Phylogenetic Study of Tad3 Evolution

Yizhu Lin

Introduction

tRNA-specific adenosine-34 deaminases (ADAT) are an essential enzymes found in Bacteria and Eukarya that catalyze the conversion of adenine (A) to inosine (I) at tRNA' s wobble position 34. I_{34} is able to wobble with adenine, cytosine, and uridine, enlarging codon recognition capacity during protein synthesis. In Eukarya, a heterodimeric form of ADAT is formed by Tad2p and Tad3p. The two subunits are sequence-related, both of which contain cytidine deaminase (CDA) motifs.[1] Interestingly, the two are separately encoded, yet function and sequence related. Previous studies presented many cases in which the genes encoding such kind of heterodimer are coevolving, exerting positive or directional selection to each other, like integratin α and β chain genes in vertebrate.[2] Therefore, here in this project, we expect the functional intimidation of Tad3p and Tad2p indicates a co-evolutionary relationship of their encoding genes. Notably, TAD3 gene is one of the 10 genes that have multiple introns in *S. cerevisiae*[3]. It contains a non-canonical branchpoint sequence 5'-AACTAAC-3' instead of 5'-TACTAAC-3', indicating an alternative-splicing-related regulation of TAD transcription. Inspired by the unique 2 intron feature, the evolution of TAD3 introns will be studied in this project. The TAD3 intron phylogenetic tree is expected to correlating with genome duplication and may point the origin of intron.

Method

1. Sequence acquirement

Saccharomyces cerevisiae Tad3p and Tad2p's amino acid sequence were obtained from *Saccharomyces* Genome Database (SGD)[4]. By Blastp *Saccharomyces cerevisiae's* Tad3p and Tad2p's amino acid sequence with fungi taxa in NCBI, we were able to obtain Tad3p and Tad2p's amino acid sequences of other fungi species. To enlarge taxa coverage, we obtained some sequences by searching NCBI with keyword "TAD3" or "TAD2" under "protein" catalog. Since we aim to compare Tad3p's phylogenetic tree with Tad2p's, only fungi species that have available sequences of both Tad3p and Tad2p were included in our study. However, we didn't find Tad3p or Tad2p's sequences of *Saccharomyces mikatae, Saccharomyces kudriavzevii, Saccharomyces paradoxus and Saccharomyces bayanus* by the two methods above. But TAD3 and TAD2 DNA sequences of these four species can be found by blastn with *Saccharomyces cerevisiae's* DNA sequences in SGD, and their sequences show high similarity with *Saccharomyces cerevisiae's* DNA sequences. So we aligned DNA sequences of these 5 species, cut off TAD3's intron sequences according to *Saccharomyces cerevisiae's* verified intron region, then translate TAD3 and TAD2 gene into protein sequences (using ApE). These DNA sequence are also used in TAD3's intron analysis.

In total, we selected 36 protein sequences of both Tad3p and Tad2p for further analysis, including 33 fungi sequences and 3 animal species as outgroup. Five *Saccharomyces* species were used in intron sequences analysis. (See Appendence for sequences used in this project)

2. Multiple protein sequences alignment

Multiple protein sequences alignment of Tad3p and Tad2p were conducted using Jalview. We aligned the sequences using each of the four major algorithms (Mafft, Muscle, Tcoffee and

Probcons) with default parameters. After alignment, columns with more than 50% gaps were trimmed for further study.

From alignment we can see that Tad3p has one highly conserved region (*S. cerevisiae* Y240~L272) and several other shorter or less conserved region. The C-terminal of Tad3p is much less conserved. All the four algorithms successfully aligned the highly conserved region consistently, except Mafft introduced 2 big gaps in *Cryptococcus neoformans*. Tcoffee seems to introduce many small gaps at the C-terminal. Mafft, Muscle and Probcons also show highly consistency at other conserved sites such as *S. cerevisiae* W140, P141, C88.

Tad2p seems to be more conserve than Tad3p. The N-terminal of Tad2p is relatively more conserved (*S. cerevisiae* M7~P167), and the four algorithms aligned this conserved region in a similar way. The C-terminal of Tad2p is much less conserved, and several species lost these C-terminal sequences completely. The alignments of Tad2p's C-terminal region from four algorithms are very different from each other. However, following tree building results show that trimming of the C-terminal unconserved region will not change the topology of phylogenetic tree. Such, we consider alignment of Tad2p from these four algorithms are equally reliable.

3. Model selection and tree building

Based on multiple alignment results, all the 4 alignments of Tad3p and Mafft alignment of Tad2p (with and without C-terminal unconserved region) were used for model selection and tree building.

Model selections were conducted by TOPALi to obtain parameters for Maximum Likelihood tree building. Models that had minimal AIC or BIC were used for PhyML tree building in Seaview. We also constructed Neighbor Join tree using Seaview.

Results of TOPALi model selections show that in all the 6 alignments, WAG+I+G model is always the one has minimal AIC and BIC. The parameters for PhyML tree building are summarized in the following table.

Protein	Alignment	pINV	alpha
Tad2p	Mafft(with C-terminal)	0.081	1.304
	Mafft(without C-terminal)	0.099	1.129
Tad3p	Mafft	0.041	1.827
	Muscle	0.041	1.767
	Tcoffee	0.042	1.71
	Probcons	0.047	1.782

4. Species tree

Species tree is from NCBI Taxomony – common tree.

5. TAD3 intron analysis

TAD3 gene sequences of five *Saccharomyces* species were aligned using Clustal in Jalview. The five sequences are highly similar to each other. Based on verified intron sequence of S. *cerevisiae,*we were able to find introns of other 4 sequences. Since the 5'splice sites, 3'splice sites and branchpoint sequences are all well aligned, we think the introns obtained by this method is reliable. Each of the two introns was cut off for tree building in Seaview.

Results

1. Comparison of trees derived from Maximum Likelihood Method and Neighbor Joining Method.

Typically, PhyML tree and NJ tree generated from same alignment have same topology. However, NJ tree fails to distinguish different branch length, while PhyML tree does better in branch length determination. For example, in the NJ tree generated from Tad3p's Mafft alignment, the *Cryptococcus neoformans var. neoformans* branch is very close to animal group, while PhyML is able to separate *Cryptococcus neoformans var. neoformans* from the animal group.

Given this result, our further analysis will based on only PhyML tree but not NJ tree.

2. Comparison of Tad2p tree with and without C-terminal unconserved sequences

Compare the Tad2p Mafft - PhyML tree with C-terminal unconserved sequences and without C-terminal unconserved sequences, we can see that both the topology and branch length of the two trees are almost the same. Since the alignments generated by four different algorithms are very similar in N-terminal conserved region, we assume that different alignment will not affect tree building results of Tad2p. We will use Tad2p's PhyML tree based on Mafft alignment for further analysis.

3. Comparison of Tad3p trees based on different alignment methods.

From results we can see that topology of these four trees is similar but still has some disagreement. Probcons and Muscle put *Saccaromycetales* group and *leotiomyceta* group together, while *Schizosaccharomyces* derived from *Ascomycota* earlier. This is consistent with species tree obtained from NCBI Taxonomy database. However, Mafft and Tcoffee put *leotiomyceta* group and *Schizosaccharomyces* group together. All of the four alignments generated a tree with a branch that has a low bootstrap when trying to distinguish these three groups. Within *Saccharomyces* group, only Mafft put *S.mikatae* and *S. paradoxus* together, while others put *S. paradoxus* closer to *S. cerevisiae*. Compare Probcons with Muscle, Muscle shows bigger bootstrap values on more branches than Probcons. Based on analysis above, we conclude that Muscle alignment is a more suitable alignment for Tad3p.

4. Comparison of Tad3p tree, Tad2p tree and species tree.

The species tree from NCBI is not a binary tree and it cannot distinguish the speciation order of most species in *Saccharomycetaceae* group except *Saccharomyces*. Tad3p tree and Tad2p tree differ a lot from each other in these species. Besides, in Tad2p tree, *Yarrowia lipolytica* is far away from other *Saccharomycetales* species. This disagrees with both species tree and Tad3p tree. Both Tad3p tree and Tad2p tree put *Candida* group within *Debaryomycetaceae*. Interestingly, both *Candida* and *Debaryomycetaceae* belong to CTG clade, which translate CTG as serine instead of leucine[5, 6]. The topology of our trees is consistent with this tRNA-related trait.

However, we cannot see obvious evidence for Tad3p and Tad2p's co-evolution.

5. TAD3 intron analysis.

Since we only have 5 species for alignment and intron sequences is relatively short, the trees generated from intron are not so robust. Trees generated by PhyML and NJ method are almost the same. Trees based on both first intron and second intron show that *S. mikatae* has longer

distance to others. And *S. mikatae* have two substitutes in first intron's BPS (CACTAAC instead of AACTAAC) and 3'SS (TAG instead of CAG).

Other TAD3 gene, such as *Candida glabrata*, *Candida albicans etc.* that closely related to *Saccharomyces* don't have introns, indicating these two introns originated with the speciation of *Saccharomyces*.

Discussion

Through this study, we generated gene trees of both Tad3p and Tad2p. Generally, Tad3p tree and Tad2p tree are consistent with species tree, and are consistent with CTG clade. However, we cannot infer Tad3p and Tad2p' co-evolution from our study, since there are no significant unique topology similarity between Tad3p tree and Tad2p tree.

As for TAD3 intron analysis, only five species and relatively short sequence length seems to be not enough for tree analysis. But we can still see that introns are much less conserved than exons, and these TAD3 introns may originate independently in *Saccharomyces*.

References

1. Novoa, E.M., et al., *A role for tRNA modifications in genome structure and codon usage.* Cell, 2012. **149**(1): p. 202-13.
2. Hughes, A.L., *Coevolution of the vertebrate integrin alpha- and beta-chain genes.* Mol Biol Evol, 1992. **9**(2): p. 216-34.
3. Hossain, M.A., C.M. Rodriguez, and T.L. Johnson, *Key features of the two-intron Saccharomyces cerevisiae gene SUS1 contribute to its alternative splicing.* Nucleic Acids Res, 2011. **39**(19): p. 8612-27.
4. Cherry, J.M., et al., *Saccharomyces Genome Database: the genomics resource of budding yeast.* Nucleic Acids Res, 2012. **40**(Database issue): p. D700-5.
5. Mallet, S., et al., *Insights into the life cycle of yeasts from the CTG clade revealed by the analysis of the Millerozyma (Pichia) farinosa species complex.* PLoS One, 2012. **7**(5): p. e35842.
6. Wang, H., et al., *A fungal phylogeny based on 82 complete genomes using the composition vector method.* BMC Evol Biol, 2009. **9**: p. 195.

