina yeast species whose genomes were available (Hittinger *et al.* 2015). We then excluded four publicly available genomes, namely, *Blastobotrys attinorum*, *B. petasosporus*, *Cephaloascus albidus*, and *C. fragrans*, which had been released under embargo and lacked a citable publication. In addition, we excluded the genomes of known hybrid species, such as *Pichia farinosa* (Louis *et al.* 2012), *Saccharomyces cerevisiae* × *S. eubayanus* syn. *S. pastorianus* (Libkind *et al.* 2011; Gibson and Liti 2015), and the wine yeast VIN7 (*S. cerevisiae* × *S. kudriavzevii*) (Borneman *et al.* 2012). For species with multiple isolates sequenced, we only included the genome of the isolate with the highest number of the "complete" genes (see below). These criteria resulted in the inclusion of genomes from 86 yeast species representing 9 of 11 major lineages of the subphylum Saccharomycotina (Hittinger *et al.* 2015). Finally, we used the genomes of 10 nonyeast fungi that are representatives of the phylum Ascomycota as outgroups. Detailed information of the nomenclature, taxonomy, and source of the 96 genomes in our study is provided in Supplemental Material, Table S1.

doi: 10.1534/g3.116.034744
Manuscript received August 18, 2016; accepted for publication September 21, 2016; published Early Online September 26, 2016.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.116.034744/-/DC1.
[1]Corresponding authors: Mycotoxin Prevention and Applied Microbiology Research Unit, National Center for Agricultural Utilization Research, Agricultural Research Service, U.S. Department of Agriculture, Peoria, IL 61604. E-mail: cletus.kurtzman@ars.usda.gov; Laboratory of Genetics, Genome Center of Wisconsin, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI 53706. E-mail: cthittinger@wisc.edu; and Department of Biological Sciences, Vanderbilt University, VU Station B 351634, Nashville, TN 37235. E-mail: antonis.rokas@vanderbilt.edu

A custom BLAST database for the genomes of the 86 yeast species

To further facilitate the use of these 86 Saccharomycotina genomes by the broader research community, we set up a custom local BLAST database using Sequenceserver, version 1.0.8 (Priyam *et al.* 2015). The database is free and publicly available through http://y1000plus.org/blast.

Assessment of genome assemblies and ortholog identification

Assessment of the 96 selected genome assemblies was performed using the BUSCO software, version 1.1b (Simão *et al.* 2015). Each individual genome was examined for the copy number of 1438 preselected genes (hereafter, referred to as BUSCO genes) that are single-copy in at least 90% of the 125 reference fungal genomes in the OrthoDB Version 7 database (www.orthodb.org) (Waterhouse *et al.* 2013). Briefly, for each BUSCO gene, a consensus protein sequence was generated from the hidden Markov model (HMM) alignment profile of the orthologous protein sequences among the 125 reference genomes using the HMMER software, version 3 (Eddy 2011). This consensus protein sequence was then used as query in a tBLASTn (Altschul *et al.* 1990; Camacho *et al.* 2009) search against each genome to identify up to three putative genomic regions, and the gene structure of each putative genomic region was predicted by AUGUSTUS (Stanke and Waack 2003). Next, the sequences of these predicted genes were aligned to the HMM alignment profile of the BUSCO gene, and the ones with alignment bit-score higher than a preset cutoff (90% of the lowest bit-score among the 125 reference genomes) were kept. If no predicted gene from a particular genome was retained after the last step, the gene was classified as "missing" from that genome. If one or more predicted genes from a genome were retained, these were further examined for their aligned sequence length in the HMM-profile alignment; predicted genes whose aligned sequence lengths were shorter than 95% of the aligned sequence lengths of genes in the 125 reference genomes were classified as "fragmented." The remaining predicted genes were classified as "complete" if only one predicted gene was present in a genome, and "duplicated" if two or more "complete" predicted genes were present in a genome. Only the sequences of single-copy "complete" genes without any in-frame stop-codon(s) were used to construct ortholog groups across the 96 genomes. We excluded the orthologous group constructed from BUSCO gene "*BUSCOfEOG7MH16D*" from our subsequent analyses because sequences of this gene consistently failed to be predicted by AUGUSTUS across the 96 genomes.

Sequence alignment, alignment trimming, and removal of spurious sequences and low-quality genes

For each ortholog group, we first translated nucleotide sequences into amino acid sequences using a custom Perl script, taking into account the differential meaning of the CUG codon in the CUG-Ser clade of yeasts whose CUG codon encodes serine, instead of leucine (Dujon 2010; Mühlhausen and Kollmar 2014; Hittinger *et al.* 2015; Riley *et al.* 2016). Next, we aligned the amino acid sequences using the E-INS-i strategy as implemented by the program MAFFT, version 7.215 (Katoh and Standley 2013), with the default gap opening penalty (–op = 1.53). We then used a custom Perl script to map the nucleotide sequences on the amino acid alignment and to generate the codon-based nucleotide alignment. Regions of ambiguous alignment in codon-based nucleotide alignments were trimmed using the trimAl software, version 1.4 (Capella-Gutierrez *et al.* 2009) with the "gappyout" option on; otherwise, default settings were assumed. Finally, the trimmed codon-based alignments were translated into trimmed amino acid alignments.

To minimize the inclusion of potentially spurious or paralogous sequences, the maximum likelihood (ML) phylogram for the trimmed

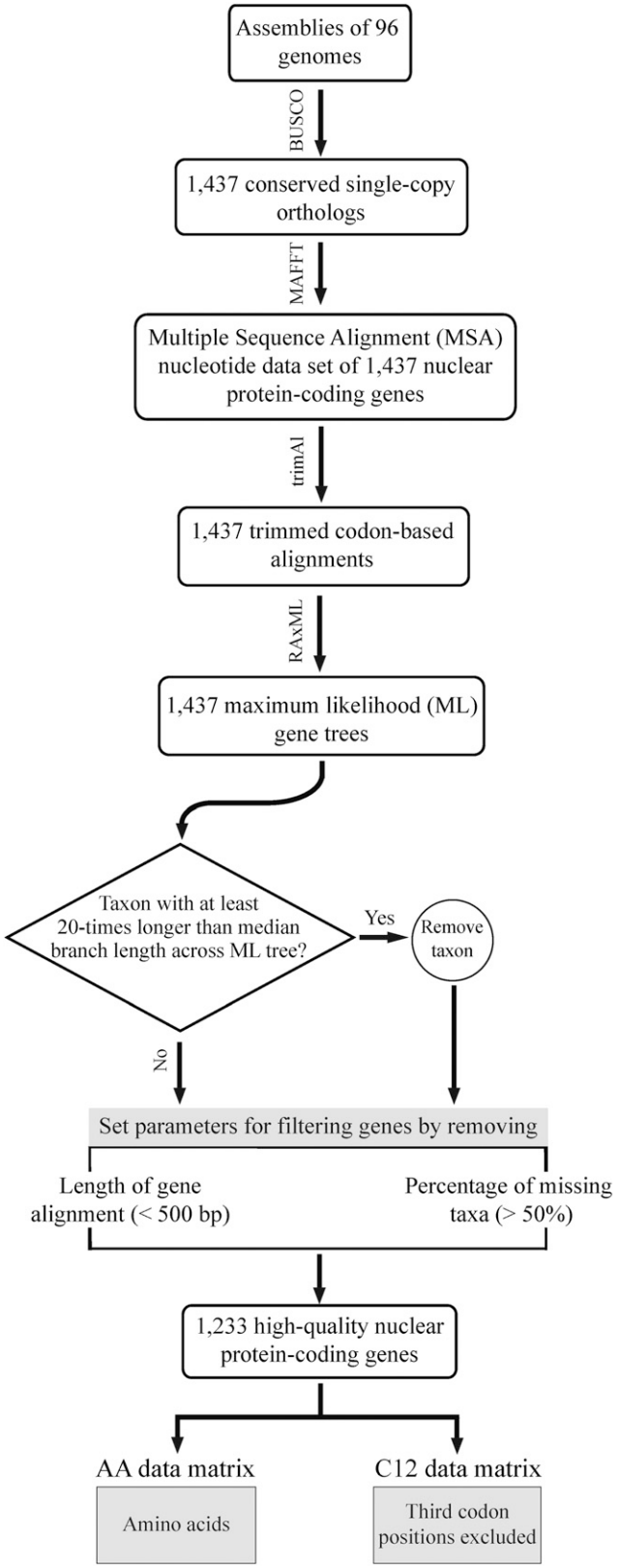

Figure 1 Workflow illustrating the steps involved in the construction of the two phylogenomic data matrices used in this study.

codon sequence alignment of each ortholog group was inferred under an unpartitioned "GTR (Tavaré 1986) + GAMMA (Yang 1994, 1996)" model as implemented in RA×ML, version 8.2.3 (Stamatakis 2014). Sequences whose terminal branch (leaf) lengths were at least 20 times longer than the median of all terminal branch lengths across the ML phylogram for a given orthologous group were excluded. In total, 49 sequences from 42 ortholog groups were removed. The resulting gene alignments were further filtered by length of trimmed gene alignment (alignments that were <500 bp in length were removed from downstream analyses) and taxon number (alignments with <50% gene occupancy, *i.e.*, that contained fewer than 48 taxa, were removed from downstream analyses).

