Supporting Information

Schoch et al. 10.1073/pnas.1117018109

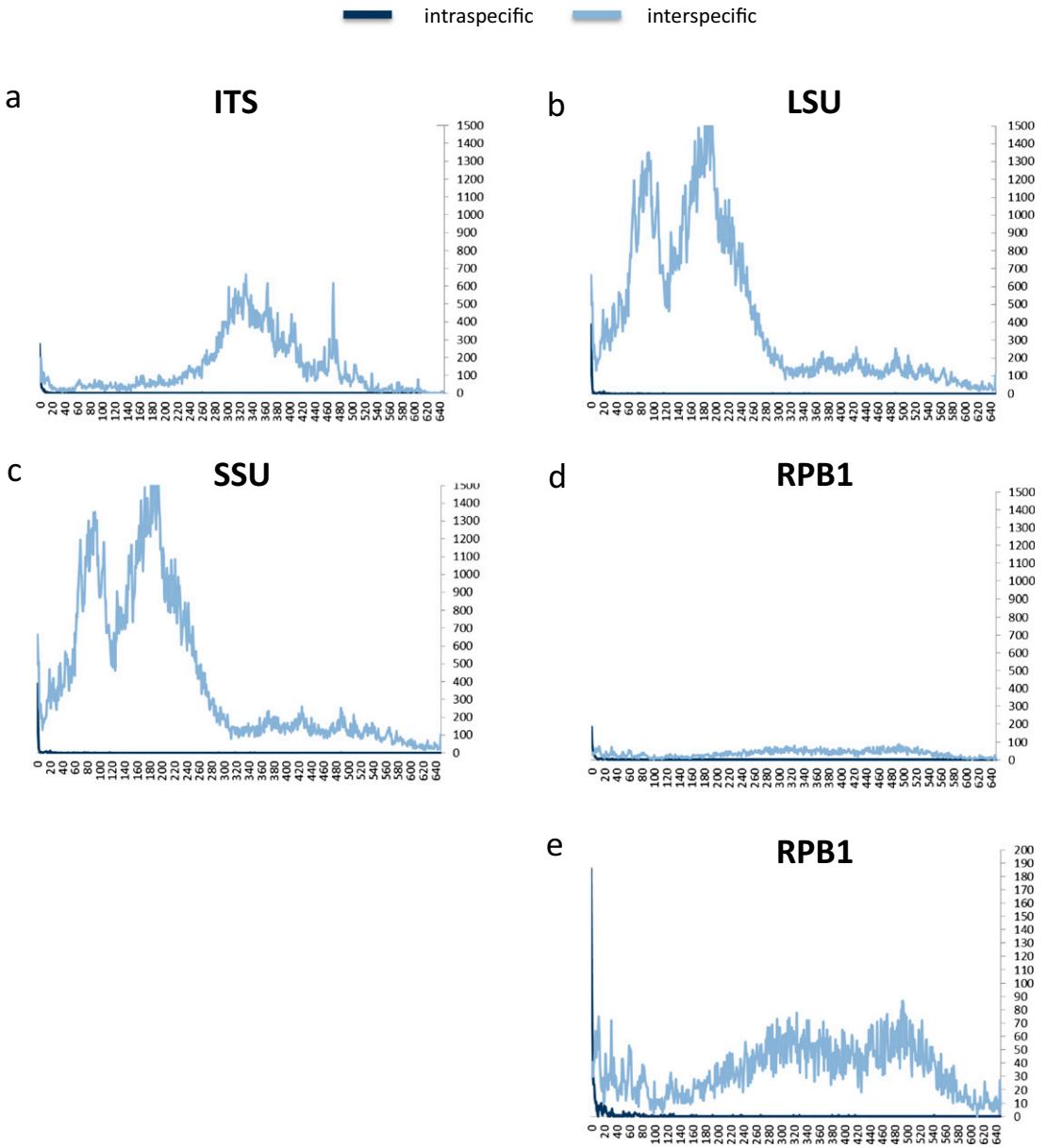


Fig. S1. Distribution of pairwise base pair differences among different barcode markers among the selected lineages from Fig. 2. Maximizing the difference between infra- and interspecific variation is an important variable to be assessed when selecting a barcode marker. To investigate and further visualize differences between sequence variation within and between species, uncorrected pairwise differences were calculated using the same datasets used for barcode gap probability of correct identification (PCI) estimates and plotted on a graph showing the variation of the four markers initially chosen for this study. Frequencies of base pair differences within and between species were recorded and are presented. A global alignment was used for these comparisons (1). Pairwise comparisons of variations within and between species are indicated. The *x* axes show numbers of base pair changes in pairwise comparisons, and the *y* axes show numbers of sequence pairs. The complete 742 strain dataset used for PCI analyses was compared for four markers. Light blue