bio-101

Extraction of Orthologs from Genome-Sequencing Data for Phylogenetic Analysis

Guan Pang[1], Feng M. Cai[2, *]

[1]Jiangsu Provincial Key Lab for Solid Organic Waste Utilization, Jiangsu Collaborative Innovation Center of Solid Organic Wastes, Educational Ministry Engineering Center of Resource-saving Fertilizers, Nanjing Agricultural University, Nanjing, China
[2]State Key Laboratory of Biocontrol, School of Ecology, Sun Yat-sen University, Shenzhen, China
[*]For correspondence: caif8@mail.sysu.edu.cn

Abstract

Homologs, including paralogs and orthologs, are genes that share sequence homologies within or between species. Determination of single-copy orthologs for phylogenomic analysis is the first step in all comparative genomic research. The current protocol provides a detailed bioinformatic pipeline from sequence data acquisition to phylogenetic reconstruction with the use of two commonly adopted tools: OrthoFinder and IQ-TREE. The protocol is demonstrated using genomic data from five fungi, including four *Trichoderma* spp. and an *Escovopsis weberi,* which served as the outgroup in the current case. Additionally, we also demonstrate a partitioned analysis for concatenated multi-locus datasets. The protocol is simple, does not require extensive bioinformatic training or special equipment, and can be easily reproduced for genome-sequencing data from other taxonomic groups.

Keywords: Gene tree, Maximum likelihood phylogeny, Molecular evolution, Orthogroup, Phylogenetics, Species tree, Substitution model

Background

With huge advances both in evolutionary theories and sequencing technologies, phylogenetic analysis is entering a new era—phylogenomics. Current methods for phylogenomic inference can generally be categorized into two types: supertree and supermatrix methods [1–3]. The former approach obtains one *supertree* by combining inferred individual gene trees, each containing information from partially overlapped sets of taxa. Alternatively, the supermatrix approach analyzes the concatenated alignment of individual genes. Unavailable genes/loci are coded as missing data in the supermatrix [2,4]. Likelihood-based reconstruction methods are particularly suited for the analysis of supermatrices. These methods consider the heterogeneity across genes referring to evolutionary rates by using partitioned-likelihood models, which allow each gene to evolve under a different substitution model. According to the total evidence principle of using all the relevant available data, it is somewhat more popular as a strategy to adopt the supermatrix method, which is also used in the current pipeline of demonstrated examples.

The two crucial steps of standard phylogenetic inference are the identification of homologous sequences and tree reconstruction. Therefore, besides the accuracy of the tree-building method, the reliability of a phylogenomic tree also largely depends on the quality of homology, that is, the determination of paralogs and orthologs within and between genomes [5]. In contrast to paralogs, which are derived from gene duplication and should thus be excluded from phylogenetic analyses, orthologs are genes that are derived from speciation events; orthology, in this case, refers to the relationship between the corresponding genes in different species. So far, the most widely used methods for orthology inference can be classified into two groups [6]. One group infers pairwise relationships between genes in two species and then to multiple species, while the other identifies complete orthogroups (OGs), which are identified as the set of genes descended from a single gene in the last common ancestor of all of the species considered [5,7].

In the current pipeline, we use OrthoFinder, a popular method for inferring OGs of protein-coding genes. Starting from gene sequences (the input files), an advantage of using this program is that, by default, it infers OGs, orthologs, the complete set of gene trees for all OGs, the rooted species tree, and all possible gene duplication events. Furthermore, it also provides extensive comparative genomics statistics [5,7]. Despite the fact that OrthoFinder generates individual gene trees and the species tree, for customized tree building we recommend IQ-TREE (IQ-TREE 2 here) for subsequent analyses. In our laboratory, when working with multiple fungal genomes, IQ-TREE runs fast and provides automatic model selection, which also includes data partitioning, an efficient search algorithm for ML trees, ultrafast bootstrapping, and more [8]. With this protocol, we aim to demonstrate effective examples of orthology inference, ortholog extraction, paralog exclusion, individual gene tree reconstruction with the ML method, and data partitioning. This is done using five fungal genomes, with the four main members belonging to the genus *Trichoderma*. *Trichoderma* spp. are among the best studied groups of filamentous fungi due to their high value in applications from agriculture to industrial enzyme production [9]. The present protocol is simple and can also be easily adopted for genomic data from other organisms.

Equipment

We explicitly assume that the user has some basic skills in working in a Linux-based operating system.

1. Linux cluster
   In the present study, we used the AuthenticAMD supercomputer, which has two nodes, each containing 32 cores (model name: AMD EPYC 7452 32-Core Processor) and 256 GB of memory in total
2. Personal computer (PC)
   We recommend using a PC with an Intel Core i7-10510U CPU or higher and at least 16 GB of RAM for sufficient post data processing

Software

1. OrthoFinder ([5,7], v2.5.4, https://github.com/davidemms/OrthoFinder)

2. IQ-TREE (8,10, v2.2.0.3, https://github.com/iqtree/iqtree2)
   Note: The required software and its dependents (including IQ-TREE, which is also integrated in OrthoFinder) should be installed properly according to the tutorials mentioned above before the analysis.

