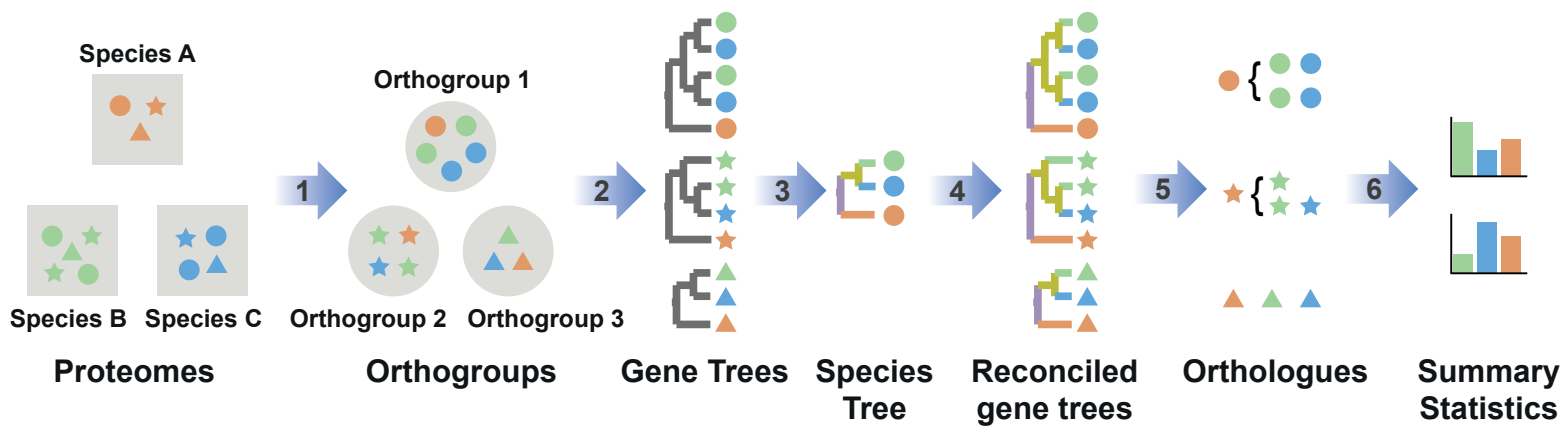


OrthoFinder Manual:

Accurate inference of orthologues and orthogroups made easy!



Dr. David Emms

david.emms@plants.ox.ac.uk

Dr. Steven Kelly

steven.kelly@plants.ox.ac.uk

June 14, 2017

What does OrthoFinder do?

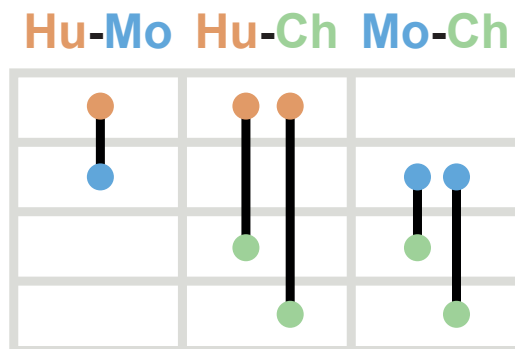
OrthoFinder is a fast, accurate and comprehensive analysis tool for comparative genomics. It finds **orthologues** and **orthogroups**, infers **gene trees** for all orthogroups and infers a **rooted species tree** for the species being analysed. OrthoFinder also provides **comprehensive statistics** for comparative genomic analyses. OrthoFinder is simple to use and all you need to run it is a set of protein sequence files (one per species) in FASTA format.

Orthogroup



Group of genes descended from single gene in LCA of group of species

Orthologues



Pairs of genes descended from single gene in LCA of pair of species

Citation

Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, *Genome Biology* 16:157

Links

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0721-2>

<https://github.com/davidemms/OrthoFinder>

Contents

1	Orthogroups, Orthologues & Paralogues	3
1.1	Why Orthogroups	3
2	Setting Up OrthoFinder	4
2.1	Set Up	4
2.2	Dependencies	4
2.2.1	BLAST+	4
2.2.2	MCL	4
2.2.3	FastME	5
2.2.4	DLCpar	5
2.3	Running OrthoFinder	5
2.4	Setup for advanced use	5
2.4.1	Trees from Multiple Sequence Alignments	5
2.4.2	Python Source Code Version	5
3	Performing Your Own OrthoFinder Analysis	6
4	Results Files	7
4.1	Results Files: Orthogroups	7
4.2	Results Files: Single-copy orthogroups & gene counts	7
4.3	Results Files: Orthogroup Statistics	7
4.4	Results Files: Orthologues	8
4.5	Results Files: Gene Trees and Species Tree	8
5	Advanced Usage	9
5.1	Controlling OrthoFinder Workflow	9
5.2	Adding Extra Species	9
5.3	Removing Species	9
5.4	Adding and Removing Species Simultaneously	9
5.5	User-specified Species Tree	9
5.6	Starting/Stopping OrthoFinder at Different Stages	11
5.7	User-speciefied MSA, Tree Inference or Sequence Seach Program	11
5.7.1	Config File	12
5.8	Inferring MSA Gene Trees	13
5.9	Parallelising OrthoFinder Algorithm (-a option)	13
5.10	Running BLAST Searches Separately	13
5.11	Using Pre-Computed BLAST Results	13
5.12	Using the Orthoxml Format	14
5.13	Regression Tests	14
6	Appendix: File Format for Pre-Computed BLAST Results	15
6.1	FASTA Files	15
6.2	BLAST Results Files	15
6.3	SequenceIDs.txt	16
6.4	SpeciesID.txt	16