Appendence

1. List of species studied

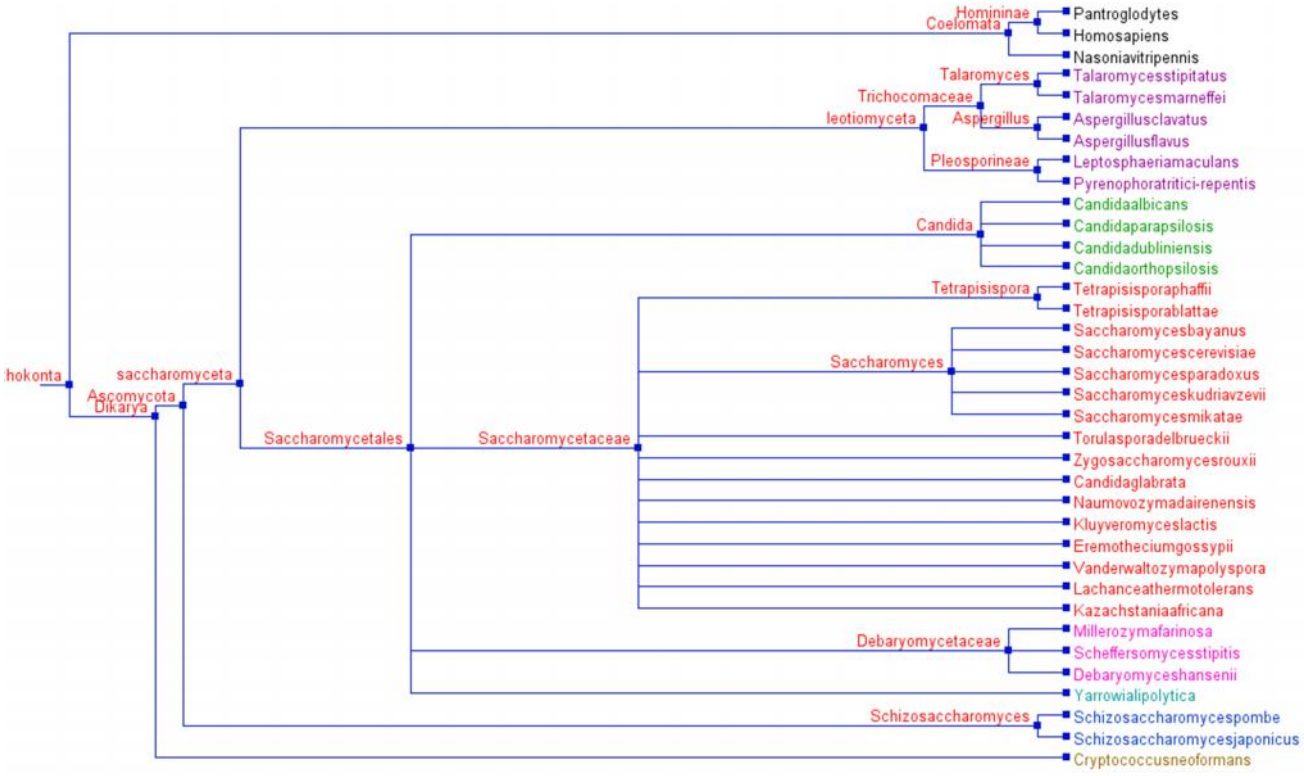
Ashbya_gossypii
Aspergillus_clavatus
Aspergillus_flavus
Candida_albicans
Candida_glabrata
Candida_dubliniensis
Candida_orthopsilosis
Candida_parapsilosis
Cryptococcus_neoformans_var._neoformans
Debaryomyces_hansenii
Homo_sapiens (outgroup)
Kazachstania_africana
Kluyveromyces_lactis
Lachancea_thermotolerans
Leptosphaeria_maculans
Millerozyma_farinosa
Nasonia_vitripennis (outgroup)
Naumovozyma_dairenensis
Pan_troglodytes (outgroup)
Penicillium_marneffei
Pyrenophora_tritici-repentis
Saccharomyces_cerevisiae
Saccharomyces_paradoxus
Saccharomyces_mikatae
Saccharomyces_kudriavzevii
Saccharomyces_bayanus

Scheffersomyces_stipitis
Schizosaccharomyces_pombe
Schizosaccharomyces_japonicus
Talaromyces_stipitatus
Tetrapisispora_blattae
Tetrapisispora_phaffii
Torulaspora_delbrueckii
Vanderwaltozyma_polyspora
Yarrowia_lipolytica
Zygosaccharomyces_rouxii

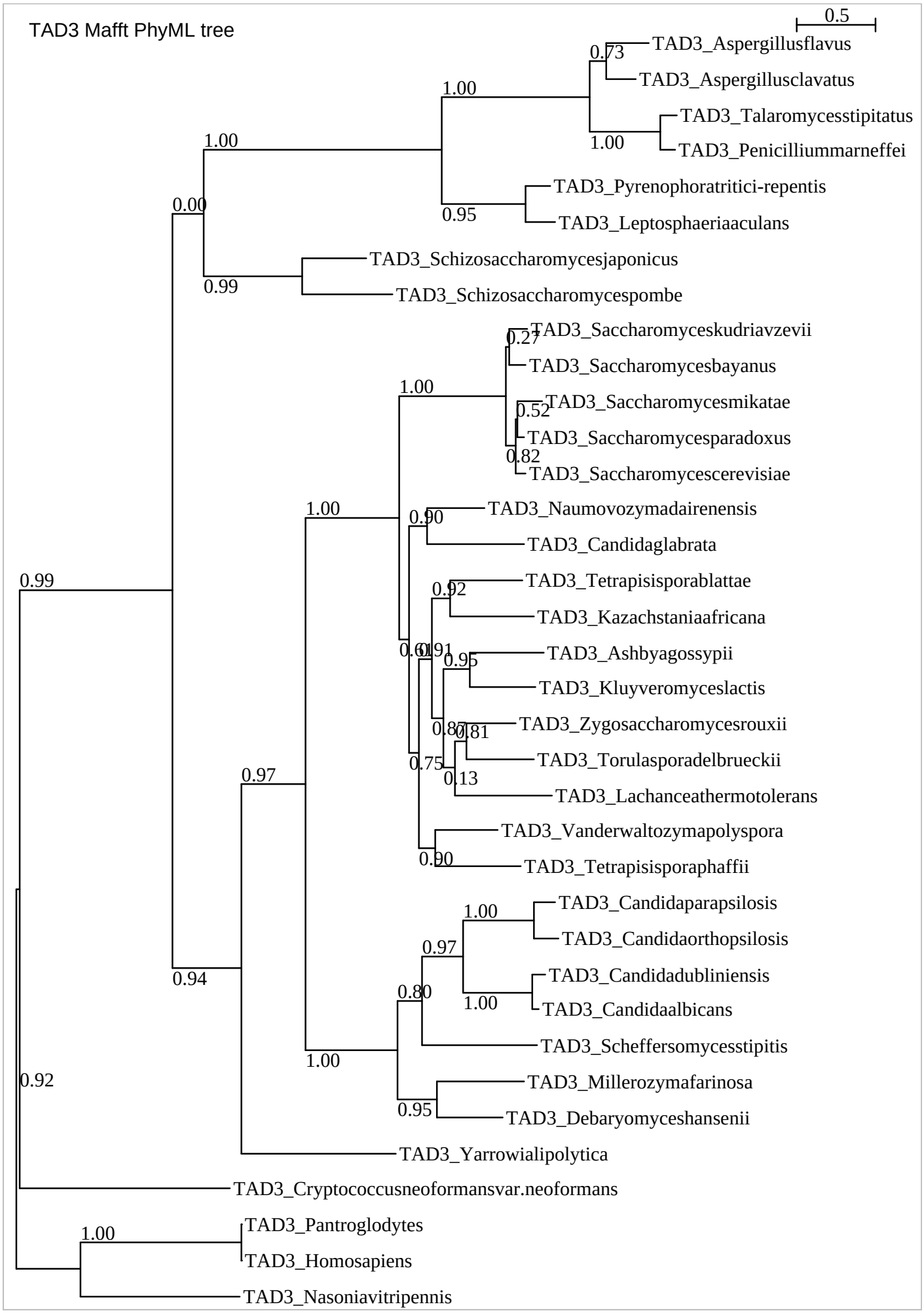
2. Sequences for alignment
   Tad3p sequences: see Tad3p_modified.fa
   Tad2p sequences: see Tad2p_modified.fa
   TAD3 gene sequences: see Tad3gene.fa

