**Article - Method**

**Title**
SHOOT: phylogenetic gene search and ortholog inference

**Authors**
Emms, D.M.¹ and Kelly, S.¹*

**Affiliations**
1) Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK.

**Corresponding Author**
Name: David Emms

Email: david.emms@plants.ox.ac.uk

Name: Steven Kelly

Email: steven.kelly@plants.ox.ac.uk

**Keywords**
Phylogenetic tree inference; sequence similarity search; orthology inference

**Abstract**
Determining the evolutionary relationships between gene sequences is fundamental to comparative biological research. However, conducting such analyses requires a high degree of technical proficiency in several computational tools including gene family construction, multiple sequence alignment, and phylogenetic inference. Here we present SHOOT, an easy to use phylogenetic search engine for fast and accurate phylogenetic analysis of biological sequences. SHOOT searches a user-provided query sequence against a database of phylogenetic trees of gene sequences (gene trees) and returns a gene tree with the given query sequence correctly grafted within it. We show that SHOOT can perform this search and placement with comparable speed to a conventional BLAST search. We demonstrate that SHOOT phylogenetic placements are as accurate as conventional multiple sequence alignment and maximum likelihood tree inference approaches. We further show that SHOOT can be used to identify orthologs with equivalent accuracy to conventional orthology inference methods. In summary, SHOOT is an accurate and fast tool for complete phylogenetic analysis of novel query sequences. An easy to use webserver is available online at www.shoot.bio.

**Introduction**
Resolving the phylogenetic relationships between biological sequences provides a framework for inferring sequence function, and a basis for understanding the diversity and evolution of life on Earth. The entry point to such phylogenetic analyses is provided by algorithms that either align or identify regions of local similarity between pairs of biological
sequences. The first implementations of such algorithms utilised global alignments to provide a basis to score similarity between sequences (Needleman and Wunsch 1970). Later, faster local alignment methods were developed (Smith and Waterman 1981), followed by the FASTA heuristic database search (Lipman and Pearson 1985) and culminating with the development of the BLAST algorithm and statistical methods for homology testing (Altschul, et al. 1990) in the 1990s. Since then, BLAST and other local alignment methods (Edgar 2010; Mirdita, et al. 2019; Buchfink, et al. 2021) have provided a critical foundation of biological science research and form the entry point to the majority of biological sequence analyses.

An under-utilised resource in BLAST and related local alignment search tools is the transitive nature of homology. Because local alignment searching methods do not store the relationships between sequences, a search of a query gene against a large database will involve carrying out many needless pairwise local alignments against numerous closely related homologs. An alternative approach would be to infer the relationships between all database sequences ahead of time using phylogenetic inference methods. These phylogenetic relationships can then be stored as part of the database facilitating the use of lighter-weight search approaches or sparse reference databases with relationships already computed. Existing methods that take these kind of approaches include TreeFam for genes within the Metazoa (Schreiber, et al. 2014) and TreeGrafter for annotating protein sequences using annotated phylogenetic trees (Tang, et al. 2019).

There are a number of advantages to taking a phylogenetic approach to sequence searching. Firstly, a gene can be rapidly assigned to its homology group, irrespective of the number of homologous genes. Secondly, false negatives are unlikely since complete homology groups can be identified securely ahead of time. This helps avoid the reduced sensitivity that results from local sequence similarity database search algorithm heuristics used to determine which sequences to consider aligning. Thirdly, phylogenetic inference methods can be used to rapidly and accurately assign the gene to its correct position within the otherwise pre-computed gene tree for its homology group. This avoids the need to evaluate gene-relatedness using e-values, which are a measure of the certainty that a pair of genes are homologous, rather than a direct evaluation of the phylogenetic relationship between genes.

Although local similarity searches such as BLAST are the primary entry point to the sequence analysis, a frequent end-goal of such analyses is to identify orthologs of the query sequence in other species. The use of phylogenetic methods is the canonical method for assessing gene relationships. Phylogenetic methods for estimating sequence similarity are more accurate that using local pairwise alignments, and critically they provide contextual information about the place of the query gene within its gene family. This includes the identification of orthologs, paralogs, and gene gain and loss within each clade in of the resultant phylogenetic tree of gene sequences. Although the similarity scores returned by local alignment methods can be used to approximate phylogenetic trees, they are not accurate and can be limited by only having alignments against a single query gene rather than alignments between sequences already in the database (Camacho, et al. 2009). Moreover, even when all pairwise similarity scores are calculated the accuracy of phylogenetic trees inferred from these scores is limited (Kelly and Maini 2013)
Here we present SHOOT, a software tool for rapidly searching a phylogenetically partitioned and structured database of biological sequences. SHOOT efficiently and accurately places query sequences directly into phylogenetic trees. In this way the phylogenetic history of the query sequence and its orthologs can be immediately visualised, interpreted, and retrieved. SHOOT is provided for use at www.shoot.bio.

Results

Pre-computed databases of phylogenetic trees allow ultra-fast phylogenetic orthology analysis of novel gene sequences

The conventional procedure for sequence orthology analysis is to first assemble a group of gene sequences which share similarity and then perform phylogenetic tree inference on this group to infer the relationships between those genes. The SHOOT algorithm was designed to make such a phylogenetic analysis feasible as a real-time search using a two-stage approach. The first stage comprises the ahead-of-time construction of a SHOOT phylogenetic database and the second stage implements the SHOOT search for a query sequence (Figure 1). The database preparation phase includes multiple automated steps including homology group inference, multiple sequence alignment, phylogenetic tree inference, and homology group profiling (see Methods). Thus, prior to database searching the phylogenetic relationships between all genes in the database are already established. Subsequent SHOOT searches exploit the fact that the alignments and trees have already been computed to enable the use of accurate phylogenetic methods for placement of query genes within pre-computed gene trees with little extra computation required.

The median time for a complete SHOOT search of a database containing 984,137 protein sequences from 78 species was 5.5 seconds using 16 cores of an Intel Xeon E5-2683 CPU for (Figure 2). This compared with 1.19 seconds for a conventional BLAST search of the same sequence set (Figure 2). However, unlike BLAST (or similar) sequence search methods, the output of a SHOOT search is not an ordered list of similar sequences but is instead a maximum likelihood phylogenetic tree with bootstrap support values inferred from a multiple sequence alignment with the query gene embedded within it. SHOOT also computes the orthologs of the query gene using phylogenetic methods.