1 Orthogroups, Orthologues & Paralogues

‘Orthologue’ is a term that applies to genes from two species. Orthologues are pairs of genes that descended from a single gene in the last common ancestor (LCA) of two species (Figure 1A & B). An orthogroup is the natural extension of the concept of orthology to groups of species. An orthogroup is the group of genes descended from a single gene in the LCA of a group of species (Figure 1A). When looking at the gene tree, the first divergence between the genes in an orthogroup is a speciation event and the same is true for orthologues.

As a result of gene duplication events, it is possible to have multiple genes from the same species when looking at either orthologues and orthogroups. In the example (Figure 1A & B), the human gene HuA has two genes that are orthologues of it in chicken, ChA1 and ChA2. Looking again at the orthogroup, we see that there are two chicken genes (Figure 1A) but only one gene from mouse and human. Some authors refer to the genes ChA1 and ChA2 as co-orthologues of HuA to emphasise the fact that there are multiple orthologues. These genes are nevertheless still orthologues and so we will usually just use this broader term. In fact, gene duplication events are so common that in addition to the one-to-many relationship implied by the term ‘co-orthologues’, there are frequently many-to-many relationships between orthologues. All of these relationships are identified by an OrthoFinder analysis.

Gene duplication events give rise to paralogues. Paralogues are pairs of genes that diverged from a single gene at a gene duplication event. The two chicken genes ChA1 and ChA2 are paralogues (Figure 1A & C). Two genes from different species can also be paralogues if they diverged from one another at a gene duplication event, although there are no examples of this in Figure 1. Since all branching events in a gene tree are either speciation events (that give rise to orthologues) or duplication events (that give rise to paralogues), any genes in the same orthogroup that are not orthologues must necessarily be paralogues.

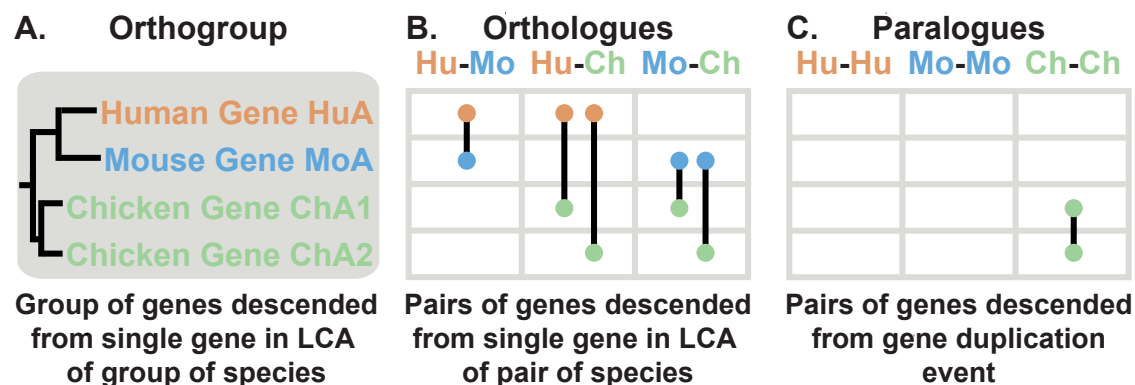


Figure 1: A hypothetical human, mouse and chicken orthogroup.

1.1 Why Orthogroups

If you followed the explanations above it will be clear that an orthogroup is just a gene family/clade of genes defined at a specific taxonomic level—namely, those genes descended from a single gene at the time of the LCA. Some may regard this definition of an orthogroup as unsatisfactory since an orthogroup can contain genes that are paralogues of one another (ChA1 is a paralogue of ChA2 in Figure 1). However, this definition of an orthogroup is the only logically consistent way of extending the concept of orthology to multiple species. If there have been gene duplication events it is not possible to create a single group of genes that contains all orthologues and only orthologues—try it with the example above!

One can still identify orthologues between the genes in each pair of species though, but the orthogroup is the correct unit of comparison when considering the group of species as a whole. In fact, one use for orthogroups is for identifying orthologues: The canonical way to identify orthologues is using a gene tree, and an orthogroup is exactly the set of genes that need to be in the gene tree in order to identify all orthologues. This is the method used by OrthoFinder.

