

Additional file 2 to:
“A fungal phylogeny based on 82 complete genomes
using the composition vector method”

Hao Wang, Zhao Xu, Lei Gao and Bailin Hao

Runtime of the CVTree

We have published a webserver for the CVTree algorithm into which the 82 fungi proteomes were integrated (<http://tlife.fudan.edu.cn/cvtree/>). We have run the program with K ranging from 3 to 10 and analyzed the performance. According to Table S1, The runtime varies with K, reaching the maximum when K=6. That runtime depends on the number of non-zero components of CV can be explained by the fact that the most time-consuming part in our current program is the harddisk I/O process reading and writing these components.

Table S1: Runtime of the CVTree. ^a the number of components of CV. ^b the mean number of non-zero components of CV. ^c the experiments were implemented on a computer of 2.3 GHz CPU and 4 G RAM

K	3	4	5	6	7	8	9	10
N ^a ($\times 10^6$)	0.008	0.16	3.2	64	1280	25600	512000	10240000
M ^b ($\times 10^6$)	0.008	0.16	3.0	19.9	12.3	4.77	3.95	3.88
Runtime ^c (min)	0.9	2.9	23.3	65.9	56.3	60.9	50.7	56.8

The best tree is obtained with K=7

The parameter K is the "resolution" controller of evolutionary information. Too small or too large K values fail to extract proper evolutionary features representing the organisms. CVTrees of K = 3 to 6 and 8 to 10 are shown in Figure S2.

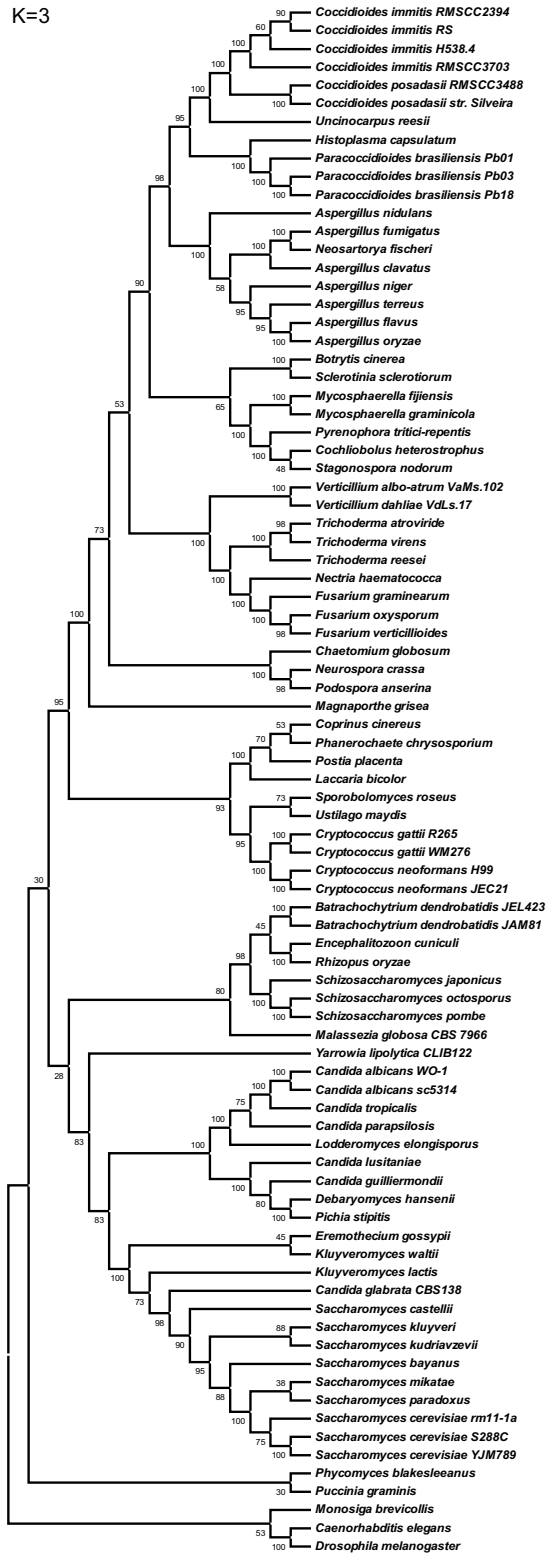
By comparing these trees to the K=7 topology (Figure 1 in the main text), we found that the bootstrap values (BPs) in trees improved with K increasing from 3 to 7 but deteriorated when K continued to increase. That is, the K=7 topology has better bootstrap values than all other trees investigated.