Procedure

The individual steps in this protocol are summarized in **Figure 1**.

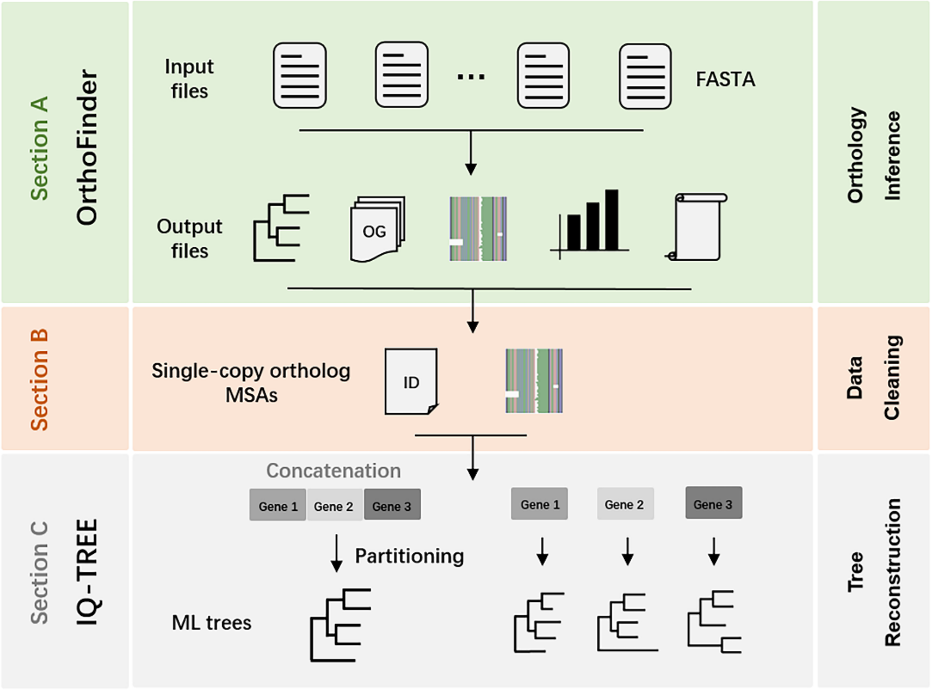


Figure 1. Bioinformatic workflow for orthology inference and maximum likelihood (ML) tree reconstruction based on genome-sequencing data. Three sections are illustrated: (A) orthology inference using OrthoFinder, (B) single-copy ortholog extraction, and (C) ML tree reconstruction (including model test and data partitioning) using IQ-TREE.

A. Orthology inference

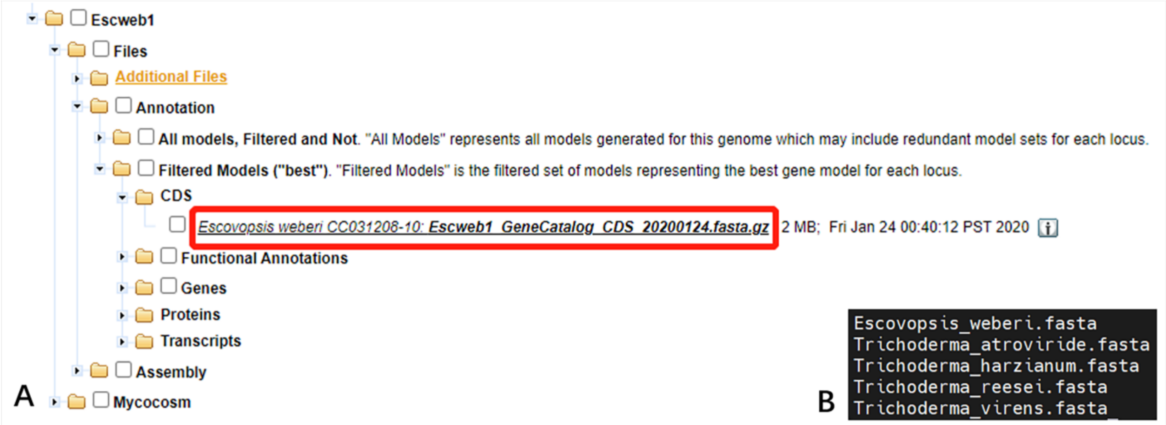
1. Prepare input data
   For given genomes on which annotation has been previously performed, download the coding sequences (CDSs) or protein sequences from the appropriate database. The current protocol uses genome datasets from five fungal strains [9,11,12] deposited in MycoCosm of the DOE Joint Genome Institute (JGI) [13]. The data resource of each fungal strain used in this protocol is shown below, which allows downloading after registration. For customized data, a minimum set of three samples (genomes) is required for such an analysis. We also recommend a minimum sequencing depth of 10× for each sample in order to obtain a sufficient OG set.