lines indicate variation between species, and dark blue lines indicate variations within species. *A*–*D* indicate all four markers with the same scale, and an additional graph (*E*) shows the largest subunit of RNA polymerase II (*RPB1*), with the *y* axis set to a maximum scale of 200-bp changes.

1. Robert VA, et al. (2011) BioloMICS Software: Biological data management, identification, classification and statistics. *Open Appl Inform J* 5:87–98.

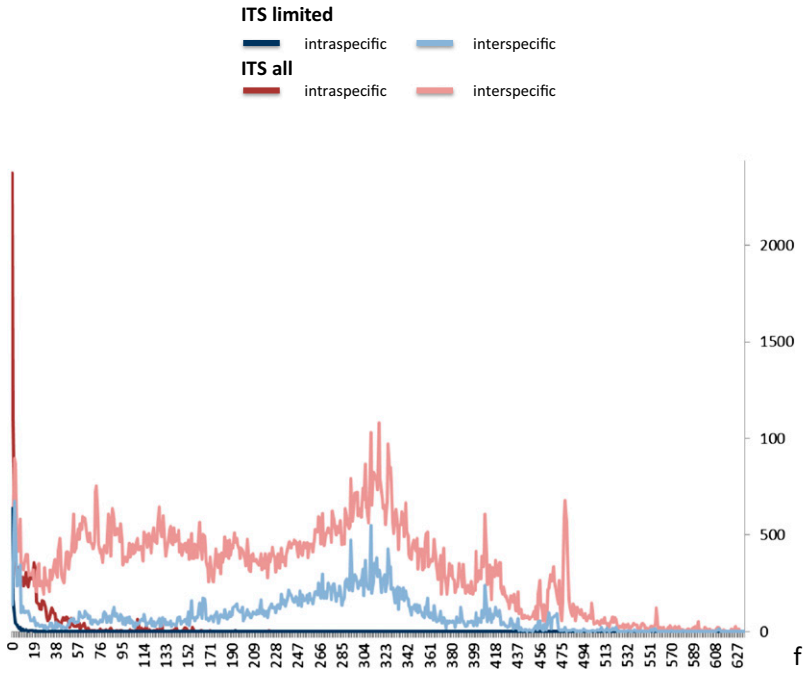


Fig. S2. Distribution of pairwise base pair differences among different barcode markers from a complete set of 3,256 strains from the fungal barcode database. The number of base pair changes seen in a complete set of 3,256 strains from the fungal barcode database after inclusion in an internal transcribed spacer (ITS) pairwise comparison. ITS all, the complete set of 3,256 strains; ITS limited, 742 strains used for the four-marker comparisons. Light and dark blue lines are as described for Fig. S1 for ITS limited. Pink and red lines apply as explained above to ITS all. The trends seen in these figures correlate very well with the barcode gap analyses in Fig. 3.