3. Multiple Sequences alignment
   See fasta files for trimmed and untrimmed alignments.

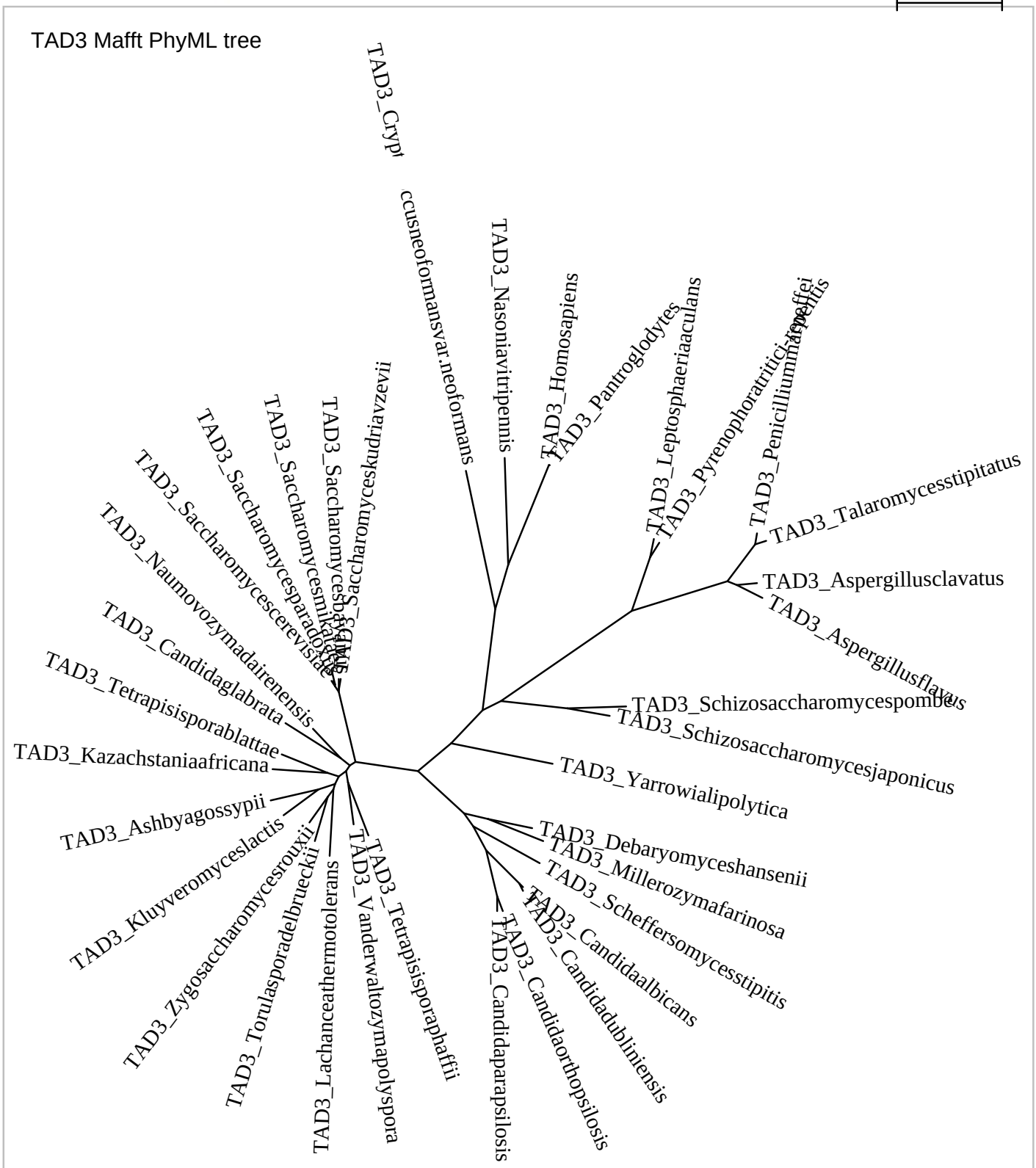
4. Species Tree from NCBI taxonomy



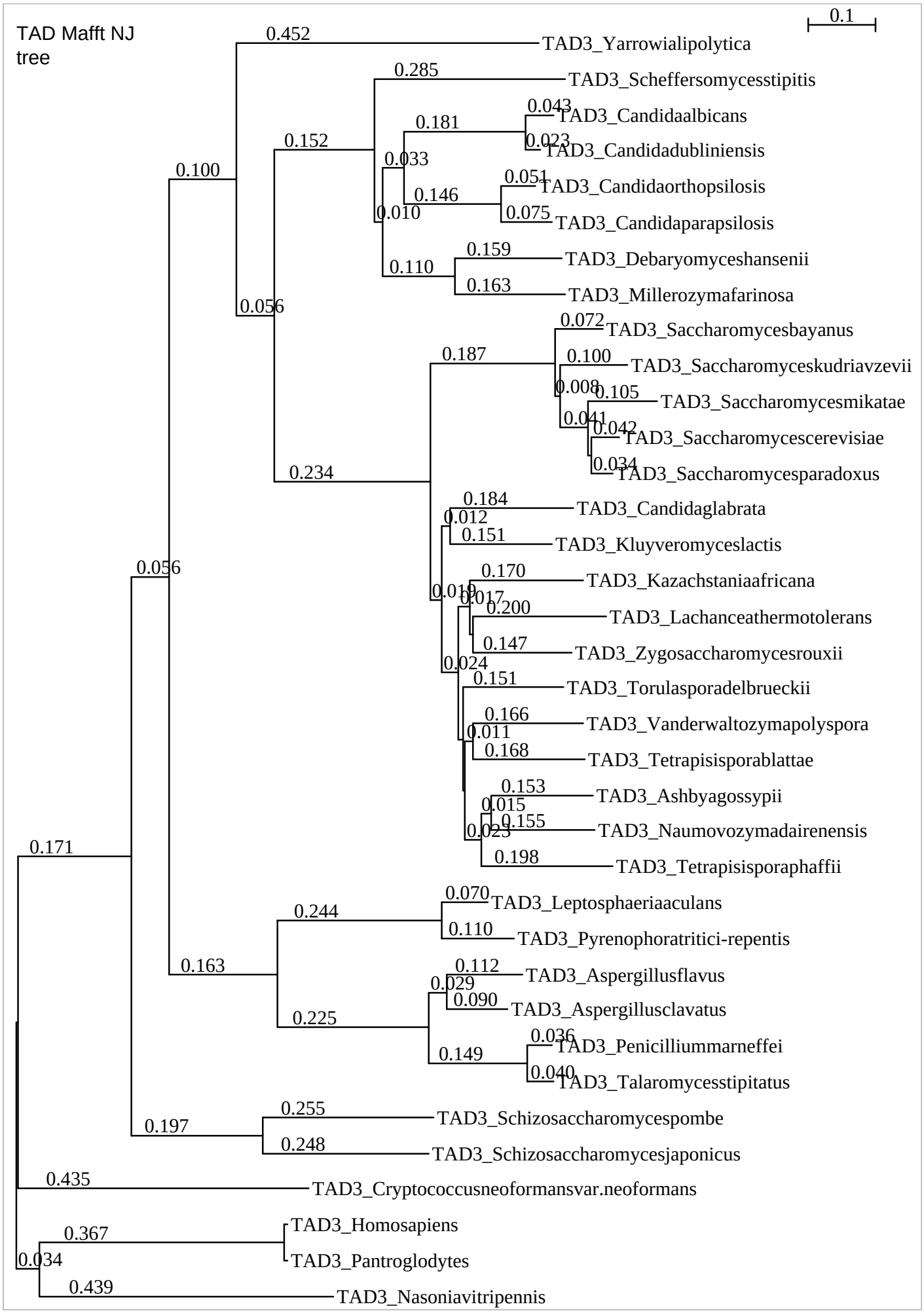
Seaview TAD3fortreeML.nwk Wed Dec 12 16:27:15 2012



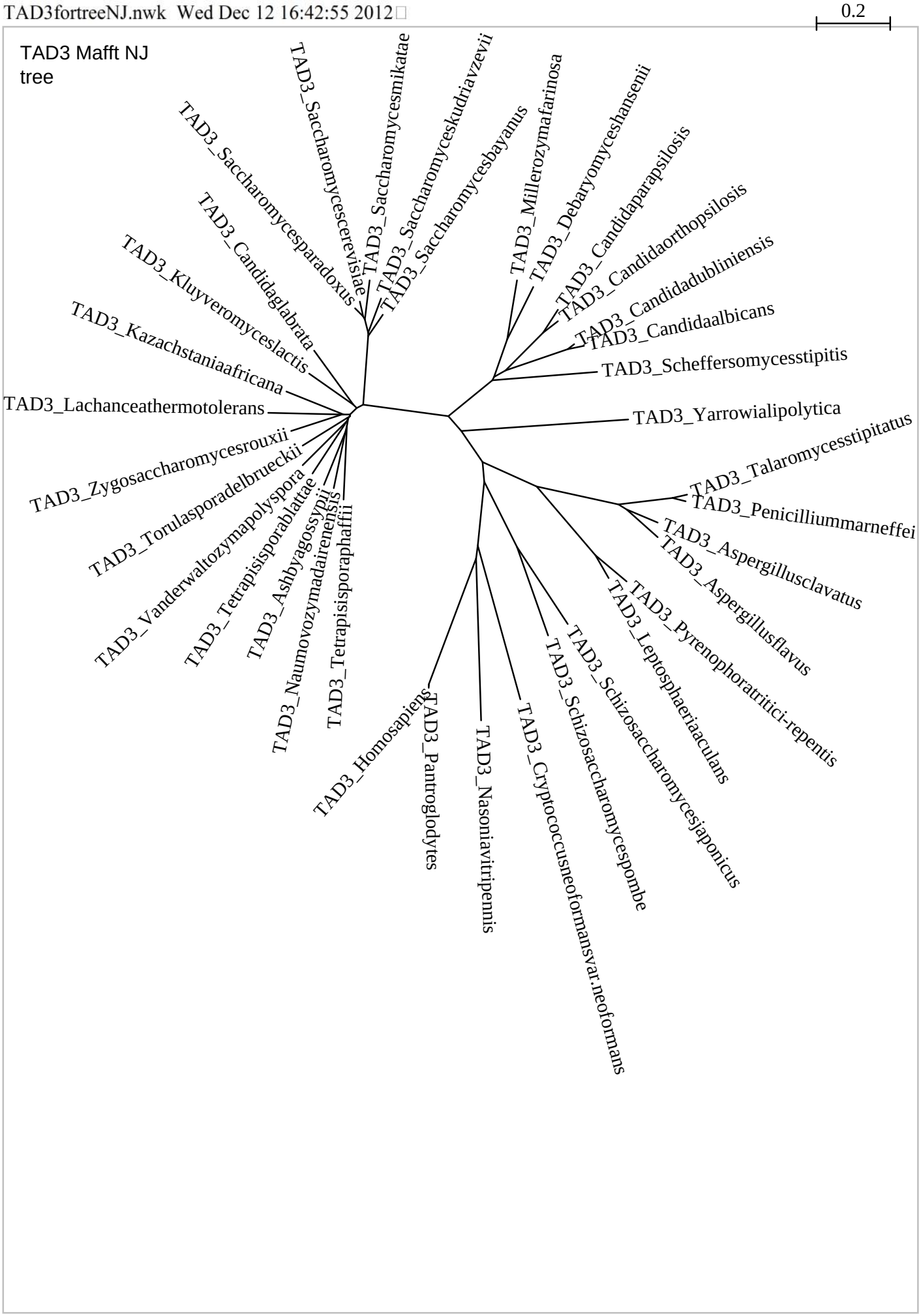
TAD3fortreeML.nwk Wed Dec 12 16:27:30 2012