   Escovopsis weberi CC031208-10:
   https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Escweb1
   Trichoderma atroviride IMI 206040:
   https://genome.jgi.doe.gov/portal/Triat2/Triat2.download.ftp.html
   Trichoderma harzianum CBS 226.95:

https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Triha1
Trichoderma reesei QM6a:
https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Trire_Chr
Trichoderma virens Gv29-8:
https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=TriviGv29_8_2

An example of the designated sequence files is illustrated with the CDS file from *Escovopsis weberi* (Figure 2A). Once the compressed files of all strains are downloaded in one folder ("test" in this protocol), use the *gunzip* command to obtain the FASTA files for each strain. Rename each FASTA file as shown in Figure 2B.

```
gunzip *.gz
```



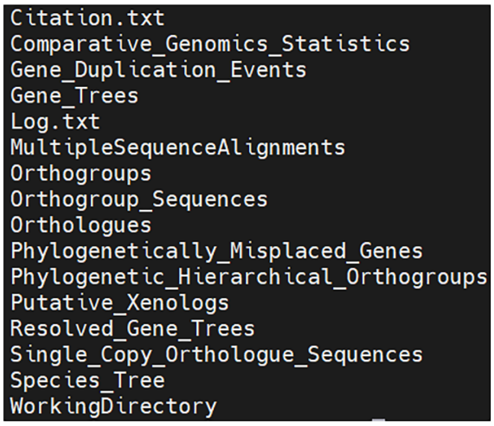
```
□ Escweb1
  □ Files
    Additional Files
    □ Annotation
      □ All models, Filtered and Not. "All Models" represents all models generated for this genome which may include redundant model sets for each locus.
      □ Filtered Models ("best"). "Filtered Models" is the filtered set of models representing the best gene model for each locus.
        CDS
          □ Escovopsis weberi CC031208-10: Escweb1_GeneCatalog_CDS_20200124.fasta.gz 2 MB; Fri Jan 24 00:40:12 PST 2020
        □ Functional Annotations
        □ Genes
        Proteins
        Transcripts
    □ Assembly
  □ Mycocosm
```

```
Escovopsis_weberi.fasta
Trichoderma_atroviride.fasta
Trichoderma_harzianum.fasta
Trichoderma_reesei.fasta
Trichoderma_virens.fasta
```

Figure 2. An overview of the architecture of genomic data deposited in MycoCosm. (A) *Escovopsis weberi* CC031208-10 (the CDS file is highlighted by a red frame). (B) List of renamed FASTA files for the five fungal genome examples in this protocol.

2. Find OGs
In this protocol, we adopted OrthoFinder to infer the orthology of protein-coding genes across multiple species. OrthoFinder does not only infer OGs; it also automatically produces several output files, such as individual gene trees, a rooted species tree, or gene duplication event results (see more in Emms and Kelly [7]). A full picture of the standard outputs can be found in Figure 3. A detailed guided tour of the other result files can be found at the following link: https://github.com/davidemms/OrthoFinder#orthofinder-results-files. In the current pipeline, we focus on the identification of the orthologous genes from a certain group of fungal species. To this end, use the following command:

```
orthofinder -f test/ -d -M msa -t 40 -a 40
```



```
Citation.txt
Comparative_Genomics_Statistics
Gene_Duplication_Events
Gene_Trees
Log.txt
MultipleSequenceAlignments
Orthogroups
Orthogroup_Sequences
Orthologues
Phylogenetically_Misplaced_Genes
Phylogenetic_Hierarchical_Orthogroups
Putative_Xenologs
Resolved_Gene_Trees
Single_Copy_Orthologue_Sequences
Species_Tree
WorkingDirectory
```

Figure 3. List of typical output files generated by running the command shown above via OrthoFinder

Note: -f is an option to specify the directory containing the working FASTA files. -d specifies the input as DNA sequences [Default = protein, p]. OrthoFinder uses MAFFT for the alignment and FastTree for the tree inference by default. In order to obtain a more accurate result, such as inference of ML trees from multiple sequence alignments (MSAs), it is also possible to use an alternative alignment or tree inference program in OrthoFinder, which usually requires a more computationally costly resource. For example, to call MUSCLE and IQ-TREE, the commands to add are -M msa -A muscle -T iqtree. The option -M msa in this case allows the inference of the species tree from a concatenated MSA of single-copy orthologous genes. However, in the case of a first analysis, FastTree is highly recommended.