The remaining 1233 ortholog groups were used to generate two data matrices: (A) a C12 data matrix that included only the first and second codon positions of every gene (third codon positions were excluded because they showed much higher variation in GC content than the first and second codon positions; Figure S1); and (B) an amino acid (AA) data matrix that included the amino acid sequences of every gene.

Phylogenetic analysis

Phylogenetic analysis was performed separately for the AA and C12 data matrices using the ML optimality criterion (Felsenstein 1981), under two different approaches: concatenation (Huelsenbeck *et al.* 1996; Rokas *et al.* 2003; Philippe *et al.* 2005) and coalescence (Edwards 2009). In the concatenation (*i.e.*, total evidence) approach, individual gene alignments are concatenated into a single data matrix and then analyzed jointly to infer the species phylogeny. Although this phylogenomic approach often yields a strongly supported phylogeny, it assumes that all individual genes have the same evolutionary history. In the coalescence approach, individual gene alignments are first used to estimate the individual gene trees, which are then used as input data to estimate the species phylogeny. Unlike the concatenation approach, the coalescence approach can efficiently account for differences in the evolutionary history among individual gene trees (Liu *et al.* 2015). However, the coalescence approach can be sensitive to errors and biases in estimating individual gene trees (Mirarab *et al.* 2015; Springer and Gatesy 2016), which in turn may mislead inference of the species phylogeny.

To infer the concatenation phylogeny for the AA data matrix, we used an unpartitioned "PROTGAMMALG" model of amino acid substitution, as 681 out of 1233 genes favored "LG (Le and Gascuel 2008) + Gamma (Yang 1994, 1996)" for rate heterogeneity among sites as best-fitting model (Figure S2). To infer the concatenation phylogeny for the C12 data matrix, we used an unpartitioned "GTR (Tavaré 1986) + GAMMA (Yang 1994, 1996)" model of nucleotide substitution. In both cases, phylogenetic reconstruction was performed using five randomized maximum parsimony trees and five random trees as starting trees in RA×ML (Stamatakis 2014). Branch support for each internode was evaluated with 100 rapid bootstrapping replicates (Stamatakis *et al.* 2008). Finally, we also used a gene-based partition scheme (1233 partitions) to separately conduct ML tree search for the AA and C12 data matrices, in which parameters of evolutionary model (amino acid; see below; DNA, GTR+G) were separately estimated for each orthologous group (-q option) in RA×ML. As the ML trees produced by the gene-based partition scheme on both data matrices were topologically identical to the ML trees produced by the unpartitioned scheme (results are deposited on the figshare repository), and the computational resources required for partitioned analyses are much greater, bootstrap support and phylogenetic signal analyses were performed using only the unpartitioned scheme.

For the coalescence-based analyses of the AA data matrix, the best-fitting model of amino acid evolution for each orthologous group was

selected using the Bayesian Information Criterion (BIC) (Schwarz 1978), as implemented in ProtTest 3.4 (Darriba *et al.* 2011). For the C12 data matrix, the GTR + GAMMA model was used to accommodate nucleotide substitution and rate heterogeneity among sites. In both cases, we inferred the best-scoring unrooted ML gene tree for every ortholog group by running 10 separate ML searches using RA×ML. Branch support for each internode was evaluated with 100 rapid bootstrapping replicates (Stamatakis *et al.* 2008). Individually estimated ML gene trees were used as input to estimate the coalescent-based phylogenies for the AA and C12 data matrices using the ASTRAL software, version 4.7.7 (Mirarab *et al.* 2014). The robustness of these phylogenies was evaluated by the multilocus bootstrap approach (Seo 2008) with 100 replicates, each of which consisted of individual gene trees each selected randomly from the set of 100 rapid bootstrapping trees available for each gene.

Finally, we used internode certainty (IC) to quantify the incongruence by considering all most prevalent conflicting bipartitions for each individual internode among individual gene trees (Salichos and Rokas 2013; Salichos *et al.* 2014; Kobert *et al.* 2016). The (partial) IC values were calculated from the set of the 1233 ML gene trees (Kobert *et al.* 2016), as implemented in RA×ML, version 8.2.3. We found that the mean IC values of our 93 internal branches in both the concatenation- and coalescence-based phylogenies inferred from the AA data matrix were slightly higher than those of 93 internal branches in both the concatenation- and coalescence-based phylogenies inferred from the C12 data matrix (mean IC values in the concatenation- and coalescence-based phylogenies from the AA data matrix are 0.404 and 0.399, respectively, whereas mean IC values in the concatenation- and coalescence-based phylogenies from the C12 data matrix were 0.368 and 0.367, respectively). Thus, phylogenetic trees based on the AA data matrix showed lower levels of incongruence than those based on the C12 data matrix, in agreement with previous phylogenomic studies at similar evolutionary depths (Rokas *et al.* 2003; Salichos and Rokas 2013; Riley *et al.* 2016; Shen *et al.* 2016). Finally, the incongruence among the 1233 individual (partial) gene trees generated by analysis of AA or C12 data was visualized in the form of a phylogenetic supernetwork [Figure S8; supernetwork by the Z-closure method with "tree size-weighted means" option implemented using the SplitsTree software, version 4.14.4 (Huson and Bryant 2006)].

Selecting subsets of genes with strong phylogenetic signal

Selecting strongly supported genes has been empirically shown to reduce incongruence among gene trees in phylogenetic analyses (Salichos and Rokas 2013; Wang *et al.* 2015). Since the AA data matrix showed lower levels of incongruence than the C12 data matrix (see above result), we examined the phylogenetic signal of individual genes using only the gene alignments of the AA data matrix. To quantify support for individual gene trees, we used two common phylogenetic measures on the AA data matrix: (1) the average bootstrap support (ABS), which was calculated using a custom Perl script, and corresponds to the average of bootstrap support values across the ML tree of a given gene; (2) the relative tree certainty (RTC), which was calculated in RA×ML, and corresponds to the average of all IC values across the ML tree of a given gene (Salichos *et al.* 2014; Kobert *et al.* 2016). IC values on each ML tree were calculated by examining the bipartitions present in the topologies generated by the 100 rapid bootstrap replicates.

Four subsets of 1233 genes in the AA data matrix were constructed based on the ABS and RTC measures, respectively: the first two subsets included the 616 genes (top 50%) having the highest ABS or RTC values, respectively; the remaining two subsets included the 308 genes (top 25%) having the highest ABS or RTC values, respectively. Since the subsets constructed using the ABS and RTC phylogenetic measures showed no significant differences in the sets of genes included (Figure S3), we used the two subsets (top 50 and 25%) based on ABS for subsequent analyses. For each subset, the concatenation phylogeny and the coalescent-based phylogeny, as well as their clade support values (bootstrap support and IC values), were separately estimated using RA×ML and ASTRAL by following the procedures described above.

Data availability

All data matrices and their resulting phylogenies have been deposited on the figshare repository at DOI: 10.6084/m9.figshare.3370675.

RESULTS AND DISCUSSION

Genome completeness

Contig or scaffold N50 (*i.e.*, the contig or scaffold size above which 50% of the total length of the sequence assembly can be found) is by far the most widely used measure to assess the quality of a genome assembly (Yandell and Ence 2012). The higher N50 is, the better the assembly generally is. Nonetheless, this value does not assess assembly completeness in terms of gene content. Thus, we assessed completeness for each of 96 fungal genomes using the BUSCO set of 1438 single-copy, conserved genes among 125 fungi (Simão *et al.* 2015). We found that the percentage of "complete" BUSCO genes among the 96 genomes ranged from 65.2 to 98.5% of 1438 fungal BUSCO genes, with the average being 94.2% (Figure 2 and Table S2 for detailed values).