2 Setting Up OrthoFinder

OrthoFinder runs on Linux and Mac, setup instructions are given below.

2.1 Set Up

1. Download the latest release from github: <https://github.com/davidemms/OrthoFinder/releases> (for this example we will assume it is OrthoFinder-1.0.6.tar.gz, change this as appropriate.)
2. In a terminal, `cd` to where you downloaded the package
3. Extract the files: `tar xzf OrthoFinder-1.0.6.tar.gz`
4. Test you can run OrthoFinder: `OrthoFinder-1.0.6/orthofinder -h`. OrthoFinder should print its ‘help’ text.

To perform an analysis OrthoFinder requires some dependencies to be installed and in the system path. Only the first two are needed to infer orthogroups and all four are needed to infer orthologues and gene trees as well. OrthoFinder is highly configurable and allows the expert user to chose any program they want for gene tree or multiple sequence alignment inference. The instructions in this section will concentrate only on the defaults, shown in bold:

1. **BLAST+** or Diamond
2. The **MCL** graph clustering algorithm
3. Either, **FastME** for distance-based tree inference.
Or, **MAFFT** and **FastTree** for multiple sequence alignment (MSA) based tree inference.
(Or, your favourite MSA and tree inference programs (see Section 5.7).
4. **DLCpar**

Brief instructions are given below although users can refer to the installation notes provided with these packages for more detailed instructions.

2.2 Dependencies

Each of the following packages provide their own detailed instructions for installation, here we give a concise guide.

2.2.1 BLAST+

NCBI BLAST+ is available in the repositories from most Linux distributions and so can be installed in the same way as any other package. For example, on Ubuntu, Debian, Linux Mint:

- `sudo apt-get install ncbi-blast+`

Alternatively, instructions are provided for installing BLAST+ on Mac and various flavours of Linux on the “Standalone BLAST Setup for Unix” page of the BLAST+ Help manual currently at <http://www.ncbi.nlm.nih.gov/books/NBK1762/>. Follow the instructions under “Configuration” in the BLAST+ help manual to add BLAST+ to the PATH environment variable.

2.2.2 MCL

The mcl clustering algorithm is available in the repositories of some Linux distributions and so can be installed in the same way as any other package. For example, on Ubuntu, Debian, Linux Mint:

- `sudo apt-get install mcl`

Alternatively it can be built from source which will likely require the ‘build-essential’ or equivalent package on the Linux distribution being used. Instructions are provided on the MCL webpage, <http://micans.org/mcl/>.

2.2.3 FastME

FastME can be obtained from <http://www.atgc-montpellier.fr/fastme/binaries.php>. The package contains a 'binaries/' directory. Choose the appropriate one for your system and copy it to somewhere in the system path e.g. '/usr/local/bin' and name it 'fastme'. I.e.:

- `sudo cp fastme-2.1.5-linux64 /usr/local/bin/fastme`

2.2.4 DLCpar

DLCpar can be downloaded from <http://compbio.mit.edu/dlcpar/> and installed as for a standard python package:

1. Download the latest version
2. Extract the package: `tar xzf dlcpar-1.0.tar.gz`
3. `cd dlcpar-1.0/`
4. `sudo python setup.py install`

2.3 Running OrthoFinder

Once the required dependencies have been installed, try running OrthoFinder on the example data:

1. `OrthoFinder-1.0.6/orthofinder -f ExampleDataset`

Assuming everything was successful OrthoFinder will end by printing the location of the results files, a short paragraph providing a statistical summary and the OrthoFinder citation. If you make use of OrthoFinder for any of your work then please cite it as this helps support future development.

If you have problems with this standalone binary version of OrthoFinder you can use the python source code version, which has a name of the form, 'OrthoFinder-1.0.6_source.tar.gz' and is available from the github 'releases tab'. See Section 2.4.2.

2.4 Setup for advanced use

The following steps are not required for the standard OrthoFinder use cases and are only needed if you want to: use Diamond as a significantly faster alternative to BLAST; infer gene trees using multiple sequence alignments; or you want to run OrthoFinder using the python source code version.

2.4.1 Trees from Multiple Sequence Alignments

To infer trees from multiple sequence alignments (instead of using the faster distance matrix approach with fastme) there are two additional dependencies which should be installed and in the system path:

1. MAFFT
2. FastTree

Alternatively, it is possible to configure OrthoFinder to use your own favourite MSA or tree inference program, see Section 5.7 for details.

2.4.2 Python Source Code Version

It is recommended that you use the standalone binaries for OrthoFinder which do not require python or scipy to be installed. However, the python source code version is available from the github 'releases' page (e.g. 'OrthoFinder-1.0.6_source.tar.gz') and requires python 2.7 and scipy to be installed. Up-to-date and clear instructions are provided here: <http://www.scipy.org/install.html>, be sure to chose a version using python 2.7. As websites can change, an alternative is to search online for "install scipy".