According to Figure S2 and Figure 1, it is obvious that both the topology and BPs become better with the K value increasing from 3 to 5 and worse when K longer than 7. Although K=5, 6 and 7 all produce reasonable topology and BPs at a glance, one observes the same improving process with the K increasing from 5 to 7 when conducting

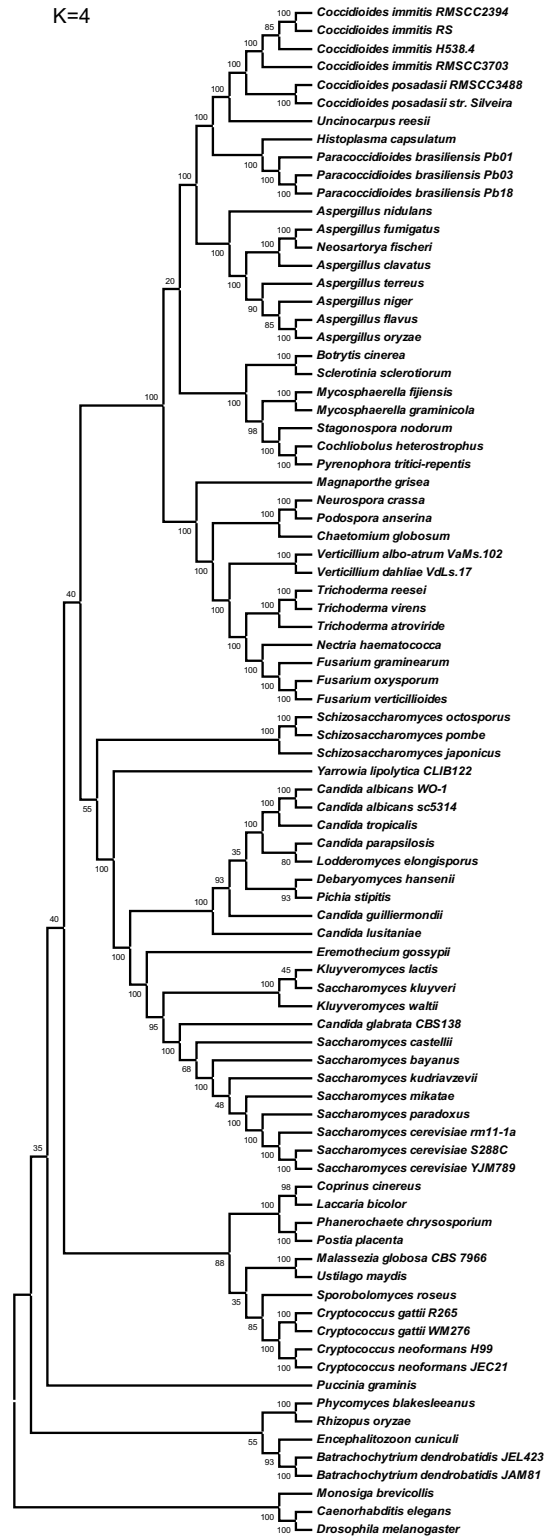
careful comparison (see Figure S2 and Figure 1): (1) At the level of Phylum, the K=5, 6, 7 trees all provides well-organized structure. The K=5 tree supports the Dikarya with a BP=93%, while the K=6 and 7 trees with BP=100%. (2) The K=5 and 6 trees give a sister relationship of the subphylum Taphrinomycotina and Saccharomycotina to the exclusion of the Pezizomycotina (see figure S2). This conflicts to, as we have mentioned in the main text, many resent researches that put the fission yeasts as the basal lineage of the Ascomycota. By contrast, the K=7 tree recovers the (Taphrinomycotina, (Saccharomycotina, Pezizomycotina)) structure with high bootstrap value (100% and 99%). (3) In the Basidiomycota, the subphylum-level relationships within the Basidiomycota in the K=5 tree is far less reasonable because organisms from the three subphyla mix together. The K=6 tree recovers the Agaricomycotina but fails to represent the monophyly of the well-supported subphyla Pucciniomycotina and Ustilaginomycotina by misplacing *Sporobolomyces roseus*. In contrast, the K=7 tree well supports (BP=100% and 99%) the three subphyla. (4) In the Saccharomycotina, the placement of *S. kudriavzevii* is modified and gains higher supports with K increasing from 5 to 7. (5) In the *Aspergillus*, the placement of *A. niger* and *A. terreus* is modified and gains higher supports with K increasing from 5 to 7.