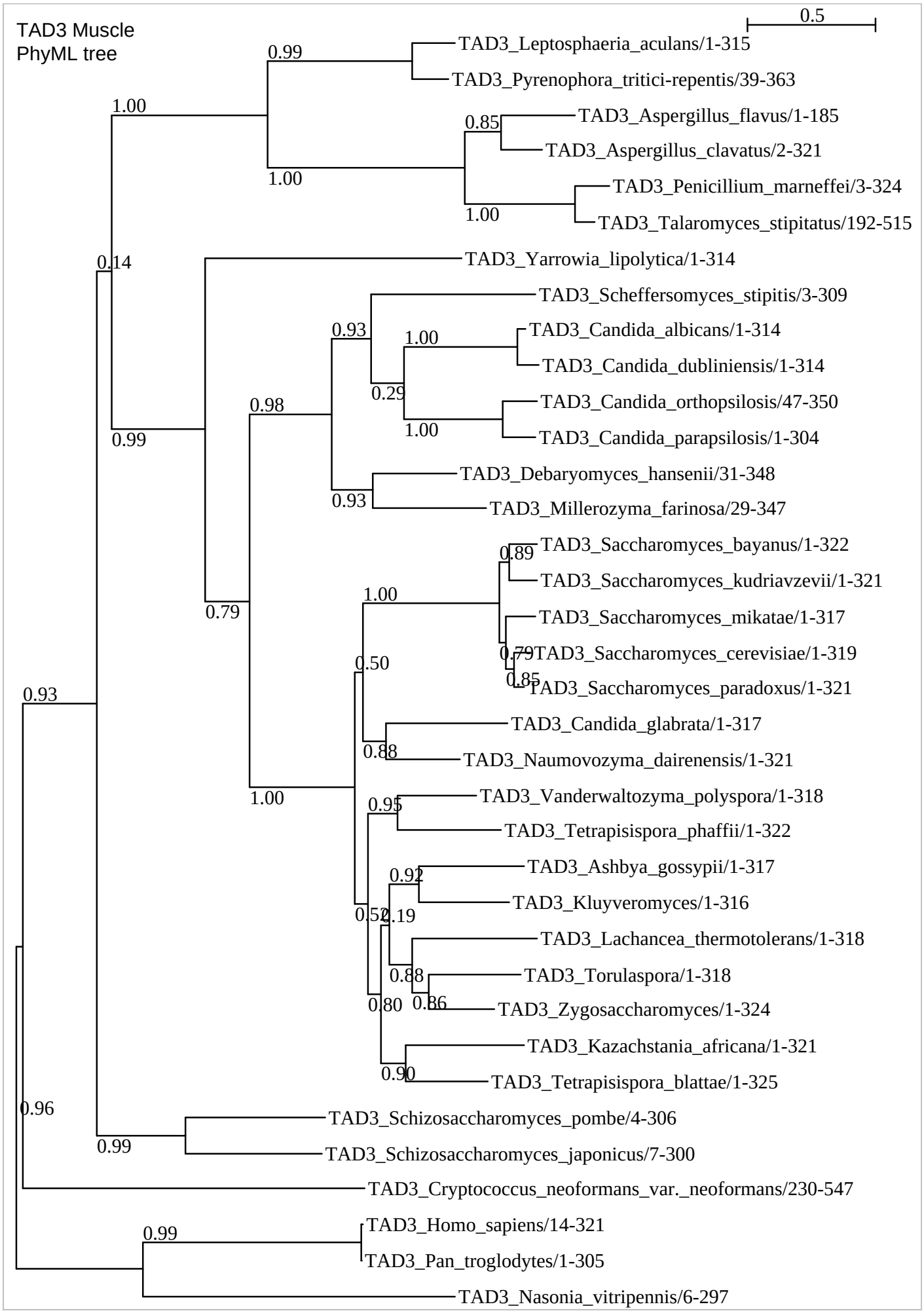
Seaview TAD3fortreeNJ.nwk Wed Dec 12 16:42:46 2012



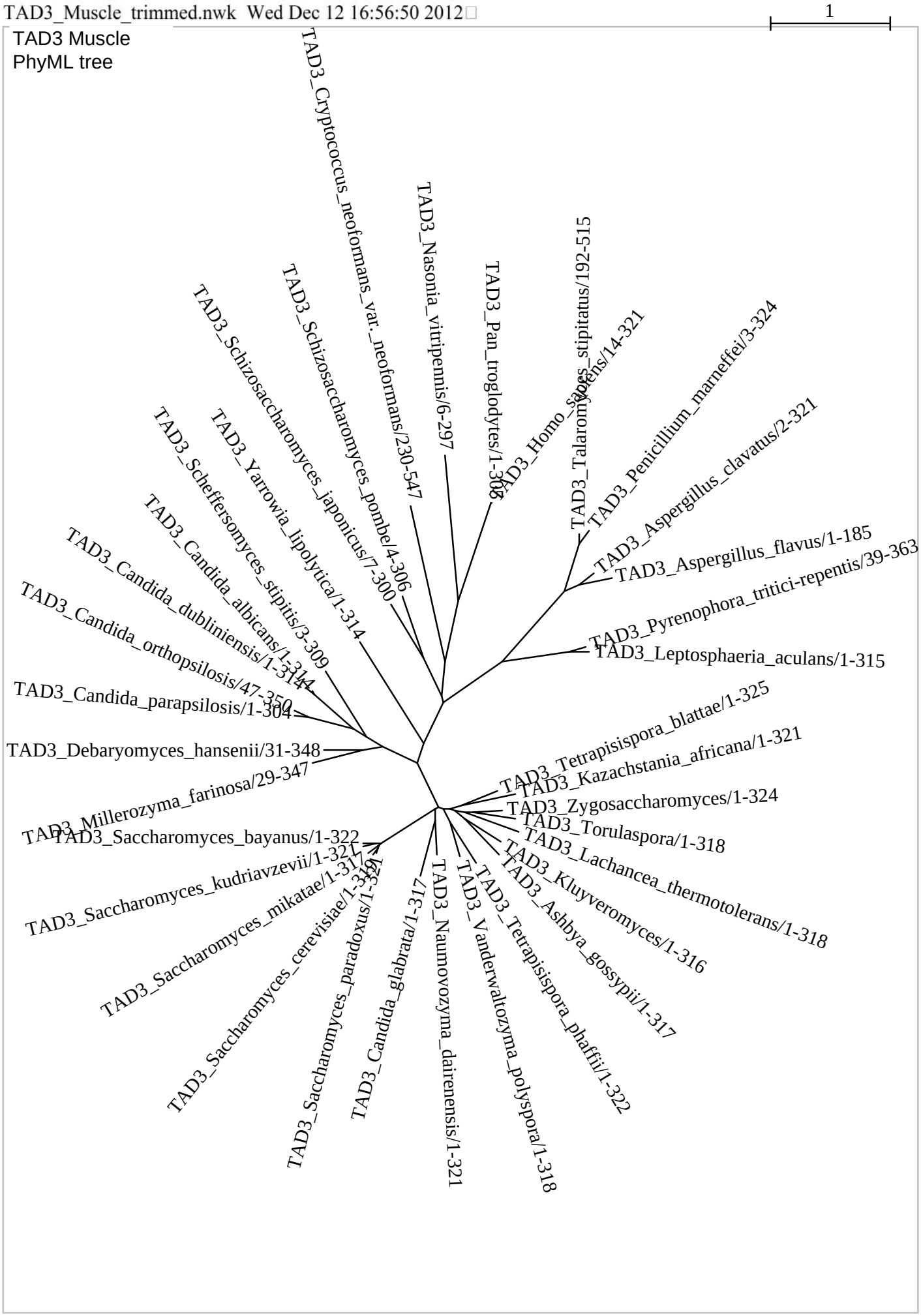
TAD3fortreeNJ.nwk Wed Dec 12 16:42:55 2012



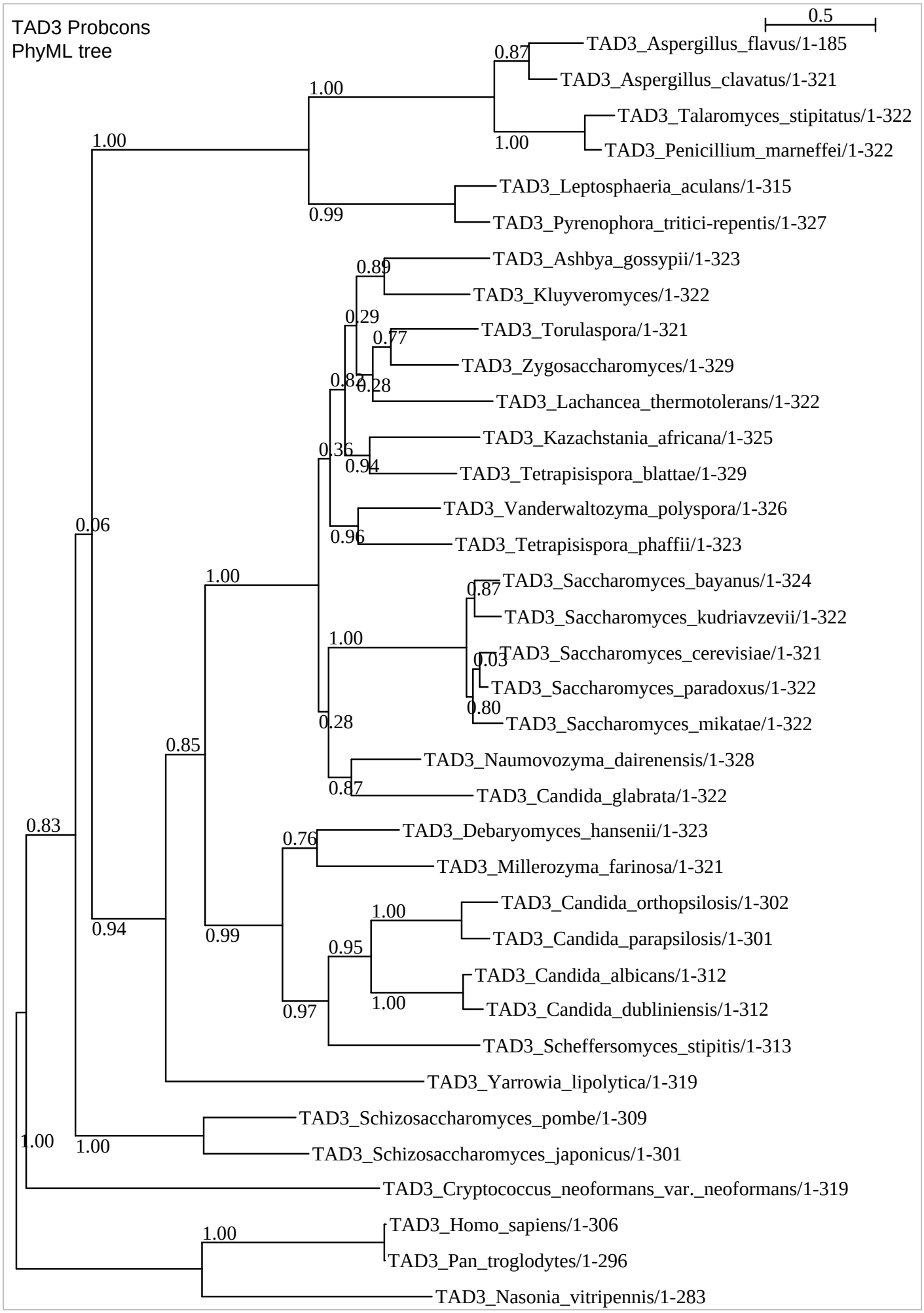
Seaview TAD3_Muscle_trimmed.nwk Wed Dec 12 16:56:33 2012



