Construct species phylogenic tree for *Saccharomyces sensu stricto* using BUSCO genes

Jing Li

1. Introduction

Phylogenic relationship is very important when we conduct research on gene origin or other evolutionary questions. Currently, we have good assembly genome for eight Saccharomyces species: *S. cerevisiae*, *S. paradoxus, S. mikatae, S. kudriavzeii, S. arboricolus, S. uravum, S.bayanus*, and *S. eubayanus*. However, their phylogenic relationship is not clear. Different papers present different phylogenic tree for *Saccharomyces sensu stricto* species. In addition, many species within the *Saccharomyces sensu stricto* clade have been found to hybridize with other species, and usually has chromosomal loss, replacement, or rearrangement within the hybrid genetic lines. *S.bayanus* is an example from *S. cerevisiae, S. uravum, and S. eubayanus* (Pérez-Través et al., 2014). But we don't know this genome is more similar to which parent.

BUSCO, also known as assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs (Robert et al., 2017). It provides all the ancient or common gene orthologs in *Saccharomycetales*. We could extract the ancient genes from each species genome using BUSCO.

Our objective is to construct a species phylogenic tree from all the BUSCO gene trees. And then compare it to the currently published phylogenetic trees, to verify how the species phylogenic tree from BUSCO genes close to other research.

2. Method

1) Data collection

Collect all the genome sequence of these eight *Saccharomyces sensu stricto* species from NCBI database. Collect all the BUSCO (Robert et al., 2017) orthologs dataset for *Saccharomycetales* from BUSCO web.

2) Predict conserved genes

Predict fungi conserved genes in eight *Saccharomyces* species through BUSCO software. BUSCO map the *Saccharomycetales* conserved gene dataset to the genome and then predict the conserved gene and protein sequence for each species.

3) Gene alignment

Extract all the conserved genes for each species, and then do multiple sequence alignment with protein sequence among these eight species by guidance (Sela et al., 2015) using MAFFT algorithm.

4) Gene tree construction

Construct gene trees for each antient gene with maximum likelihood approach using RAxML program (Stamatakis et al., 2014). Using JTT model for amino acid sequence with 100 replicates bootstrap analysis to search for the best-scoring ML tree.

5) Species tree construction

Combine all the gene trees and then generate a species tree using the ASTRAL (Mirarab et al., 2014) coalescent-based species tree estimation program.

3. Results

1) BUSCO result

The total amount of *Saccharomycetales* ancient genes is 1,711 in *Saccharomyces sensu stricto*. However, not every ancient gene in all eight species since not completed genome assembly, or lost in the evolution history. Hence, only 1,223 genes exist in all eight species. Figure 1 shows the proportion of the existing or missing ancient genes in each genome.

2) ML gene trees

We constructed gene trees using the amino acid sequences, since the sequence is more flexible than nucleotide sequence. The maximum likelihood gene trees were analyzed by RAxML using JTT model with 100 bootstrap replicates. In the 1,771 BUSCO genes, 1,662 genes exiting in at least two species. Within these 1,662 best score ML trees, the maximum likelihood is range from -9642.5 to -324.0. The distribution is left skewed, which means most gene trees have high maximum likelihood. Figure 2 shows the distribution of the maximum likelihood in each best core gene trees. Exploring the 1,223 gene trees with all species, we found 38 different types topological trees. Some topological gene trees only include 1 gene, the most commontopological gene tree include 470 genes. All the gene trees are shown in the supplemental material.

3) ASTRAL species tree

ASTRAL is statically consistent under the multi-species coalescent model, and is useful for handling incomplete lineage sorting. We combine the 1,662 gene trees together, and the obtain a species tree from ASTRAL, which is shown in figure 3.

4. Discussion

From RAxML maximum likelihood analysis, we get 38 different topological trees. Comparing our gene trees to the final species tree, no single gene tree has the same topological shape as the species tree. Gene trees and species trees can be incongruent for many reasons: 1) genes can have unequal rates of evolution; 2) gene loss and gene duplication are common; 3) gene flow can occur between lineages after their separation; 4) recombination between neighboring regions can also lead to species phylogenies and gene histories that do not match. The most common gene tree is only account for about 30% of all genes. However, these gene trees and species tree still have some common part. For example, in species tree and most gene trees, the relation between *S. cerevisiae* and *S. paradoxus* are closed than other, which also find between *S. uravum, S. bayanus*. We can also speculate the gene emergence, change or lost according each gene trees.