SHOOT is more accurate than BLAST in identifying the closest related gene sequence

A leave-one-out analysis was conducted to test SHOOT’s ability to find the most closely related gene sequence in a given database. Here a set of 1000 test cases was randomly sampled from the UniProt Reference Proteomes database. Each test case consisted of a pair of genes sister to each other with 100% bootstrap support in a maximum likelihood gene tree. One member of the test pair was arbitrarily designated the “query sequence” and the other gene was designated “the expected closest gene” i.e. the gene that should be identified by a search method as the most similar gene in the database. To provide a comparison, BLAST (Camacho, et al. 2009) was also tested on the same dataset. The set of query genes were searched against the database and each method was scored on whether or not the closest/best scoring gene in each search result was “the expected closest gene”. If either method was not correct, they were secondarily scored on whether the position of retrieved top hit was within two branches of the correct position in the original reference phylogenetic tree. The tests showed that SHOOT identified “the expected closest
gene” as the most closely related gene in 94.2% of cases and was within at most 2 branches from the correct position in 96.5% of cases (Figure 4A). For comparison, BLAST correctly identified the “the expected closest gene” as the most similar gene sequence in 88.4% of cases and its best hits were at most 2 branches from “the expected closest gene” in 94.2% of cases. Thus, SHOOT is better able to identify the closest related gene to a given query gene in a given database and can be used as an alternative to BLAST for this purpose.

**SHOOT has high accuracy in identifying orthologs of the query gene**

A frequent goal of sequence similarity searches is to identify orthologs of the query gene in other species. As stated above, local similarity search tools such as BLAST do not do this. Instead, they return a list of genes that should be subject to multiple sequence alignment and phylogenetic inference in order to infer the orthology relationships between genes. The phylogenetic tree returned by SHOOT provides the evolutionary relationships between genes inferred from multiple sequence alignment and maximum likelihood tree inference allowing orthologs and paralogs to be identified. SHOOT also automatically identifies orthologs and colours the genes in the tree according to whether they are orthologs or paralogs (Figure 3), as identified using the species overlap method (Huerta-Cepas, et al. 2008; Mi, et al. 2010), which has been shown to be an accurate method for automated orthology inference (Altenhoff, et al. 2020). The tree viewer also supports a zoom functionality to view a progressively larger or smaller clade of genes around the query gene. An image of the tree can be downloaded, the tree can also be exported in Newick format, and the FASTA file of protein sequences in the tree can be downloaded to support further downstream analyses.

The leave-one-out analysis for phylogenetic placement was used to test the accuracy of SHOOT orthology inference. The list of orthologs from the conventional phylogenetic approach and the SHOOT approach were compared. This revealed that orthologs inferred using SHOOT are as accurate as those inferred using conventional approaches, agreeing with those from the conventional phylogenetic approach with a precision of 99.2% and a recall of 98.3% (Figure 4B).

**Curated databases place the gene in the context of model species and key events in the gene’s evolution**

The initial release of SHOOT includes phylogenetic databases for Metazoa, Fungi, Plants, Bacteria & Archaea, and also the UniProt Quest for Orthologs (QfO) reference proteomes, which cover all domains of cellular life (Supplementary Tables 1-5). To maximise the utility of the gene trees to a wide range of researchers, the species within the databases have been chosen to contain model species, species of economic or scientific importance, and species selected because of their key location within the evolutionary history covered by the database. Each database also contains multiple outgroup species to allow robust rooting of the set of gene trees. As an example, Figure 5 shows the phylogeny for the initial release version of the metazoan database, highlighting the taxonomic groups of the included species. Although a number of databases are provided on the SHOOT webserver, the SHOOT command line tool has been designed so that databases can be compiled from any species set.
Discussion

SHOOT is a phylogenetic search engine for analysis of biological sequences. It has been designed to take a user-provided query sequence and return a phylogenetic orthology analysis of that sequence using a database of reference organisms. We show that SHOOT can perform this search and analysis with comparable speed to a typical sequence similarity search and thus SHOOT is provided as a phylogenetically informative alternative to BLAST, and as a general-purpose sequence search algorithm for analysis and retrieval of related biological sequences.

Local similarity or profile-based search methods such as BLAST (Camacho, et al. 2009), DIAMOND (Buchfink, et al. 2021) or MMseqs (Steinegger and Soding 2017) have a wide range of uses across the biological and biomedical sciences. The near-ubiquitous utility of these methods has led to them being referred to as the Google of biological research. However, one of the most frequent use cases of these searches is to identify orthologs of a given query sequence. Due to the frequent occurrence of gene duplication and loss, orthologs are often indistinguishable from paralogs in the results of local similarity searches. This is because a given query sequence can have none, one, or many orthologs in a related species. Accordingly, the sequences identified by local similarity searching methods will be an unknown mixture of orthologs and paralogs (Dalquen and Dessimoz 2013). The problem of distinguishing orthologs from paralogs can be partially mitigated by a reciprocal best hit search, but with low sensitivity (Dalquen and Dessimoz 2013). Phylogenetic methods are required to correctly distinguish orthologs from paralogs as they are readily able to distinguish sequence similarity (branch length) and evolutionary relationships (the topology of the tree). SHOOT was designed to provide the accuracy and information of a phylogenetic analysis with the speed and simplicity of a local sequence similarity search. By pre-computing the within-database sequence relationships, SHOOT can perform an individual search in a comparable time to BLAST, but instead of a list SHOOT provides a full maximum-likelihood phylogenetic tree as a result enabling immediate phylogenetic interrogation of the sequence search results.

A standard phylogenetic approach to identifying orthologs of a query gene is to begin a local sequence similarity search or profile search (HMMER (Eddy 2011), MMseqs (Steinegger and Soding 2017)). Frequently, an e-value cut-off is applied to identify a set of similar sequences for subsequent phylogenetic analysis. Because e-values (and their constituent bit-scores) are imperfectly correlated with evolutionary relatedness, the set of similar sequences meeting the search threshold will often be missing some genes as well as often including genes that should not be present. A systematic study using HMMER found that for all n genes from an orthogroup clade to pass an e-value threshold, on average the threshold would have to be set such that 1.8n genes in total met the threshold (Emms and Kelly 2020). i.e. an additional 80% of genes needed to be included, on average, to ensure the orthogroup was complete (Emms and Kelly 2020). Thus, unless a very lenient search is used, genes will be incorrectly absent from the final tree. This can lead to incorrect rooting and subsequent mis-interpretation even by phylogenetic experts (Emms and Kelly 2020). Thus, even for bespoke phylogenetic analyses, it is better to use phylogenetic methods to first select the clade of genes of interest. SHOOT supports this by inferring the tree for the entire family of detectable homologs. The use of trees for complete sets of homologs, together
with OrthoFinder’s robust rooting algorithm (Emms and Kelly 2019), avoids the problem of mis-rooting and misinterpretation of a tree inferred for a more limited set of genes. Also, by using OrthoFinder clustering approach (Emms and Kelly 2015, 2019), hits missed for a single sequence are also corrected by multiple hits identified for its homologs. This “phylogenetic gene selection workflow” is supported by SHOOT’s web interface, which allows a clade of genes to be selected and the protein sequences for just this clade to be downloaded for downstream user analyses.

In summary, SHOOT was designed to be as easy to use as BLAST, but to provide phylogenetically resolved results in which the query sequence is correctly placed in a phylogenetic tree. In this way the phylogenetic history of the query sequence and its orthologs can be immediately visualised, interpreted, and retrieved.

**Materials and Methods**

**Database preparation**

SHOOT consists of a database preparation program and a database search program. The database preparation program takes as input the results of an OrthoFinder (Emms and Kelly 2019) analysis of a set of proteomes.