Only 10 of the 96 fungal genomes had <90% of 1438 fungal BUSCO genes, with the assemblies of *Hanseniaspora valbyensis* (62.2%) and *H. uvarum* (64.5%) having the lowest coverage "complete" BUSCO genes, with ~510 BUSCO genes either "missing," "fragmented," or "duplicated" in each of the two genomes. After performing GO term enrichment analysis of 365 genes missing in both *H. valbyensis* and *H. uvarum* with the *S. cerevisiae* GOSlim annotations using the Cytoscape plugin BinGO (Shannon *et al.* 2003; Maere *et al.* 2005), we found eight significantly overrepresented GO terms (Biological process: mitochondrion organization, mRNA processing, RNA splicing, peroxisome organization; Cellular component: endomembrane system, endoplasmic reticulum, peroxisome, Golgi apparatus; adjusted *P*-value ≤0.05; see Table S4A). Five of the eight overrepresented GO terms are also found in analysis of all 1437 BUSCO genes (Table S4B). Since the N50 values for the assemblies of the two *Hanseniaspora* genomes are much higher (*H. valbyensis*, N50 = 332,595 bp; *H. uvarum*, N50 = 251,359 bp) than the N50 (83,174 bp) of one (*H. vineae*) of their closest relatives, their low coverage was unlikely due to lower quality of the genome assemblies. Rather, the *Hanseniaspora* genomes may be missing genes from specific functional categories or alternatively these genes were not detected due to the accelerated evolutionary rates of these genomes. This inference is consistent with the observation that the ancestral internode length leading to the two *Hanseniaspora* genomes was the longest in the yeast phylogeny (Figure 3), an observation also reported in the original genome study for *H. valbyensis* (Riley *et al.* 2016).

Data matrix completeness

Following orthology identification, alignment trimming, and removal of spurious sequences and low-quality genes, we retained 1233 of the original 1438 orthologous groups from the 96 genomes (see *Materials and Methods*). None of the genomes of the 96 species had all 1233 orthologous groups, but 88 had >1000 orthologous groups each (Figure 2 and Table S2). The percentage of gene occupancy in the remaining eight species ranged from 60.5% (746/1233) to 78.6% (969/1233). Two recent studies showed that nonrandom

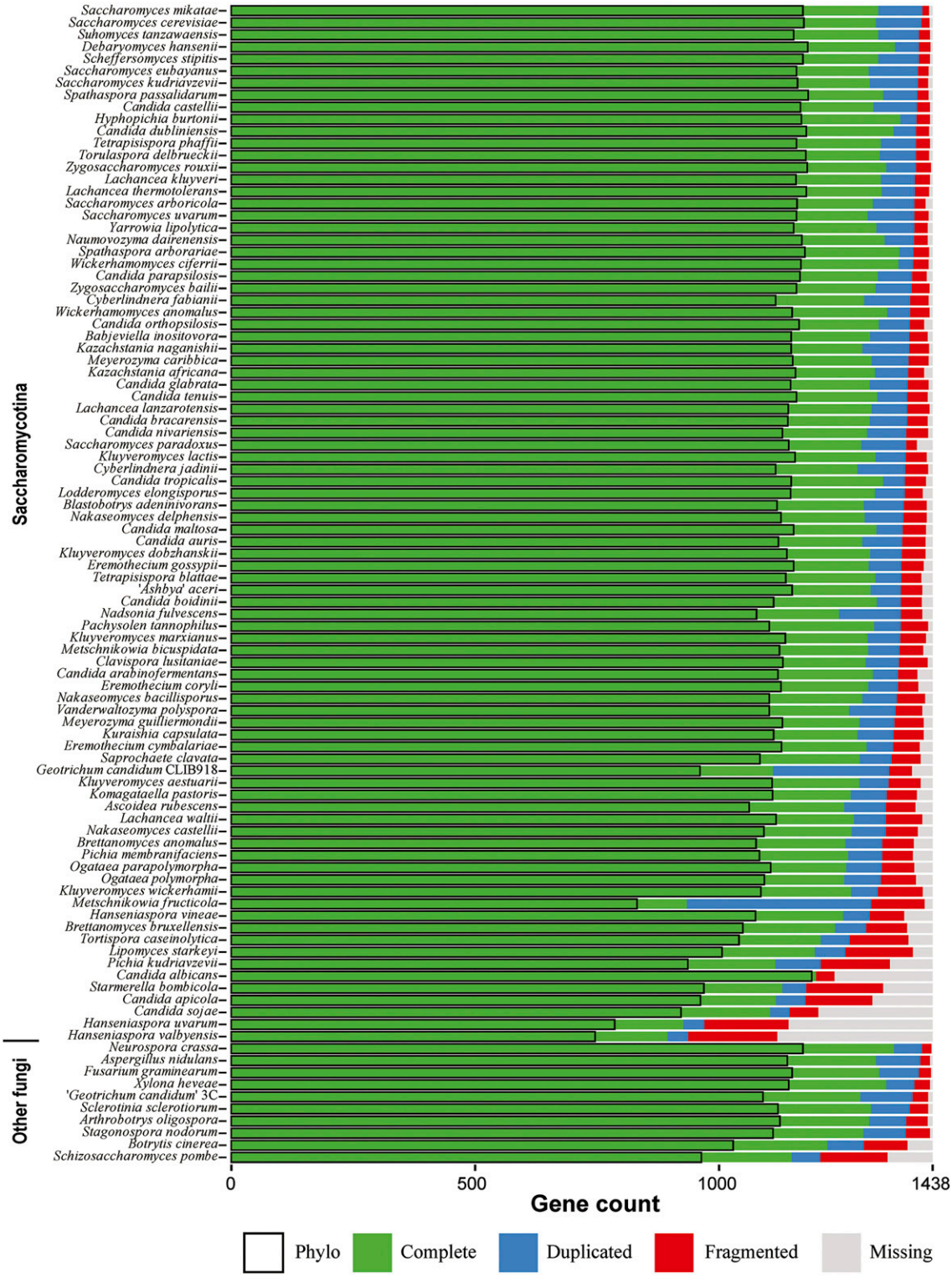

Figure 2 Genomic quality assessment of the 86 yeast and 10 outgroup fungal genomes used in this study. The bar plot next to each species indicates the fractions of BUSCO genes that are present or missing in each genome. "Complete": fraction of single-copy, full-length genes; "Duplicated": fraction of multiple-copy, full-length genes; "Fragmented": fraction of genes with a partial sequence; "Missing": fraction of genes not found in the genome; "Phylo": fraction of single-copy "Complete" genes used to construct the phylogenomic data matrices. Yeast species are arranged along the Y-axis in ascending order of the total number of "Fragmented" and "Missing" genes. The exact value of quality assessment of each species can be found in Table S2.