In these eight species, only *S. bayanus* is a hybrid, which is also known as *Saccharomyces bayanus var. bayanus*. The species tree we got from this project shows that it is more closed to *S. uravum* (*Saccharomyces bayanus var. uravum*). *S. bayanus* should inherit more genes from *S. uravum* than other two species. Moreover, *S. bayanus* may only got a small part of gene from *S. cerevisiae*, since their distance is larger than other two species.

There is no paper shows the phylogenic relation among all eight species, but some papers showed the phylogenic tree for seven species except *S. bayanus* (Borneman et al., 2015). The species tree is a little different with our result, which is shown in figure 4. In this paper, *S. eubayanus* and *S. uravum* are in the same clade which is in the same hierarchical level as another clade with other 5 species. However, our result shows that *S. eubayanus* is the most ancient species among these 7 species, and then is *S. uvarum* (except *S. bayanus*). *S. eubayanus* and *S. uravum* are not in the same clade. But the relation among other species are the same. One possible reason is that we use protein sequence for this study, not nucleotide sequence. Protein sequence is more flexible in the alignment because of the degeneracy of codons.

In conclusion, we can better understand the phylogenic relation in *Saccharomyces sensu stricto* through both the gene trees and species tree. Furthermore, we can explore the gene evolutional way for some specific genes, such as orphan genes.

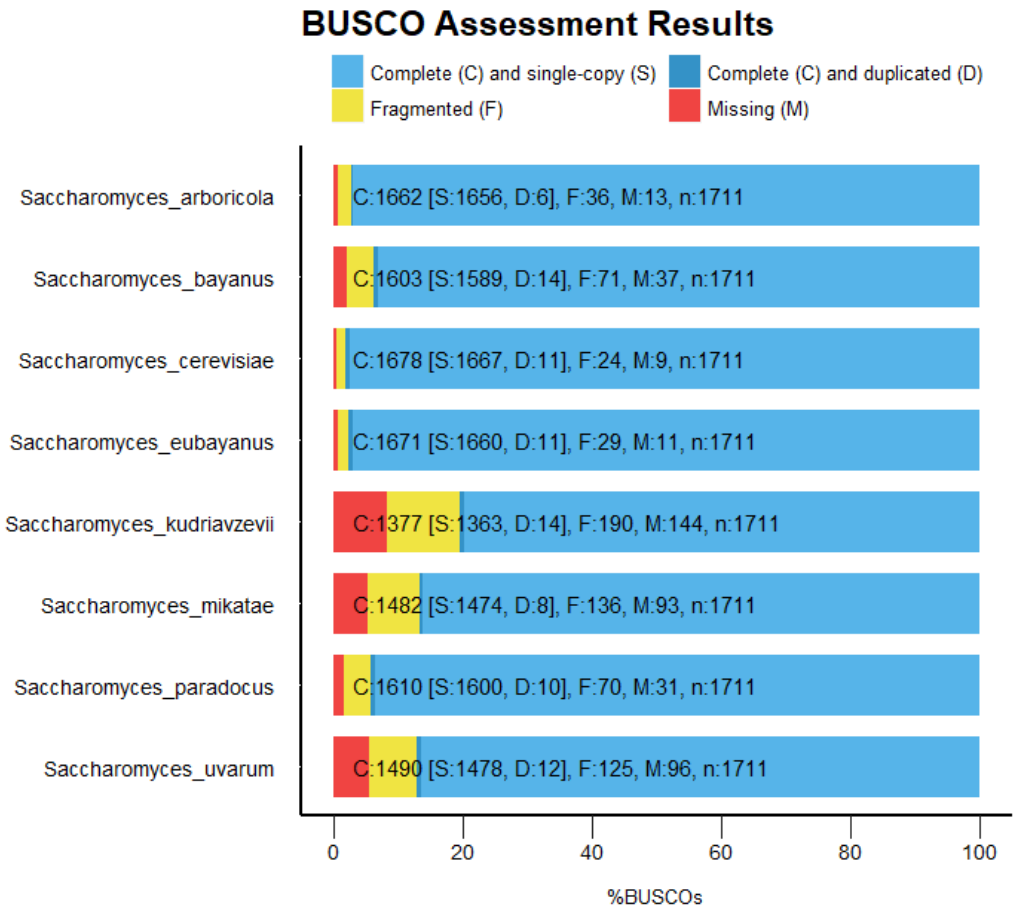


Figure 1. BUSCO analysis for eight genomes. The light blue bar means the proportion of genes with complete and single-copy sequence match to the antient *Saccharomycetales* genes. Dark blue bar means the proportion of genes with complete and duplicate-copy sequence. Yellow bar means the proportion of genes with fragmented sequence. Red bar means the proportion of antient genes missing in the species.