To prepare phylogenetic databases for the SHOOT website, the OrthoFinder version 3.0 option, “-c1”, was used to cluster genes into groups consisting of all homologs, rather than the default behaviour which is to split homologous groups at the level of orthogroups. The advantage of the creating complete homologous groups is that their gene trees show an expanded evolutionary history of those genes, including ancient gene duplication events linking gene families, rather than only reaching back to the last common ancestor of the included species. This differs from a default OrthoFinder orthogroup analysis, for which the partitioning of genes into taxonomically comparable orthogroups groups is the priority. OrthoFinder-inferred rooted gene trees for these homolog groups are computed using MAFFT (Nakamura, et al. 2018) and IQ-TREE (Minh, et al. 2020) by using the additional options “-M msa -A mafft -T iqtree -s species_tree.nwk”, where “species_tree.nwk” was the rooted species tree for the included species. For IQ-TREE, the best fitting evolutionary model was tested for using “-m TEST” and bootstrap replicates performed using “-bb 1000”. The tree inference with IQ-TREE for a small number of the largest trees has not yet been completed (1st September 2021) and these are temporarily replaced with trees inferred with FastTree (Supplementary Table 6). An up-to-date version of this table is available at https://shoot.bio/faq.

The OrthoFinder results were converted to a SHOOT database in two steps: splitting of large trees and creation of the DIAMOND profiles database for assigning novel sequences to their correct gene tree. Large trees are split since the time requirements for adding a sequence to an MSA for a homologous group and for adding a sequence to its tree can grow super-linearly in the size of the group, leading to needlessly long runtimes. It was found that DIAMOND could instead be used to assign a gene to its correct subtree and then phylogenetic placement could be applied to assign the gene to its correct position within the subtree (Figure 4).

The script “split_large_tree.py” was used to split any tree larger than 2500 genes into subtrees of no more than 2500 genes each. Each subtree tree also contained an outgroup
gene, from outside the clade in the tree for that subtree, which was required for the later sequence search stage. For each tree that was split into subtrees, a super-tree was also created by the script of the phylogenetic relationships linking the subtrees. For each subtree, the script extracted the sub-MSA for later use. This subtree size of 2500 genes was chosen as it is the approximate upper limit tree size for which SHOOT could place a novel query gene in the tree in 15 seconds. This was judged to be a reasonable wait for users of the website to receive the tree for their query sequence. For the databases provided by the SHOOT website, between 2 and 40 of the largest trees were split into subtrees.

The script “create_shoot_db.py” was used to create a DIAMOND database of “profiles” for each unsplit tree or each subtree. A profile here refers to a set of representative sequences that best describe the sequence variability within a homologous group. These profiles are used to assign a novel query sequence to the correct tree or subtree. The representative sequences for a gene tree are selected using k-means clustering applied to the MSA corresponding to that (sub)tree using the python library Scikit-learn (Pedregosa, et al. 2011). For each cluster, the sequence closest to the centroid is chosen as a representative. For a homologous group of size N genes, k=N/10 representative sequences are used, with a minimum of min(20, N) representative sequences. This ensures that large and diverse homologous groups have sufficient representative sequences in the assignment database.

**Database search**

A query sequence is searched against the profiles database using DIAMOND (Buchfink, et al. 2021) with default sensitivity and an e-value cut-off of $10^{-3}$. If no hit is found, a second search is performed with the "--ultra-sensitive" setting. The top hitting sequence is used to assign the gene to the correct tree or subtree. The query gene is added to the pre-computed alignment using the MAFFT "--add" option and a phylogenetic tree is computed from this alignment using the precomputed tree for the reference alignment using EPA-ng (Barbera, et al. 2019) and gappa (Czech, et al. 2019).

If the gene is added to a subtree then the tree is rooted on the outgroup sequence for that subtree. The outgroup is then removed from the subtree and the subtree is grafted back into the original larger tree, using the supertree to determine the overall topology. This method provides the accuracy of phylogenetic analysis to place the gene in its correct position within the subtree while at the same time providing the user with the full gene history for the complete homologous group given by the supertree, which was calculated in full in the earlier database construction phase. All tree manipulations by SHOOT are performed using the ETE Toolkit (Huerta-Cepas, et al. 2016).

**Curated databases**

For the Plants database, the protein sequences derived from primary transcripts were downloaded from Phytozome (Goodstein, et al. 2012). The Uniport Reference Proteomes database was constructed using the 2020 Reference Proteomes (Altenhoff, et al. 2020). For the Fungi and Metazoa databases the proteomes were downloaded from Ensembl (Howe, et al. 2021) and the longest transcript variant of each gene was selected as a representative of that gene using OrthoFinder’s “primary_transcripts.py” script (Emms and Kelly 2019). The Bacterial and Archaeal database proteomes were downloaded from UniProt (UniProt 2021). The parallelisation of tasks in the preparation of the databases was performed using GNU parallel (Tange 2011).
Accuracy validation & performance

The UniProt Reference Proteomes database was used for validation of the SHOOT phylogenetic placements using a leave-one-out test. As this database covers the greatest phylogenetic range (covering all domains of life), its homologous groups contain the greatest sequence variability, and it provides the severest test of the accuracy of SHOOT. Test cases were constructed by selecting 1000 ‘cherries’ (pairs of genes sister to one another) with 100% bootstrap support from gene trees with median bootstrap support of at least 95%. The use of cherries allowed BLAST to be tested alongside SHOOT. This test was possible for BLAST since it would only have to identify a single closest gene, rather than having to identify a gene as the sister gene to a whole clade of genes (as SHOOT is designed to be able to do). The bootstrap support criteria ensured that the correct result was known with high confidence so that both methods could be assessed accurately. To ensure an even sampling of test cases, at most one test case was extracted from any one gene tree. Both the BLAST and SHOOT databases were completely pruned of the 1000 test cases. The correct result for each test case was recorded in two forms: the name of the gene that was the expected sister to the test gene, and the original IQ-TREE containing both the test gene and its expected sister gene. The test preparation also created a pruned tree for each test case, which had the test sequence removed.

Both methods were scored on whether they exactly identified the unique expected sister gene. If either method was not exactly correct, they were secondarily scored on whether the position of the assignment/best hit was within two branches of the correct position in the original tree. To assess this, a results tree was constructed to represent the result of the search. For SHOOT, this was the gene tree it returned, with the query gene grafted into its predicted position. For BLAST, the query gene was inserted into the pruned tree for that test case as the sister to the gene identified as the best hit by BLAST. In 996 of the 1000 test cases BLAST identified a single unique best hit. In 4 cases it identified multiple identical best hit sequences according to both e-value and bit score. In these cases, the gene was inserted into the tree as sister to the clade containing the equal closest hits. This would correspond to SHOOT inserting the query gene as sister to a clade of genes rather than a single closest gene. The distance between the expected sister gene and the actual sister gene (clade) could then easily be calculated since it is the Robinson-Foulds distance (Robinson and Foulds 1981) between the results tree and original tree from IQ-TREE.