TAD3_Muscle_trimmed.nwk Wed Dec 12 16:56:50 2012

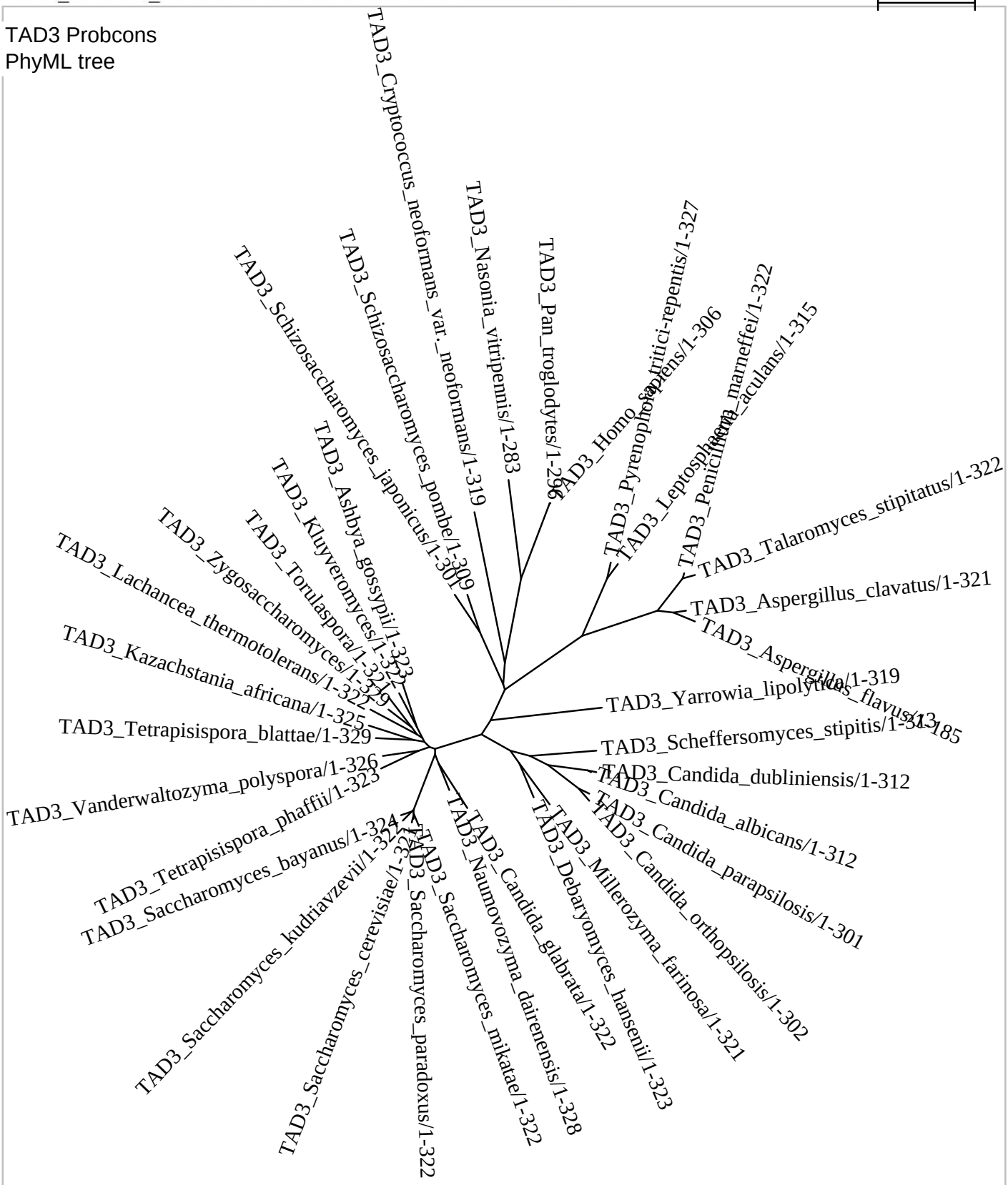


Seaview TAD3_Probcons_trimmedML.nwk Wed Dec 12 16:59:33 2012

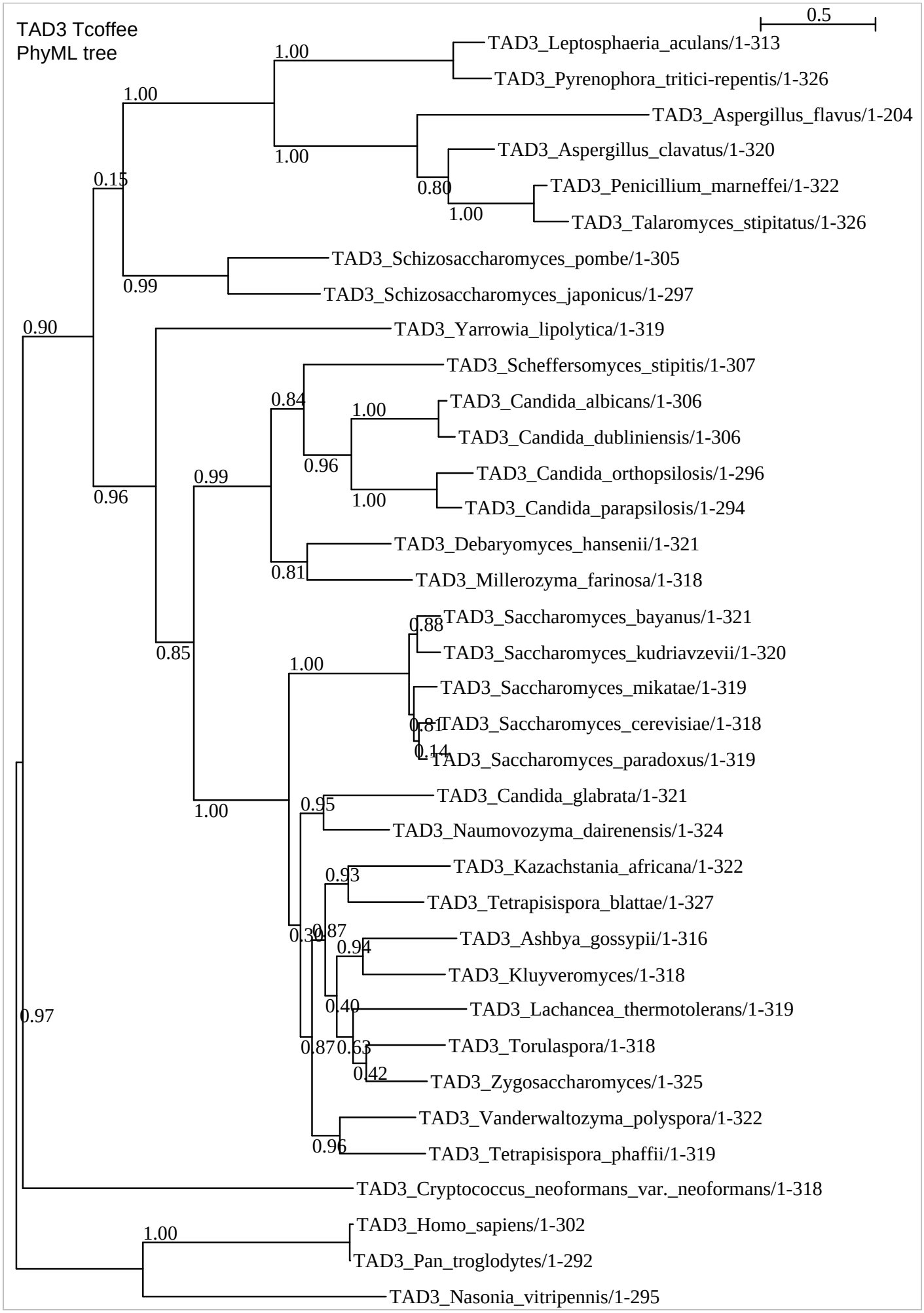


TAD3_Probcons_trimmedML.nwk Wed Dec 12 16:59:45 2012