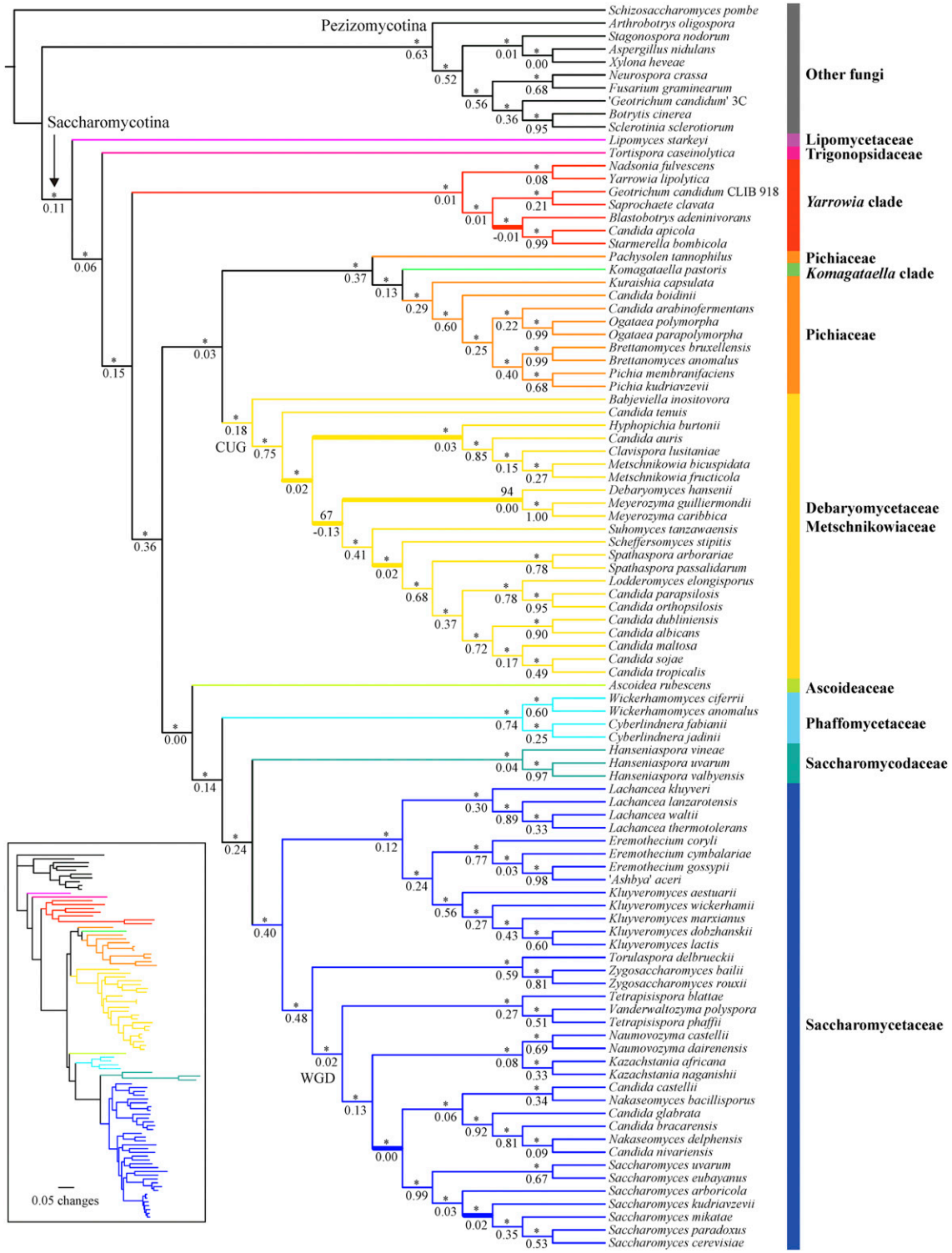

Figure 3 The phylogenetic relationships of Saccharomycotina yeasts inferred from the concatenation-based analysis of a 1233 single-copy BUSCO gene amino acid (AA) data matrix. The ML phylogeny was reconstructed based on the concatenation amino acid data matrix (609,899 sites) under an unpartitioned LG + GAMMA substitution model using RA×ML version 8.2.3 (Stamatakis 2014). Branch support values near internodes are indicated as bootstrap support value (above) and internode certainty (below), respectively. * indicates bootstrap support values ≥95%. Thicker branches show conflicts between concatenation-based phylogeny (Figure 3) and coalescence-based phylogeny (Figure 4). Note, branch lengths on the ML tree are given in the *inset* at the bottom left.

bias in missing data might be potentially problematic for phylogenetic inference (Hosner *et al.* 2016; Xi *et al.* 2016), particularly for coalescence-based phylogenetic inference (Hovmöller *et al.* 2013). Interestingly, the placements for the eight low gene-coverage species in our study were robust in both concatenation and coalescent analyses, as well as in the two data matrices and subsampling analyses (see results below), suggesting that the impact of missing data in this study is negligible.

Among the 1233 orthologous groups, the percentage of sequence occupancy ranged from 50 to 100%, with an average value of 90%; 1084 orthologous groups displayed at least 80% sequence occupancy, and 24 contained sequences from all 96 species (Figure S4 and Table S3). In addition, gene alignment lengths ranged from 501 to 14,562 bp, with an average length of 1484 bp (Figure S5 and Table S3). The AA and C12 data matrices contained a total of 609,899 and 1,219,798 sites, respectively.

A genome-wide yeast phylogeny

All phylogenetic analyses consistently separated the 10 nonyeast outgroup taxa (nine Pezizomycotina and one Taphrinomycotina species) from the Saccharomycotina yeasts (Figure 3, Figure 4, Figure S6, and Figure S7). Surprisingly, the genomes of two yeast isolates purported to be from the same species (*Geotrichum candidum* isolates CLIB 918 and 3C) were placed into two different clades; *G. candidum* CLIB 918 (Morel *et al.* 2015) was nested within Saccharomycotina yeasts, whereas "*G. candidum*" 3C (Polev *et al.* 2014) was nested within the Pezizomycotina outgroup. Given its phylogenetic placement, we infer that the genome sequence of the isolate "*G. candidum*" 3C represents a misidentified Pezizomycotina species.

Most internodes in the concatenation phylogenies inferred from the AA and C12 data matrices received high bootstrap support values that were ≥95% (AA, 91 out of 93 internodes; C12, 88 out of 93 internodes) (Figure 3 and Figure S6). Similar to the results of the concatenation approach, the two coalescence-based phylogenies were also robustly supported, with 86/93 internodes (AA data matrix) and 86/93 internodes (C12 data matrix) showing 95% or greater bootstrap support (Figure 4 and Figure S7). There were five conflicting internodes between the AA and C12 concatenation-based phylogenies and two conflicting internodes between the AA and C12 coalescence-based phylogenies. In addition, comparison of the concatenation-based phylogeny to the coalescence-based phylogeny showed eight topological differences in the phylogenies inferred using the AA data matrix and five in the phylogenies inferred using the C12 data matrix. These topological differences are discussed below.

Stable and conflicted internodes

Of the nine major Saccharomycotina lineages, eight were monophyletic in all analyses; the only exception was the family Pichiaceae. This lineage was paraphyletic because *Komagataella pastoris*, which belongs to the *Komagataella* clade, groups within it (Figure 3). Overall, the family Lipomycetaceae was resolved as the earliest-branching lineage of Saccharomycotina yeasts, followed by the family Trigonopsidaceae and the *Yarrowia* clades (Figure 3). A clade consisting of the family Pichiaceae, the CUG-Ser clade, and the *Komagataella* clade was well supported. The family represented by the most genome sequences, the Saccharomycetaceae, was recovered as the sister group to the family Saccharomycodaceae. This Saccharomycetaceae/Saccharomycodaceae clade was sister to the Phaffomycetaceae. These relationships were mostly recovered in two multigene studies (Kurtzman and Robnett 2013; Mühlhausen and Kollmar 2014) and one phylogenomic study based on 25 yeast genomes (Riley *et al.* 2016), and they are broadly consistent with the most recent views of the yeast phylogeny (Hittinger *et al.* 2015). Finally, eight of the nine major Saccharomycotina lineages were robustly and strongly supported in both concatenation- and coalescence-based analyses. The only exception was the family Ascoideaceae, whose support values under concatenation analyses were high (AA: BS = 95%; C12: BS = 97%) but were much lower under coalescence-based analyses (AA: BS = 53%; C12 = 45%). This finding is consistent with the instability in the placement of *Ascoidea rubescens* inferred from multiple analyses of different data matrices in the original genome study (Riley *et al.* 2016). Thus, although this family was consistently recovered as the closest relative of the Saccharomycetaceae/Saccharomycodaceae/Phaffomycetaceae clade in our analyses (Figure 3, Figure 4, Figure S6, and Figure S7), its current placement in the yeast phylogeny should be considered tenuous and unresolved.

To quantify incongruence between the 1233 orthologous groups across the 93 internodes of the yeast phylogeny, we used the 1233 individual ML gene trees to calculate IC values (Salichos and Rokas 2013; Salichos *et al.* 2014; Kobert *et al.* 2016). Our results showed that most of the internodes in concatenation- and coalescence-based phylogenies inferred from AA and C12 data matrices had IC values >0 (Figure 3, Figure 4, Figure S6, and Figure S7), suggesting that those relationships were recovered by the majority of 1233 genes. For the aforementioned internodes that were incongruent between approaches (concatenation *vs.* coalescence) or data matrices (AA *vs.* C12), their IC values were often <0 (Figure 3, Figure 4, Figure S6, and Figure S7). Examination of the phylogenetic supernetworks built from the AA and C12 data matrices also suggests that the degree of incongruence (visualized as the degree of reticulation in Figure S8) among gene trees was negatively correlated with IC values. These conflicting internodes occurred within the WGD/allopolyploidization clade (Wolfe and Shields 1997; Marcet-Houben and Gabaldón 2015) in the family Saccharomycetaceae (three topological differences), the CUG-Ser clade (four topological differences), and the *Yarrowia* clade (one topological difference).

Within the WGD clade, the concatenation phylogenies identified the *Nakaseomyces* clade as the sister group to the genus *Saccharomyces* (Figure 3 and Figure S6), a relationship supported by several rare genomic changes (Scannell *et al.* 2006) and phylogenomic analysis (Shen *et al.* 2016), whereas the phylogenies inferred from the coalescence-based approach strongly supported the *Kazachstania*/*Naumovozyma* clade as the sister group to the genus *Saccharomyces* (Figure 4 and Figure S7), a relationship supported by phylogenomic (Salichos and Rokas 2013) and multigene (Mühlhausen and Kollmar 2014) analyses. The monophyly of yeasts for the WGD clade was recovered by the concatenation- and coalescence-based phylogenies on the AA data matrix, as well as by the concatenation-based phylogeny on the C12 data matrix, consistent with recent studies (Dujon 2010; Salichos and Rokas 2013; Wolfe *et al.* 2015). In contrast, the coalescence-based phylogeny on the C12 data matrix weakly supported a paraphyly of the WGD clade, in which a ZT clade composed of the genus *Zygosaccharomyces* and the genus *Torulaspora* nested within the WGD clade in agreement with the results from multigene studies (Kurtzman and Robnett 2013; Mühlhausen and Kollmar 2014). Interestingly, a recent examination of 516 widespread orthologs from 25 yeast genomes inferred that the WGD event was the result of an allopolyploidization between the KLE clade (the genus *Kluyveromyces*, the genus *Lachancea*, and the genus *Eremothecium*) and the ZT clade (Marcet-Houben and Gabaldón 2015), providing a potential explanation for the observed instability of the WGD clade. A sister relationship between *S. arboricola* and *S. kudriavzevii* was only recovered by the coalescence-based phylogeny inferred from the AA data matrix (Figure 4); this relationship contrasts with all other phylogenies inferred in this study, as well as those obtained in most published studies