Each of the 1000 test cases was run using 16 cores of an Intel Xeon E5-2683 CPU and the runtime recorded (Figure 2).

SHOOT website

The tree visualisation is provided by the phylotree.js library (Shank, et al. 2018). The SHOOT website is implemented in JavaScript and Bootstrap and using the Flask web framework.

Acknowledgements

SK is a Royal Society University Research Fellow. This work was supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement number 637765.
References
Altenhoff AM, Garrayo-Ventas J, Cosentino S, Emms D, Glover NM, Hernandez-Plaza A, Nevers Y, Sundesha V, Szklarczyk D, Fernandez JM, et al. 2020. The Quest for Orthologs benchmark service and consensus calls in 2020. Nucleic Acids Research 48:W538-W545.
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215:403-410.
Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A. 2019. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Systematic Biology 68:365-369.
Buchfink B, Reuter K, Drost HG. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods 18:366-368.
Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. Bioinformatics 10:421.
Czech L, Barbera P, Stamatakis A. 2019. Methods for automatic reference trees and multilevel phylogenetic placement. Bioinformatics 35:1151-1158.
Dalquen DA, Dessimoz C. 2013. Bidirectional Best Hits Miss Many Orthologs in Duplication-Rich Clades such as Plants and Animals. Genome Biology and Evolution 5:1800-1806.
Eddy SR. 2011. Accelerated Profile HMM Searches. Plos Computational Biology 7.
Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-2461.
Emms DM, Kelly S. 2020. Benchmarking Orthogroup Inference Accuracy: Revisiting Orthobench. Genome Biology and Evolution 12:2258-2266.
Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20.
Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16.
Goodstein DM, Shu SQ, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research 40:D1178-D1186.
Huerta-Cepas J, Bueno A, Dopazo JQ, Gabaldon T. 2008. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Research 36:D491-D496.
Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution 33:1635-1638.
Kelly S, Maini PK. 2013. DendroBLAST: Approximate Phylogenetic Trees in the Absence of Multiple Sequence Alignments. Plos One 8.
Lipman DJ, Pearson WR. 1985. Rapid and sensitive protein similarity searches. Science 227:1435-1441.
Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Research 38:D204-D210.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37:1530-1534.
Mirdita M, Steinegger M, Soding J. 2019. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics 35:2856-2858.
Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 34:2490-2492.
Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48:443-453.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12:2825-2830.

Robinson DF, Foulds LR. 1981. Comparison of Phylogenetic Trees. Mathematical Biosciences 53:131-147.

Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. 2014. TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Research 42:D922-D925.

Shank SD, Weaver S, Kosakovsky Pond SL. 2018. phyloTree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. Bmc Bioinformatics 19:276.

Smith TF, Waterman MS. 1981. Identification of Common Molecular Subsequences. Journal of Molecular Biology 147:195-197.

Steinegger M, Soding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026-1028.

Tang H, Finn RD, Thomas PD. 2019. TreeGrafter: phylogenetic tree-based annotation of proteins with Gene Ontology terms and other annotations. Bioinformatics 35:518-520.

Tange O. 2011. GNU Parallel - The Command-Line Power Tool. ;login: The USENIX Magazine 36:42-47.

UniProt C. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49:D480-D489.
**Figures**

*Figure 1*

**A. Database preparation**

![Diagram of database preparation stage]

**B. Novel sequence search**

![Diagram of novel sequence search stage]

*Figure 1.* The workflow for the two separate stages of SHOOT: **A)** The database preparation stage. **B)** The sequence search stage. MSA, multiple sequence alignment. HG, homologous group. Individual shapes represent individual protein sequences.
Figure 2. Violin plot of the runtime. Violin plots show the time taken by BLAST and SHOOT to search an individual gene against the same database of 984,137 protein sequences from 78 species. The runtimes for 1000 randomly sampled searches are shown.
Figure 3. An example gene tree and orthologs table returned by SHOOT. Here, the UniProt Reference Proteomes database was searched using a for a query gene sequence labelled “Duck_gene_X”. This corresponds to the Duck protein ENSAPLP00000002788, which is not included in the database.
**Figure 4.** Measures of search and ortholog inference accuracy. **A)** The accuracy of SHOOT and BLAST at identifying the closest related database gene to a randomly selected query sequence. **B)** The agreement in orthologs inferred for a query gene using the species-overlap method with SHOOT phylogenetic trees vs. conventional phylogenetic trees.
Figure 5. The phylogeny for the species in the Metazoan dataset.
## Supplementary Table 1: UniProt 2020 Reference Proteomes – Species list

| Domain         | Species                                |
|----------------|----------------------------------------|
| Archaea        | Halobacterium salinarum                |
| Archaea        | Korarchaeum cryptofilum                |
| Archaea        | Methanocaldococcus jannaschii          |
| Archaea        | Methanosarcina acetivorans             |
| Archaea        | Nitrosopumilus maritimus               |
| Archaea        | Saccharolobus solfataricus             |
| Archaea        | Thermococcus kodakarensis              |
| Bacteria       | Aquifex aeolicus                       |
| Bacteria       | Bacillus subtilis                      |
| Bacteria       | Bacteroides thetaiotaomicron           |
| Bacteria       | Bradyrhizobium diazoefficiens          |
| Bacteria       | Chlamydia trachomatis                  |
| Bacteria       | Chloroflexus aurantiacus               |
| Bacteria       | Deinococcus radiodurans                |
| Bacteria       | Dictyoglomus turgidum                  |
| Bacteria       | Escherichia coli                       |
| Bacteria       | Fusobacterium nucleatum                |
| Bacteria       | Geobacter sulfurreducens               |
| Bacteria       | Gloeobacter violaceus                  |
| Bacteria       | Helicobacter pylori                    |
| Bacteria       | Leptospira interrogans                 |
| Bacteria       | Mycobacterium tuberculosis             |
| Bacteria       | Mycoplasma genitalium                  |
| Bacteria       | Neisseria meningitidis                 |
| Bacteria       | Pseudomonas aeruginosa                 |
| Bacteria       | Rhodopirellula baltica                 |
| Bacteria       | Streptomyces coelicolor                |
| Bacteria       | Synechocystis sp.                      |
| Bacteria       | Thermodesulfovibrio yellowstonii       |
| Bacteria       | Thermotoga maritima                    |
| Eukaryota      | Anopheles gambiae                      |
| Eukaryota      | Arabidopsis thaliana                   |
| Eukaryota      | Batrachochytrium dendrobatidis         |
| Eukaryota      | Bos taurus                             |
| Eukaryota      | Branchiostoma floridae                 |
| Eukaryota      | Caenorhabditis elegans                 |
| Eukaryota      | Candida albicans                       |
| Eukaryota      | Canis lupus familiaris                 |
| Eukaryota      | Chlamydomonas reinhardtii              |
| Eukaryota      | Ciona intestinalis                     |
| Eukaryota      | Cryptococcus neoformans                |
| Eukaryota      | Danio rerio                            |
| Eukaryota      | Dictyostelium discoideum               |
| Eukaryota      | Drosophila melanogaster                |
| Eukaryota      | Gallus gallus                         |
| Eukaryota      | Giardia intestinalis                   |
| Eukaryota      | Gorilla gorilla gorilla                |
| Eukaryota          | Helobdella robusta  |
|--------------------|---------------------|
| Eukaryota          | Homo sapiens        |
| Eukaryota          | Ixodes scapularis   |
| Eukaryota          | Leishmania major    |
| Eukaryota          | Lepisosteus oculatus|
| Eukaryota          | Monodelphis domestica|
| Eukaryota          | Monosiga brevicollis|
| Eukaryota          | Mus musculus        |
| Eukaryota          | Nematostella vectensis|
| Eukaryota          | Neosartorya fumigata|
| Eukaryota          | Neurospora crassa   |
| Eukaryota          | Oryza sativa subsp. japonica |
| Eukaryota          | Oryzias latipes     |
| Eukaryota          | Pan troglodytes     |
| Eukaryota          | Paramecium tetraurelia |
| Eukaryota          | Phaeosphaeria nodorum|
| Eukaryota          | Physcomitrella patens|
| Eukaryota          | Phytophthora ramorum|
| Eukaryota          | Plasmodium falciparum|
| Eukaryota          | Puccinia graminis   |
| Eukaryota          | Rattus norvegicus   |
| Eukaryota          | Saccharomyces cerevisiae |
| Eukaryota          | Schizosaccharomyces pombe |
| Eukaryota          | Sclerotinia sclerotiorum |
| Eukaryota          | Thalassiosira pseudonana |
| Eukaryota          | Tribolium castaneum |
| Eukaryota          | Trichomonas vaginalis|
| Eukaryota          | Ustilago maydis     |
| Eukaryota          | Xenopus tropicalis  |
| Eukaryota          | Yarrowia lipolytica |
| Eukaryota          | Zea mays            |
**Supplementary Table 2: Fungi species list**