3 Performing Your Own OrthoFinder Analysis

Performing a complete OrthoFinder analysis is simple:

1. Download the amino acid sequences, in FASTA format, for the species you want to analyse. If you have the option, it is best to use a version containing a single representative/longest transcript-variant for each gene.
2. Optionally, you may want to rename the files to something simple since the filenames will be used as species identifiers in the results. E.g if you were using the ‘Homo_sapiens.GRCh38.pep.all.fa’ file you could rename it to ‘Homo_sapiens.fa’ or ‘Human.fa’.
3. Place the FASTA files all in a single directory.
4. To perform a complete OrthoFinder analysis requires just one command:
`orthofinder -f fasta_files_directory [-t number_of_threads]`

The argument ‘`number_of_threads`’ is an optional argument to specify the number of parallel threads to use for the BLAST searches, tree inference and reconciliation. As the BLAST queries can be a time-consuming step it is best to use at least as many BLAST processes as there are CPUs on the machine.

The OrthoFinder run will finish by printing the location of the results files, a short paragraph providing a descriptive statistical summary and the OrthoFinder citation. If you make use of OrthoFinder for any of your work then please cite it as this helps justify OrthoFinder support and future development. The OrthoFinder results files are described in Section 4.

4 Results Files

A standard OrthoFinder run produces a set of files describing the orthogroups, orthologues and gene trees for the set of species being analysed. Their locations are given at the end of an OrthoFinder run.

4.1 Results Files: Orthogroups

OrthoFinder generates the main orthogroup file, **Orthogroups.csv**, and two supporting files:

- **Orthogroups.csv** is a tab separated text file. Each row contains the genes belonging to a single orthogroup. The genes from each orthogroup are organized into columns, one per species.
- **Orthogroups_UnassignedGenes.csv** is a tab separated text file that is identical in format to Orthogroups.csv but contains all of the genes that were not assigned to any orthogroup.
- **Orthogroups.txt** (legacy format) is a second file containing the orthogroups described in the Orthogroups.csv file but using the OrthoMCL output format.

4.2 Results Files: Single-copy orthogroups & gene counts

Count-based orthogroup information is provided in:

- **SingleCopyOrthogroups.txt** contains a list of the orthogroups containing exactly one gene per species. Such orthogroups are very useful since they allow easy comparison across species. For example, alignments of single-copy orthogroups are used for almost all species tree inference methods.
- **Orthogroups.GeneCount.csv** gives the number of genes from each species in each orthogroup.

4.3 Results Files: Orthogroup Statistics

The statistics calculated from the orthogroup analysis provide the basis for any comparative genomics analysis. They are easily plotted and can also be used for quality control.

- **Statistics_Overall.csv** is a tab separated text file giving useful statistics from the orthogroup analysis.
- **Statistics_PerSpecies.csv** is a tab separated text file giving many of the same statistics as the 'Statistics_Overall.csv' file but on a species-by-species basis.
- **Orthogroups_SpeciesOverlaps.csv** is a tab separated text file containing a matrix of the number of orthogroups shared by each species-pair (i.e. the number of orthogroups which contain at least one gene from each of the species-pairs).

Most of the terms in the files **Statistics_Overall.csv** and **Statistics_PerSpecies.csv** are self-explanatory, the remainder are defined below.

- Species-specific orthogroup: An orthogroups that consist entirely of genes from one species.
- G50: The number of genes in the orthogroup such that 50% of genes are in orthogroups of that size or larger.
- O50: The smallest number of orthogroups such that 50% of genes are in orthogroups of that size or larger.
- Single-copy orthogroup: An orthogroup with exactly one gene (and no more) from each species. These orthogroups are ideal for inferring a species tree and many other analyses.
- Unassigned gene: A gene that has not been put into an orthogroup with any other genes.

4.4 Results Files: Orthologues

The orthologues spreadsheets are contained in sub-directories, one per species. Within these directories is one spreadsheet per species-pair giving all the inferred orthologues between those two species. The spreadsheets contain one column for the genes from one species and one column for genes from the other species. Orthologues can be one-to-one, one-to-many or many-to-many depending on the gene duplication events since the orthologues diverged (see Section 1 for more details). Each set of orthologues is cross-referenced to the orthogroup that contains them.

4.5 Results Files: Gene Trees and Species Tree

The gene trees for each orthogroup and the rooted species tree are in newick format and can be viewed using programs such as Dendroscope (<http://dendroscope.org/>) or FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>).

5 Advanced Usage

OrthoFinder provides a number of options to allow you to incrementally add and remove species and control other aspects of the analysis.