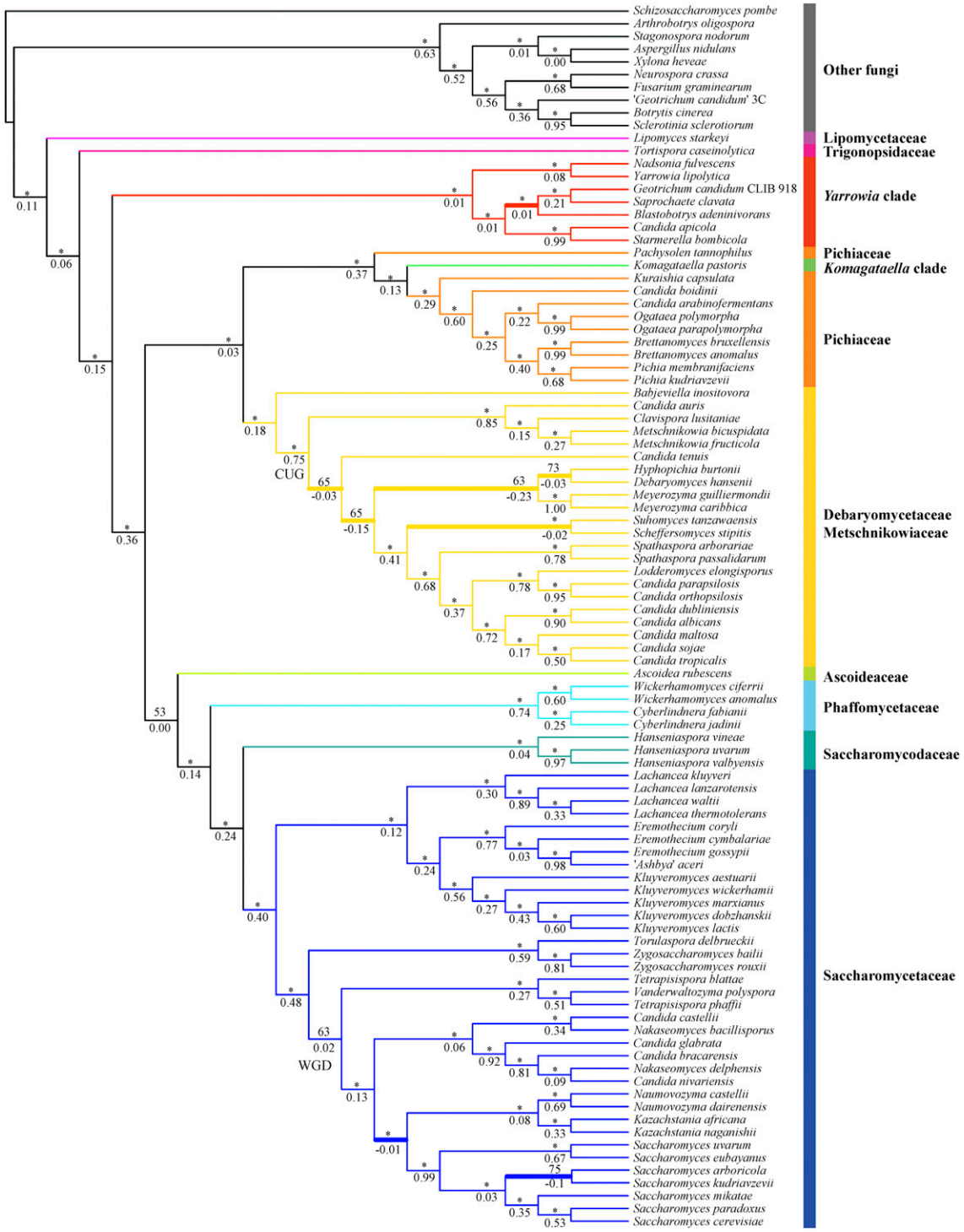

Figure 4 The phylogenetic relationships of Saccharomycotina yeasts inferred from the coalescence-based analysis of a 1233 single-copy BUSCO gene amino acid (AA) data matrix. The coalescence-based phylogeny estimation was conducted using ASTRAL version 4.7.7 (Mirarab *et al*. 2014). Branch support values near internodes are indicated as bootstrap support value (above) and internode certainty (below), respectively. * indicates bootstrap support values ≥95%. Thicker branches show conflicts between coalescence-based phylogeny (Figure 4) and concatenation-based phylogeny (Figure 3).

(Scannell *et al*. 2011; Hittinger 2013; Liti *et al*. 2013; Mühlhausen and Kollmar 2014; Hittinger *et al*. 2015).

Within the CUG-Ser clade, *Suhomyces* (*Candida*) *tanzawaensis* was strongly recovered as sister to *Scheffersomyces stipitis* by both methods on the C12 data matrix, as well as by the coalescence-based phylogeny on the AA data matrix. In contrast, the concatenation-based phylogeny using the AA data matrix strongly supported a sister relationship between *Su. tanzawaensis* and a clade composed of *Sc. stipitis*, the genus

Spathaspora, and some of the yeasts within the CUG-Ser clade, in agreement with the results of the original study describing the *Su. tanzawaensis* genome (Riley *et al.* 2016). The closest relatives of the genus *Meyerozyma* were either *Debaryomyces hansenii*, *Candida tenuis*, or a clade containing *D. hansenii* and *Hyphopichia burtonii*, depending on the analysis and data matrix considered.

Finally, within the *Yarrowia* clade, the concatenation phylogenies inferred from AA and C12 data matrices placed the *Candida apicola/Starmerella bombicola* clade as the sister group to *Blastobotrys (Arxula) adeninivorans* (Figure 3 and Figure S6), whereas the coalescence-based phylogenies inferred from AA and C12 data matrices supported *B. adeninivorans* as sister to a clade containing *Saprochaete clavata* and *G. candidum* isolate CLIB 918 (Figure 4 and Figure S7). Finally, the concatenation- and coalescence-based phylogenies inferred from AA and C12 data matrices consistently recovered a sister group of *Nadsonia fulvescens* and *Yarrowia lipolytica*, but its resolution in coalescence-based phylogenies inferred from C12 data matrices received weak median bootstrap support (BS = 65%). This group was not recovered in the previous multigene study of these taxa (Kurtzman and Robnett 2013).

Selecting genes with strong phylogenetic signal reduces incongruence

To examine whether the use of genes with strong phylogenetic signal could reduce incongruence among individual gene trees (Salichos and Rokas 2013; Wang *et al.* 2015), we constructed two AA data matrices containing the 308 (top 25%) or 616 (top 50%) ortholog groups showing the highest average bootstrap values in their bootstrap consensus gene trees and reconstructed their phylogenies by concatenation and coalescence. The IC values of most internodes in both of the concatenation-based (ML) phylogeny (all orthologous groups, average IC value = 0.41; top 50%, average IC value = 0.55; top 25%, average IC value = 0.64) and the coalescence-based phylogeny (all orthologous groups, average IC value = 0.40; top 50%, average IC value = 0.54; top 25%, average IC value = 0.65) increased (Figure 5), suggesting that the use of genes with strong phylogenetic signal decreased the amount of incongruence in the yeast phylogeny.

In agreement with the IC results, there were fewer topological differences between the concatenation- and coalescence-based phylogenies in the two reduced data matrices relative to the full data matrix (five topological differences instead of eight) (Figure 6 and Figure S9). These five remaining conflicting internodes occurred within the CUG-Ser clade and the *Yarrowia* clade (Figure 6 and Figure S9). Specifically, the concatenation phylogeny inferred from the top 50% data matrix was topologically identical with that inferred from the complete data matrix, albeit more weakly supported. Furthermore, unlike the coalescence-based phylogeny recovered from the complete data matrix, the coalescence-based phylogeny from the top 50% data matrix recovered the *Nakaseomyces* clade as the sister group to the genus *Saccharomyces* (Figure 6). For the top 25% data matrix, both concatenation- and coalescence-based phylogenies supported this sister relationship. However, they consistently supported the paraphyly of the WGD clade (Figure S9), which was not recovered in concatenation- and coalescence-based phylogenies inferred from the top 50% data matrix but was observed in previous multigene studies (Kurtzman and Robnett 2013; Mühlhausen and Kollmar 2014).

In summary, 75/83 internodes in this 86-taxon phylogeny of the Saccharomycotina yeasts are robust to different approaches (concatenation *vs.* coalescence) and phylogenomic data matrices (AA *vs.* C12), while 72 internodes are highly supported and the remaining 11 internodes are still unresolved or equivocal (Figure 7). After comparing our results to those of the most recent consensus view of the yeast phylogeny (Hittinger *et al.* 2015) and to those of the most recent phylogenomic study (Riley *et al.* 2016), we found that 14 of our strongly supported internodes are new to our study (Hittinger *et al.* 2015; Riley *et al.* 2016) (Figure 7).