Optionally, while OrthoFinder is designed to require minimal computation, it can be tailed by using the -t and -a options to suit the computational and data sources. -t <int.> indicates the number of threads for sequence search, MSA, and tree inference [Default is the number of cores on the machine], and -a <int.> indicates the number of parallel analysis threads for internal RAM-intensive tasks [Default = 1]. Note that an extra step of adding a [Speciesidentifier] at the head of each FASTA sequence is required if partitioning tree reconstruction is subsequently needed for the phylogenetic analysis (Section C). To do this, write the following command before running OrthoFinder.

```
For file in *.fasta; do sed -i 's/>/>Speciesidentifier_/g' $file; done
```

B. Single-copy orthologous gene extraction

Orthologs are genes that derive from speciation events, in contrast to paralogs, which derive from gene duplication and should thus be excluded from phylogenetic analyses [6], as mentioned above. In the current pipeline, to obtain the orthologous gene sequences from multiple genomes for subsequent analyses such as gene tree reconstruction (Section C), only the results exported in the folders of [Single_Copy_Orthologue_Sequences] and [MultipleSequenceAlignments] are needed.

1. Extract the IDs of single-copy OGs
   The [Single_Copy_Orthologue_Sequences] folder contains FASTA files of each OG. These OGs ideally contain one gene per species at most, from which paralogs have been excluded. Here is an example of extracting the IDs of single-copy OGs from the designated FASTA files to a txt file as an output (namely [singlecopyOG_idlist.txt].

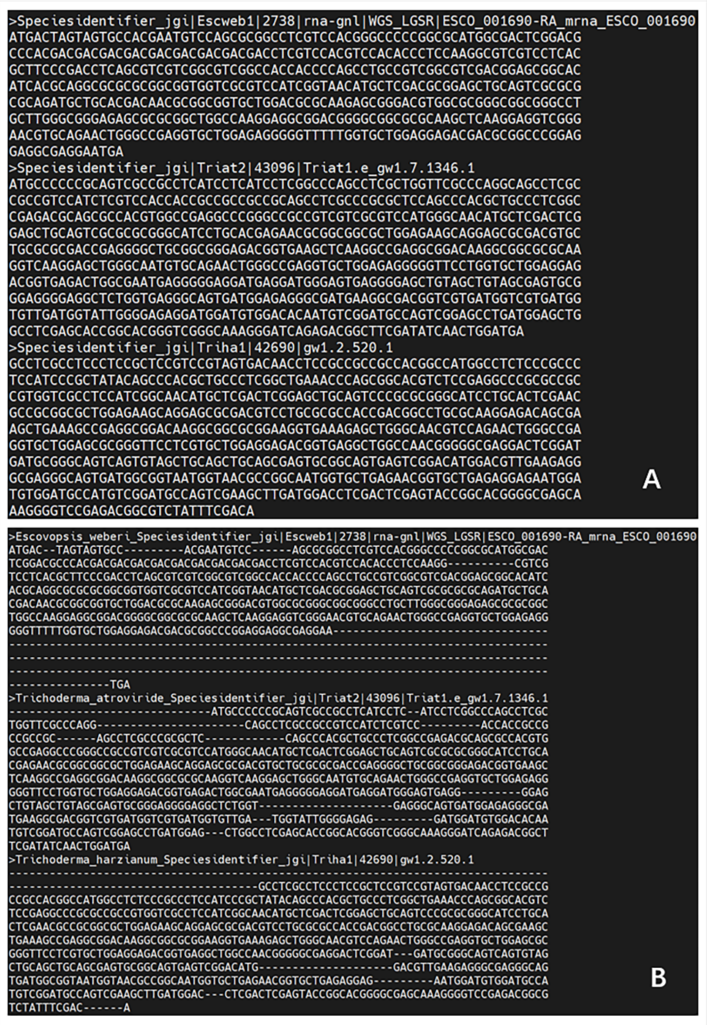
```
ls -l Single_Copy_Orthologue_Sequences | awk '{print $9}' | awk NF > singlecopyOG_idlist.txt
```

2. Retrieve multiple sequence alignments of single-copy OGs
   In the current protocol, we take advantage of ready MSAs of each OG generated in Section A. Thus, with the ID list of single-copy OGs obtained in the last step, the aligned and trimmed FASTA files of each single-copy OG can be easily retrieved from the results sorted in the folder of [MultipleSequenceAlignments]. To do so, use the following command:

```
mkdir singlecopyOG_seqs
```

```
length=`awk 'END{print NR}' ./singlecopyOG_idlist.txt`; for x in $(seq 1 $length); do y=`awk "{if (NR
== $x) print }" ./singlecopyOG_idlist.txt`; cp ./MultipleSequenceAlignments/$y* ./singlecopyOG_seqs;
done
```