From the trees showed in Figure S2 and Figure 1, one can easily see a trend that both bootstrap values and the quality of tree topology become better and better when K increases from 3 to 7 but worse and worse when K increases from 7 to 10. Here we must explain what we mean by “the quality of tree topology”: in the traditional SSU rRNA or few-gene based whole-genome phylogeny, the quality of a tree is judged mostly by stability and self-consistent arguments (by statistical resampling, e.g., bootstrapping or Jack-knifing). However, we take the century-long taxonomy, especially the recent results based on SSU rRNA analysis as EXPERIMENTAL facts and view the CVTree as THEORETICAL construction. The quality of a tree is judged by how close THEORY agrees with EXPERIMENT. We choose K=7 not only based on its best BP values but also on its best quality (in the above sense).

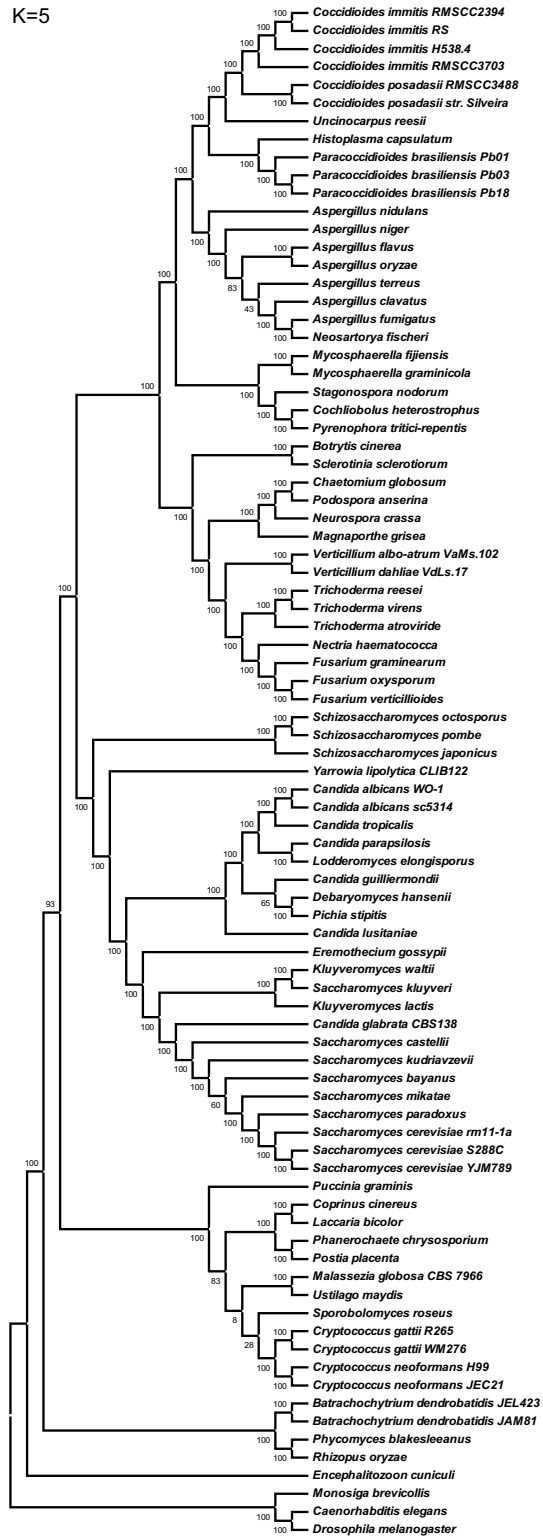
K=3