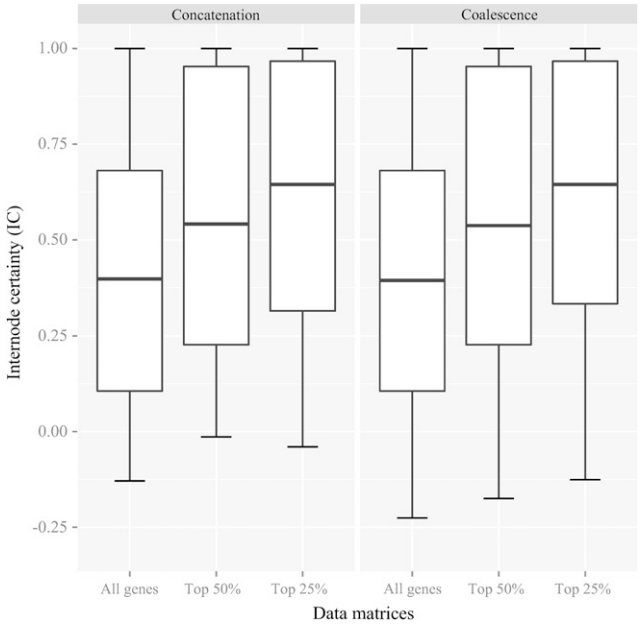

Figure 5 Comparison of the distributions of internode certainty (IC) values across three data matrices and two approaches. The three data matrices are: "All genes," all 1233 genes in the AA data matrix; "Top 50%," AA data matrix including the 616 genes with the strongest phylogenetic signal; "Top 25%," AA data matrix including the 305 genes with the strongest phylogenetic signal (see *Materials and Methods*). For each data matrix, the set of individual ML gene trees is used to calculate (partial) internode certainty (IC) values for all internodes in the concatenation-based ML phylogeny (left panel) and the coalescence-based ASTRAL phylogeny (right panel), respectively. Rectangles in the boxplot denote 1st and 3rd quartiles. Horizontal thick bars represent mean IC values.

The 11 remaining unresolved or equivocal internodes (Figure 7) are placed at different evolutionary depths and in different lineages. Therefore, different strategies might be potentially helpful to resolve some of them, including but not limited to an increase in the density of taxonomic sampling (Zwickl and Hillis 2002; Heath *et al.* 2008), the adoption of different types of data such as rare genomic changes (RGCs) (Rokas and Holland 2000; Scannell *et al.* 2006; Butler *et al.* 2009; Polzin and Rokas 2014), and the use of mixture substitution models that better account for across-site heterogeneity [*e.g.*, CAT model implemented in PhyloBayes (Lartillot and Philippe 2004; Lartillot *et al.* 2009); currently, using this model in phylogenomic data matrices such as ours is computationally prohibitive]. Finally, we note that some of these internodes may represent genuine polytomies, in other words cases in which an ancestral yeast lineage split into more than two descendant lineages at approximately the same time (Rokas and Carroll 2006).

Conclusions

Twenty years ago, the genome sequence of *S. cerevisiae* (Goffeau *et al.* 1996) ushered yeast biology into the age of genomics.

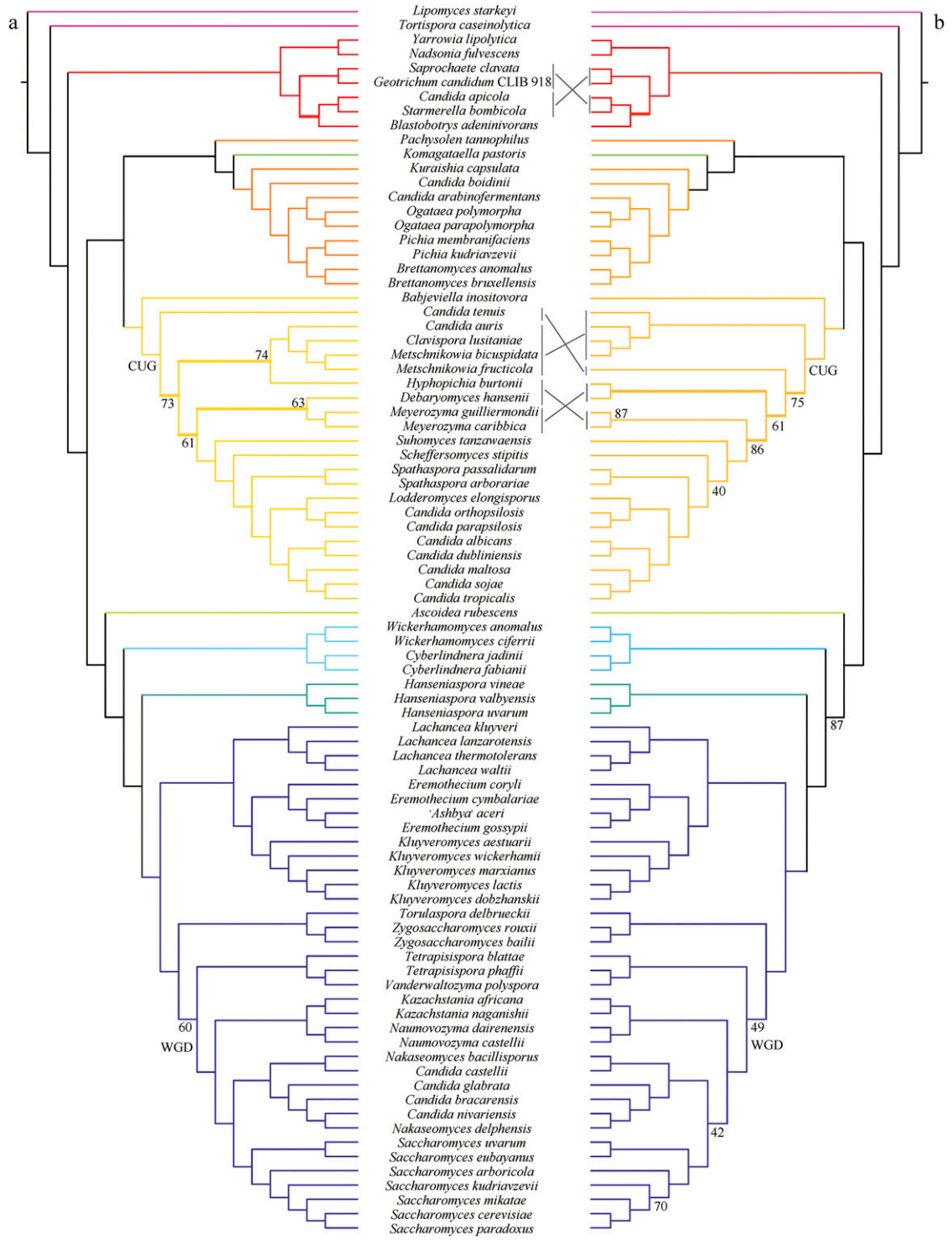

Figure 6 Conflicts in the phylogenetic relationships of Saccharomycotina yeasts inferred from the concatenation-based (A) and coalescence-based (B) analysis of the 616 genes in the AA data matrix whose bootstrap consensus gene trees had the highest average bootstrap support (top 50%). Branch support values near internodes are indicated as bootstrap support values (internodes without designation have values ≥95%). Thicker branches show conflicts between coalescence-based phylogeny and concatenation-based phylogeny. Note that outgroup taxa are not shown.

Although we still lack genomic data from most of the known yeast biodiversity, the public availability of dozens of yeast genomes provided us with an opportunity to examine the quality of genomic data presently available for the lineage (Figure 2) and infer the backbone phylogeny of the Saccharomycotina (Figure 3 and Figure 4). With several large-scale efforts to sample yeast biodiversity currently underway, such as the 1002 Yeast Genomes Project focusing on *S. cerevisiae* (http://1002genomes.u-strasbg.fr), the