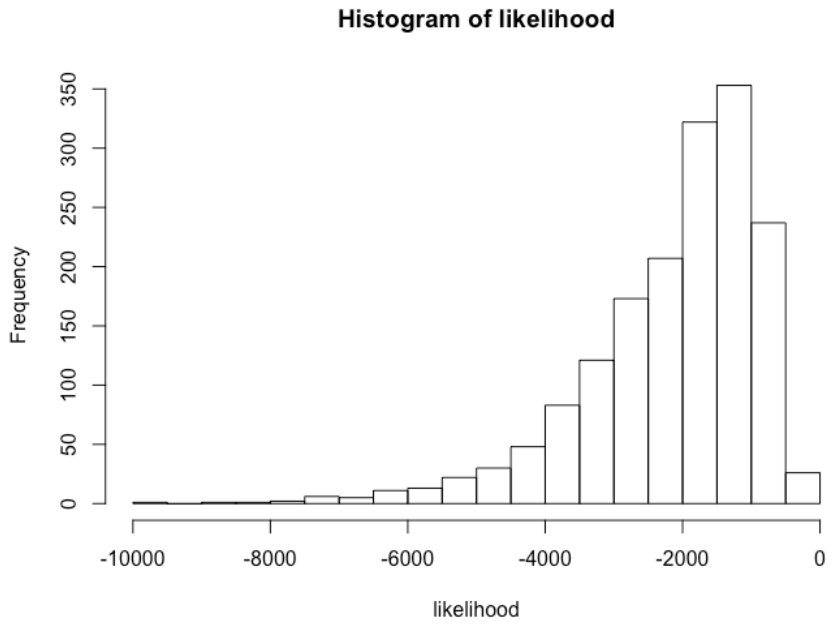


Figure 2. Histogram of likelihood for 1,662 best scoring ML gene trees. X-axis is the likelihood of gene trees, y-axis is the frequency of the likelihood for every 500 breaks. The distribution is left-skewed.

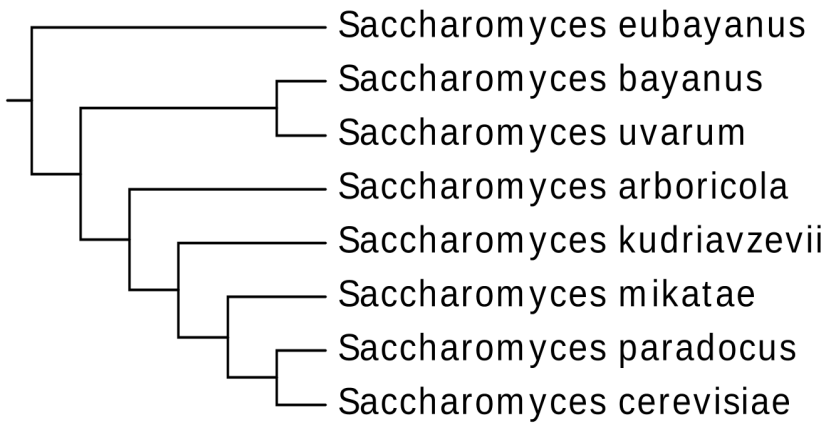


Figure 3. ASTRAL species tree for 8 species in *Saccharomyces sensu stricto*. The tree obtained by combined 1,662 best scoring ML gene trees, and ignore the branch length.