5.1 Controlling OrthoFinder Workflow

The OrthoFinder workflow can be controlled so as to stop or restart the analysis at different steps. It also allows gene trees to be inferred using distance matrices or multiple sequence alignments and alternative programs to be used in place of BLAST, MAFFT, FastTree etc. An overview of these options is given in Figure 5.1 and the main options are described in Section 5.6. Configuration instructions for using alternative programs is given in 5.7.

5.2 Adding Extra Species

OrthoFinder allows you to add extra species without re-running the previously computed BLAST searches:

- `orthofinder -b previous_orthofinder_directory -f new_fasta_directory`

This will add each species from the `new_fasta_directory` to existing set of species, reuse all the previous BLAST results, perform only the new BLAST searches required for the new species and recalculate the orthogroups. The `previous_orthofinder_directory` is the OrthoFinder ‘WorkingDirectory/’ containing the file ‘SpeciesIDs.txt’.

5.3 Removing Species

OrthoFinder allows you to remove species from a previous analysis. In the ‘WorkingDirectory/’ from a previous analysis there is a file called ‘SpeciesIDs.txt’. Comment out any species to be removed from the analysis by placing a ‘#’ character at the start of the line containing the species to be removed and then run OrthoFinder using:

- `orthofinder -b previous_orthofinder_directory`

where `previous_orthofinder_directory` is the OrthoFinder ‘WorkingDirectory/’ containing the file ‘SpeciesIDs.txt’.

5.4 Adding and Removing Species Simultaneously

The previous two options can be combined, comment out the species to be removed as described above and use the command:

- `orthofinder -b previous_orthofinder_directory -f new_fasta_directory`

5.5 User-specified Species Tree

The inference of orthologues is performed using gene-tree—species-tree reconciliation. The inference is therefore affected by the species-tree used, although the reconciliation process used does make it relatively robust to small differences in the tree topology. OrthoFinder will infer the species-tree automatically but if you know the correct, rooted species-tree you can request that OrthoFinder use it using the “-s” option:

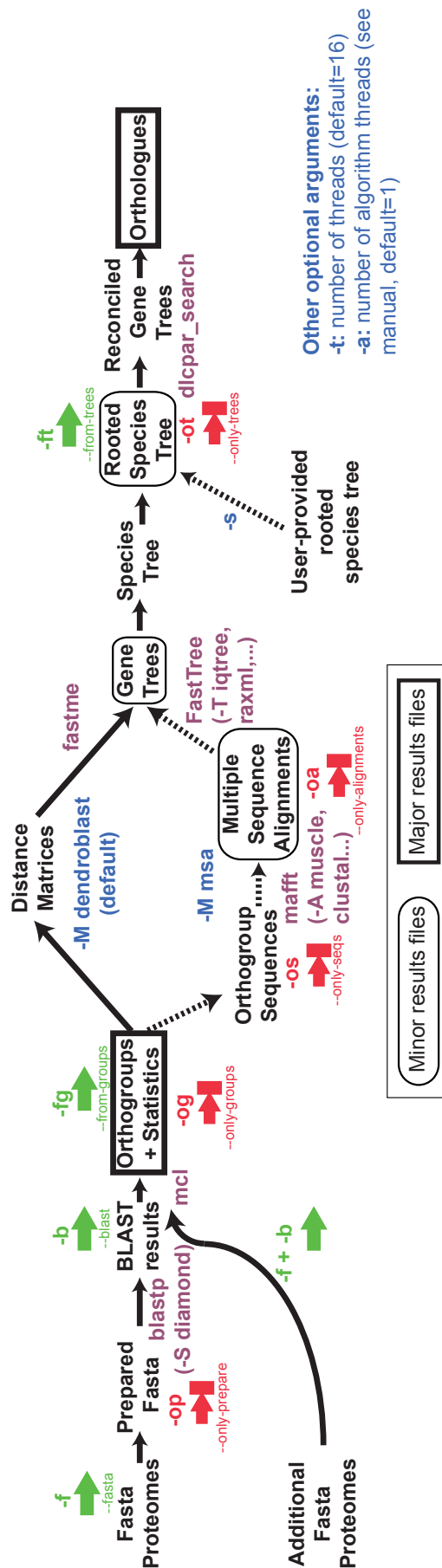
- `orthofinder -f fasta_dir -s species_tree`

A particularly handy use case is to reperform just the final orthologue inference step of the OrthoFinder analysis using an edited species-tree. This is useful if you want to see the effect of a different species-tree topology or rooting on the orthologues that are inferred. This allows you to skip all the previous steps (the orthogroups and gene-trees are unaffected by the species-tree used):

Controlling the OrthoFinder Analysis

If you just want to run a full analysis automatically,
use: 'orthofinder -f fasta_dir'

Control where OrthoFinder starts
Command-line switch takes directory containing files as argument.
Control where OrthoFinder stops (e.g. -ot = 'only up to & including trees')
Optional arguments
Programs required for each step (alternatives)



Example commands:

-f <fasta_dir>

Perform a complete OrthoFinder analysis on the proteomes contained in **fasta_dir**, use the default dendroblast method to infer gene trees.