   Note: An example (OG0000819) comparing the data in the folder of [Single_Copy_Orthologue_Sequences] and the output files (see [singlecopyOG_seqs] generated in this step) is given in Figure 4. Importantly, although the majority of species tree inferences are recommended to be restricted to one-to-one orthologous sequences that are present in all species in the analysis, realistically, such groups of sequences are rare in real biological datasets, especially for distant taxa. Thus, such cases are only available if gene duplication or loss has not occurred during the divergence of that gene family [7]. To address this challenge, the updated OrthoFinder2 has been designed to leverage the sequence data from all genes via a new algorithm: Species Tree from All Genes (STAG). STAG was developed to allow species tree inference for sample sets with few or no complete sets of one-to-one orthologs present in all species of interest using the most closely related genes within single-copy or multi-copy OGs [7].



```
>Speciesidentifier_jgi|Escweb1|2738|rna-gnl|WGS_LGSR|ESCO_001690-RA_mrna_ESCO_001690
ATGACTAGTAGTGCCACGAATGTCCAGCGCGGCCTCGTCCACGGGCCCCCGGCGCATGGCGACTCGGACG
CCCACGACGACGACGACGACGACGACGACGACCTCGTCCACGTCCACACCCTCCAAGGCGTCGTCCTCAC
GCTTCCCGACCTCAGCGTCGTCGGCGTCGGCCACCACCCCAGCCTGCCGTCGGCGTCGACGGAGCGGCAC
ATCACGCAGGCGCGCGCGGCGGTGGTCGCGTCCATCGGTAACATGCTCGACGCGGAGCTGCAGTCGCGCG
CGCAGATGCTGCACGACAACGCGGCGGTGCTGGACGCGCAAGAGCGGGACGTGGCGCGGGCGGCGGGCCT
GCTTGGGCGGGAGAGCGCGCGGCTGGCCAAGGAGGCGGACGGGGCGGCGCGCAAGCTCAAGGAGGTCGGG
AACGTGCAGAACTGGGCCGAGGTGCTGGAGAGGGGGTTTTTGGTGCTGGAGGAGACGACGCGGCCCGGAG
GAGGCGAGGAATGA
>Speciesidentifier_jgi|Triat2|43096|Triat1.e_gw1.7.1346.1
ATGCCCCCGCAGTCGCCGCCTCATCCTCATCCTCGGCCCAGCCTCGCTGGTTCGCCCAGGCAGCCTCGC
CGCCGTCCATCTCGTCCACCACCGCCGCCGCCGCAGCCTCGCCCGCGCTCCAGCCCACGCTGCCCTCGGC
CGAGACGCAGCGCCACGTGGCCGAGGCCCGGGCCGCCGTCGTCGCGTCCATGGGCAACATGCTCGACTCG
GAGCTGCAGTCGCGCGCGGGCATCCTGCACGAGAACGCGGCGGCGCTGGAGAAGCAGGAGCGCGACGTGC
TGCGCGCGACCGAGGGGCTGCGGCGGGAGACGGTGAAGCTCAAGGCCGAGGCGGACAAGGCGGCGCGCAA
GGTCAAGGAGCTGGGCAATGTGCAGAACTGGGCCGAGGTGCTGGAGAGGGGGTTCCTGGTGCTGGAGGAG
ACGGTGAGACTGGCGAATGAGGGGGAGGATGAGGATGGGAGTGAGGGGAGCTGTAGCTGTAGCGAGTGCG
GGAGGGGAGGCTCTGGTGAGGGCAGTGATGGAGAGGGCGATGAAGGCGACGGTCGTGATGGTCGTGATGG
TGTTGATGGTATTGGGGAGAGGATGGATGTGGACACAATGTCGGATGCCAGTCGGAGCCTGATGGAGCTG
GCCTCGAGCACCGGCACGGGTCGGGCAAAGGGATCAGAGACGGCTTCGATATCAACTGGATGA
>Speciesidentifier_jgi|Triha1|42690|gw1.2.520.1
GCCTCGCCTCCCTCCGCTCCGTCCGTAGTGACAACCTCCGCCGCCGCCACGGCCATGGCCTCTCCCGCCC
TCCATCCCGCTATACAGCCCACGCTGCCCTCGGCTGAAACCCAGCGGCACGTCTCCGAGGCCCGCGCCGC
CGTGGTCGCCTCCATCGGCAACATGCTCGACTCGGAGCTGCAGTCCCGCGCGGGCATCCTGCACTCGAAC
GCCGCGGCGCTGGAGAAGCAGGAGCGCGACGTCCTGCGCGCCACCGACGGCCTGCGCAAGGAGACAGCGA
AGCTGAAAGCCGAGGCGGACAAGGCGGCGCGGAAGGTGAAAGAGCTGGGCAACGTCCAGAACTGGGCCGA
GGTGCTGGAGCGCGGGTTCCTCGTGCTGGAGGAGACGGTGAGGCTGGCCAACGGGGGCGAGGACTCGGAT