| Species                        | Species                        | Species                        |
|--------------------------------|--------------------------------|--------------------------------|
| Agaricus bisporus              | Cryptococcus neoformans        | Rhizoctonia solani             |
| Amanita muscaria               | Encephalitozoon intestinalis   | Rhizopus delemar               |
| Aspergillus fumigatus          | Enterocytozoon bieneusi        | Saccharomyces cerevisiae       |
| Aspergillus nidulans           | Fusarium oxysporum             | Schizosaccharomyces pombe      |
| Batrachochytrium salamandrivors| Magnaporthe oryzae             | Sclerotinia sclerotiorum       |
| Blumeria graminis              | Mortierella elongata           | Spizellomyces punctatus        |
| Botrytis cinerea               | Neurospora crassa              | Ustilago maydis                |
| Candida albicans               | Phaeosphaeria nodorum          | Yarrowia lipolytica            |
| Colletotrichum graminicola     | Puccinia graminis              | Zymoseptoria tritici           |

**Outgroup**

| Species                        | Species                        | Species                        |
|--------------------------------|--------------------------------|--------------------------------|
| Caenorhabditis elegans         | Homo sapiens                   | Dictyostelium discoideum       |
| Drosophila melanogaster        | Monosiga brevicollis           |                                |
Supplementary Table 3: Metazoan species list

| Amphimedon queenslandica | Danio rerio | Octopus bimaculoides |
|--------------------------|-------------|----------------------|
| Anolis carolinensis | Daphnia magna | Oncorhynchus mykiss |
| Anopheles gambiae | Drosophila melanogaster | Ornithorhyncus anatinus |
| Apis mellifera | Gadus morhua | Oryzias latipes |
| Astatotilapia calliptera | Gallus gallus | Pan troglodytes |
| Bombyx mori | Glossina morsitans | Petromyzon marinus |
| Bos taurus | Helobdella robusta | Phascolarctos cinereus |
| Branchiostoma lanceolatum | Homo sapiens | Poecilia formosa |
| Bubol bubo | Ixodes scapularis | Rattus norvegicus |
| Caenorhabditis elegans | Latimeria chalumnae | Schistosoma mansoni |
| Callithrix jacchus | Lepisosteus oculatus | Strongylocentrotus purpuratus |
| Callorhinchus milii | Leptobrachium leishanense | Tetraodon nigroviridis |
| Canis familiaris | Mnemiopsis leidyi | Thelohanellus kitauei |
| Chrysemys picta | Monodelphis domestica | Trichinella spiralis |
| Ciona intestinalis | Musculus | Trichoplax adhaerens |
| Corvus moneduloides | Nematostella vectensis | Xenopus tropicalis |
| Amphimedon queenslandica | Danio rerio | Octopus bimaculoides |

Outgroup

| Dictyostelium discoideum | Phaeosphaeria nodorum | Schizosaccharomyces pombe |
|--------------------------|----------------------|--------------------------|
| Monosiga brevicollis | Saccharomyces cerevisiae | |
### Supplementary Table 4: Plants species list

| Amborella trichopoda | Glycine max | Picea glauca |
|----------------------|-------------|--------------|
| Anthoceros punctatus | Gossypium raimondii | Pinus sylvestris |
| Aquilegia coerulea   | Hordeum vulgare | Prunus persica |
| Arabidopsis thaliana | Manihot esculenta | Selaginella moellendorfii |
| Azolla filiculoides  | Marchantia polymorpha | Setaria italica |
| Brassica oleracea    | Micromonas spRCC299 | Solanum lycopersicum |
| Chara braunii        | Musa acuminata | Spirodela polyrhiza |
| Chlamydomonas reinhardtii | Oryza sativa | Triticum aestivum |
| Eucalyptus grandis   | Ostreococcus lucimarinus | Volvox carteri |
| Gingko biloba        | Physcomitrella patens | Zea mays |

### Outgroup

| Chondrus crispus | Chondrus crispus | Chondrus crispus |
Supplementary Table 5: Bacterial & Archaeal strains list