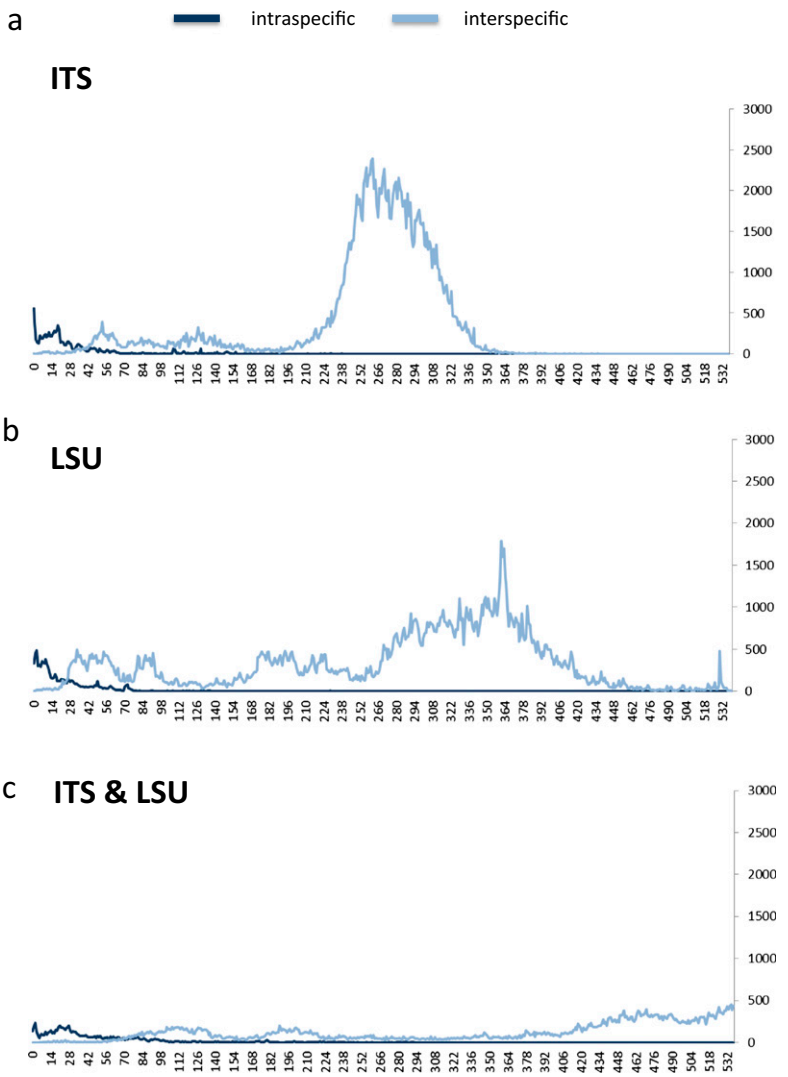


Fig. S3. Distribution of pairwise base pair differences among different barcode markers within *Glomeromycota*. The same analysis as in Fig. S1 was performed but applied to 606 sequences from 42 species of *Glomeromycota*. Variation within a species is indicated in dark blue, and variation between species is in light blue. The scale of the *x* axes indicates single base differences, and the scale on the *y* axes indicates numbers of pairwise comparisons. (*A*) ITS and (*B*) long subunit rRNA gene (LSU) are compared as well as (*C*) a combination of the two. The higher level of variation within species is reflected by the higher initial values for the intraspecific variation.

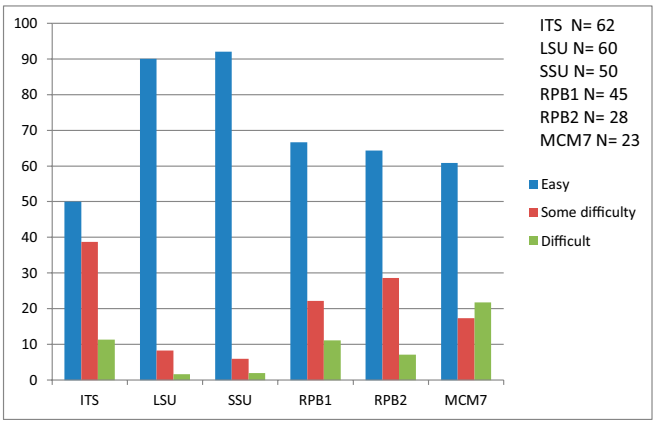
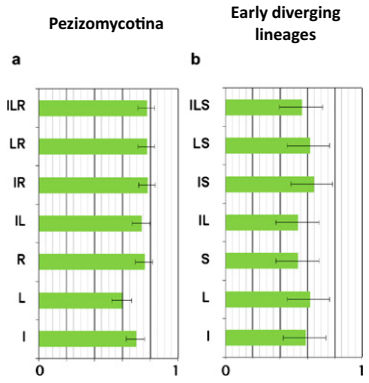
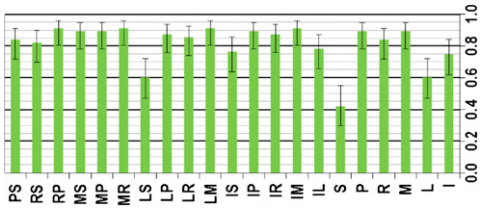


Fig. S4. A breakdown of PCR success according to survey responses among study contributors. The majority of respondents ranked all genes attempted as easy to edit and align, but some difficulty was noted for ITS; the three protein-coding genes with a gene encoding a minichromosome maintenance protein (*MCM7*) were considered the most difficult to edit and align (22% of respondents).



