OrthoFinder3

OrthoFinder identifies orthogroups, infers gene trees for all orthogroups, and analyzes the gene trees to identify the rooted species tree. The method subsequently identifies all gene duplication events in the complete set of gene trees, and analyses them at both gene tree and species tree level. OrthoFinder further analyzes all of this phylogenetic information to identify the complete set of orthologs between all species, and provide extensive comparative genomics statistics.

Table of contents

A single PDF with all documentation and tutorial is available here

Installation

Install in conda (recommended)

The easiest way to install OrthoFinder is using conda.

```
conda create -n of3_env python=3.10
conda activate of3_env
conda install orthofinder
```

Alternatively, you could install via github, or download the source code and install locally.

Install via github

```
python3 -m venv of3_env
. of3_env/bin/activate
pip install git+https://github.com/OrthoFinder/OrthoFinder.git
```

Install locally from source code

The following commands provide three ways to download the source code of OrthoFinder locally into a directory named `OrthoFinder`.

```
# Download via git
git clone https://github.com/OrthoFinder/OrthoFinder.git
# or download the zipfile and unzip it into OrthoFinder
mkdir OrthoFinder && wget -qO- https://github.com/OrthoFinder/OrthoFinder/archive/refs/heads/maste
# or download the tar.gz and unzip it into OrthoFinder
mkdir OrthoFinder && wget -qO- https://github.com/OrthoFinder/OrthoFinder/releases/download/v3.0.1
```

Next, you can run the following commands to install OrthoFinder inside the of3_env virtural environment.

```
cd OrthoFinder
python3 -m venv of3_env && . of3_env/bin/activate
pip install .
```

Whether you've installed OrthoFinder directly from GitHub or downloaded and set it up locally, the OrthoFinder package will only be available within the `of3_env` virtual environment. This avoids potential conflicts with Python dependencies.

To deactivate the virtual environment when you are finished, run:

```
deactivate
```

Test your installation

Once you have installed OrthoFinder, you can print the help information and version, and test it on the example data.

```
orthofinder --help # Print out help informatioin
orthofinder --version # Check the version
orthofinder -f ExampleData # Test OrthoFinder on an example dataset
```

Uninstalling

To uninstall on conda:

```
conda deactivate
conda remove -n of3_env --all
```

To remove the virtual environment where OrthoFinder is installed:

```
deactivate
cd ..
rm -rf OrthoFinder
```

Simple Usage

Run OrthoFinder on FASTA format proteomes in `<dir>`

```
orthofinder [options] -f <dir>
```

OrthoFinder requires one FASTA file for each species. Each file should contain the complete set of protein sequences from that species' genome, with a single representative sequence for each gene.

If your files have multiple transcript variants for each gene, then we provide a script `primary_transcripts.py` to extract the longest variant per gene. This script should be run on your files prior to running OrthoFinder;

```
for f in *fa ; do python primary_transcript.py $f ; done
```

Advanced Usage - Scaling to Thousands of Species

If you are analysing >100 species, we reccommend that you use the scalable implementation.

Add the files for 64 species into one directory `<core>` Add the remaining files into another directory `<additional>`

First, run OrthoFinder on the subset of 64 species

```
orthofinder [options] -f <core>
```

Then, add the additional species to the results of the core run

```
orthofinder [options] --assign <additional> --core <Results_Dir>
```

To choose which 64 species to include in the core, aim to capture a broad range of the evolutionary diversity of your species.

Note that this alternative way of running OrthoFinder requires that the core species are run using the multiple sequence alignment option. You cannot add additional species to OrthoFinder results that were run with the `-M dendroblast` option, which was the default for OrthoFinder2

Command-line options

Command-line options for OrthoFinder

Command-line options

Adding additional species

Parameter	Description
`--assign <dir1> --core <dir2>`	Assign species from `<dir1>` to existing orthogroups in `<dir2>`.

Method choices

Parameter	Description	Default	Options
`-M`	Method for gene tree inference.	`msa`	`dendroblast`, `msa`
`-S`	Sequence search program	`diamond`	`blast`, `diamond`, `diamond_ultra_sens`, `blastp`, `mmseqs`, `blastn`
`-A`	MSA program, requires `-M msa`	`famsa`	`famsa`, `mafft`, `muscle`
`-T`	Tree inference method, requires `-M msa`	`fasttree`	`fasttree`, `fasttree_fastest`, `raxml`, `iqtree`
`-I`	MCL inflation parameter	`1.2`	`1-10`

Input options

Parameter	Description
`-d`	Input is DNA sequences.
`-s`	User-specified rooted species tree.

Output options

Parameter	Description
`-X`	Don't add species names to sequence IDs.
`-n <txt>`	Name to append to the results directory.
`-o <txt>`	Specify a non-default results directory.

Parallel processing options

Parameter	Description	Default
`-t`	Number of parallel sequence search threads.	`All available`
`-a`	Number of parallel analysis threads.	`16 or t/8 (whichever is lower)`

Workflow stopping options

Parameter	Description
`-op`	Stop after preparing input files for BLAST.

Workflow restart options

Parameter	Description
`-b <dir>`	Start OrthoFinder from pre-computed BLAST results in `<dir>`.

Other options

Parameter	Description
`-1`	Only perform one-way sequence search.
`-z`	Don't trim MSAs (columns ≥ 90% gap, min. alignment length 500).
`-y`	Split paralogous clades below the root of a HOG into separate HOGs.
`-h`	Print this help text.
`-v`	Print version.

Output files

A standard OrthoFinder run produces a set of files describing the orthogroups, orthologs, gene trees, resolve gene trees, the rooted species tree, gene duplication events, and comparative genomic statistics for the set of species being analysed. These files are located in an intuitive directory structure.