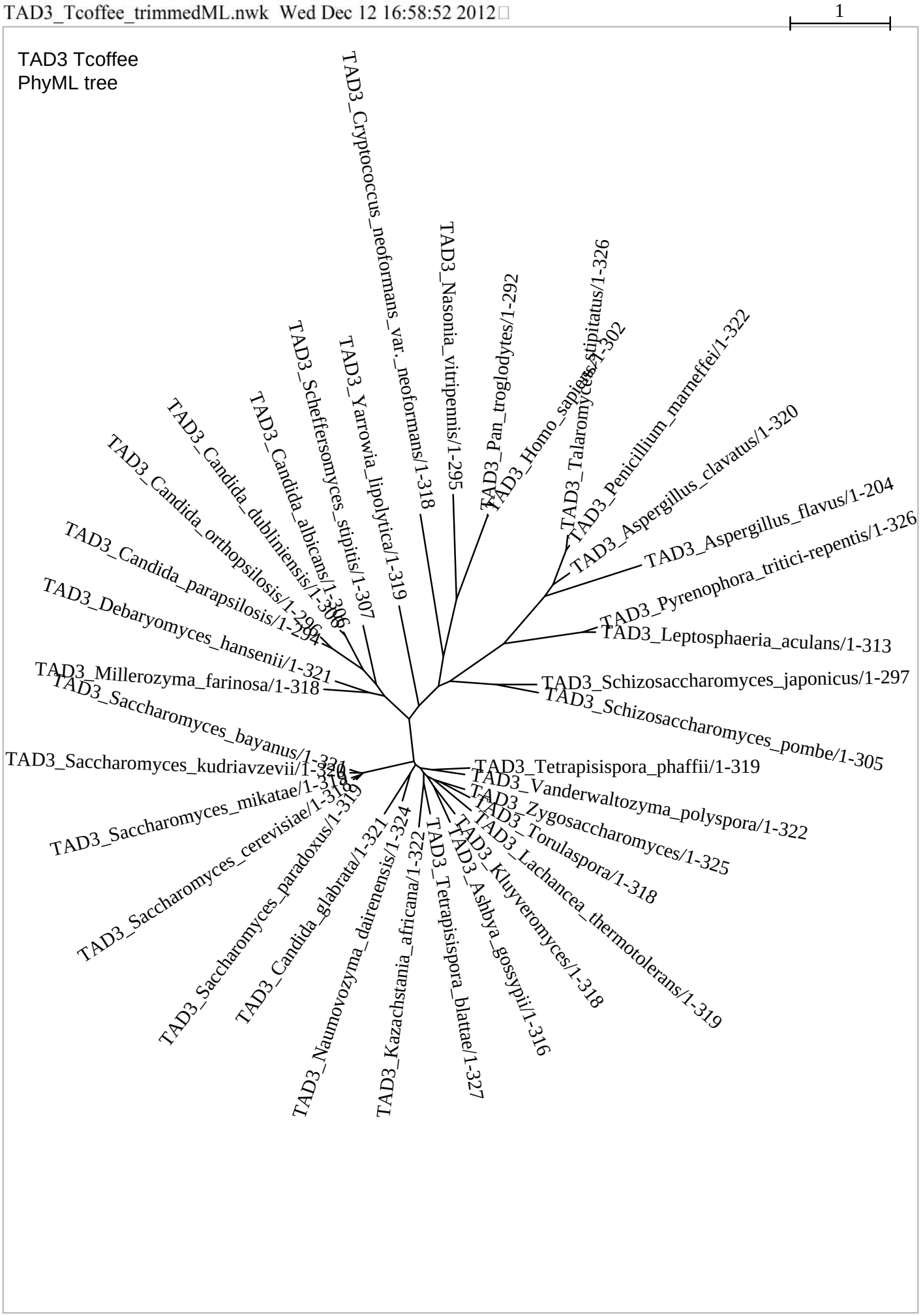
TAD3 Probcons
PhyML tree



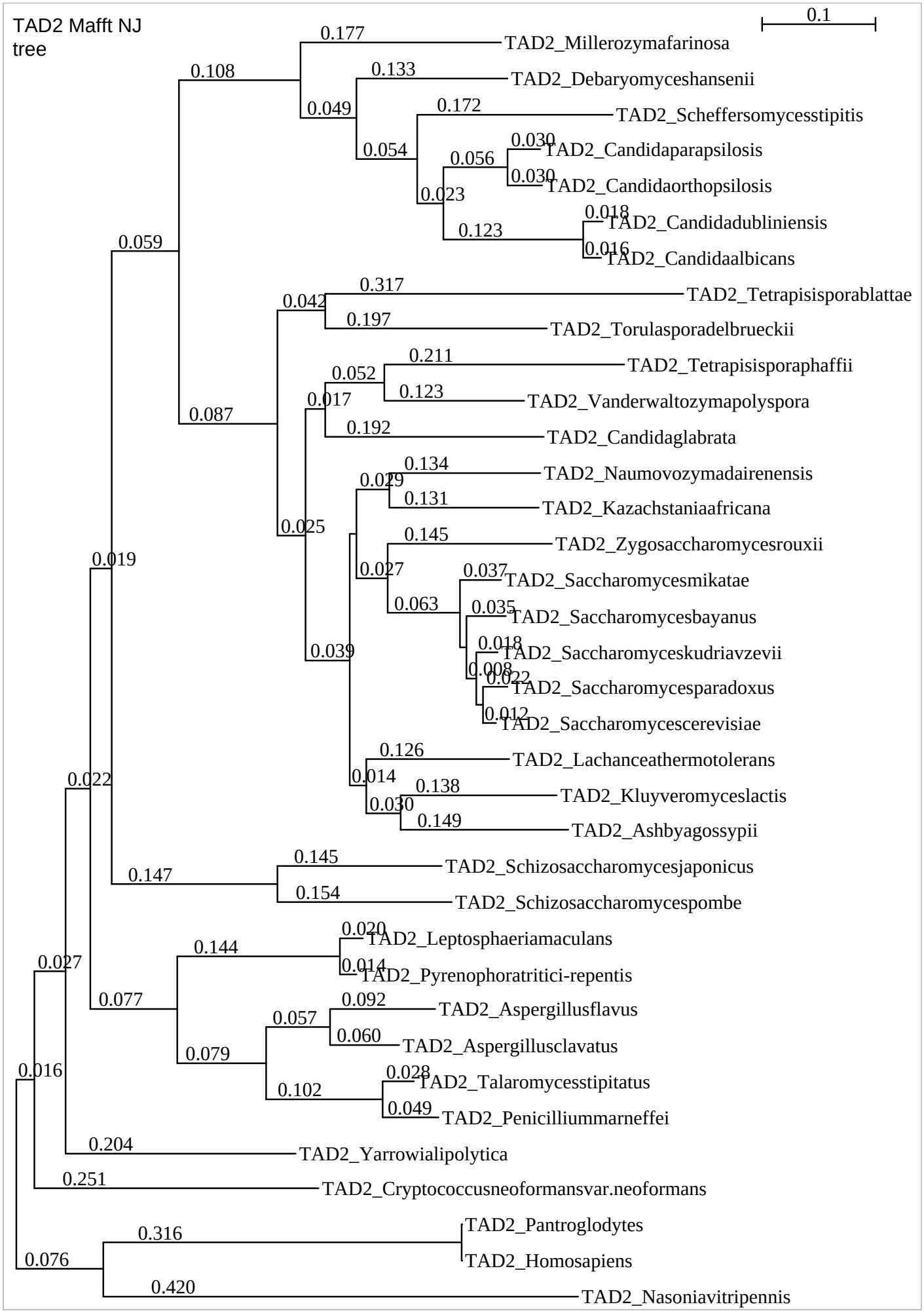
Seaview TAD3_Tcoffee_trimmedML.nwk Wed Dec 12 16:58:40 2012



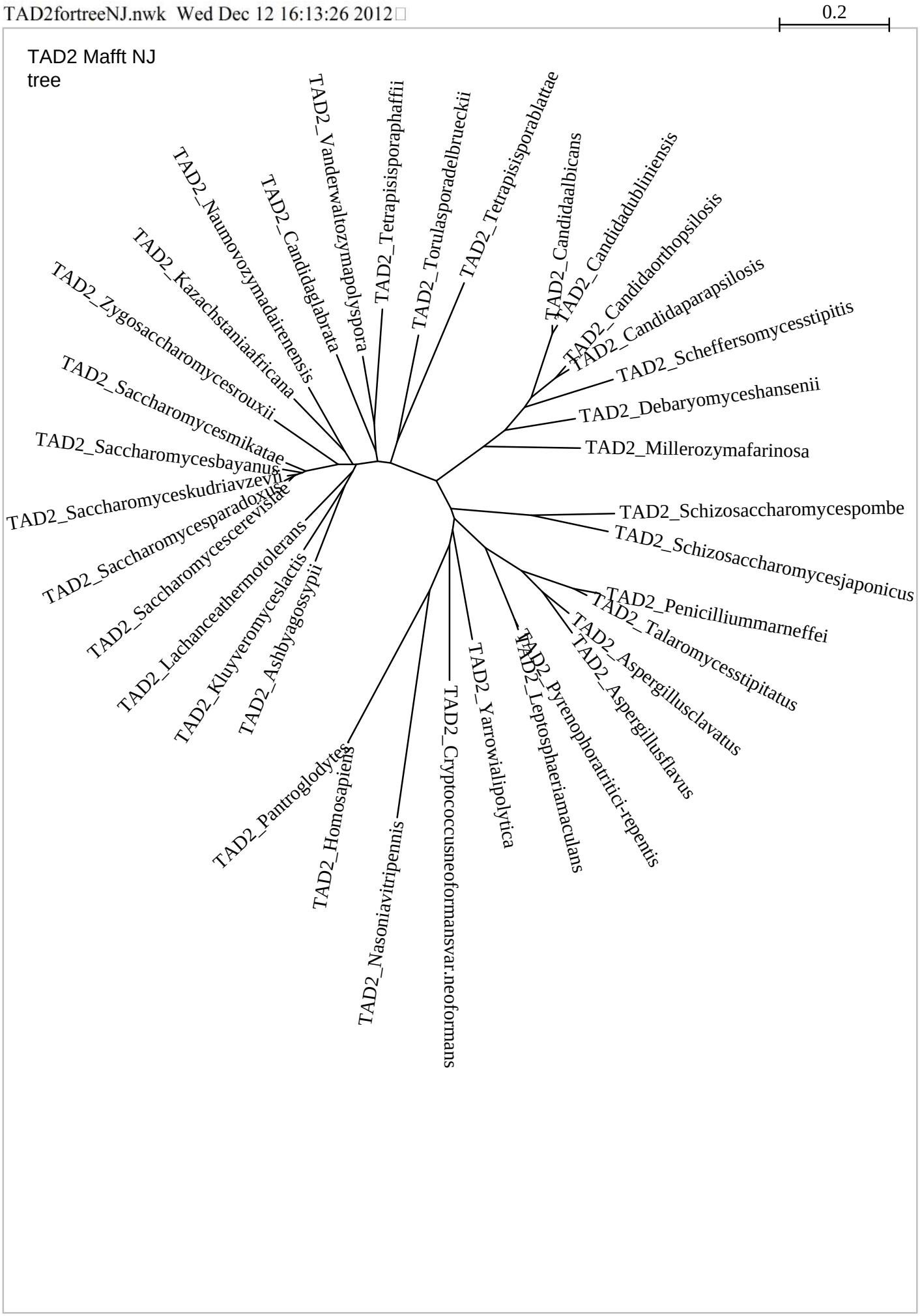
TAD3_Tcoffee_trimmedML.nwk Wed Dec 12 16:58:52 2012