GATGCGGGCAGTCAGTGTAGCTGCAGCTGCAGCGAGTGCGGCAGTGAGTCGGACATGGACGTTGAAGAGG
GCGAGGGCAGTGATGGCGGTAATGGTAACGCCGGCAATGGTGCTGAGAACGGTGCTGAGAGGAGAATGGA
TGTGGATGCCATGTCGGATGCCAGTCGAAGCTTGATGGACCTCGACTCGAGTACCGGCACGGGGCGAGCA
AAGGGGTCCGAGACGGCGTCTATTTCGACA
```

```
>Escovopsis_weberi_Speciesidentifier_jgi|Escweb1|2738|rna-gnl|WGS_LGSR|ESCO_001690-RA_mrna_ESCO_001690
ATGAC--TAGTAGTGCC---------ACGAATGTCC------AGCGCGGCCTCGTCCACGGGCCCCCGGCGCATGGCGAC
TCGGACGCCCACGACGACGACGACGACGACGACGACGACCTCGTCCACGTCCACACCCTCCAAGG----------CGTCG
TCCTCACGCTTCCCGACCTCAGCGTCGTCGGCGTCGGCCACCACCCCAGCCTGCCGTCGGCGTCGACGGAGCGGCACATC
ACGCAGGCGCGCGCGGCGGTGGTCGCGTCCATCGGTAACATGCTCGACGCGGAGCTGCAGTCGCGCGCGCAGATGCTGCA
CGACAACGCGGCGGTGCTGGACGCGCAAGAGCGGGACGTGGCGCGGGCGGCGGGCCTGCTTGGGCGGGAGAGCGCGCGGC
TGGCCAAGGAGGCGGACGGGGCGGCGCGCAAGCTCAAGGAGGTCGGGAACGTGCAGAACTGGGCCGAGGTGCTGGAGAGG
GGGTTTTTGGTGCTGGAGGAGACGACGCGGCCCGGAGGAGGCGAGGAA--------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
---------------TGA
>Trichoderma_atroviride_Speciesidentifier_jgi|Triat2|43096|Triat1.e_gw1.7.1346.1
-------------------------------ATGCCCCCGCAGTCGCCGCCTCATCCTC--ATCCTCGGCCCAGCCTCGC
TGGTTCGCCCAGG----------------------CAGCCTCGCCGCCGTCCATCTCGTCC---------ACCACCGCCG
CCGCCGC------AGCCTCGCCCGCGCTC------------CAGCCCACGCTGCCCTCGGCCGAGACGCAGCGCCACGTG
GCCGAGGCCCGGGCCGCCGTCGTCGCGTCCATGGGCAACATGCTCGACTCGGAGCTGCAGTCGCGCGCGGGCATCCTGCA
CGAGAACGCGGCGGCGCTGGAGAAGCAGGAGCGCGACGTGCTGCGCGCGACCGAGGGGCTGCGGCGGGAGACGGTGAAGC
TCAAGGCCGAGGCGGACAAGGCGGCGCGCAAGGTCAAGGAGCTGGGCAATGTGCAGAACTGGGCCGAGGTGCTGGAGAGG
GGGTTCCTGGTGCTGGAGGAGACGGTGAGACTGGCGAATGAGGGGGAGGATGAGGATGGGAGTGAGG---------GGAG
CTGTAGCTGTAGCGAGTGCGGGAGGGGAGGCTCTGGT--------------------GAGGGCAGTGATGGAGAGGGCGA
TGAAGGCGACGGTCGTGATGGTCGTGATGGTGTTGA---TGGTATTGGGGAGAG---------GATGGATGTGGACACAA
TGTCGGATGCCAGTCGGAGCCTGATGGAG---CTGGCCTCGAGCACCGGCACGGGTCGGGCAAAGGGATCAGAGACGGCT
TCGATATCAACTGGATGA
>Trichoderma_harzianum_Speciesidentifier_jgi|Triha1|42690|gw1.2.520.1
--------------------------------------------------------------------------------
-------------------------------------GCCTCGCCTCCCTCCGCTCCGTCCGTAGTGACAACCTCCGCCG
CCGCCACGGCCATGGCCTCTCCCGCCCTCCATCCCGCTATACAGCCCACGCTGCCCTCGGCTGAAACCCAGCGGCACGTC
TCCGAGGCCCGCGCCGCCGTGGTCGCCTCCATCGGCAACATGCTCGACTCGGAGCTGCAGTCCCGCGCGGGCATCCTGCA
CTCGAACGCCGCGGCGCTGGAGAAGCAGGAGCGCGACGTCCTGCGCGCCACCGACGGCCTGCGCAAGGAGACAGCGAAGC
TGAAAGCCGAGGCGGACAAGGCGGCGCGGAAGGTGAAAGAGCTGGGCAACGTCCAGAACTGGGCCGAGGTGCTGGAGCGC
GGGTTCCTCGTGCTGGAGGAGACGGTGAGGCTGGCCAACGGGGGCGAGGACTCGGAT---GATGCGGGCAGTCAGTGTAG
CTGCAGCTGCAGCGAGTGCGGCAGTGAGTCGGACATG--------------------GACGTTGAAGAGGGCGAGGGCAG
TGATGGCGGTAATGGTAACGCCGGCAATGGTGCTGAGAACGGTGCTGAGAGGAG---------AATGGATGTGGATGCCA
TGTCGGATGCCAGTCGAAGCTTGATGGAC---CTCGACTCGAGTACCGGCACGGGGCGAGCAAAGGGGTCCGAGACGGCG
TCTATTTCGAC------A
```