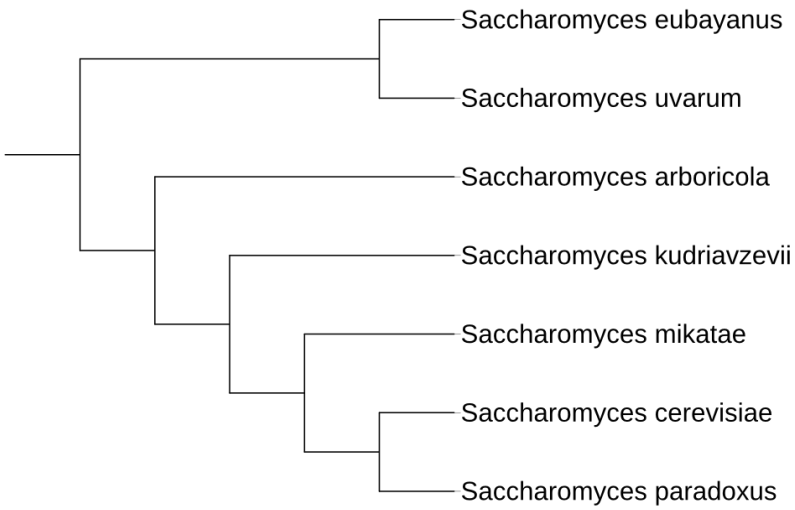


Figure 4. Species tree for 7 species in *Saccharomyces sensu stricto* from Borneman et al., 2015.

5. Reference

Borneman, Anthony R., and Isak S. Pretorius. "Genomic insights into the Saccharomyces sensu stricto complex." Genetics 199.2 (2015): 281-291.

Mirarab, Siavash, Rezwana Reaz, Md S. Bayzid, Théo Zimmermann, M. Shel Swenson, and Tandy Warnow. "ASTRAL: genome-scale coalescent-based species tree estimation." Bioinformatics 30, no. 17 (2014): i541-i548.

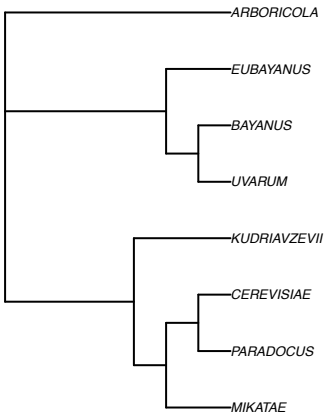
Pérez-Través, Laura, et al. "On the complexity of the Saccharomyces bayanus taxon: hybridization and potential hybrid speciation." PLoS One 9.4 (2014): e93729.

Sela, Itamar, Haim Ashkenazy, Kazutaka Katoh, and Tal Pupko. "GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters." Nucleic acids research 43, no. W1 (2015): W7-W14.

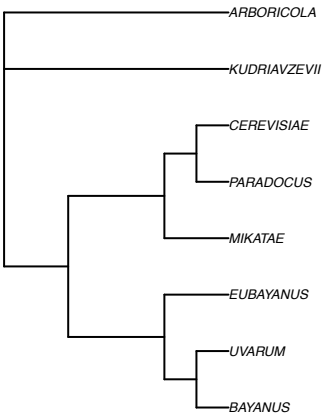
Stamatakis, Alexandros. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." Bioinformatics 30, no. 9 (2014): 1312-1313.

Waterhouse, Robert M., et al. "BUSCO applications from quality assessments to gene prediction and phylogenomics." Molecular biology and evolution 35.3 (2017): 543-548.

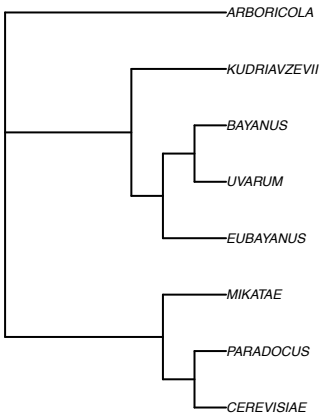
Topological tree of Group1; Group size: 470



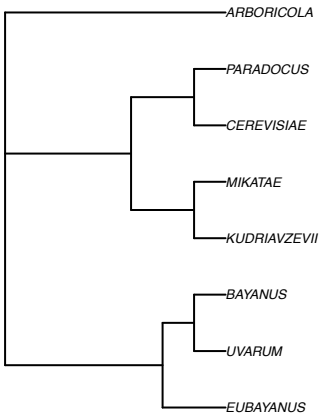
Topological tree of Group2; Group size: 141



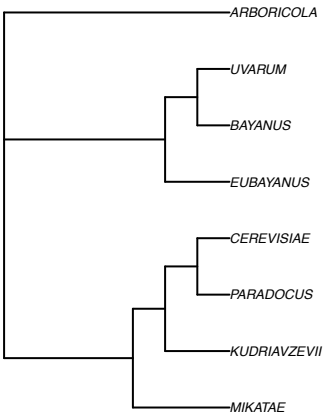
Topological tree of Group3; Group size: 127



