Fig S1: Generation of the global key: technical description

Two distinct sequence datasets were used to generate global key clusters:

1. sequences from the UNITE database (UNITE dataset: 12 667 sequences);
2. fungal rDNA ITS sequences retrieved from GenBank (INSD dataset: 323 513 sequences) using the following search string

```
((("Fungi"[ORGN] AND (140[SLEN] : 3000[SLEN])) AND
(((ITS1[titl] OR ITS2[titl]) OR 5.8S[titl]) OR "internal
transcribed spacer"[titl] OR "internal transcribed
spacers"[titl] OR "ITS 1" [titl] OR "ITS 2"[titl])) NOT
"Uncultured Neocallimastigales"[ORGN]
```

Step 1: quality filtering

Initial quality filtering (sequences flagged as "low quality" or "chimeric" on the PlutoF workbench (Abarenkov et al., 2010b)) discarded 64 and 9 365 sequences from the UNITE (Abarenkov et al., 2010a) and INSD (Benson et al., 2006) datasets respectively.

Step 2: fungal ITS extractor

For the remaining 326 751 sequences ITS1 and ITS2 were separated using the fungal ITS extractor (Nilsson et al. 2010). Sequences without ITS2 region (61 475), sequences containing more than 3 ambiguous (N) nucleotides in the ITS2 region (2 591), and sequences with questionable suitability for the global key by manual inspection (62) were excluded from further analysis.

Step 3: USEARCH clustering (clustering step 1)

ITS2 regions for the 262 623 sequences surviving the cleaning step were submitted to USEARCH v6.0.307 (Edgar, 2010) analysis for clustering on 80% similarity threshold with the following command

```
usearch –clusterfast infile.fasta –id 0.80 –centroids
centroids_out.fasta –uc clusters_out.uc
```

Clustering produced 7 470 clusters and 4 902 singletons, 1 046 sequences having length < 32 nucleotides were discarded by the program.

Step 4: aligning clusters

All clusters were aligned using the multiple sequence alignment program MAFFT v6.833b (Katoh et al., 2002) with the following parameters

number of sequences in cluster <= 200: `mafft-linsi`
200 < number of sequences in cluster <= 750: `mafft --retree 2 --maxiterate 3`
number of sequences in cluster > 750: `mafft`

Sequence alignment was carried out separately for full-length ITS and ITS2 sequences. Sequence ordering in mafft alignment is stored in the database for viewing purposes.

Step 5: blastclust clustering (clustering step 2)

Both full-length ITS and ITS2 sequence clusters from clustering step 1 (UCL clusters) were clustered further using blastclust version 2.2.22 (Altschul et al. 1997) on different similarity tresholds (97-99%) using the following program parameters

```
blastclust -i infile.fasta -S 97 [97.5, 98, 98.5, 99] -L 0.85 -a 8 -e
F -o outfile -p F
```

Step 6: choosing representative sequences for clustering step 2 (SH) clusters

Representative sequences for all SH clusters (full-length ITS and ITS2 region on different similarity tresholds) were calculated using the following procedure

1. consensus sequence for each SH cluster was generated by USEARCH program with the following command

   ```
   usearch -cluster_fast infile.fasta -consout consensus.fasta -id
   0.80
   ```

2. consensus sequence was blasted against all sequences in the same cluster for finding out best match among “true” sequences using megaBLAST version 2.2.23 (Zhang et al., 2000) with the following parameters

   ```
   megablast -W 8 -r 2 -q -3 -G 5 -E 2 -v 1 -b 1 -m 8 -i
   consensus.fasta -d cluster_db
   ```