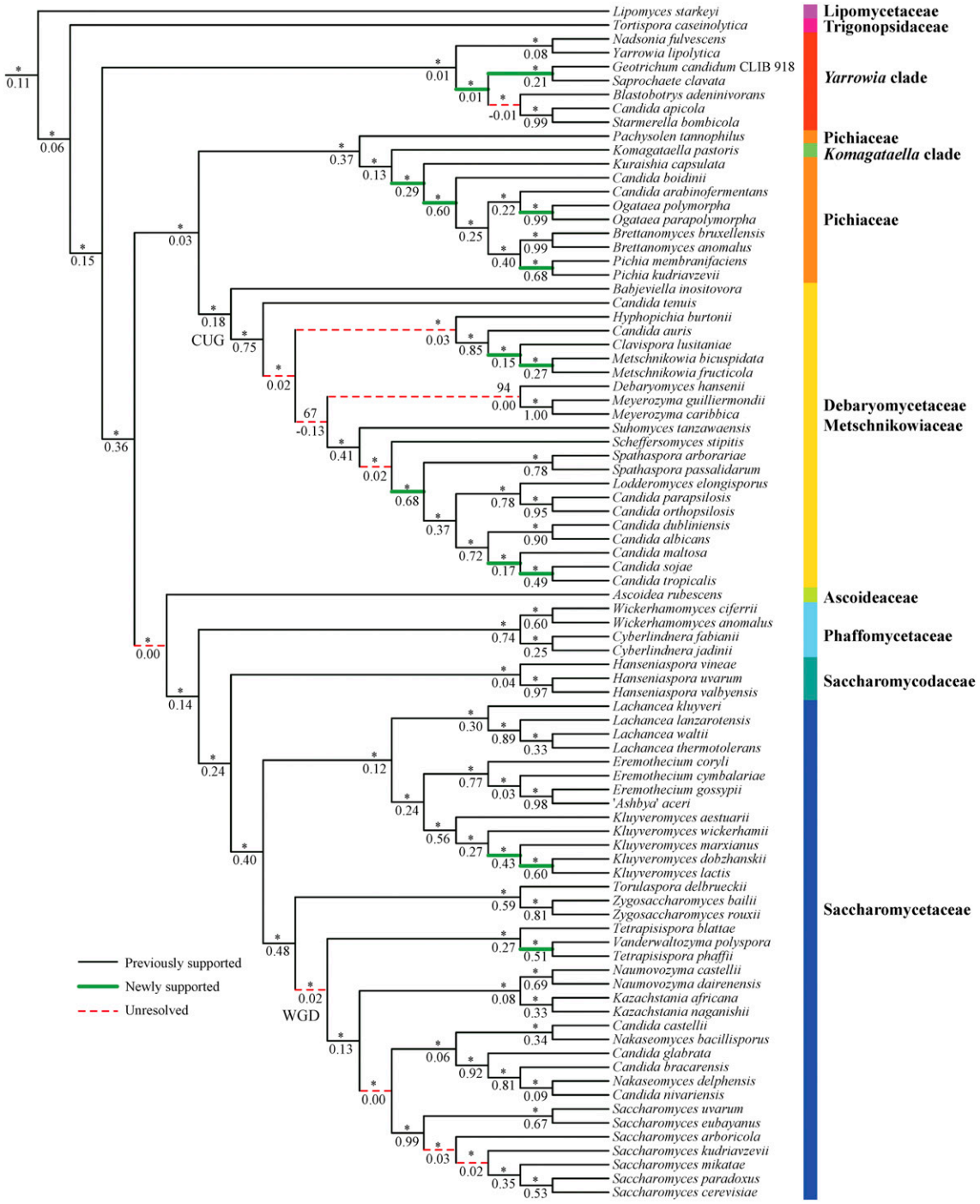

Figure 7 Supported and unresolved internodes in phylogeny of Saccharomycotina yeasts. Branch support values near internodes are indicated as bootstrap support value (above) and internode certainty (below), respectively. * indicates bootstrap support values ≥95%. Solid lines indicate internodes that are robustly and highly supported by different approaches (concatenation, coalescence) and phylogenomic data matrices (AA, C12). Internodes reported as resolved in the most recent consensus view of yeast phylogeny (Hittinger *et al.* 2015) or the most recent yeast phylogenomic study (Riley *et al.* 2016) are labeled as black solid lines, whereas those that are new to this study are labeled as thicker green solid lines. Dashed red lines indicate internodes that show conflict or are weakly supported; we consider such internodes to be unresolved or equivocal. Note that the topology shown is the same as that shown in Figure 3 but with the outgroup taxa removed.

iGenolevures Consortium (http://gryc.inra.fr), and the Y1000+ Project focusing on sequencing the genomes of all known species of the subphylum Saccharomycotina (http://y1000plus.org), the phylogenomic analyses reported in this study, including the classification of branches into resolved and unresolved (Figure 7), provide a robust roadmap for future comparative research across the yeast subphylum, while highlighting clades in need of further scrutiny.

ACKNOWLEDGMENTS

We thank Thomas W. Jeffries, Meredith Blackwell, and the Department of Energy (DOE) Joint Genome Institute for releasing several genome sequences through MycoCosm prior to their formal publication (Riley *et al.* 2016) and Abigail Lind for help with the GO term enrichment analysis. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the United States Department of Agriculture (USDA). USDA is an equal opportunity provider and employer. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University and of the UW-Madison Center for High Throughput Computing. This work was supported by the National Science Foundation (DEB-1442113 to A.R.; DEB-1442148 to C.T.H.), in part by the DOE Great Lakes Bioenergy Research Center (DOE Office of Science BER DE-FC02- 07ER64494), the USDA National Institute of Food and Agriculture (Hatch project 1003258 to C.T.H.), and the National Institutes of Health (NIAID AI105619 to A.R.). C.T.H. is an Alfred Toepfer Faculty Fellow, supported by the Alexander von Humboldt Foundation. C.T.H. is a Pew Scholar in the Biomedical Sciences, supported by the Pew Charitable Trusts.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Borneman, A. R., B. A. Desany, D. Riches, J. P. Affourtit, A. H. Forgan *et al.*, 2012 The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii* origins. FEMS Yeast Res. 12: 88–96.

Butler, G., M. D. Rasmussen, M. F. Lin, M. A. S. Santos, S. Sakthikumar *et al.*, 2009 Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature 459: 657–662.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Capella-Gutierrez, S., J. M. Silla-Martinez, and T. Gabaldon, 2009 TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada, 2011 ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27: 164–165.

Dujon, B., 2010 Yeast evolutionary genomics. Nat. Rev. Genet. 11: 512–524.

Eddy, S. R., 2011 Accelerated profile HMM searches. PLOS Comput. Biol. 7: e1002195.

Edwards, S. V., 2009 Is a new and general theory of molecular systematics emerging? Evolution. 63: 1–19.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17: 368–376.

Fitzpatrick, D. A., M. E. Logue, J. E. Stajich, and G. Butler, 2006 A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol. Biol. 6: 99.

Gibson, B., and G. Liti, 2015 *Saccharomyces pastorianus*: genomic insights inspiring innovation for industry. Yeast 32: 17–27.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.*, 1996 Life with 6000 Genes. Science 274: 546–567.

Hall, C., and F. S. Dietrich, 2007 The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. Genetics 177: 2293–2307.

Heath, T. A., S. M. Hedtke, and D. M. Hillis, 2008 Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46: 239–257.

Hittinger, C. T., 2013 Saccharomyces diversity and evolution: a budding model genus. Trends Genet. 29: 309–317.

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

Hittinger, C. T., A. Rokas, F.-Y. Bai, T. Boekhout, P. Gonçalves *et al.*, 2015 Genomics and the making of yeast biodiversity. Curr. Opin. Genet. Dev. 35: 100–109.

Hosner, P. A., B. C. Faircloth, T. C. Glenn, E. L. Braun, and R. T. Kimball, 2016 Avoiding missing data biases in phylogenomic inference: an empirical study in the Landfowl (Aves: Galliformes). Mol. Biol. Evol. 33: 1110–1125.

Hovmöller, R., L. L. Knowles, and L. S. Kubatko, 2013 Effects of missing data on species tree estimation under the coalescent. Mol. Phylogenet. Evol. 69: 1057–1062.

Huelsenbeck, J. P., J. J. Bull, and C. W. Cunningham, 1996 Combining data in phylogenetic analysis. Trends Ecol. Evol. 11: 152–158.

Huson, D. H., and D. Bryant, 2006 Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23: 254–267.

James, T. Y., F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter *et al.*, 2006 Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443: 818–822.

Katoh, K., and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30: 772–780.

Kobert, K., L. Salichos, A. Rokas, and A. Stamatakis, 2016 Computing the internode certainty and related measures from partial gene trees. Mol. Biol. Evol. 33: 1606–1617.

Kurtzman, C. P., and C. J. Robnett, 1994 Orders and families of ascosporogenous yeasts and yeast-like taxa compared from ribosomal RNA sequence similarities, pp. 249–258 in *Ascomycete Systematics: Problems and Perspectives in the Nineties*, edited by Hawksworth, D. L.. Plenum Press, New York.

Kurtzman, C. P., and C. J. Robnett, 1998 Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie van Leeuwenhoek 73: 331–371.

Kurtzman, C. P., and C. J. Robnett, 2003 Phylogenetic relationships among yeasts of the "*Saccharomyces* complex" determined from multigene sequence analyses. FEMS Yeast Res. 3: 417–432.

Kurtzman, C. P., and C. J. Robnett, 2007 Multigene phylogenetic analysis of the *Trichomonascus*, *Wickerhamiella* and *Zygoascus* yeast clades, and the proposal of *Sugiyamaella* gen. nov. and 14 new species combinations. FEMS Yeast Res. 7: 141–151.