-fg <orthogroups_dir> -ot

Infer gene trees for the orthogroups in **orthogroups_dir**, the rooted species tree and the all orthologues (use dendroblast for gene trees).

-f <fasta_dir> -b <previous_blast_results_dir> -M msa -oa

Reinfer orthogroups by adding the species from **fasta_dir** to species in **previous_blast_results_dir** and infer MSAs for each orthogroup.

-f <fasta_dir> -t 64 -M msa

Perform a complete OrthoFinder analysis on the proteomes contained in **fasta_dir**, use gene trees inferred from **multiple sequence alignments** and 64 threads.

Figure 2: The options controlling the OrthoFinder workflow

- **orthofinder -ft orthologues_results_dir -s species_tree**

The species-tree should be a rooted binary tree in Newick format, any branch lengths are ignored. The species names should match the names of the input fasta files containing the genes for that species with the filename extension removed. For an example see the species-tree produced by running OrthoFinder on any dataset, "SpeciesTree.rooted.txt". For the example dataset a suitable species-tree would look like this:

```
((Mycoplasma_hyopneumoniae:0.474995,Mycoplasma_agalactiae:0.456095),
(Mycoplasma_genitalium:0.434881,Mycoplasma_gallisepticum:0.412451));
```

5.6 Starting/Stopping OrthoFinder at Different Stages

The main argument to the OrthoFinder program ("**-f**", "**-b**", "**-fg**" or "**-ft**") controls at what point an OrthoFinder analysis is (re)started. The main workflow is:

FASTA Files $\xrightarrow{-f}$ BLAST Search Results $\xrightarrow{-b}$ Orthogroups $\xrightarrow{-fg}$ Gene Trees $\xrightarrow{-ft}$ Orthologues.

and the captions above the arrows show the starting point corresponding to each of the arguments (from FASTA, from BLAST, from groups & from trees). In each case the argument should be followed by the name of the directory containing the relevant files as explained below.

- **-f fasta_dir**: Directory containing the fasta files.
E.g. `orthofinder -f /home/david/ExampleDataset`.
- **-b blast_results_dir**: Directory containing the Blast*.txt result files.
E.g. `orthofinder -b /home/david/ExampleDataset/Results_Nov30/WorkingDirectory/`.
- **-fg orthogroup_results_dir**: Directory containing the orthogroup results files, "Orthogroups.csv".
E.g. `orthofinder -fg /home/david/ExampleDataset/Results_Nov30/`.
- **-ft orthologues_results_dir**: Directory containing the orthologues results including the "Gene_Trees" directory.
E.g. `orthofinder -ft /home/david/ExampleDataset/Results_Nov30/Orthologues_Nov30`.

Note, the "**-f**" and "**-b**" arguments can be combined to allow new species to be added to an analysis without needing to redo any of the (time-consuming) all-versus-all BLAST searches that OrthoFinder has already performed, see Section 5.2 for details.

The "**-og**" (only groups) option can be used to perform an analysis that only goes as far as the orthogroup inference and does not infer gene trees and orthologues. The option does not take any arguments:

- **orthofinder -f /home/david/ExampleDataset -og**

The "**-op**" (only prepare) option is used to only prepare the fasta files prior to the BLAST search and is described in Section 5.10.

5.7 User-specified MSA, Tree Inference or Sequence Search Program

OrthoFinder allows you to use your favourite program for sequence searches (in place of BLAST), MSA or tree inference. For example, you can use Diamond as a significantly faster alternative to BLAST. The command line arguments are used:

- **-S search_program**: Program to use for all-versus-all searches instead of BLAST.
- **-A msa_program**: Program to use for MSA inference (requires "**-M msa**" option).
- **-T tree_program**: Program to use for tree inference (requires "**-M msa**" option).

The options available for each of these arguments can be seen by calling "`orthofinder -h`" to display the help file. To add a program that is not currently supported you just need to add an entry to the `config.json` file in the same directory as the orthofinder executable and it will automatically appear in the OrthoFinder help file. It should be straight forward to follow the examples already contained in the config file but for a description of the file follows.

5.7.1 Config File

The `config.json` file is in the standard, 'human-readable' json format. An example is given for “muscle” below and another slightly more complicated example for “iqtree”. The iqtree version has an extra entry “output_filename” since IQ-Tree automatically names the output file rather than allowing the user to specify what filename to use for the output.

```
"muscle":{
  "program_type": "msa",
  "cmd_line": "muscle -in INPUT -out OUTPUT"
},

"iqtree":{
  "program_type": "tree",
  "cmd_line": "iqtree -s INPUT -pre PATH/IDENTIFIER > /dev/null",
  "output_filename": "PATH/IDENTIFIER.treefile"
},
```

To add an new MSA or tree inference program you need to specify:

- The name that will be used to refer to it on the OrthoFinder command line (muscle or iqtree for the examples)
- The “program_type”, options are msa, tree and search.
- The “cmd_line”, to be used to call the program.
- The “output_filename” that the program will name the multiple sequence alignment or tree. This is only required if the program does not allow you to specify what filename it should use.