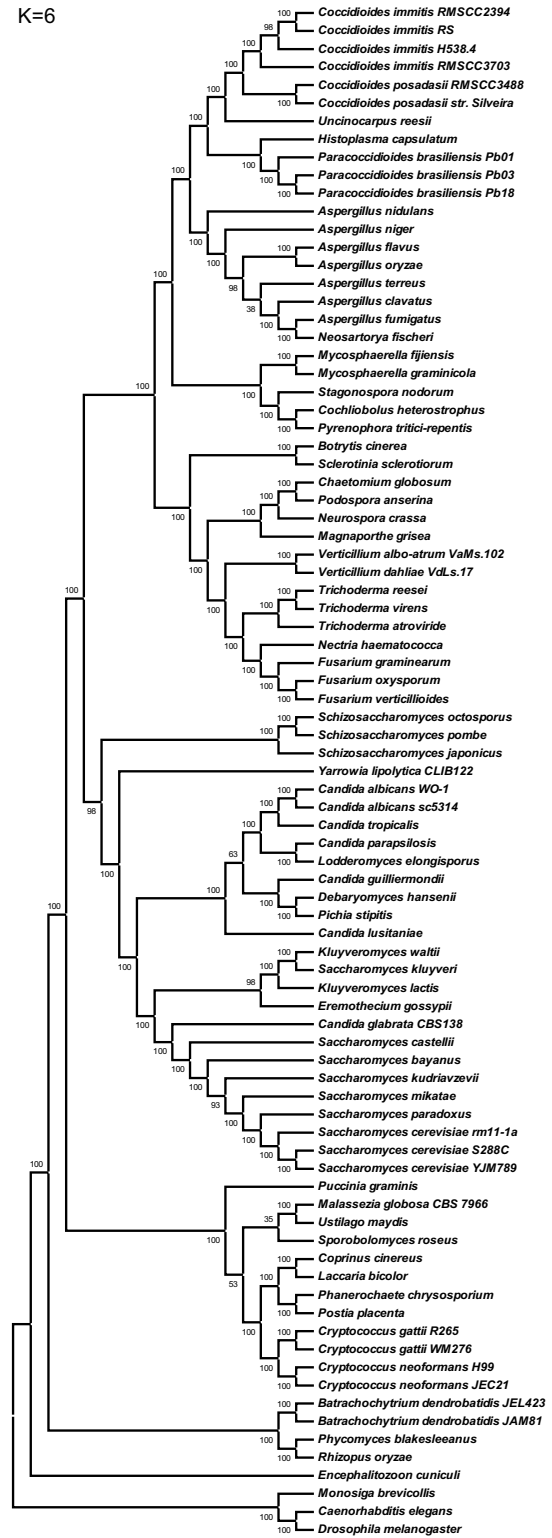
K=4



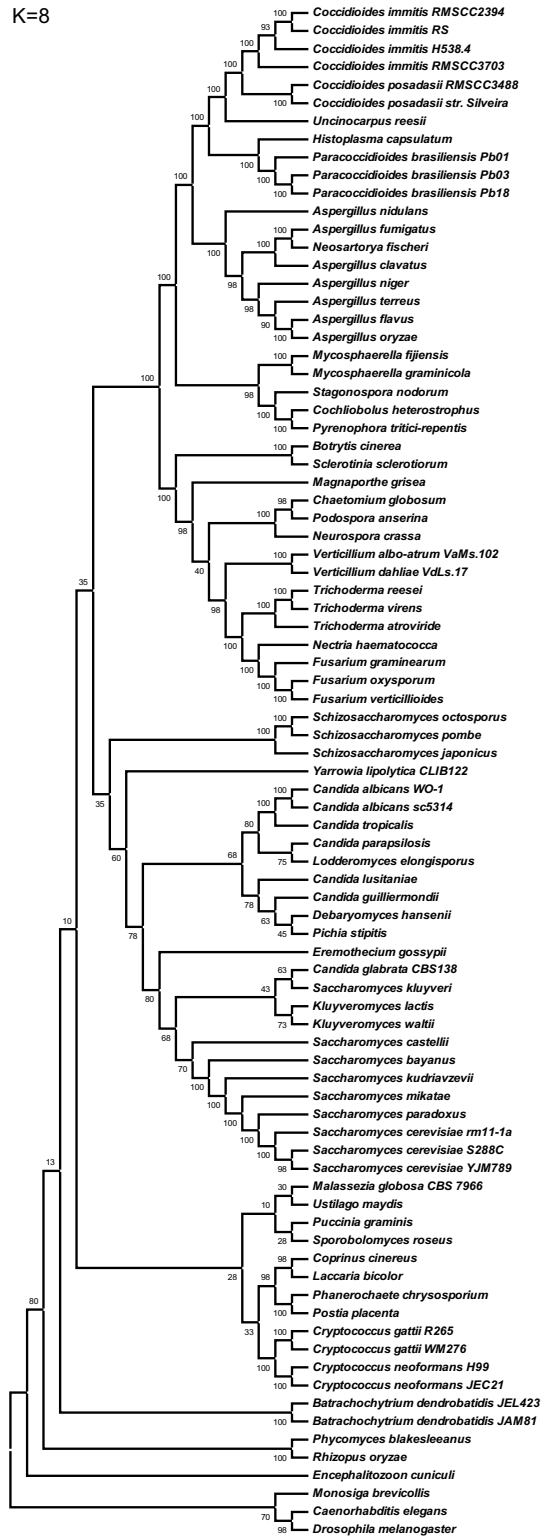
K=5



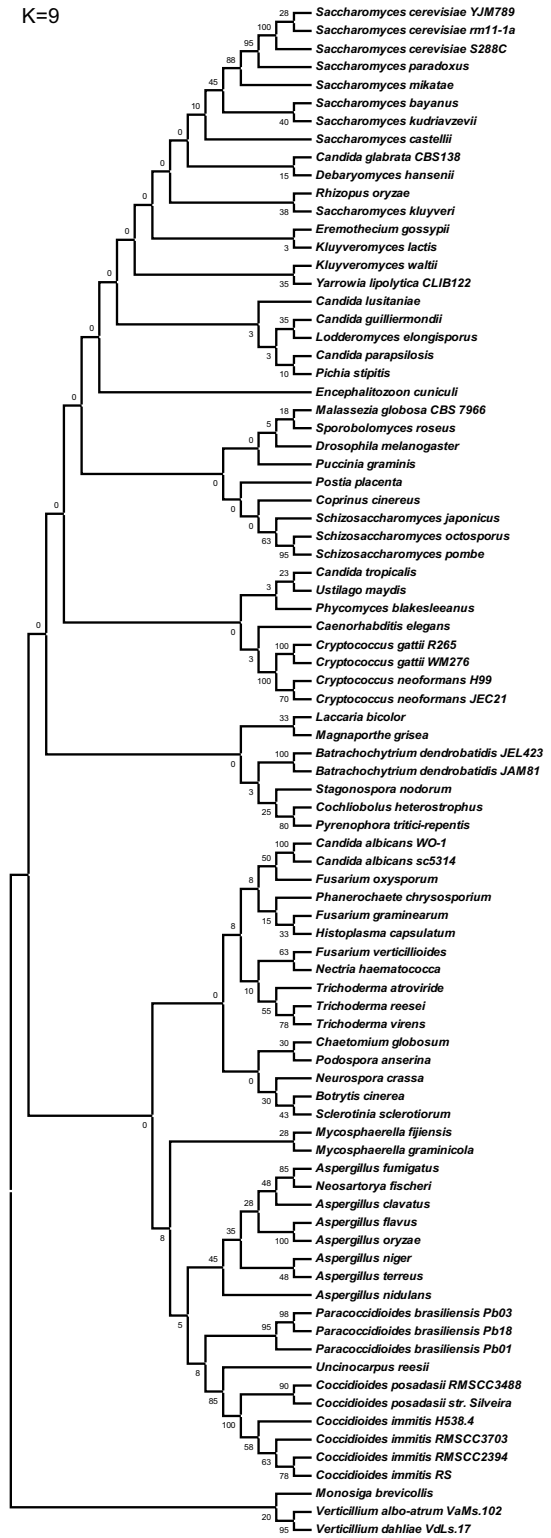
K=6



K=8



K=9



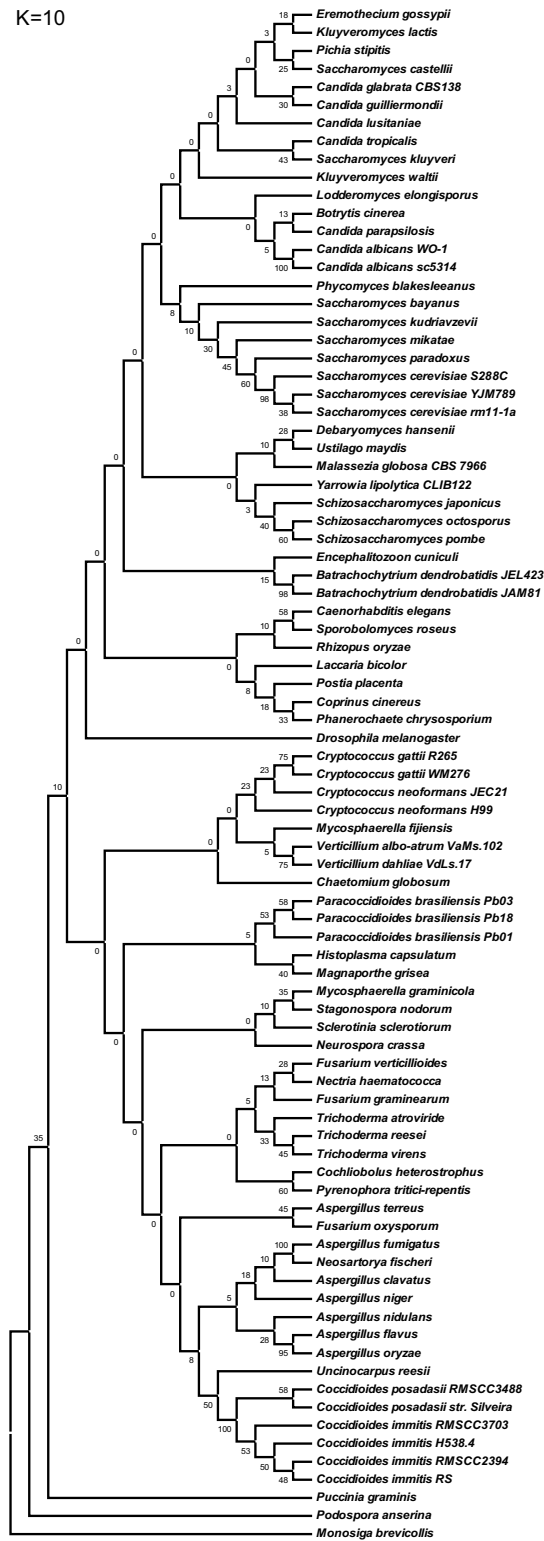


Figure S2: CVTrees with K=3 to 6 and 8 to 10. In each tree, BPs (100 bootstrap replicates) are reported as percentages.