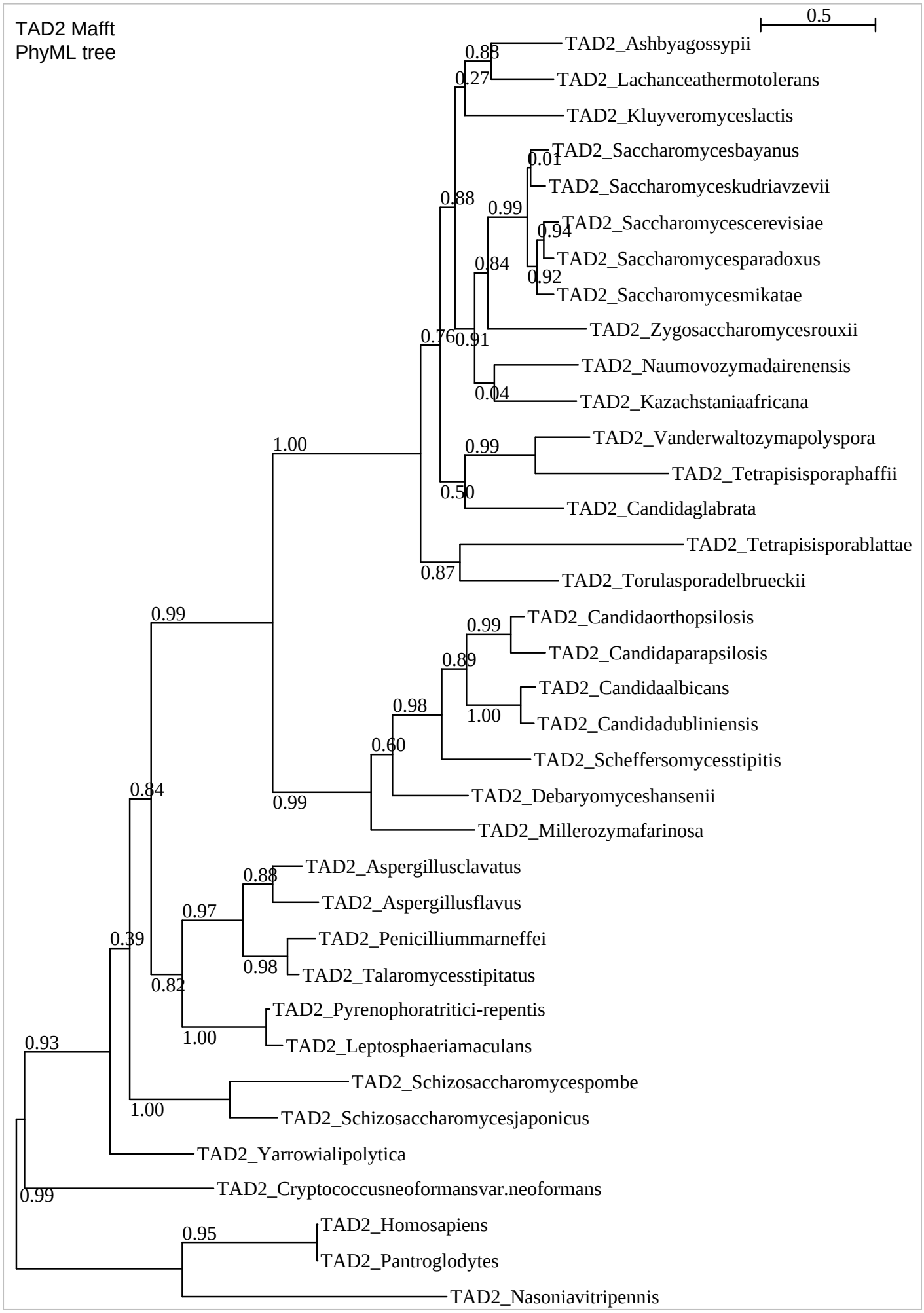
Seaview TAD2fortreeNJ.nwk Wed Dec 12 16:12:51 2012



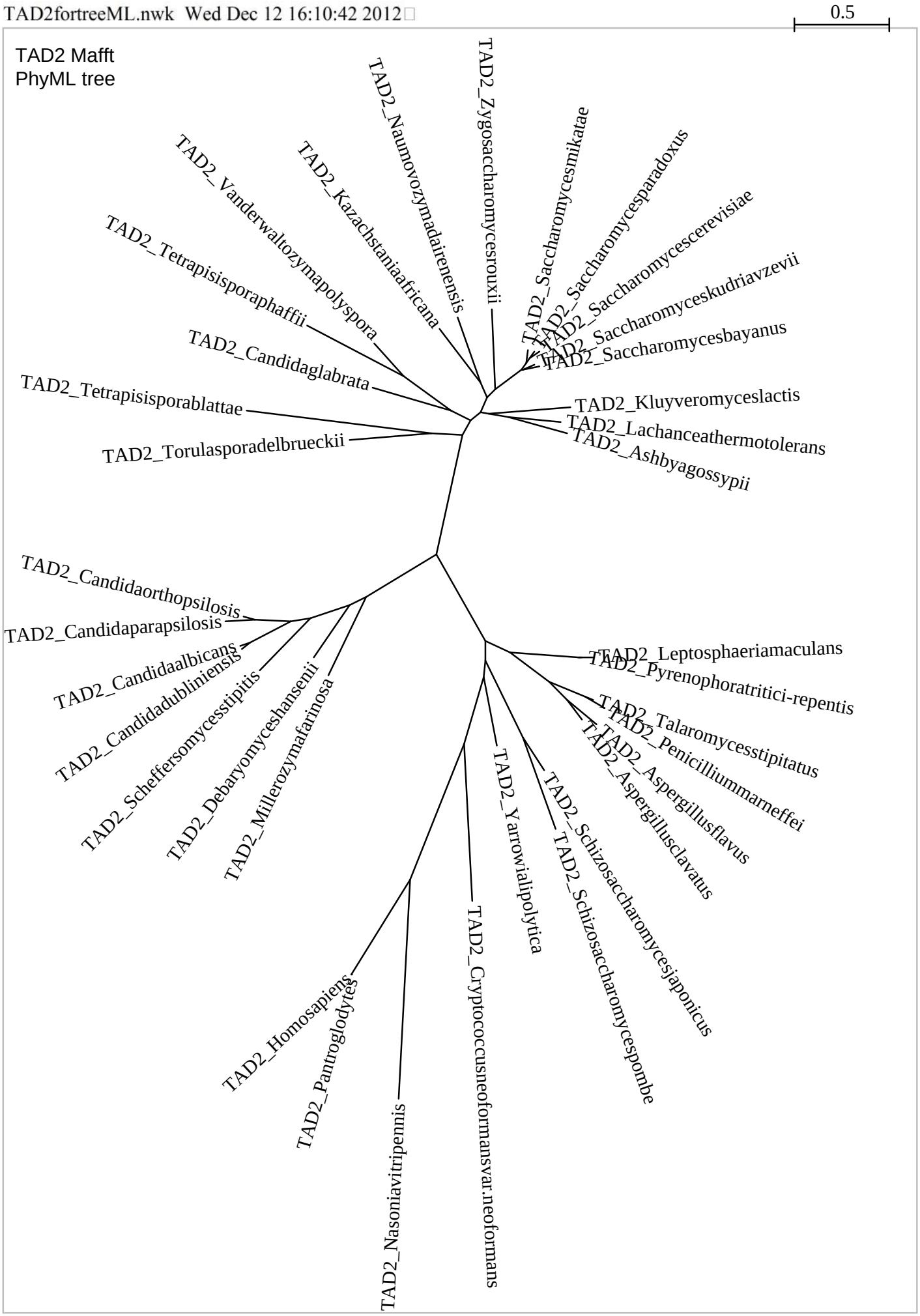
TAD2fortreeNJ.nwk Wed Dec 12 16:13:26 2012