Barcode gap PCI, three marker analyses

Fig. S5. Barcode gap probability of identification for expanded sets of strains comparing only three barcode markers. Barcode gap PCI for the *Pezizomycotina* and early diverging lineages using three-marker datasets. The plots show the combinations of barcode markers investigated on the *y* axes. I, ITS; L, LSU; R, *RPB1*; S, small subunit rRNA gene. The *x* axes show the bar code gap PCI estimates for (*A*) *Ascomycota: Pezizomycotina* (179 species) and (*B*) early diverging lineages (34 species). The error bars indicate 95% confidence intervals for the bar code gap PCI estimate.



Barcode gap PCI, six marker analysis

Fig. S6. Barcode gap probability of identification for a six-marker dataset including three protein-coding markers. Barcode gap PCI for a six-marker dataset. The plots show the combinations of barcode markers investigated on the x axis. I, ITS; L, LSU; M, *MCM7*; P, *RPB2*; R, *RPB1*; S, small subunit rRNA gene. The y axis shows the barcode gap PCI estimate for the six-marker dataset (55 species). The error bars indicate 95% confidence intervals for the PCI estimate.

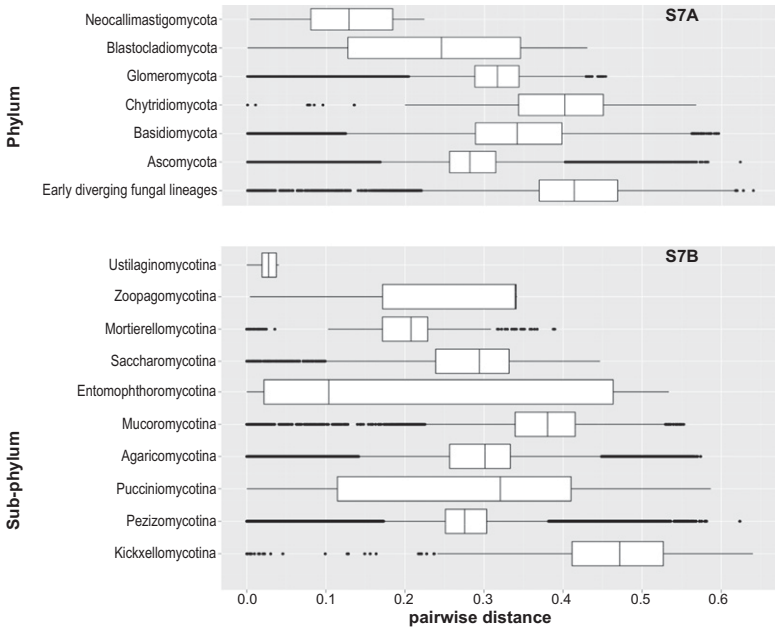


Fig. S7. Intraspecific pairwise distances within each phylum and subphylum in an expanded set of ITS sequences. The ITS dataset with 2,896 sequences from this study was analyzed with R, excluding a number of sequences analyzed in Fig. S2 because of short lengths. The analysis was performed (1) to generate box plots for all possible intraspecific pairwise distances within each phylum (*A*) and subphylum (*B*). The line in the middle of the box is the median, the left and right sides of the box are the 25th and 75th percentiles, respectively, and the whiskers are 1.5 times the interquantile range above and below the box limits. The dots are outliers (i.e., beyond ±2.7 SD). In all three figures, the variation among the pairwise difference within each taxon reflects the different characteristics and variation of ITS in each fungal group. Groups with narrow amounts of variation, especially those groups clustered near zero on the *x* axis, may not function well with ITS as a barcode. Those groups with broader ranges and medians with higher values on the *x* axis should function better with ITS as a barcode, although variation in species concepts and sampling density may obscure the significance of the apparent differences.