Full details on the output files and directories can be found in the PDF here. The directories that are useful for most users are;

`/Orthogroups` - Orthogroups.tsv is the main orthogroup file. Each row contains the genes belonging to a single orthogroup. The genes from each orthogroup are organized into columns, one per species. - Orthogroups.txt is a text file with each line showing the genes in a single orthogroup. It differs from Orthogroups.tsv in that it doesn't show the species which each gene belongs to. - Orthogroups.GeneCount.tsv is a tab separated text file that contains counts of the number of genes for each species in each orthogroup. - Orthogroups_SingleCopyOrthologues.txt is a list of orthogroups that contain exactly one gene per species - Orthogrouops_UnassignedGenes.tsv is a tab separated text file that contains all of the genes that were not assigned to any orthogroup.

`/Phylogenetic_Hierarchical_Orthogroups` - Each file is a phylogenetic hierarchical orthogroup (HOG) for a different node of the species tree - Each row of a file contain the genes belonging to a single orthogroup - Each species is represented by a single column

`/Orthologues` - Each species has a sub-directory that in turn contains a file for each pairwise species comparison, listing the orthologs between that species pair.

`/Comparative_Genomics_Statistics` - Files containing summary statistics across all orthogroups, as well as comparisons between each pair of species

`/Resolved_Gene_Trees` - A rooted phylogenetic tree inferred for each orthogroup with 4 or more sequences and resolved using the OrthoFinder hybrid species-overlap/duplication-loss coalescent model.

`/Species_Tree` - SpeciesTree_rooted.txt = A species tree inferred using ASTRAL-Pro. - SpeciesTree_rooted_node_lab = The same tree, but with nodes labels instead of support values. This labelled version is useful for interpreting and analysing the results of the gene duplication analyses.

`/Gene_Duplication_Events` - `Duplications.tsv` has a row for each gene duplication event, with information on orthogroup in which it occured, the species that contain the duplicated gene, the node in the species tree on which the gene duplication event occured, and the support score for the gene duplication event. - `SpeciesTree_Gene_Duplications_0.5_Support.txt` provides a summation of the above duplications over the branches of the species tree.

`/Orthogroup_Sequences` - A FASTA file for each orthogroup giving the amino acid sequences for each gene in the orthogroup.

Latest additions

The current version of OrthoFinder has several major changes comapred to OrthoFinder version 2 (Emms & Kelly 2019)

New workflow for scalability

The `--core --assign` workflow uses SHOOT to create profiles for previously computed orthogroups, and adds new genes to these orthogroups without requiring a costly all-versus-all sequence search. Genes that cannot be assigned using the SHOOT approach are analysed using a standard OrthoFinder workflow.

Phylogenetic Hierarchical Orthogroups

OrthoFinder has now extended its phylogenetic approach to orthogroups, allowing orthogroups to be defined for each node within the species tree. This significantly increases the accuracy of orthogroups, and enables users to perform orthogroup analyses for any clade of species in the species tree.

Citation

The manuscript “OrthoFinder: scalable phylogenetic orthology inference for comparative genomics” is available as a preprint on biorxiv.

Emms & Kelly (2015) introduced the orthogroup inference method.

Emms & Kelly (2019) introduced the phylogenetic inference of orthologs, including rooted gene and species trees, and gene duplication events.

Emms & Kelly (2017) introduced the STRIDE method to root an unrooted species tree.

Emms & Kelly (2017) introduced the STAG method of species tree inference.

Meet the team

OrthoFinder was developed by David Emms & Steve Kelly

Current members of the OrthoFinder team:

Yi Liu, Jonathan Holmes, Laurie Belcher

Beginner tutorial for using OrthoFinder

This tutorial will cover;

1) Downloading OrthoFinder
2) Downloading input files for OrthoFinder
3) Running OrthoFinder
4) Exploring the results of OrthoFinder

All these steps will be done on the command line so that you can just copy and paste the commands yourself. If you are not familiar with the command line there are many online tutorials and reference pages, here is a nice short one that covers the basics: https://www.techspot.com/guides/835-linux-command-line-basics/

1) Downloading OrthoFinder

There are two main ways of getting OrthoFinder. You can either use conda, or you can install it directly from github. Installing directly from github will always give you the latest version, but you might have to manually install other software that OrthoFinder is dependent on, and it can be trickier to troubleshoot if you aren't familiar with the command line. Conda automates the installation process and handles all dependencies, making it very
beginner-friendly.

To install via conda, we first need to install miniconda. Follow the instructions here https://docs.anaconda.com/miniconda/
We then need to run these commands

```
conda config --add channels defaults conda config --add channels bioconda conda config --add channels conda-forge
conda create -n orthofinder
conda activate orthofinder
conda install orthofinder
```

If you are on one of the newer Macs with the new chips (M1/M2/M3), you will need to follow a few extra steps to use conda
https://towardsdatascience.com/how-to-manage-conda-environments-on-an-apple-silicon-m 1-mac-1e29cb3bad12

To install directly from github, we need to run these commands

```
python3 -m venv of3_env
. of3_env/bin/activate
pip install git+https://github.com/OrthoFinder/OrthoFinder.git
```