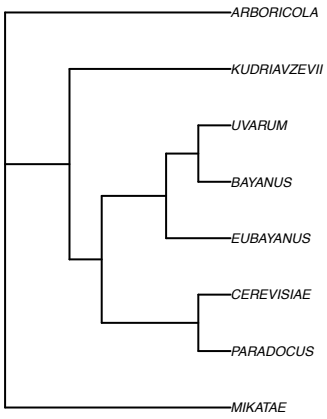
Topological tree of Group4; Group size: 31



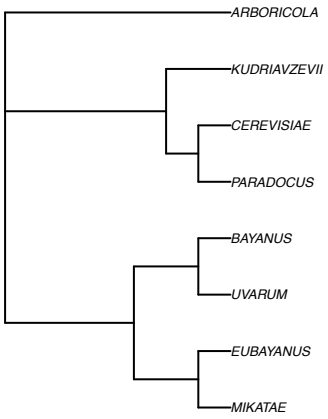
Topological tree of Group5; Group size: 27



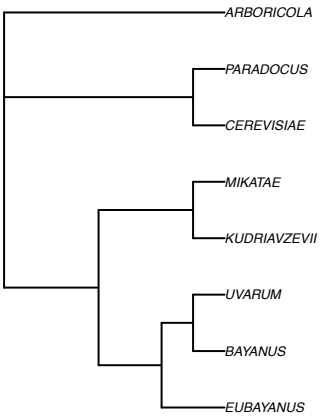
Topological tree of Group6; Group size: 23



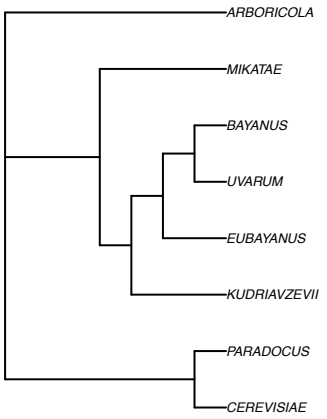
Topological tree of Group7; Group size: 32



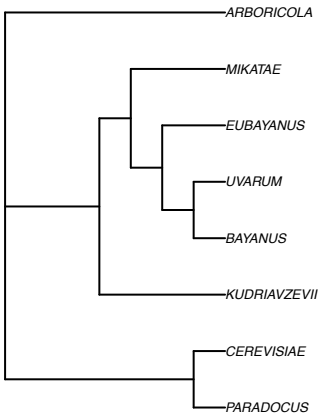
Topological tree of Group8; Group size: 7



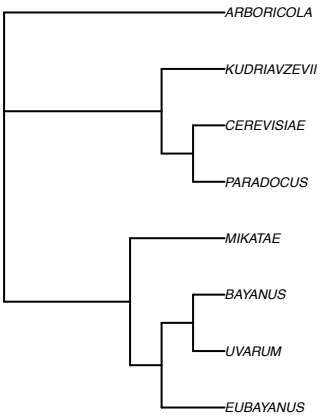
Topological tree of Group9; Group size: 9



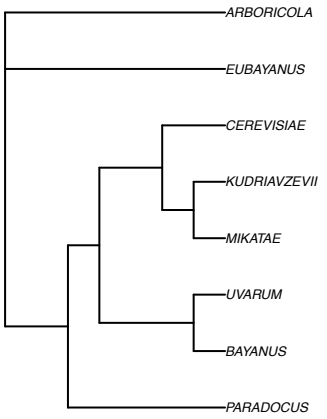
Topological tree of Group10; Group size: 8



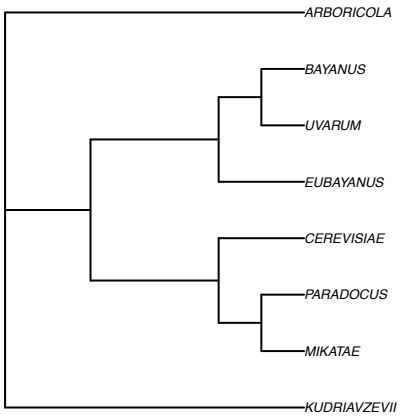
Topological tree of Group11; Group size: 8