1. R Development Core Team (2011) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).

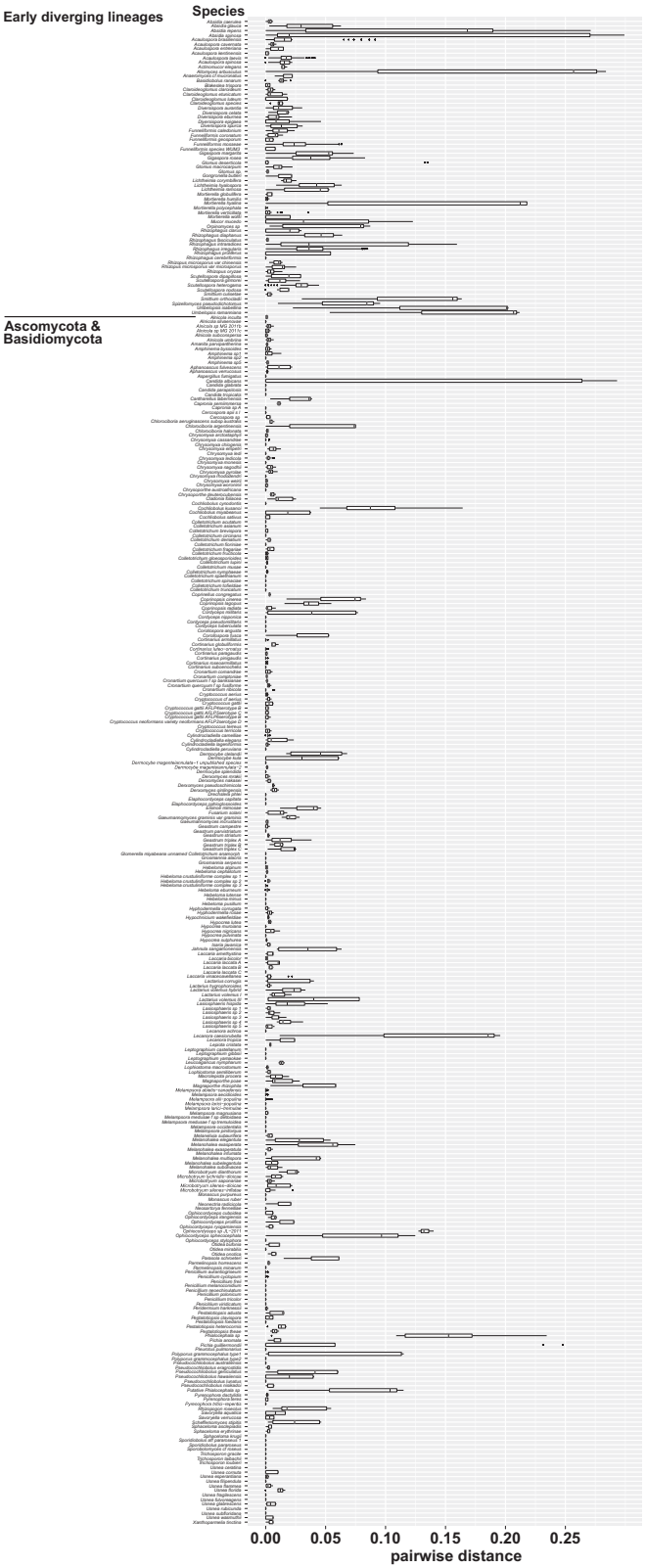


Fig. S8. Intraspecific pairwise distances within each species in an expanded set of ITS sequences. The ITS dataset with 2,896 sequences from this study was analyzed as in Fig. S7 but on species level. Only species with three strains or more were considered. The line in the middle of the box is the median, the left and right sides of the box are the 25th and 75th percentiles, respectively, and the whiskers are 1.5 times the interquantile range above and below the box limits. The dots are outliers (i.e., beyond ±2.7 SD).