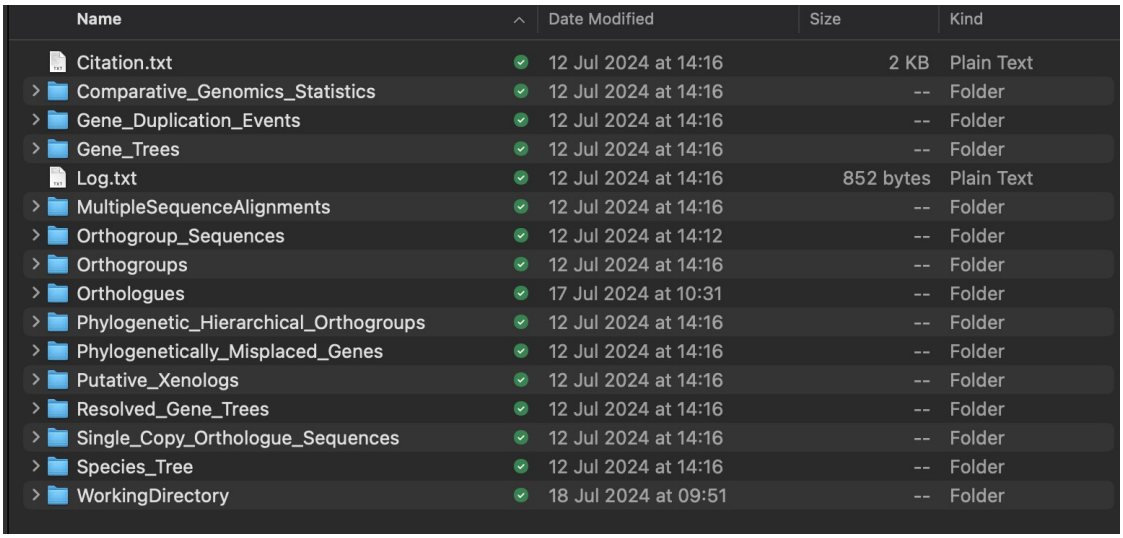
You can test that OrthoFinder has been installed by printing its help file
```orthofinder -h```, which will print all of the command line options
```

You can test that OrthoFinder is working correctly by running it on the example dataset, which you can download from our github
https://github.com/OrthoFinder/OrthoFinder/

```orthofinder -f ExampleData/

OrthoFinder will print lots of information to the command line as it runs. If you get an error message, the best way to troubleshoot is to just google the error message. You can also look on the github issues page for OrthoFinder
https://github.com/OrthoFinder/OrthoFinder/

When OrthoFinder has finished running, it will generate a folder containing the output, with the folder named according to today's date.
'ExampleData/OrthoFinder/Results_Oct1
1' The folder looks like this:



| Name | Date Modified | Size | Kind |
|---|---|---|---|
| Citation.txt | 12 Jul 2024 at 14:16 | 2 KB | Plain Text |
| Comparative_Genomics_Statistics | 12 Jul 2024 at 14:16 | -- | Folder |
| Gene_Duplication_Events | 12 Jul 2024 at 14:16 | -- | Folder |
| Gene_Trees | 12 Jul 2024 at 14:16 | -- | Folder |
| Log.txt | 12 Jul 2024 at 14:16 | 852 bytes | Plain Text |
| MultipleSequenceAlignments | 12 Jul 2024 at 14:16 | -- | Folder |
| Orthogroup_Sequences | 12 Jul 2024 at 14:12 | -- | Folder |
| Orthogroups | 12 Jul 2024 at 14:16 | -- | Folder |
| Orthologues | 17 Jul 2024 at 10:31 | -- | Folder |
| Phylogenetic_Hierarchical_Orthogroups | 12 Jul 2024 at 14:16 | -- | Folder |
| Phylogenetically_Misplaced_Genes | 12 Jul 2024 at 14:16 | -- | Folder |
| Putative_Xenologs | 12 Jul 2024 at 14:16 | -- | Folder |
| Resolved_Gene_Trees | 12 Jul 2024 at 14:16 | -- | Folder |
| Single_Copy_Orthologue_Sequences | 12 Jul 2024 at 14:16 | -- | Folder |
| Species_Tree | 12 Jul 2024 at 14:16 | -- | Folder |
| WorkingDirectory | 18 Jul 2024 at 09:51 | -- | Folder |

We'll discuss how to interpret and analyse these files and folders later on, in the 'Exploring the results' section of the tutorial.

### 2) Downloading input files for OrthoFinder

OrthoFinder requires as input the amino acid sequences for all the protein coding genes in your species of interest.

For this tutorial, we will assume that you have your files ready.

We provide a separate detailed tutorial for getting input files for OrthoFinder

### 3) Running OrthoFinder
```

You can now run OrthoFinder

First, you have to open a terminal and navigate to the directory where your files

are. You can now run OrthoFinder on your proteomes.

```orthofinder -f primary_transcripts ```

That's it! OrthoFinder will print updates on its progress to the terminal, and tell you when it's finished. If you get an error message, the best first step is to google the error message. You can also head over to the issues page on our github.

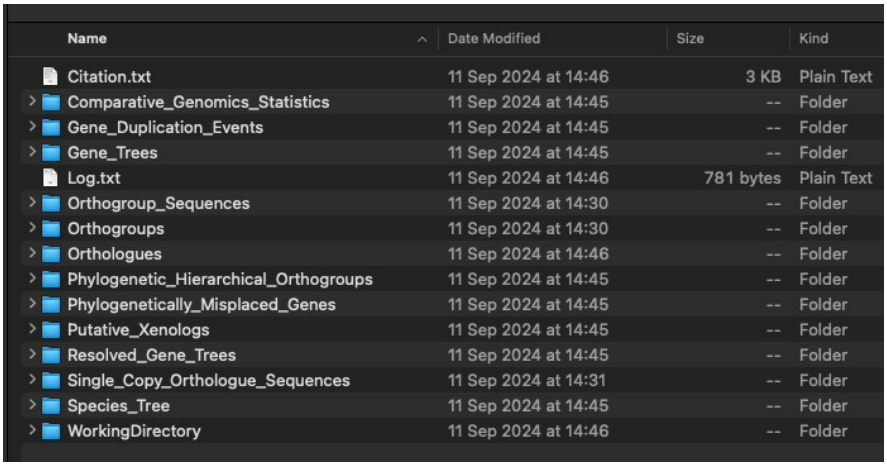
The command above will run OrthoFinder on default settings. To see what options you might want to adjust for your own data, check out our github page, and the advanced tutorial below

4) Exploring the results of OrthoFinder

In the last tutorial, we downloaded proteomes, pre-processed the files, and ran OrthoFinder on them.

Now, we are going to explore the results

OrthoFinder creates a results directory named OrthoFinder inside the proteome directory, and puts the results here. My results directory looks like this:



<u>Step 1: Quality Control</u>
Before we start diving into the orthogroups, it would behove us to check the quality of the OrthoFinder run. We want to make sure that most genes across all species have been assigned to orthogroups, and that the species tree looks realistic.
```