| UniProt proteome | NCBI taxon | Name in SHOOT | Selection |
|------------------|------------|---------------|-----------|
| UP000000425      | 122586     | Neisseria_meningitidis | QfO UniProt ref. prot. |
| UP000000429      | 85962      | Helicobacter_pylori | QfO UniProt ref. prot. |
| UP000000431      | 272561     | Chlamydia_trachomatis | QfO UniProt ref. prot. |
| UP000000536      | 69014      | Thermococcus_kodakarensis | QfO UniProt ref. prot. |
| UP000000554      | 64091      | Halobacterium_salinarum | QfO UniProt ref. prot. |
| UP000000557      | 251221     | Gloeobacter_violaceus | QfO UniProt ref. prot. |
| UP000000577      | 243231     | Geobacter_sulfurreducens | QfO UniProt ref. prot. |
| UP000000625      | 83333      | Escherichia_coli | QfO UniProt ref. prot. |
| UP000000718      | 289376     | Thermodesulfovibrio_yellowstonii | QfO UniProt ref. prot. |
| UP000000792      | 436308     | Nitrosopumilus_maritimus | QfO UniProt ref. prot. |
| UP000000798      | 224324     | Aquifex_aeolicus | QfO UniProt ref. prot. |
| UP000000805      | 243232     | Methanocaldococcus_jannaschii | QfO UniProt ref. prot. |
| UP000000807      | 243273     | Mycoplasma_genitalium | QfO UniProt ref. prot. |
| UP000001025      | 243090     | Rhodopirellula_baltica | QfO UniProt ref. prot. |
| UP000001408      | 189518     | Leptospira_interrogans | QfO UniProt ref. prot. |
| UP000001414      | 226186     | Bacteroides_thetaiotaomicron | QfO UniProt ref. prot. |
| UP000001425      | 1111708    | Synechocystis_Kazusa | QfO UniProt ref. prot. |
| UP000001570      | 224308     | Bacillus_subtilis | QfO UniProt ref. prot. |
| UP000001584      | 83332      | Mycobacterium_tuberculosis | QfO UniProt ref. prot. |
| UP000001686      | 374847     | Korarchaeum_cryptofilum | QfO UniProt ref. prot. |
| UP000001973      | 100226     | Streptomyces_coelicolor | QfO UniProt ref. prot. |
| UP000001974      | 273057     | Saccharolobus_solfataricus | QfO UniProt ref. prot. |
| UP000002008      | 324602     | Chloroflexus_aurantiacus | QfO UniProt ref. prot. |
| UP000002438      | 208964     | Pseudomonas_aeruginosa | QfO UniProt ref. prot. |
| UP000002487      | 188937     | Methanosarcina_acetivorans | QfO UniProt ref. prot. |
| UP000002521      | 190304     | Fusobacterium_nucleatum | QfO UniProt ref. prot. |
| UP000002524      | 243230     | Deinococcus_radionudans | QfO UniProt ref. prot. |
| UP000002526      | 224911     | Bradyrhizobium_diazoefficiens | QfO UniProt ref. prot. |
| UP0000007719     | 515635     | Dictyoglomus_turgidum | QfO UniProt ref. prot. |
| UP000000813      | 243274     | Thermotoga_maritima | QfO UniProt ref. prot. |
| UP000000265      | 272620     | Klebsiella_pneumoniae | Highly cited |
| UP000000579      | 71421      | Haemophilus_influenzae | Highly cited |
| UP000000580      | 262316     | Mycolicibacterium_paratuberculosis | Highly cited |
| UP000000584      | 243277     | Vibrio_cholerae | Highly cited |
| UP000000586      | 171101     | Streptococcus_pneumoniae | Highly cited |
| UP000000588      | 242619     | Porphyromonas_gingivalis | Highly cited |
| UP000000609      | 272624     | Legionella_pneumophila | Highly cited |
| UP000000799      | 192222     | Campylobacter_jejuni | Highly cited |
| UP000000813      | 176299     | Agrobacterium_fabrum | Highly cited |
| UP000000815      | 632        | Yersinia_pestis | Highly cited |
| UP000000817      | 169963     | Listeria_monocytogenes | Highly cited |
| UP000000818      | 195102     | Clostridium_perfringens | Highly cited |
| UP000001006      | 623        | Shigella_Flexneri | Highly cited |
| UP000001014      | 99287      | Salmonella_typhimurium | Highly cited |
| UP000001978      | 272563     | Clostridoides_difficile | Highly cited |
| UP000002196      | 272623     | Lactobacillus_lactis | Highly cited |
| UP000002256      | 395491     | Rhizobium_leguminosarum | Highly cited |
| UP000006381 | 272621 | Lactobacillus_acidophilus | Highly cited |
|------------|--------|---------------------------|--------------|
| UP000007477 | 871585 | Acinetobacter_calcoaceticus | Highly cited |
| UP000008315 | 529507 | Proteus_mirabilis | Highly cited |
| UP000008816 | 93061  | Staphylococcus_aureus | Highly cited |
| UP000014594 | 1260356 | Enterococcus_faecalis | Highly cited |
| UP000075229 | 140    | Borrelia_hermsii | Highly cited |
| UP000198289 | 615    | Serratia_marcescens | Highly cited |
| UP000289336 | 1528098 | Rickettsiales_bacterium | Mitochondrion relative |
| UP000180235 | 1188229 | Gloeomargarita_lithophora | Chloroplast relative |
| UP000000543 | 279808  | Staphylococcus_haemolyticus | Phylo. sampling |
| UP000000547 | 167879  | Colwellia_psychrerythraea | Phylo. sampling |
| UP000001169 | 272569  | Haloarcula_marismortui | Phylo. sampling |
| UP000001361 | 883    | Desulfobulbus_vulgaris | Phylo. sampling |
| UP000001362 | 243159  | Acidithiobacillus_ferrooxidans | Phylo. sampling |
| UP000001961 | 64471   | Synechococcus_CC9311 | Phylo. sampling |
| UP000002208 | 546414  | Deinococcus_deserti | Phylo. sampling |
| UP000002257 | 395965  | Methylclostridium_silvestris | Phylo. sampling |
| UP000002386 | 471223  | Geobacillus_WCH70 | Phylo. sampling |
| UP000002457 | 521011  | Methanospirillum_palustris | Phylo. sampling |
| UP000002495 | 235279  | Helicobacter_hepaticus | Phylo. sampling |
| UP000003277 | 742743  | Dialister_succinatiphilus | Phylo. sampling |
| UP000003415 | 469616  | Fusobacterium_mortiferum | Phylo. sampling |
| UP000003446 | 661087  | Olsenella_F0356 | Phylo. sampling |
| UP000003855 | 665956  | Subdoligranulum_4-3-54A2FAA | Phylo. sampling |
| UP000003981 | 621372  | Paenibacillus_D14 | Phylo. sampling |
| UP000004073 | 1105031 | Clostridium_MSTE9 | Phylo. sampling |
| UP000004090 | 428127  | Absiella_dolichum | Phylo. sampling |
| UP000004259 | 246199  | Ruminococcus_albus | Phylo. sampling |
| UP000004478 | 1225176 | Cecembia_lonarensis | Phylo. sampling |