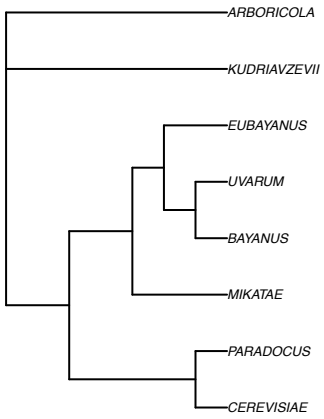
Topological tree of Group12; Group size: 4



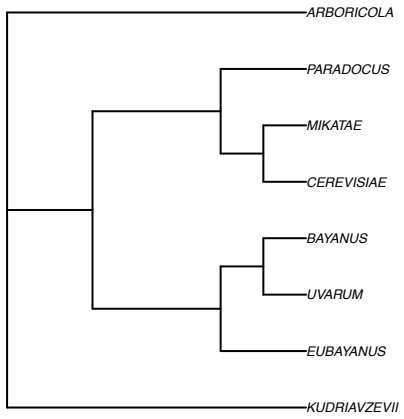
Topological tree of Group13; Group size: 9



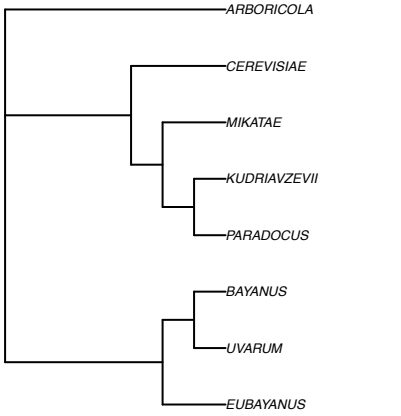
Topological tree of Group14; Group size: 17



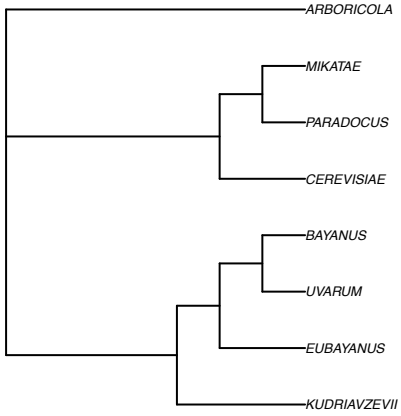
Topological tree of Group15; Group size: 9



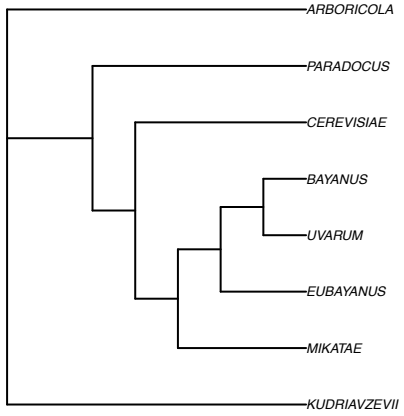
Topological tree of Group16; Group size: 5



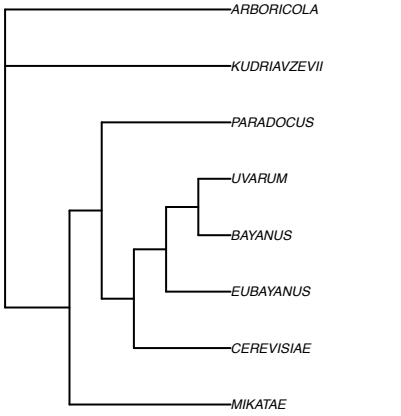
Topological tree of Group17; Group size: 8



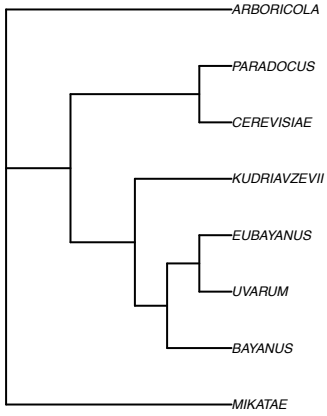
Topological tree of Group18; Group size: 2



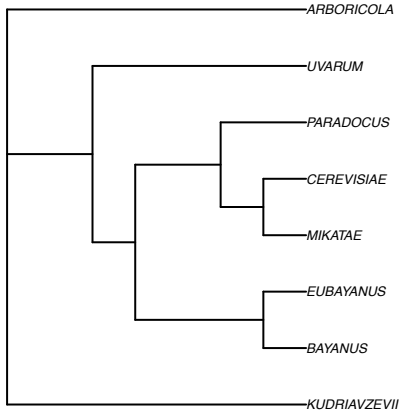
Topological tree of Group19; Group size: 16