Open the file Statistics_Overall.tsv from the folder ‘Comparative_Genomics_Statistics’. This file can be opened in spreadsheet software like Microsoft Excel, or in a text editor like Notepad.

On the 5th line, we can see the ‘Percentage of genes in orthogroups’, which in my case is 95.7%.

	A	B
1	Number of species	6
2	Number of genes	107980
3	Number of genes in orthogroups	103302
4	Number of unassigned genes	4678
5	Percentage of genes in orthogroups	95.7
6	Percentage of unassigned genes	4.3

A good rule of thumb is that this number should be >80%. If not, you are likely missing some orthology relationships that actually exist. The best way to fix this would be better species sampling.

Now open the file ‘Statistics_PerSpecies’, from the same folder. This file gives us the % of genes in each species that are assigned to orthogroups, rather than the percentage for all genes across species.

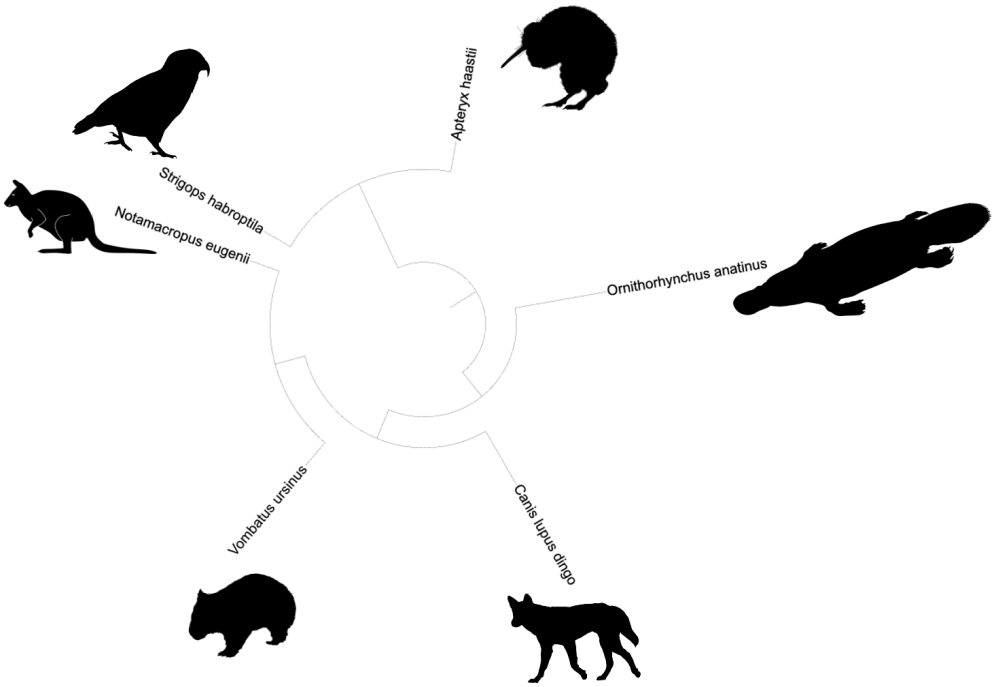
You can see here that we capture most genes across all species.

	A	B	C	D	E	F	G
1		Apteryx_haastii	Canis_lupus_dingo	Notamacropus_eugenii	Ornithorhynchus_anatinus	Strigops_habroptila	Vombatus_ursinus
2	Number of genes	16674	21360	15290	17418	16037	21201
3	Number of genes in orthogroups	15675	20093	14748	16824	15474	20488
4	Number of unassigned genes	999	1267	542	594	563	713
5	Percentage of genes in orthogroups	94	94.1	96.5	96.6	96.5	96.6
6	Percentage of unassigned genes	6	5.9	3.5	3.4	3.5	3.4

The lowest percentage is the kiwi (*A. haastii*), but we still managed to assign 94% of its genes to orthogroups. The key message here is that it’s always a good idea to look at this information before you start interpreting your results. If the numbers were too low for one species, we might want to consider sampling more species to fill in the long evolutionary divergence between species (e.g. something in between a Kiwi and a Kakapo, such as a Hoatzin).

One more useful thing to do before we really start to dive in is to look at the species tree. Go to the website https://itol.embl.de/, and click ‘Upload a tree’.
You can then drag and drop the tree file, which is in ‘‘Species_Tree/SpeciesTree_rooted.txt’

You will now see the phylogenetic tree that OrthoFinder has produced. I have annotated my version with icons PhyloPic, so that we can see what is going on



We now want to do some common-sense checking that everything appears to be in order, and we aren't rewriting the history of life on earth. With our six species, this tree looks exactly as we would expect.

If the tree doesn't look correct, then this won't impact orthogroup inference, but will affect our measures of gene duplication, and might affect our assignment of orthologs and paralogs within an orthogroup. If you need to, you can run OrthoFinder with your own species tree (use the -s option).