| UP000004870 | 638300  | Cardiobacterium_hominis | Phylo. sampling |
| UP000005262 | 768704  | Desulfosporosinus_meridiei | Phylo. sampling |
| UP000006229 | 1131455 | Mycoplasma_canis | Phylo. sampling |
| UP000006415 | 857290  | Scardovia_wigginsiae | Phylo. sampling |
| UP000006556 | 370438  | Pelotomaculum_thermopropionicum | Phylo. sampling |
| UP000006743 | 557723  | Haemophilus_parasuis | Phylo. sampling |
| UP000007271 | 1185325 | Lactobacillus_coryniformis | Phylo. sampling |
| UP000007753 | 452662  | Sphingobium_japonicum | Phylo. sampling |
| UP000007995 | 997888  | Bacteroides_fiber | Phylo. sampling |
| UP000008204 | 41431   | Rippkaea_orientalis | Phylo. sampling |
| UP000008212 | 243275  | Treponema_denticola | Phylo. sampling |
| UP000008308 | 263358  | Micromonospora_maris | Phylo. sampling |
| UP000008701 | 290317  | Chlorobium_phaeobacteroides | Phylo. sampling |
| UP000009044 | 634177  | Komagataebacter_medellinensis | Phylo. sampling |
| UP000009154 | 1112204 | Gordonia_polyisoprenivorans | Phylo. sampling |
| UP000011615 | 1230457 | Haloterrigena_limicola | Phylo. sampling |
| UP000011728 | 931276  | Clostridium_saccharoperbutylicum | Phylo. sampling |
| Accession | Length | Species/Genus | Category |
|-----------|--------|--------------|----------|
| UP000013232 | 1123367 | Thauera_linaloolentis | Phylo. sampling |
| UP000017993 | 1262970 | Subdoligranulum_CAG314 | Phylo. sampling |
| UP000018014 | 1262708 | Bacillus_CAG988 | Phylo. sampling |
| UP000018042 | 1262875 | Eggerthella_CAG209 | Phylo. sampling |
| UP000018237 | 1262989 | Firmicutes_bacterium | Phylo. sampling |
| UP000018329 | 1262693 | Alistipes_CAG268 | Phylo. sampling |
| UP000018361 | 1263102 | Prevotella_copri | Phylo. sampling |
| UP000018415 | 1341679 | Acinetobacter_indicus | Phylo. sampling |
| UP000019028 | 1239307 | Sodalis_praecaptivus | Phylo. sampling |
| UP000019082 | 1302241 | Cutibacterium_acnes | Phylo. sampling |
| UP000019222 | 1224164 | Corynebacterium_vitaeruminis | Phylo. sampling |
| UP000019267 | 1276246 | Spiroplasma_culicicola | Phylo. sampling |
| UP000019278 | 1356164 | Paucilactobacillus_wasatchensis | Phylo. sampling |
| UP000019288 | 1462526 | Virgibacillus_massiliensis | Phylo. sampling |
| UP000020878 | 1346853 | Novosporangium_imitans | Phylo. sampling |
| UP000028780 | 156978 | Corynebacterium_imitans | Phylo. sampling |
| UP000028875 | 1454005 | Candidatus_Accumulibacter | Phylo. sampling |
| UP000030960 | 561184 | Phylomicrobium_alba | Phylo. sampling |
| UP000031057 | 1348853 | Novosporangium_haematiophilum | Phylo. sampling |
| UP000031627 | 1410383 | Candidatus_Tachikawaea | Phylo. sampling |
| UP000032279 | 1335616 | Paucilactobacillus_wasatchensis | Phylo. sampling |
| UP000032287 | 137591 | Weissella_cibaria | Phylo. sampling |
| UP000033511 | 43662 | Pseudoalteromonas_piscicida | Phylo. sampling |
| UP000035114 | 1628212 | Chromobacterium_LK11 | Phylo. sampling |
| UP000036921 | 1581033 | Bacillus_FJAT-21945 | Phylo. sampling |
| UP000037870 | 270918 | Salegentibacter_mishustinae | Phylo. sampling |
| UP000037870 | 1109412 | Brenneria_goodwinii | Phylo. sampling |
| UP000039571 | 1736540 | Aeromicrobium_Root472D3 | Phylo. sampling |
| UP000041587 | 1736232 | Arthrobacter_Leaf69 | Phylo. sampling |
| UP000044377 | 1736381 | Aureimonas_Leaf454 | Phylo. sampling |
| UP000044377 | 1736189 | Aureimonas_Leaf454 | Phylo. sampling |
| UP000045182 | 1698267 | Candidate_MSBL1-archaeon | Phylo. sampling |
| UP000045824 | 1891921 | Piscirickettsia_litoralis | Phylo. sampling |
| UP000045824 | 1891921 | Bosea_RAC05 | Phylo. sampling |
| UP000046629 | 1891921 | Butyrivibrio_proteoclasticus | Phylo. sampling |
| Accession | Score | Name                  | Method               |
|-----------|-------|-----------------------|----------------------|
| UP000184455 | 1855338 | Nitrosospira_Nsp11 | Phylo. sampling      |
| UP000184520 | 634436  | Mariseminitalea_aggregata | Phylo. sampling    |
| UP000186096 | 58117    | Microbispora_rosea | Phylo. sampling      |
| UP000186602 | 1261349  | Roseburia_sp499 | Phylo. sampling      |
| UP000187327 | 1883416  | Halomonas_sp1513 | Phylo. sampling      |
| UP000187995 | 1805827  | Rhodococcus_MTMM3W5 | Phylo. sampling      |
| UP000190286 | 745368   | Gemmiger_formicilis | Phylo. sampling      |
| UP000191905 | 1873176  | Pseudaminobacter_manganicus | Phylo. sampling |
| UP000192042 | 1325564  | Nitrospira_japonica | Phylo. sampling      |
| UP000193006 | 199441   | Alkalihalobacillus_krulwichiae | Phylo. sampling |
| UP000193136 | 1969733  | Geothermobacter_EPR-M | Phylo. sampling    |
| UP000194216 | 1985172  | Sphingomonas_IBVSS2 | Phylo. sampling      |
| UP000195076 | 1932621  | Nostoc_T09 | Phylo. sampling      |
| UP000195529 | 1965622  | Megasphaera_An286 | Phylo. sampling      |
| UP000195781 | 1232426  | Collinsella_massiliensis | Phylo. sampling |
| UP000197446 | 431059   | Pelomonas_puraquae | Phylo. sampling      |
| UP000198953 | 46177    | Nonomuraea_pusilla | Phylo. sampling      |
| UP000199067 | 1780377  | Coriobacteriaceae_bacterium | Phylo. sampling |
| UP000199242 | 1141221  | Chryseobacterium_taihuense | Phylo. sampling |
| UP000199432 | 1882749  | Opitutus_GAS368 | Phylo. sampling      |
| UP000199671 | 332524   | Actinomyces_ruminicola | Phylo. sampling   |
| UP000199705 | 551996   | Mucilaginibacter_gossypii | Phylo. sampling |
| UP000199768 | 1881066  | Phyllobacterium_YR620 | Phylo. sampling |
| UP000199802 | 1965654  | Lachnoclostridium_An76 | Phylo. sampling |
| UP000202922 | 1524263  | Confluentimicrobium_lipolyticum | Phylo. sampling |
| UP000215509 | 554312   | Paenibacillus_rigui | Phylo. sampling      |
| UP000216308 | 1383851  | Halorubrum_halodurans | Phylo. sampling   |
| UP000217076 | 83401    | Roseospirillum_parvum | Phylo. sampling    |
| UP000217289 | 1294270  | Melittangium_boletus | Phylo. sampling     |
| UP000219434 | 442709   | Flavimobilis_soli | Phylo. sampling      |