Figure 4. Example of sequence data (OG0000819). Sequence data before (from the folder [Single_Copy_Orthologue_Sequences] (A) and after alignment ([singlecopyOG_seqs] (B).

C. Reconstruction of individual gene trees with ML method

If a single gene tree for some specific gene(s) of interest is required, ML tree reconstruction via IQ-TREE is recommended in this protocol. An example for using the single-copy OG MSAs (OG0000819) obtained in Section B is given below.

1. Select substitution model and start ML tree reconstruction
   Phylogenetic tree reconstruction methods such as the ML method usually start with model selection, in which the program searches for the best-fit model of sequence evolution of the available data [14]. IQ-TREE has been developed to support a wide range of substitution models for DNA, protein, codon, and also morphological data by integrating ModelFinder in it [8]. Start the tree reconstruction by entering:

   ```
   iqtree2 -s OG0000819.fa -B 1000
   ```

   With this command, IQ-TREE calls ModelFinder by default. If only the best-fit model is required without doing the tree reconstruction, then run:

```
iqtree2 -s OG0000819.fa -m MF
```

For multiple MSAs, such as all of the single-copy OG MSAs obtained in [singlecopyOG_seqs] of Section B, write a loop as follows:

```
for file in *.fa; do iqtree2 -s $file -B 1000; done
```

Note: -s is used to specify the name of the MSA file. IQ-TREE also supports other input file formats such as NEXUS, CLUSTALW, and PHYLIP. -m is used to specify the name of the model if known beforehand (for example, -m TIM+F). -B is used to specify the number of bootstrap replicates, and 1000 is recommended. IQ-TREE writes several output files, such as .iqtree, .treefile, and .log files (see more at http://www.iqtree.org/doc/). In the current protocol, the .treefile is the designated file for tree visualization. The .log file records the entire run, including the best-fit model used in the tree reconstruction. ModelFinder integrated in IQ-TREE computes the loglikelihoods of an initial parsimony tree for different models and automatically chooses the model that minimizes the Bayesian Information Criterion (BIC) score if no customized command is given. Otherwise, -AIC or -AICc can be used to change model selection according to the Akaike Information Criterion (AIC) or the corrected Akaike Information Criterion (AICc), respectively. Note that IQ-TREE prevents loss of data by overwriting [8]. If a rerun is needed, add the -redo option at the end of the command line.

2. Run partitioned analysis for multi-locus alignments
IQ-TREE also allows combining sub-alignments from different MSAs (e.g., the MSAs in the folder of [singlecopyOG_seqs] generated in Section B). Here, an example is given below using three of the single-copy OG MSAs prepared in a NEXUS input file (i.e., threeOG.nex). As every gene within each OG has its own unique ID, it is essential to rename all of the gene IDs for each taxon (refer to the [Speciesidentifier]).

```
for file in singlecopyOG_seqs/*.fa; do awk -F "_Speciesidentifier" '{print $1}' ${file##/*} > ${file##/*}.renamed; done
```

```
iqtree2 -s threeOG.nex -p threeOG.nex -B 1000
```

Note: -p allows each partition to have its own evolution rate. Note that with the use of this command, ModelFinder implements a greedy strategy that starts with the full partition model and then merges every two loci until the model fit does not increase further [15]. This causes considerable computational burden. Therefore, partitioning tree reconstruction should be used as a precaution for large datasets. For preparing concatenated multi-locus datasets in a NEXUS file, programs such as PhyloSuite [16], SequenceMatrix [17], or some other software providing similar functions are recommended.

Data analysis

The resulting ML tree (in NEWICK format) from IQ-TREE can be visualized by any supported tree viewers on a PC, such as Figtree (v1.4.4, https://github.com/rambaut/figtree/releases) and iTOL (https://itol.embl.de/). **Figure 5** shows an example of an annotated tree. The results obtained using the current protocol showed a similarly robust phylogenomic relationship for the fungi tested as the ones previously published [9,12].