Step 2: Interpreting results

Now that we are happy with our OrthoFinder run, we can start diving into the results.

Orthologues

We will start by finding orthologues of a gene that we are interested in.

We will focus on the gene ENSVURG00010002700.1 in wombats, which is an olfactory receptor. Let's find out what its orthologues are in the Tammar wallaby.

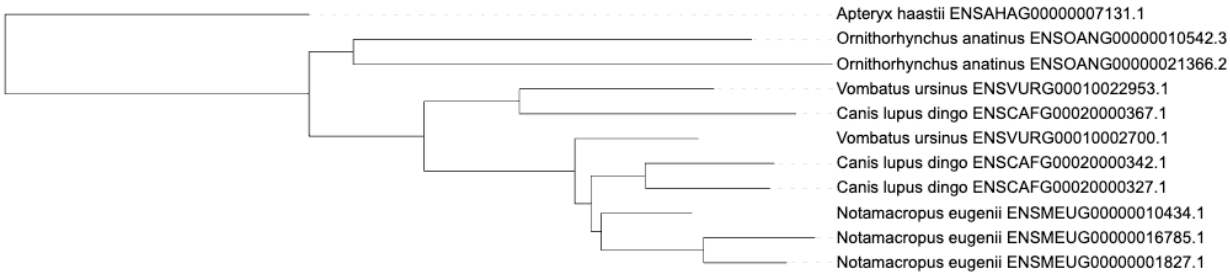
In the Orthologues directory there is a sub-directory for each species. Open 'Orthologues/Orthologues_Vombatus_ursinus/Vombatus_ursinus__v__Notamacropus_eugeni i.tsv', in a spreadsheet program (specifying that it's tab-delimited if necessary). The file has three columns, "Orthogroup", "Vombatus_ursinus", and "Notamacropus_eugenii". Find 'ENSVURG00010002700.1' in the table, I can see that the gene is in orthogroup OG0000365 and that it has three orthologues in wallabies: ENSMEUG00000016785.1, ENSMEUG00000001827.1, ENSMEUG00000010434.1

Gene trees

Next, we are going to look at the gene tree to see how these orthologues arose. OrthoFinder infers orthlologues from 'resolved' gene trees using a Duplication-Loss-Coalescence analysis to identify the more parsimonious interpretation of the tree (see the OrthoFinder2 paper for more details).

All of the gene trees are in one file (Resolved_Gene_Trees /Resolved_Gene_Trees.txt). Each line of the file contains the ID of an orthogroup (e.g. OG0000365:), followed by the gene tree for that orthogroup. To find the tree for certain orthogroup, just search for the orthogroup ID.

We are going to view the tree for OG0000365 on itol https://itol.embl.de/



Looking at the gene tree, we can see that there have been several gene duplications in the lineage leading to wallabies (Notamacropus). This has resulted in a one-to-three orthology relationship, i.e. all three of the wallaby genes are equally related to the wombat gene ENSVURG00010002700.1. It's often the case that orthology relationships aren't one-to-one, and it's important to know this—you don't want to spend months doing experiments on 'the orthologue' only to find out later there are actually three!

Gene duplications

Having the gene trees means that OrthoFinder can identify all gene duplication events that occurred. There is a folder called 'Gene_Duplication_Events' that has two files that allow us to explore duplications. Let's first open 'Gene_Duplication_Events/SpeciesTree_Gene_Duplications_0.5_Support.txt' in itol.

Go into the 'Advanced' tab on the Control Panel and select 'Display' next to 'Node IDs:' to see the node labels

This gives a summary of gene duplication events. Each node shows the node name followed by an underscore and then the number of well-supported gene duplication events mapped to each node in the species tree. Gene-duplication events are considered 'well-supported' if at least 50% of the descendant species have retained both copies of the duplicated gene. For the common ancestor of the mammals, N2,

there were 812 of these well-supported gene duplication events. The numbers after the species names are the number of ‘terminal’ duplications that map to that species, rather than an internal node of the species tree. We can see the full list of gene duplication events in the file ‘Gene_Duplication_Events/Duplications.tsv’. Here are just a few lines from the file:

	A	B	C	D	E	F	G
1	Orthogroup	Species Tree	Gene Tree Node	Support	Type	Genes 1	Genes 2
2	OG0000000	N1	n0	1	Non-Terminal	Strigops_habroptila_ENSSH	Strigops_habroptila_ENSSHB
3	OG0000000	Strigops_hab	n2	1	Terminal	Strigops_habroptila_ENSSH	Strigops_habroptila_ENSSHB
4	OG0000000	Strigops_hab	n3	1	Terminal	Strigops_habroptila_ENSSH	Strigops_habroptila_ENSSHB
5	OG0000000	Strigops_hab	n4	1	Terminal	Strigops_habroptila_ENSSH	Strigops_habroptila_ENSSHB
6	OG0000000	Strigops_hab	n5	1	Terminal	Strigops_habroptila_ENSSH	Strigops_habroptila_ENSSHB

Each gene duplication event is cross-referenced to the species tree node, and the node in the gene tree. It also lists the genes descended from each of the two copies arising from the gene duplication event. We can check this out for our wombat olfactory receptor orthologues.