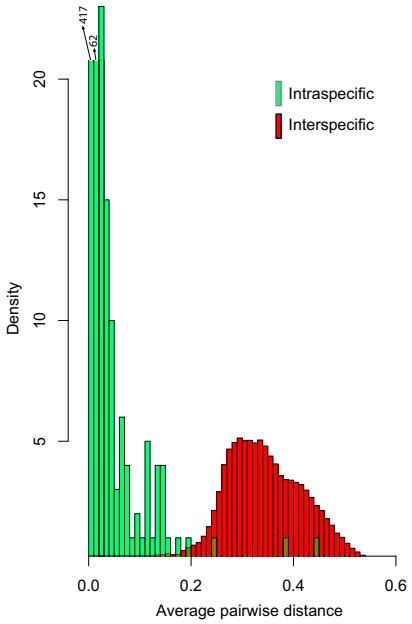


Fig. S9. Barcode gap analysis for an expanded set of ITS sequences covering all *Fungi*. Barcode gap analysis for 2,896 ITS sequences generated in this study, excluding a number of sequences analyzed in Fig. S2 because of short lengths. The averages were derived with the methodology described in the work by Robideau et al. (1), except that computation and plots shown here were done with R (2) instead of Statistic Analysis Software.

1. Robideau GP, et al. (2011) DNA barcoding of oomycetes with cytochrome c oxidase subunit I and internal transcribed spacer. *Mol Ecol Resour* 11:1002–1011.
2. R Development Core Team (2011) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).

Table S1. PCR conditions and primers used for the majority of amplifications

	Forward	Reverse	Final concentration	PCR protocol
LSU (LR0R-LR5)	5′-ACCCGCTGAACTTAAGC-3′	5′-TCCTGAGGGAAACTTCG-3′	0.2 μM each	95 °C for 10 min; 35 cycles of 95 °C for 15 s, 48 °C for 30 s, and 72 °C for 1.5 s; 72 °C for 7 min; and 4 °C on hold
RPB1 (RPB1-Af, RPB1-Ac-RPB1-Cr)	5′-GARTGYCCDGGDCAYTTYGG-3′	5′-CCNGCDATNTCRTTRTCCATRTA-3′	1 μM each	95 °C for 10 min; 35 cycles of 95 °C for 15 s, 52 °C for 30 s, and 72 °C for 1.5 s; 72 °C for 7 min; and 4 °C on hold
SSU (NS1, NS4)	5′-GTAGTCATATGCTTGTCTC-3′	5′-CTTCCGTCAATTCCTTTAAG-3′	0.4 μM each	95 °C for 10 min; 35 cycles of 95 °C for 15 s, 52 °C for 30 s, and 72 °C for 1.5 s; 72 °C for 7 min; and 4 °C on hold
ITS (ITS5, ITS4)	5′-TCCTCCGCTTATTGATATGC-3′	5′-GGAAGTAAAAGTCGTAACAAGG-3′	0.2 μM each	95 °C for 10 min; 35 cycles of 95 °C for 15 s, 52 °C for 30 s, and 72 °C for 1.5 s; 72 °C for 7 min; and 4 °C on hold
M13F-20	5′-GTAAAACGACGGCCAGTG-3′		1.6 pmol per 10 μL sequencing reaction	NA
M13R-27		5′-GGAAACAGCTATGACCATG-3′	1.6 pmol per 10 μL sequencing reaction	NA
RPB2 (fRPB2-5F-RPB2-7R)	5′-GAYGAYMGWGATCAYTTYGG-3′	5′-CCCATWGCYTGCTTMCCCAT-3′	1 μM each	95 °C for 10 min; 35 cycles of 95 °C for 15 s, 52 °C for 30 s, and 72 °C for 1.5 s; 72 °C for 7 min; and 4 °C on hold
MCM7 (Mcm7-709for, Mcm7-1348rev)	5′-ACIMGIGTITCVGAYGTHAARCC-3′	5′-GAYTTDGCIACICCIGGRTCWCCCAT-3′	1 μM each	94 °C for 10 min; 38 cycles of 94 °C for 45 s, 56 °C for 50 s, 72 °C for 1 min, and 72 °C for 5 min

ITS, internal transcribed spacer; LSU, long subunit rRNA gene; *MCM7*, gene encoding a minichromosome maintenance protein; *RPB1*, largest subunit of RNA polymerase II; *RPB2*, second largest subunit of RNA polymerase II; SSU, small subunit rRNA gene.

Other Supporting Information Files

Dataset S1 (XLS)
SI Appendix (DOC)