Kurtzman, C. P., and C. J. Robnett, 2013 Relationships among genera of the *Saccharomycotina* (*Ascomycota*) from multigene phylogenetic analysis of type species. FEMS Yeast Res. 13: 23–33.

Kurtzman, C. P., and M. Suzuki, 2010 Phylogenetic analysis of ascomycete yeasts that form coenzyme Q-9 and the proposal of the new genera *Babjeviella*, *Meyerozyma*, *Millerozyma*, *Priceomyces*, and *Scheffersomyces*. Mycoscience 51: 2–14.

Kurtzman, C. P., C. J. Robnett, and E. Basehoar-Powers, 2008 Phylogenetic relationships among species of *Pichia*, *Issatchenkia* and *Williopsis* determined from multigene sequence analysis, and the proposal of *Barnettozyma* gen. nov., *Lindnera* gen. nov. and *Wickerhamomyces* gen. nov. FEMS Yeast Res. 8: 939–954.

Kurtzman, C. P., J. W. Fell, and T. Boekhout, 2011 *The Yeasts: A Taxonomic Study*. Elsevier Science, New York.

Lartillot, N., and H. Philippe, 2004 A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21: 1095–1109.

Lartillot, N., T. Lepage, and S. Blanquart, 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25: 2286–2288.

Le, S. Q., and O. Gascuel, 2008 An improved general amino acid replacement matrix. Mol. Biol. Evol. 25: 1307–1320.

Liang, D., X. X. Shen, and P. Zhang, 2013 One thousand two hundred ninety nuclear genes from a genome-wide survey support lungfishes as the sister group of tetrapods. Mol. Biol. Evol. 30: 1803–1807.

Libkind, D., C. T. Hittinger, E. Valério, C. Gonçalves, J. Dover *et al.*, 2011 Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. Proc. Natl. Acad. Sci. USA 108: 14539–14544.

Lin, Z., and W. H. Li, 2011 Expansion of hexose transporter genes was associated with the evolution of aerobic fermentation in yeasts. Mol. Biol. Evol. 28: 131–142.

Liti, G., A. N. N. Ba, M. Blythe, C. A. Müller, A. Bergström *et al.*, 2013 High quality de novo sequencing and assembly of the *Saccharomyces arboricolus* genome. BMC Genomics 14: 69.

Liu, L., Z. Xi, S. Wu, C. C. Davis, and S. V. Edwards, 2015 Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360: 36–53.

Liu, Y., J. W. Leigh, H. Brinkmann, M. T. Cushion, N. Rodriguez-Ezpeleta *et al.*, 2009 Phylogenomic analyses support the monophyly of Taphrinomycotina, including *Schizosaccharomyces* fission yeasts. Mol. Biol. Evol. 26: 27–34.

Louis, V. L., L. Despons, A. Friedrich, T. Martin, P. Durrens *et al.*, 2012 *Pichia sorbitophila*, an interspecies yeast hybrid, reveals early steps of genome resolution after polyploidization. G3 (Bethesda) 2: 299–311.

Maere, S., K. Heymans, and M. Kuiper, 2005 BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21: 3448–3449.

Marcet-Houben, M., and T. Gabaldón, 2015 Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. PLoS Biol. 13: e1002220.

Medina, E. M., G. W. Jones, and D. A. Fitzpatrick, 2011 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. 73: 116–133.

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson *et al.*, 2014 ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30: i541–i548.

Mirarab, S., M. S. Bayzid, B. Boussau, and T. Warnow, 2015 Response to Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree." Science 350: 171.

Morel, G., L. Sterck, D. Swennen, M. Marcet-Houben, D. Onesime *et al.*, 2015 Differential gene retention as an evolutionary mechanism to generate biodiversity and adaptation in yeasts. Sci. Rep. 5: 11571.

Mühlhausen, S., and M. Kollmar, 2014 Molecular phylogeny of sequenced *Saccharomycetes* reveals polyphyly of the alternative yeast codon usage. Genome Biol. Evol. 6: 3222–3237.

Nguyen, N. H., S.-O. Suh, C. J. Marshall, and M. Blackwell, 2006 Morphological and ecological similarities: wood-boring beetles associated with novel xylose-fermenting yeasts, *Spathaspora passalidarum* gen. sp. nov. and *Candida jeffriesii* sp. nov. Mycol. Res. 110: 1232–1241.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot, 2005 Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36: 541–562.

Polev, D. E., K. S. Bobrov, E. V. Eneyskaya, and A. A. Kulminskaya, 2014 Draft genome sequence of *Geotrichum candidum* strain 3C. Genome Announc. 2: e00956–14.

Polzin, K., and A. Rokas, 2014 Evaluating rare amino acid substitutions (RGC_CAMs) in a yeast model clade. PLoS One 9: e92213.

Priyam, A., B. J. Woodcroft, V. Rai, A. Munagala, I. Moghul *et al.*, 2015 Sequenceserver: a modern graphical user interface for custom BLAST databases. bioRxiv http://biorxiv.org/lookup/doi/10.1101/033142.

Riley, R., S. Haridas, K. H. Wolfe, M. R. Lopes, C. T. Hittinger *et al.*, 2016 Comparative genomics of biotechnologically important yeasts. Proc. Natl. Acad. Sci. USA 113: 9882–9887.

Rokas, A., and S. B. Carroll, 2006 Bushes in the tree of life. PLoS Biol. 4: e352.

Rokas, A., and P. W. H. Holland, 2000 Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. 15: 454–459.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll, 2003 Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798–804.

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

Salichos, L., A. Stamatakis, and A. Rokas, 2014 Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31: 1261–1271.

Scannell, D. R., K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe, 2006 Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440: 341–345.

Scannell, D. R., O. A. Zill, A. Rokas, C. Payen, M. J. Dunham *et al.*, 2011 The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces sensu stricto* genus. G3 (Bethesda) 1: 11–25.

Schwarz, G., 1978 Estimating the dimension of a model. Ann. Stat. 6: 461–464.

Seo, T.-K., 2008 Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25: 960–971.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang *et al.*, 2003 Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13: 2498–2504.

Shen, X.-X., L. Salichos, and A. Rokas, 2016 A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol. Evol. 8: 2565–2580.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Slot, J. C., and A. Rokas, 2010 Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. Proc. Natl. Acad. Sci. USA 107: 10136–10141.

Song, S., L. Liu, S. V. Edwards, and S. Wu, 2012 Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. USA 109: 14942–14947.

Springer, M. S., and J. Gatesy, 2016 The gene tree delusion. Mol. Phylogenet. Evol. 94: 1–33.

Stamatakis, A., 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

Stamatakis, A., P. Hoover, and J. Rougemont, 2008 A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57: 758–771.

Stanke, M., and S. Waack, 2003 Gene prediction with a hidden markov model and a new intron submodel. Bioinformatics 19(Suppl 2): ii215–ii225.

Sugiyama, J., K. Hosaka, and S.-O. Suh, 2006 Early diverging Ascomycota: phylogenetic divergence and related evolutionary enigmas. Mycologia 98: 996–1005.

Tavaré, S., 1986 Some probabilistic and statistical problems in the analysis of DNA sequences, pp. 57–86 in *Lectures on Mathematics in the Life Sciences*, edited by Miura, R. M.. American Mathematical Society, Providence, RI.

Taylor, J. W., and M. L. Berbee, 2006 Dating divergences in the fungal tree of life: review and new analyses. Mycologia 98: 838–849.

Wang, Y., X. Zhou, D. Yang, and A. Rokas, 2015 A genome-scale investigation of incongruence in culicidae mosquitoes. Genome Biol. Evol. 7: 3463–3471.

Waterhouse, R. M., F. Tegenfeldt, J. Li, E. M. Zdobnov, and E. V. Kriventseva, 2013 OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 41: D358–D365.

Wenger, J. W., K. Schwartz, and G. Sherlock, 2010 Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from *Saccharomyces cerevisiae*. PLoS Genet. 6: e1000942.

Whelan, N., K. M. Kocot, L. L. Moroz, and K. M. Halanych, 2015 Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. USA 112: 5773–5778.

Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter *et al.*, 2014 Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. USA 111: E4859–E4868.

Wolfe, K. H., and D. C. Shields, 1997 Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.

Wolfe, K. H., D. Armisén, E. Proux-Wera, S. S. ÓhÉigeartaigh, H. Azam *et al.*, 2015 Clade- and species-specific features of genome evolution in the Saccharomycetaceae. FEMS Yeast Res. 15: fov035.

Xi, Z., L. Liu, J. S. Rest, and C. C. Davis, 2014 Coalescent *vs.* concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. 63: 919–932.

Xi, Z., L. Liu, and C. C. Davis, 2016 The impact of missing data on species tree estimation. Mol. Biol. Evol. 33: 838–860.

Yandell, M., and D. Ence, 2012 A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13: 329–342.

Yang, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39: 306–314.

Yang, Z., 1996 Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11: 367–372.

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

Communicating editor: J. C. Fay