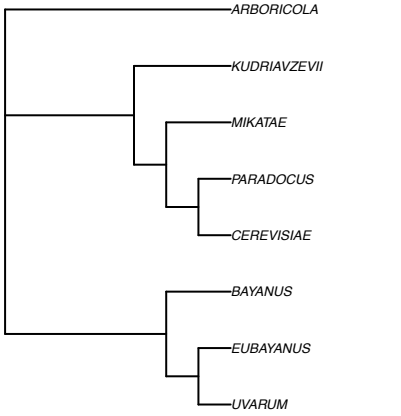
Topological tree of Group20; Group size: 9



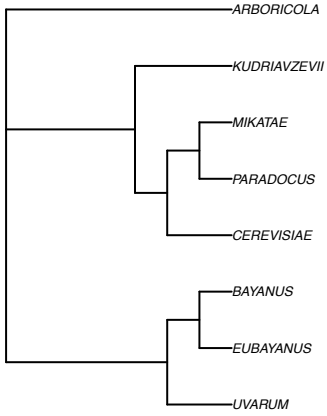
Topological tree of Group21; Group size: 1



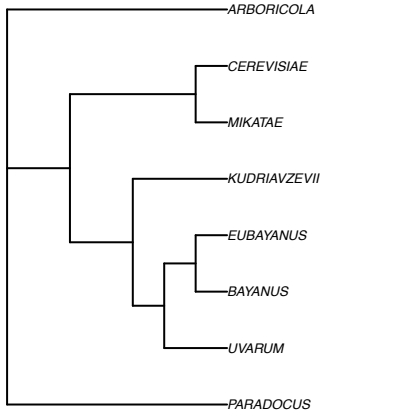
Topological tree of Group22; Group size: 3



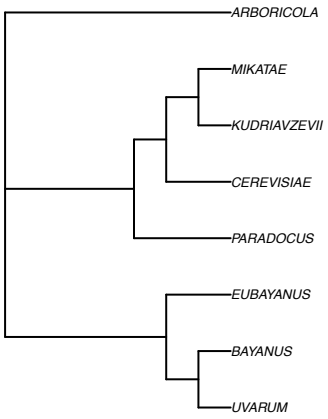
Topological tree of Group23; Group size: 4



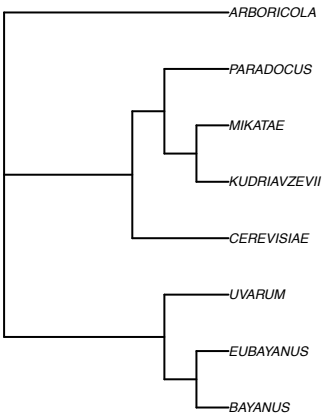
Topological tree of Group24; Group size: 4



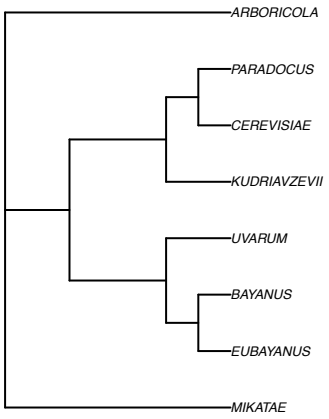
Topological tree of Group25; Group size: 10



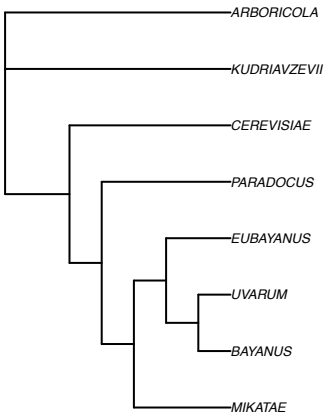
Topological tree of Group26; Group size: 12



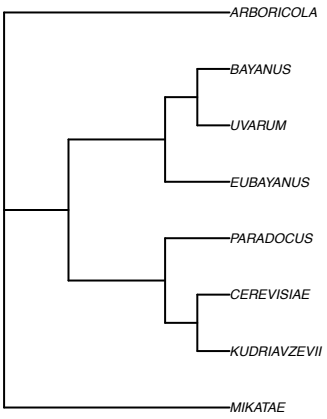
Topological tree of Group27; Group size: 1