6375	OG0000365	N3	n2	0.66666667	Non-Terminal	Vombatus_ursinus_ENSVURG0	Canis_lupus_dingo_ENSCAFG00020000367.1, Vombat
6376	OG0000365	Canis_lupus_	n5	1	Terminal	Canis_lupus_dingo_ENSCAFG0	Canis_lupus_dingo_ENSCAFG00020000342.1
6377	OG0000365	Notamacrop	n6	1	Terminal	Notamacropus_eugenii_ENSME	Notamacropus_eugenii_ENSMEUG00000010434.1
6378	OG0000365	Notamacrop	n7	1	Terminal	Notamacropus_eugenii_ENSME	Notamacropus_eugenii_ENSMEUG00000016785.1
6379	OG0000365	Ornithorhync	n9	1	Terminal	Ornithorhynchus_anatinus_EN	Ornithorhynchus_anatinus_ENSOANG00000010542.3

These events are also summarised by orthogroup and by species tree node in the files Duplications_per_Orthogroup.tsv and Duplications_per_Species_Tree_Node.tsv which are both in the directory Comparative_Genomics_Statistics/.

Orthogroups

Often we’re interested in group-wise species comparisons, that is comparisons across a clade of species rather than between a pair of species. The generalisation of orthology to multiple species is the orthogroup. Just like orthologues are the genes descended from a single gene in the last common ancestor of a pair of species an orthogroup is the set of genes descended from a single gene in a group of species. Each gene tree from OrthoFinder, for example the one above, is for one orthogroup. The orthogroup gene tree is the tree we need to look at if we want it to include all pairwise orthologues. And even though some of the genes within an orthogroup can be paralogs of one another, if we tried to take any genes out then we would also be removing orthologs too.

So if we want to do a comparison of the ‘equivalent’ genes in a set of species, we need to do the comparison across the genes in an orthogroup. The orthogroups are in the file Orthogroups/Orthogroups.tsv. This table has one orthogroup per line and one species per column and is ordered from largest orthogroup to smallest.

Hierarchical Orthogroups

OrthoFinder3 also infers hierarchical orthogroups for each node in the species tree. A file equivalent to Orthogroups/Orthogroups.tsv is available for each node in ‘/Phylogenetic_Hierarchical_Orthogroups’. You can compare the node number (e.g. N3) to the species tree, to see which species will be included.

Orthogroup sequences

For each orthogroup there is a FASTA file in Orthogroup_Sequences/ which contains the sequences for the genes in that orthogroup.

Other results files

We have now covered all of the main output files that will be useful to most users, but OrthoFinder also outputs much more useful information! A full description of the output files is available below.

There are also some useful community tools that allow interactive viewing of results, such as OrthoBrowser
https://orthobrowserexamples.netlify.app/

Advanced tutorial for using OrthoFinder3

OrthoFinder3 provides a new workflow to assign new genes from new species to an already inferred set of orthogroups for a smaller, core group of species. If you are analysing >100 species, we reccommend that you use this scalable implementation.

To do this effectively, we need to pick a good set of core species. If we are running OrthoFinder on some bees, some moths, and some flies, we want to build our core orthogroups on a phylogenetically diverse set of those species. If instead we chose randomly and ended up using all moths as our core species, we might end up with less accurate orthogroups, and a longer runtime.

If you have a phylogenetic tree, you can use tools like Phylogenetic Diversity Analyzer to pick a representative set (http://www.cibiv.at/software/pda/). Or, you could just pick a diverse set yourself. We recommend that you pick 64 species for your core set.

First, run OrthoFinder on the subset of 64 species

orthofinder [options] -f <core>

Then, add the additional species to the results of the core run

orthofinder [options] --assign <additional> --core <Results_Dir>

This workflow also makes it really quick and easy to add new species to previous OrthoFinder runs. For example, if your research group works on various species of Angiosperms you might collectively share a core OrthoFinder results folder with a phylogenetically diverse set of species, which individual researchers could easily add any new species to.

An important note is that this workflow requires multiple sequence alignments, so unfortunately you cannot use it to add to OrthoFinder2 results that were run with the default
-M dendroblast option.

Using Outgroups

You can make orthogroup inference more accurate by including outgroup species. You just need to make sure that you use the correct N*.tsv file in Phylogenetic_Hierarchical_Orthogroups to look at the orthogroups (use species tree to discover which one)

Advanced usage

Manually installing dependencies

To run OrthoFinder3 on default setting, you will need **diamond, famsa, fasttree, MCL**

If you install OrthoFinder using a recommended method, you shouldn't need to install any of the dependencies manually, but in case you do;

Diamond
https://github.com/bbuchfink/diamond

FAMSA
https://github.com/refresh-bio/FAMSA

FastTree
http://www.microbesonline.org/fasttree/#Install

MCL
https://github.com/micans/mcl

Running BLAST searches separately

The '-op' option will prepare the files in the format required by

OrthoFinder and print the set of BLAST commands to run.

```
orthofinder -f fasta_files_directory -op
```

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 (see below)

Use pre-computed BLAST

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 '-op' 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 '-op' option to prepare the files first (see Section "Running BLAST Searches Separately")

Downloading input files for OrthoFinder

Let's imagine we want to perform a phylogenomic analysis across a set of species: Dingo (Canis lupus dingo), Great spotted kiwi (Apteryx haastii), Kakapo (Strigops habroptila), Platypus (Ornithorhynchus anatinus), Tammar wallaby (Notamacropus eugenii), and Common wombat (Vombatus ursinus).

OrthoFinder requires as input the amino acid sequences for all the protein coding genes in your species of interest. In this section of the tutorial we will discuss how to get these files for the species that you want to analyse. We will cover three major websites for getting this data, Ensembl, NCBI, and Phytozome.

We recommend **Ensembl** if you want to analyse eukaryotic species, especially vertebrates and model organisms.