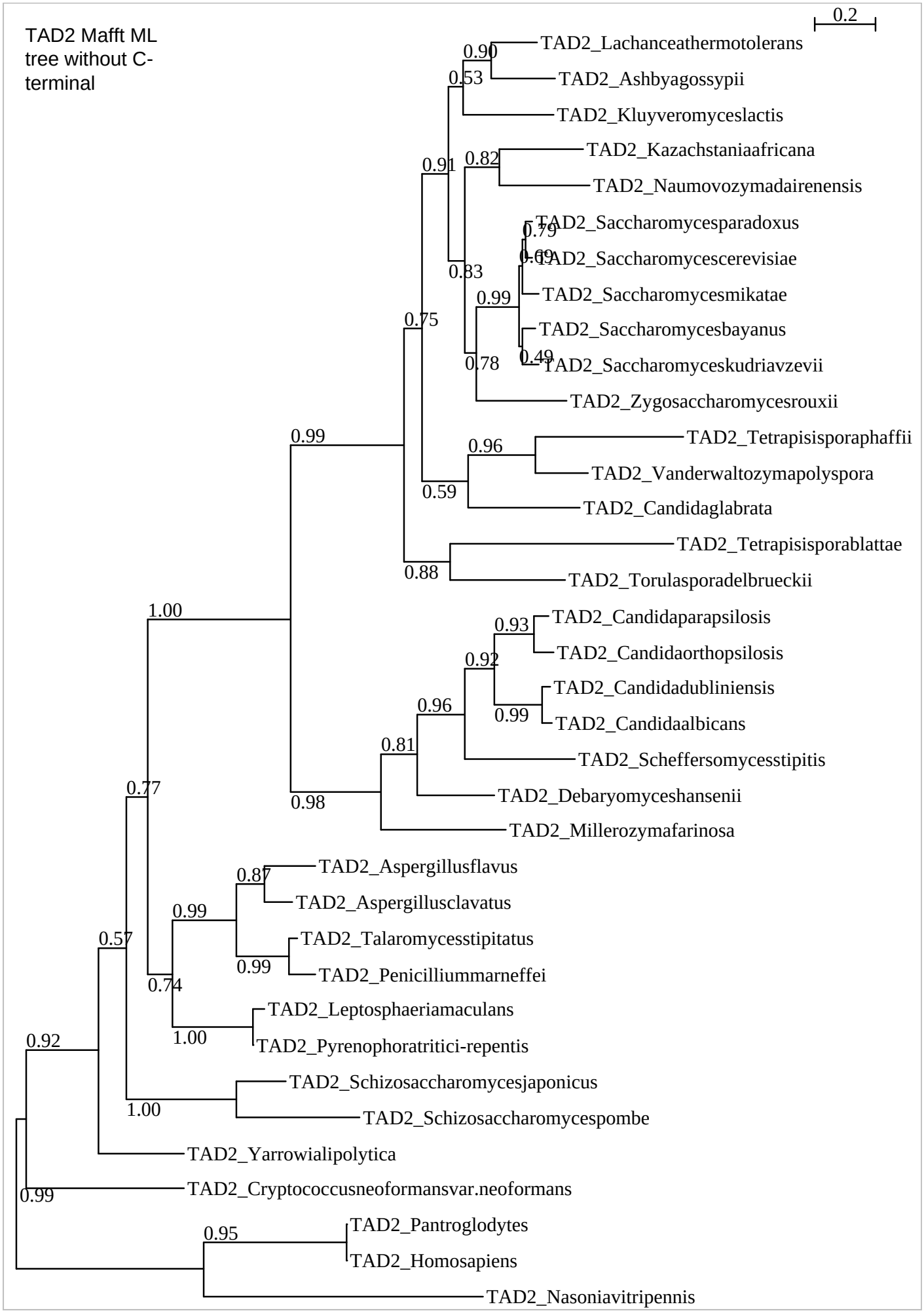
Seaview TAD2fortreeML.nwk Wed Dec 12 16:09:40 2012



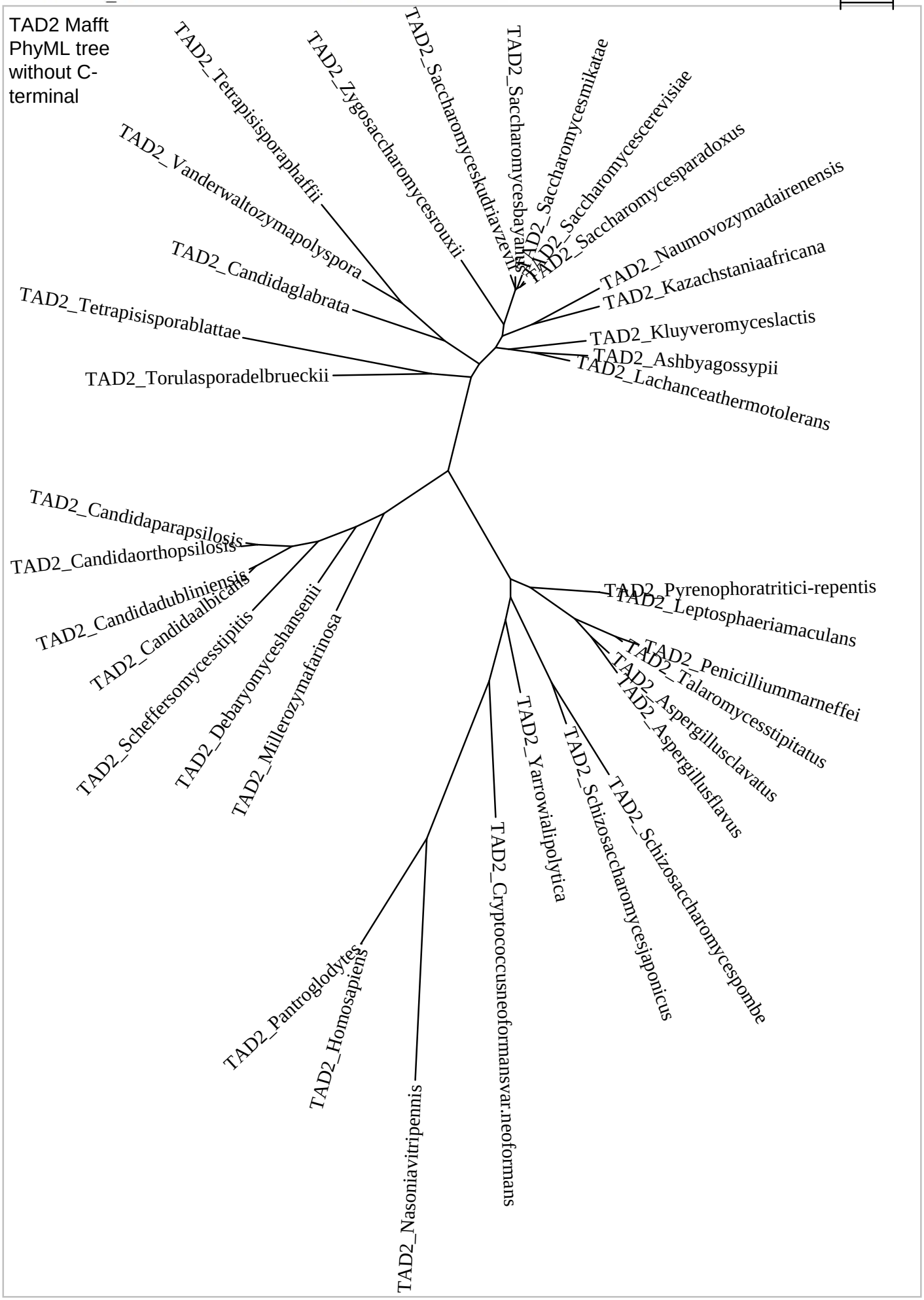
TAD2fortreeML.nwk Wed Dec 12 16:10:42 2012



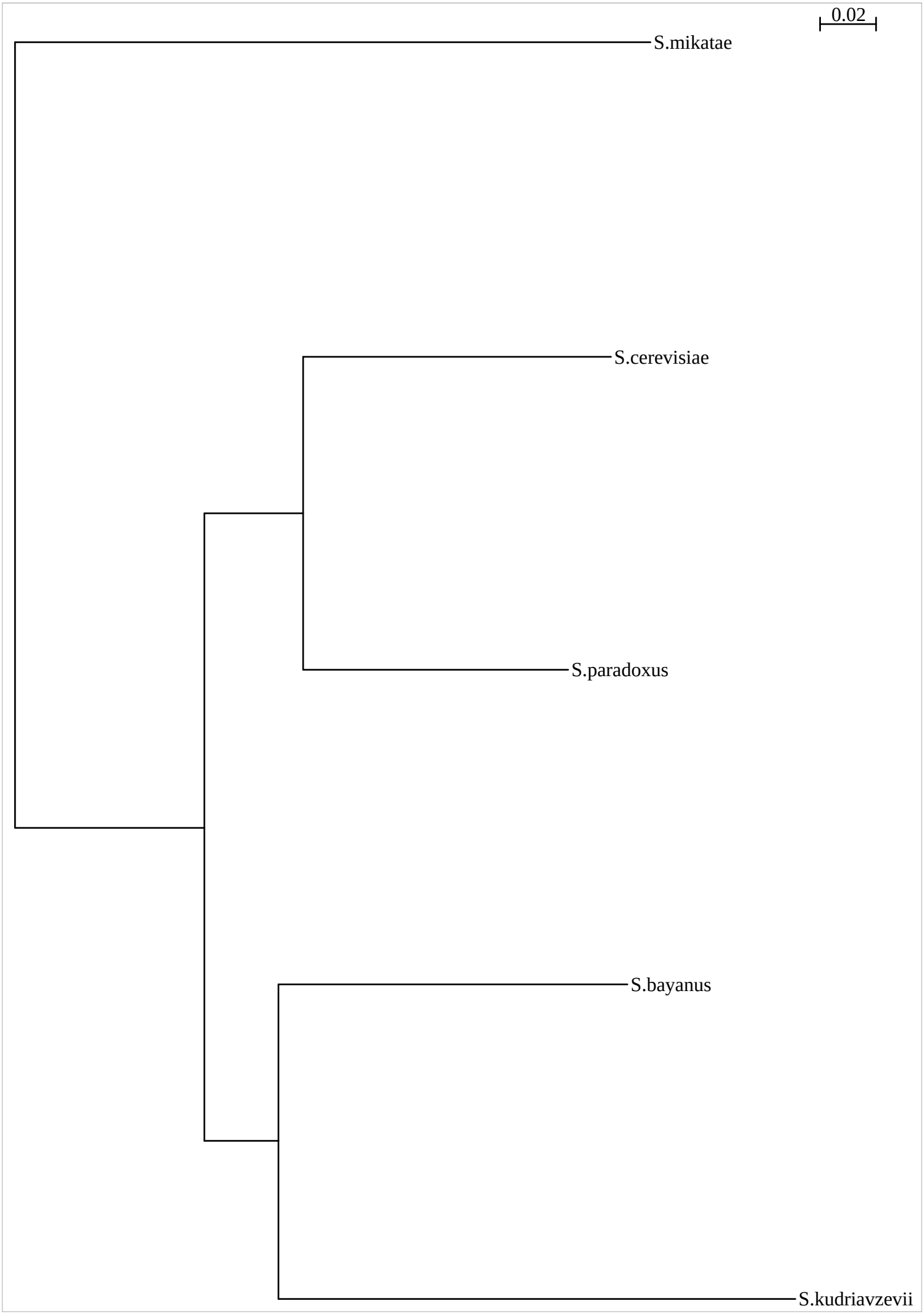
Seaview TAD2fortree_trimmoreML.nwk Wed Dec 12 16:23:24 2012



TAD2fortree_trimmoreML.nwk Wed Dec 12 16:23:57 2012



Seaview firstintron.nwk Wed Dec 12 22:42:55 2012



Seaview secondintron.nwk Wed Dec 12 22:43:39 2012