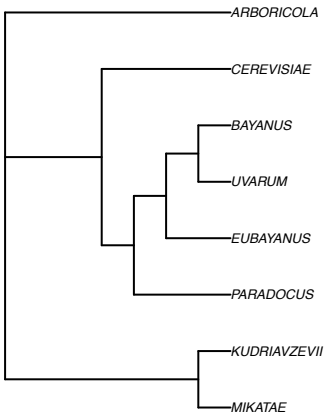
Topological tree of Group28; Group size: 2



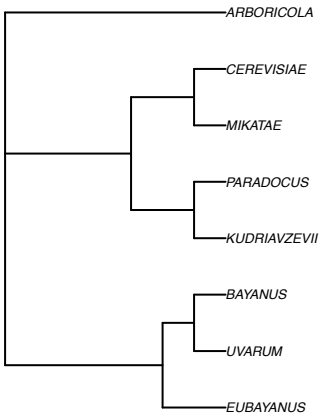
Topological tree of Group29; Group size: 5



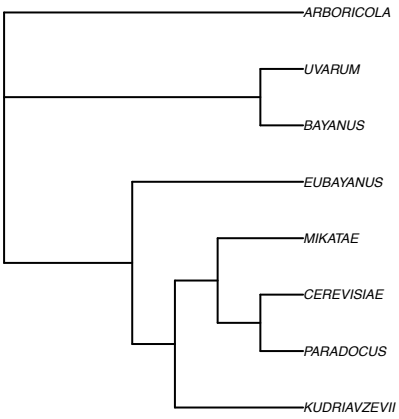
Topological tree of Group30; Group size: 2



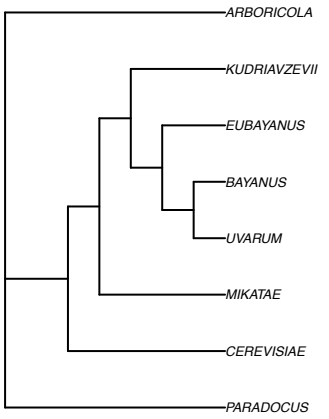
Topological tree of Group31; Group size: 2



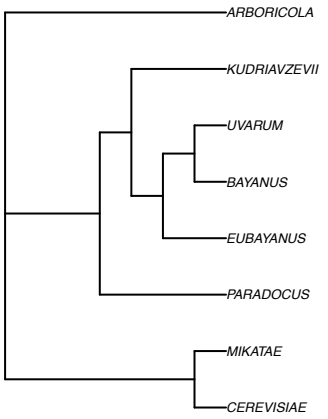
Topological tree of Group32; Group size: 3



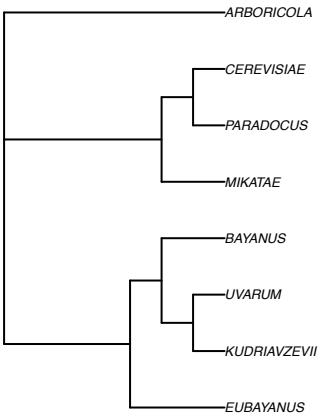
Topological tree of Group33; Group size: 2



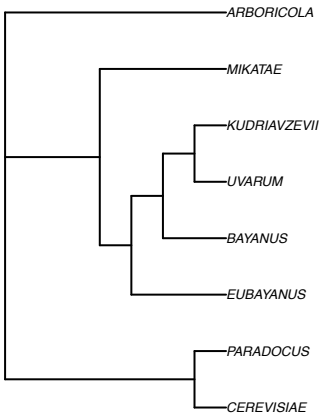
Topological tree of Group34; Group size: 5



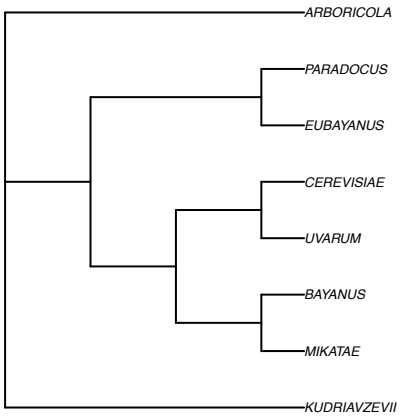
Topological tree of Group35; Group size: 1



Topological tree of Group36; Group size: 2



Topological tree of Group37; Group size: 1



Topological tree of Group38; Group size: 1