You can use the variables:

- INPUT: The full path of the input filename (fasta file of sequences for and msa method, multiple sequence alignment fasta file for tree method)
- BASENAME: Just the filename without the directory path. A number of methods use this to name the output file automatically. If this is the case then use the BASENAME variable to specify what the “output_filename” will be.”,
- PATH : Path to the directory containing the input file
- OUTPUT: The user specified output filename without any directory path.
- IDENTIFIER: A name generated by OrthoFinder to uniquely identify the orthogroup (a number of methods use this to name the output file automatically, see RAxML command for an example). Not applicable for “program_type” “search”.
- DATABASE: For the search program_type, for use in the search_cmd. The full path of the database to search against.

For a sequence search program (i.e. an alternative to BLAST), an example entry would look like this:

```
"diamond":{
  "program_type": "search",
  "db_cmd": "diamond-sse2 makedb --in INPUT -d OUTPUT",
  "search_cmd": "diamond-sse2 blastp -d DATABASE -q INPUT -o OUTPUT --sensitive -p 1 --quiet -e 0.
},
```

The fields specific to this are:

- “db_cmd”: The command line to create a sequence database to search against (replacing makeblastdb).
- “search_cmd”: The command line used to search against a blast database (replacing blastp).

5.8 Inferring MSA Gene Trees

This replaces the functionality previously provided by the `trees_from_MSA` utility.

To infer gene trees for each orthogroup using multiple sequence alignments use the option `"-M msa"`. This will use MAFFT (MAFFT LINSI for orthogroups with fewer than 500 sequences) to generate the multiple sequence alignments and FastTree to generate the gene trees. Both of these programs need to be installed and in the system path. See section 5.7 for details on using different programs for multiple sequence alignment or tree inference.

5.9 Parallelising OrthoFinder Algorithm (-a option)

There are two separate options for controlling the parallelisation of OrthoFinder. The `'-t'` option should always be used whereas RAM requirements may affect whether you use the `'-a'` option or not.

1. `'-t number_of_threads'`: This option should always be used. It makes the BLAST searches, the tree inference and gene-tree reconciliation run in parallel. These are all highly-parallelisable and the BLAST searches in particular are by far the most time-consuming task. You should use as many threads as there are cores available.
2. `'-a number_of_orthofinder_threads'` The remainder of the algorithm, beyond these highly-parallelisable tasks, is relatively fast and efficient and so this option has less overall effect. It is most useful when running OrthoFinder using pre-calculated BLAST results since the time savings will be more noticeable in this case. Using this option will also increase the RAM requirements (see below).

RAM availability is an important consideration when using the `'-a'` option. Each thread loads all BLAST hits between one species and all sequences in all other species. To give some very approximate numbers, each thread might require:

- 0.02 GB per species for small genomes (e.g. bacteria)
- 0.04 GB per species for larger genomes (e.g. vertebrates)
- 0.2 GB per species for even larger genomes (e.g. plants)

I.e. running an analysis on 10 vertebrate species with 5 threads for the OrthoFinder algorithm (`-a 5`) might require $10 \times 0.04 = 0.4$ GB per thread and so $5 \times 0.4 = 2$ GB of RAM in total. If you have the BLAST results already then the total size of all the `Blast*.0.txt` files gives a good approximation of the memory requirements per thread. Additionally, the speed at which files can be read is likely to be the limiting factor when using more than 5-10 threads on current architectures so you may not see any increases in speed beyond this.

5.10 Running BLAST Searches Separately

The `'-p'` option will prepare the files in the format required by OrthoFinder and print the set of BLAST commands that need to be run.

- `orthofinder -f fasta_files_directory -p`

This is useful if you want to manage the BLAST searches yourself. For example, you may want to distribute them across multiple machines. Once the BLAST searches have been completed the orthogroups can be calculated using the `'-b'` command as described in Section 5.11.

5.11 Using Pre-Computed BLAST Results

It is possible to run OrthoFinder with pre-computed BLAST results provided they are in the correct format. They can be prepared in the correct format using the `'-p'` command and, equally, the files from a previous OrthoFinder run are also in the correct format to rerun using the `'-b'` option. The command is simply:

- `orthofinder -b directory_with_processed_fasta_and_blast_results`

If you are running the BLAST searches yourself it is strongly recommended that you use the ‘-p’ option to prepare the files first (see Section 5.10). Should you need to prepare them manually, the required files and their formats are described in the appendix (for example, if you already have BLAST search results from another source and it will take too much computing time to redo them).