We recommend **NCBI** if you want to analyse prokaryotic

species. We recommend **Phytozome** if you want to analyse

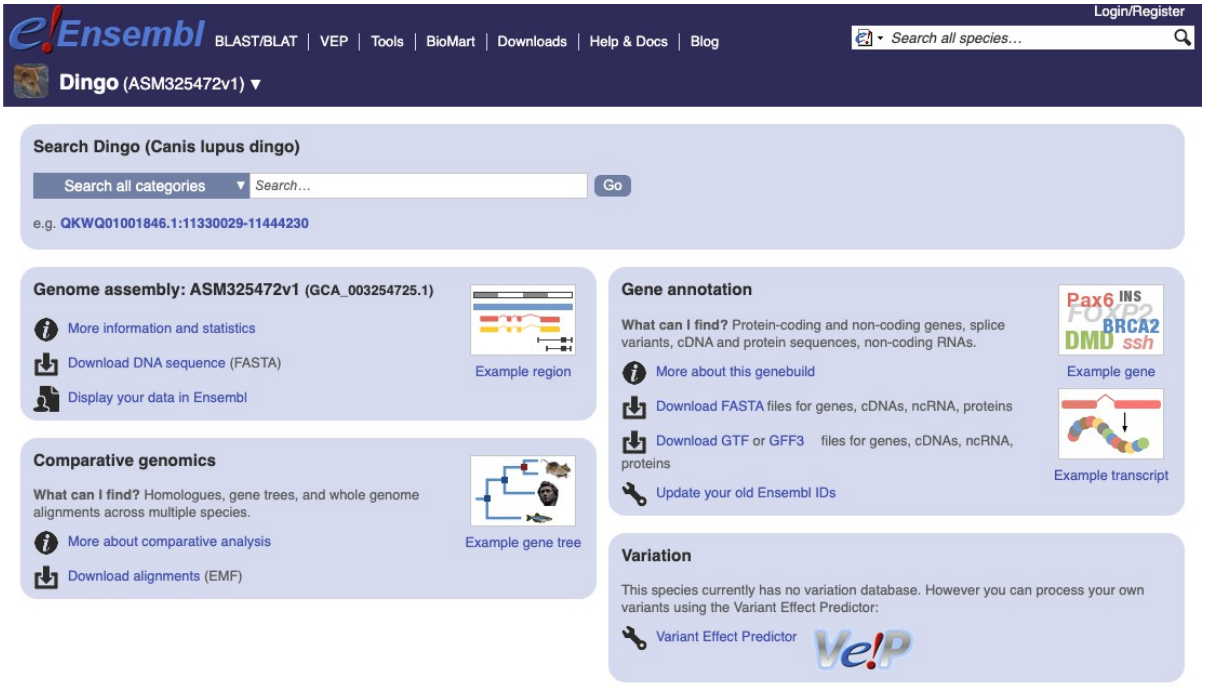
plants.

A key consideration when getting input data for OrthoFinder is gene transcripts. When you download the amino acid sequences for a genome, you might have multiple sequences for the same gene. This is because a single gene can produce multiple transcripts, which each might produce a different protein. You can see an example of transcripts in this table of the Actin gene http://www.ensembl.org/Homo_sapiens/Gene/Splice?db=core;g=ENSG00000075624;r=7:55 27151-5563784.
If we ran OrthoFinder on these raw files it would take much longer than necessary, and could lower the accuracy. We therefore want to extract just the longest transcript variant for each gene. OrthoFinder provides scripts to do this, which we will learn how to use later.

Getting data from Ensembl

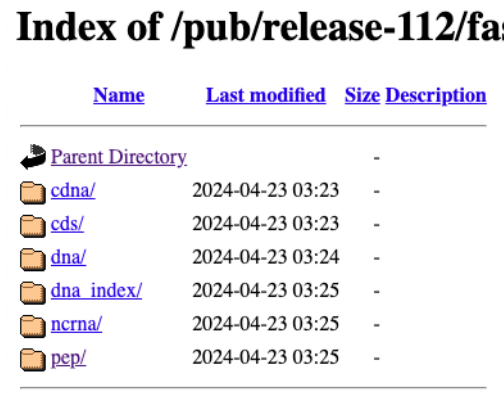
First, we go to the ensembl webpage with the list of species
https://www.ensembl.org/info/about/species.html
You can also use the new beta website, which is updated more regularly
https://beta.ensembl.org/species-selector

We can then search for our first species (Canis lupus dingo), and click on the link This will take us to the ensembl page for that species



On the right hand side under the heading ‘Gene annotation’ we can press the link to ‘Download FASTA’ files for the genome.

This will take us to a webpage with some folders.



Index of /pub/release-112/fasta/canis_lupus_dingo

Name	Last modified	Size	Description
Parent Directory		-	
cdna/	2024-04-23 03:23	-	
cds/	2024-04-23 03:23	-	
dna/	2024-04-23 03:24	-	
dna_index/	2024-04-23 03:25	-	
ncrna/	2024-04-23 03:25	-	
pep/	2024-04-23 03:25	-	

We need to click on the ‘pep/’ folder, and then click on the file that ends in .pep.all.fa.gz

Index of /pub/release-112/fasta/canis_lupus_dingo/pep

Name	Last modified	Size	Description
Parent Directory		-	
CHECKSUMS	2024-03-05 10:31	136	
Canis_lupus_dingo.ASM325472v1.pep.abinitio.fa.gz	2024-02-14 01:14	13M	
Canis_lupus_dingo.ASM325472v1.pep.all.fa.gz	2024-02-14 00:47	8.6M	
README	2024-02-14 01:14	2.4K	

We can then repeat this step for our other species, and place the files that are

downloaded into a folder on our computer. I have named my folder ‘Proteomes’.