| UP000222106 | 638953   | Georgenia_soli | Phylo. sampling      |
| UP000223878 | 2058137  | Polaribacter_ALD11 | Phylo. sampling      |
| UP000223889 | 1250229  | Ulvibacter_MAR-2010-11 | Phylo. sampling   |
| UP000223535 | 2029108  | Bacillus_UMB0899 | Phylo. sampling      |
| UP0002236356 | 2067550  | Clostridium_chh4-2 | Phylo. sampling      |
| UP0002236731 | 797291   | Sphingobacterium_lactis | Phylo. sampling |
| UP0002238164 | 75385   | Micropyrina_glycogenica | Phylo. sampling |
| UP0002238375 | 1469603  | Spirosoma_oryzae | Phylo. sampling      |
| UP0002243063 | 1245526  | Pseudomonas_guangdongensis | Phylo. sampling |
| UP0002243494 | 2020948  | Romboutsia_maritimum | Phylo. sampling    |
| UP0002244224 | 589035   | Gemmobacter_caeni | Phylo. sampling      |
| UP0002245108 | 2108523  | Lawsonibacter_asaccharolyticus | Phylo. sampling |
| UP0002245507 | 2201891  | Nocardoides_silvaticus | Phylo. sampling    |
| UP0002245623 | 2173179  | Microbacterium_4-13 | Phylo. sampling      |
| UP0002245926 | 2202825  | Methylobacterium_durans | Phylo. sampling |
| UP0002247832 | 670078   | Arthrobacter_livingstonensis | Phylo. sampling |
| UP0002249065 | 2230885  | Roseicella_frigidaeris | Phylo. sampling    |
| Accession   | Accession 2  | Name                          | Sample Type       |
|-------------|-------------|-------------------------------|-------------------|
| UP000250434 | 1804986     | Amycolatopsis_albispora       | Phylo. sampling   |
| UP000252733 | 989         | Marinilabilia_salmonicolor    | Phylo. sampling   |
| UP000253318 | 1931232     | Marinitenerispora_sediminis   | Phylo. sampling   |
| UP000254875 | 2211104     | Paraburkholderia_lacunae      | Phylo. sampling   |
| UP000260665 | 2184758     | Rhodoferax_IMCC26218          | Phylo. sampling   |
| UP000265971 | 1825976     | Neorhizobium_NCHU2750         | Phylo. sampling   |
| UP000266860 | 1630648     | Novosphingobium_MD-1          | Phylo. sampling   |
| UP000269803 | 2485200     | Frondihabitans_PhB188         | Phylo. sampling   |
| UP000273083 | 1329262     | Mobilisporobacter_senegalensis| Phylo. sampling   |
| UP000275325 | 2495580     | Sphingomonas_TF3              | Phylo. sampling   |
| UP000276437 | 1930071     | Rhodoferax(IMCC26218)         | Phylo. sampling   |
| UP000282084 | 2072        | Saccharothrix_australiensis   | Phylo. sampling   |
| UP000287188 | 2014872     | Dictyobacter_kobayashii       | Phylo. sampling   |
| UP000287890 | 2507159     | Clostridium_JN-9              | Phylo. sampling   |
| UP000288096 | 45657       | Desulfitobacter_firmatenuis   | Phylo. sampling   |
| UP000288291 | 2495899     | Lactobacillus_vulgaris        | Phylo. sampling   |
| UP000288967 | 2501295     | Thiothrix_endosymbiont        | Phylo. sampling   |
| UP000289784 | 2137479     | Sphingomonas_Composti          | Phylo. sampling   |
| UP000292120 | 2528630     | Aquabacterium_KMB7            | Phylo. sampling   |
| UP000294096 | 2510646     | Loktanella_IMCC34160          | Phylo. sampling   |
| UP000294498 | 1539049     | Dinghuibacter_silviterrae     | Phylo. sampling   |
| UP000295707 | 1537524     | Thiobacillus_longum           | Phylo. sampling   |
| UP000297351 | 2561925     | Brevundimonas_S30B            | Phylo. sampling   |
| UP000306069 | 2040651     | Campylobacter_12-5580         | Phylo. sampling   |
| UP000307244 | 2571272     | Pedobacter_RP-3-15            | Phylo. sampling   |
| UP000307467 | 343240      | Thiobacillus_longum           | Phylo. sampling   |
| UP000307507 | 2565924     | Flavobacterium_CC-CTC003      | Phylo. sampling   |
| UP000307657 | 2565367     | Lacinutrix_CAUsp-1491         | Phylo. sampling   |
| UP000315440 | 2527991     | Pseudobutyrophilus_maris      | Phylo. sampling   |
| UP000316225 | 384678      | Paracoccus_sulfuroxidans      | Phylo. sampling   |
| UP000316304 | 2528004     | Novibacillus_australis        | Phylo. sampling   |
| UP000318165 | 92402       | Mycoplasma_equirhinis         | Phylo. sampling   |
| UP000318431 | 1036180     | Massilia_lurida               | Phylo. sampling   |
| UP000318566 | 2768454     | Streptomyces_SLBN-118         | Phylo. sampling   |
| UP000319173 | 713054      | TM7_phyllum                   | Phylo. sampling   |
| UP000322791 | 2606448     | Hymenobacter_KIGAM108         | Phylo. sampling   |
| UP000324880 | 1948890     | Rhodobacterales_bacterium     | Phylo. sampling   |
| UP000325372 | 2613842     | Wenzhouxiangella_W260         | Phylo. sampling   |
| UP000326711 | 2487892     | Corynebacterium_LMM-1652      | Phylo. sampling   |
| UP000326944 | 2590022     | Sulfurimonas_GYSZ1            | Phylo. sampling   |
| UP000347955 | 2653936     | Tetradsphaera_F2B08           | Phylo. sampling   |
| UP00041772  | 2650774     | Bifidobacterium_LMGsp-31471   | Phylo. sampling   |
| UP000462055 | 2650748     | Actinomadura_LD22             | Phylo. sampling   |
| UP000474632 | 2710884     | Parapsitilimonas_SNGA-6       | Phylo. sampling   |
| UP000476210 | 343235      | Methanotrophic_endosymbiont   | Phylo. sampling   |
| UP000477884 | 2703788     | Edaphobacter_12200R-103       | Phylo. sampling   |
| UP000481552 | 2706104     | Streptomyces_SID8455          | Phylo. sampling   |
| UP000500686 | 754515      | Mycoplasma_ES2806-GEN         | Phylo. sampling   |
| UP000502894 | 2708020     | Leucobacter_TUM19329          | Phylo. sampling   |
| UP000503441 | 2714933     | Leucobacter_HDW9A             | Phylo. sampling   |
Supplementary Table 6: Unfinished IQTrees temporarily replaced with FastTree trees

| Database            | Fungi | Metazoa | Plants | QfO UniProt | Bacteria & Archaea |
|---------------------|-------|---------|--------|-------------|--------------------|
| Trees in database   | 9115  | 10516   | 10617  | 17124       | 16156              |
| FastTree tree IDs    | N/A   | 0-30    | 0, 1, 3| 0           | 0, 1               |