5.12 Using the Orthoxml Format

Orthogroups can be output in XML using the (bulky) orthoxml format. This is requested by adding ‘-x speciesInfoFilename’ to the command used to call orthofinder, where speciesInfoFilename should be the filename (including the path if necessary) of a user-prepared file providing the information about the species that is required by the orthoxml format. This file should contain one line per species and each line should contain the following 5 fields separated by tabs:

1. **FASTA filename:** the filename (without path) of the FASTA file for the species described on this line
2. **species name:** the name of the species
3. **NCBI Taxon ID:** the NCBI taxon ID for the species
4. **source database name:** the name of the database from which the FASTA file was obtained (e.g. Ensembl)
5. **database FASTA filename:** the name given to the FASTA file by the database (e.g. Homo_sapiens.NCBI36.52.pep.all.fa)

As an example, a single line of the file could look like this (where each field has been separated by a tab rather than just spaces):

```
HomSap.fa Homo sapiens 36 Ensembl Homo_sapiens.NCBI36.52.pep.all.fa
```

Information on the orthoxml format can be found here: http://orthoxml.org/0.3/orthoxml_doc_v0.3.html

5.13 Regression Tests

A set of regression tests are included in the directory ‘Tests’ available from the github repository. They can be run by calling the script ‘test_orthofinder.py’. They currently require version 2.2.28 of NCBI BLAST and the script will exit with an error message if this is not the case.

6 Appendix: File Format for Pre-Computed BLAST Results

If you want to run the BLAST searches outside of OrthoFinder and you have not already computed them then by far the best option is to use the ‘prepare’ option, ‘-p’. This will prepare the files and you can then run the BLAST searches in whatever way you wish and then run OrthoFinder on them using the ‘-b’ option. If you already have a set of BLAST search results that you want to convert into the format that OrthoFinder uses then the details are given below.

The files that must be in `directory_with_processed_fasta_and_blast_results` are:

- a FASTA file for each species
- a BLAST results file for each species pair
- SequenceIDs.txt
- SpeciesIDs.txt

Examples of the format required for the files can be seen by running OrthoFinder on the supplied ‘ExampleDataset’ and looking in the ‘WorkingDirectory/’ created. A description is given below.

6.1 FASTA Files

```
Species0.fa
Species1.fa
...
```

Within each FASTA file the accessions for the sequences should be of the form ‘x-y’ where x is the species ID number, matching the number in the filename and y is the sequence ID number, which starts from 0 within each species. So the first few lines of start of ‘Species0.fa’ would look like:

```
>0_0
MFAPRGK...

>0_1
MFAVYAL...

>0_2
MTTIID...
```

And the first few lines of start of ‘Species1.fa’ would look like:

```
>1_0
MFAPRGK...

>1_1
MFAVYAL...

>1_2
MTTIID...
```

6.2 BLAST Results Files

For each species pair ‘x’, ‘y’ there should be a BLAST results file ‘Blastx.y.txt’ where x is the index of the query FASTA file and y is the index of the species used for the database. Similarly, there should be a BLAST results file ‘Blasty.x.txt’ where y is the index of the query FASTA file and x is the index of the species used for the database. The tabular BLAST output format 6 should be used. The query and hit IDs in the BLAST results files should correspond to the IDs in the FASTA files.

Aside, reducing BLAST computations: Note that since the BLAST queries are by far the most computationally expensive step, considerable time could be saved by only performing $\frac{n(n+1)}{2}$ of the

species versus species BLAST queries instead of n^2 , where n is the number of species. This would be done by only searching ‘Speciesx.fa’ against the BLAST database generated from ‘Speciesy.fa’ if $x \leq y$. The results would give the file ‘Blastx_y.txt’ and then this file could be used to generate the ‘Blasty_x.txt’ file by swapping the query and hit sequence on each line in the results file. This should have only a small effect on the generated orthogroups.

6.3 SequenceIDs.txt

The SequenceIDs.txt give the translation from the IDs of the form x_y to the original accessions. An example line would be:

```
0_42: gi|290752309|emb|CBH40280.1|
```

The IDs should be in order, i.e.

```
0_0: gi|290752267|emb|CBH40238.1|
0_1: gi|290752268|emb|CBH40239.1|
0_2: gi|290752269|emb|CBH40240.1|
...
...
1_0: gi|284811831|gb|AAP56351.2|
1_1: gi|284811832|gb|AAP56352.2|
...
```

6.4 SpeciesID.txt

The SpeciesIDs.txt file gives the translation from the IDs for the species to the original FASTA file, e.g.:

```
0: Mycoplasma_agalactiae.faa
1: Mycoplasma_gallisepticum.faa
2: Mycoplasma_genitalium.faa
3: Mycoplasma_hyopneumoniae.faa
```