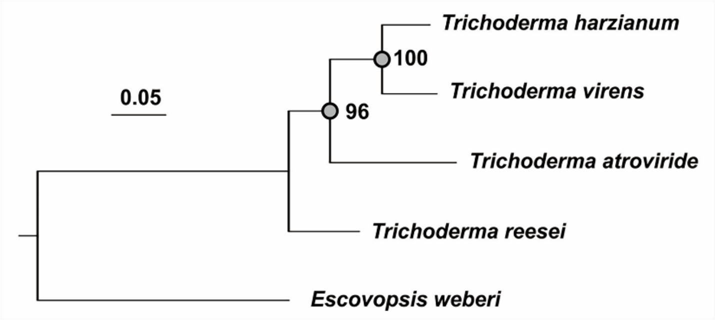


Figure 5. Maximum likelihood (ML) phylogenetic tree constructed based on the concatenated sequence matrix of three randomly selected single-copy orthologs in [singlecopyOG_seqs]. Individual models TN+F+G4, TIM3+F+G4, and TPM3u+F+G4 were selected according to BIC for the three orthologs OG0000545, OG0002657, and OG0002676, respectively. The IQ-TREE ultrafast bootstrap values are represented by the nodes (ultrafast bootstrap, N = 1000).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (32102473) and the Fundamental Research Funds for the Central Universities (KYQN2022046).

Competing interests

The authors declare no competing financial interest.

References

1. Bininda-Emonds, O. R. P., Gittleman, J. L. and Steel, M. A. (2002). The (Super)Tree of Life: Procedures, Problems, and Prospects. *Annu Rev Ecol Syst.* 33(1): 265–289. https://doi.org/10.1146/annurev.ecolsys.33.010802.150511
2. Delsuc, F., Brinkmann, H. and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.* 6(5): 361–375. https://doi.org/10.1038/nrg1603
3. Sanderson, M. J., McMahon, M. M. and Steel, M. (2011). Terraces in Phylogenetic Tree Space. *Science* (1979) 333(6041): 448–450. https://doi.org/10.1126/science.1206357
4. Chernomor, O., von Haeseler, A. and Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst Biol.* 65(6): 997–1008. https://doi.org/10.1093/sysbio/syw037
5. Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16(1): e1186/s13059–015–0721–2. https://doi.org/10.1186/s13059-015-0721-2
6. Debray, K., Marie-Magdelaine, J., Ruttink, T., Clotault, J., Foucher, F. and Malécot, V. (2019). Identification and assessment of variable single-copy orthologous (SCO) nuclear loci for low-level phylogenomics: a case study in the genus Rosa (Rosaceae). *BMC Evol Biol.* 19(1): e1186/s12862–019–1479–z. https://doi.org/10.1186/s12862-019-1479-z
7. Emms, D. M. and Kelly, S. (2018). OrthoFinder: phylogenetic orthology inference for comparative genomics: *Genome Biol.* 238 https://doi.org/10.1101/466201
8. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. and Minh, B. Q. (2014). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol.* 32(1): 268–274.

https://doi.org/10.1093/molbev/msu300

9. Kubicek, C. P., Steindorff, A. S., Chenthamara, K., Manganiello, G., Henrissat, B., Zhang, J., Cai, F., Kopchinskiy, A. G., Kubicek, E. M., Kuo, A., et al. (2019). Evolution and comparative genomics of the most common Trichoderma species. *BMC Genomics.* 20(1): 485. https://doi.org/10.1186/s12864-019-5680-7
10. Minh, B. Q., Schmidt, H., Chernomor, O., Schrempf, D., Woodhams, M., von Haeseler, A. and Lanfear, R. (2019). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol *Biol Evol.* 37(5):1530–1534. https://doi.org/10.1101/849372
11. de Man, T. J. B., Stajich, J. E., Kubicek, C. P., Teiling, C., Chenthamara, K., Atanasova, L., Druzhinina, I. S., Levenkova, N., Birnbaum, S. S. L., Barribeau, S. M., et al. (2016). Small genome of the fungus *Escovopsis weberi*, a specialized disease agent of ant agriculture. *Proc Natl Acad Sci U S A.* 113(13): 3567–3572. https://doi.org/10.1073/pnas.1518501113
12. Druzhinina, I. S., Chenthamara, K., Zhang, J., Atanasova, L., Yang, D., Miao, Y., Rahimi, M. J., Grujic, M., Cai, F., Pourmehdi, S., et al. (2018). Massive lateral transfer of genes encoding plant cell wall-degrading enzymes to the mycoparasitic fungus Trichoderma from its plant-associated hosts. *PLos Genet.* 14(4): e1007322. https://doi.org/10.1371/journal.pgen.1007322
13. Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., et al. (2013). MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Re.s* 42: D699–D704. https://doi.org/10.1093/nar/gkt1183
14. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.* 14(6): 587–589. https://doi.org/10.1038/nmeth.4285
15. Lanfear, R., Calcott, B., Ho, S. Y. W. and Guindon, S. (2012). PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Mol Biol Evol.* 29(6): 1695–1701. https://doi.org/10.1093/molbev/mss020
16. Zhang, D., Gao, F., Li, W. X., Jakovlić, I., Zou, H., Zhang, J. and Wang, G. T. (2018). PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol Ecol Resour.* 20(1):348–355. https://doi.org/10.1101/489088
17. Vaidya, G., Lohman, D. J. and Meier, R. (2011). SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics.* 27(2): 171–180. https://doi.org/10.1111/j.1096-0031.2010.00329.x