The files are stored as .gz compressed files to save space, but we now need to expand the files so that we can access the sequences. You can either double-click on them all, or use the command line

```gunzip *.gz```

We’ll use a script provided with OrthoFinder to extract just the longest transcript variant per gene and run OrthoFinder on these files:

You can find the script here
https://github.com/OrthoFinder/OrthoFinder/blob/master/tools/primary_transcript.py

To run the script, first place it in the Proteomes folder. Then, open the command line and navigate to the Proteomes folder (using cd). You can then use the following command to run the script

```for f in *fa ; do python primary_transcript.py $f ; done```

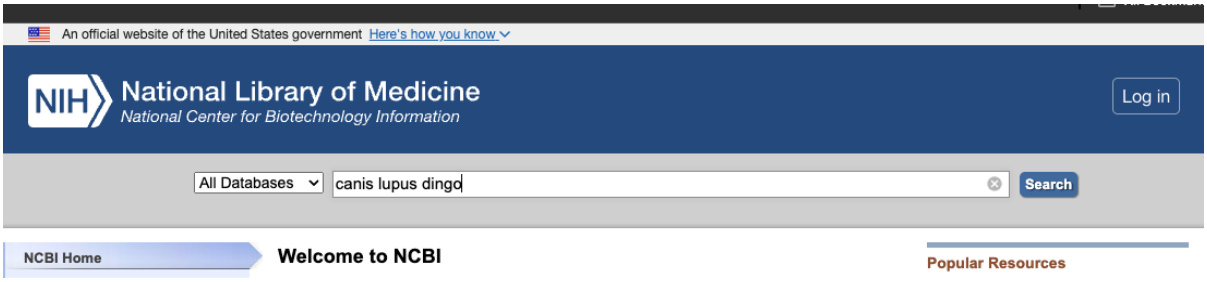
The script will generate a new folder called ‘primary_transcripts’, which contains our files.

Shortening the filenames is a good idea as it keeps the results tidy as the filenames are used to refer to the species, e.g. I shortened it to Canis_lupus_dingo.fa.

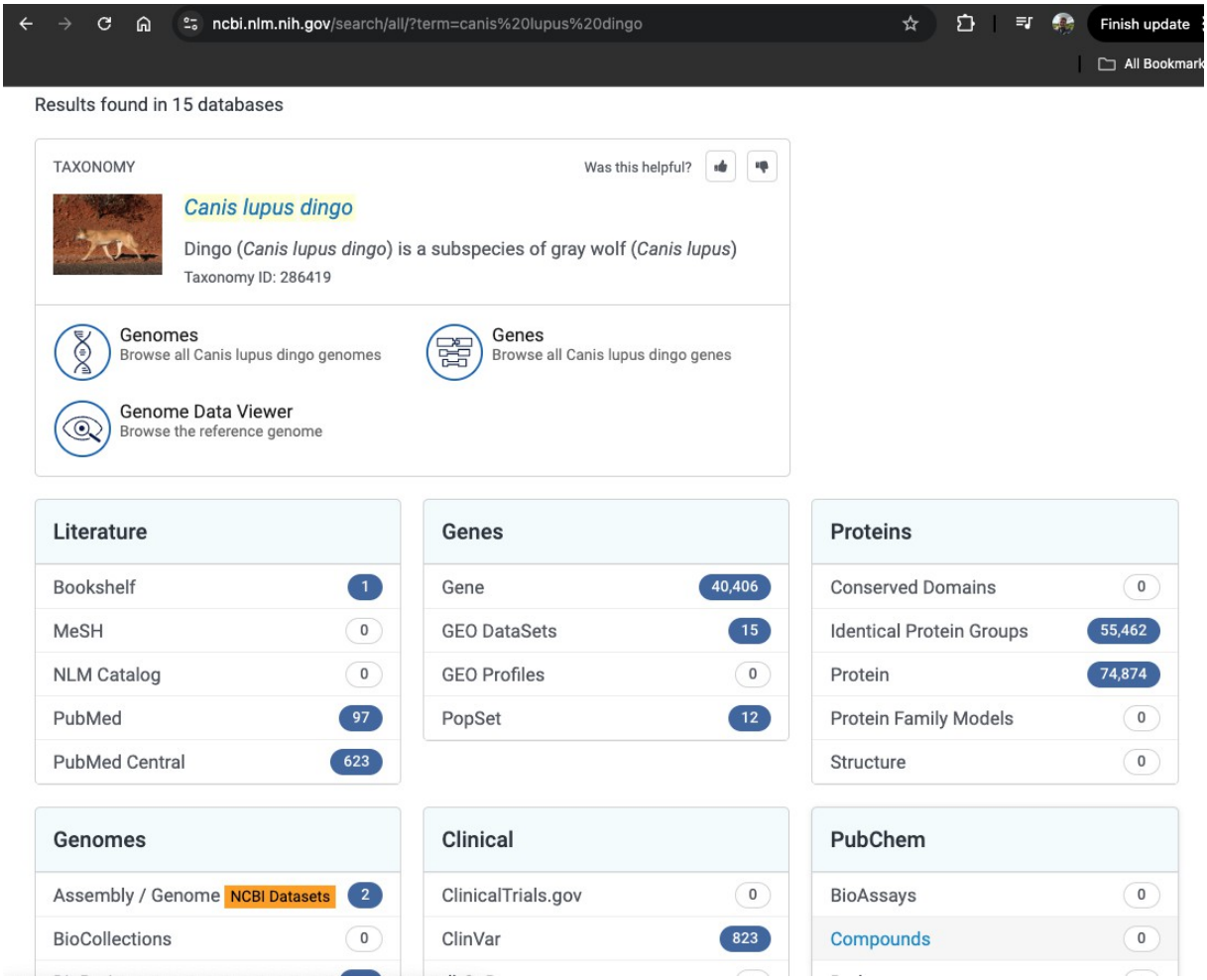
Our data is now ready for OrthoFinder. You can skip to Section 3 – Running OrthoFinder, or check out the below guides on getting input data from other sources

Getting data from NCBI

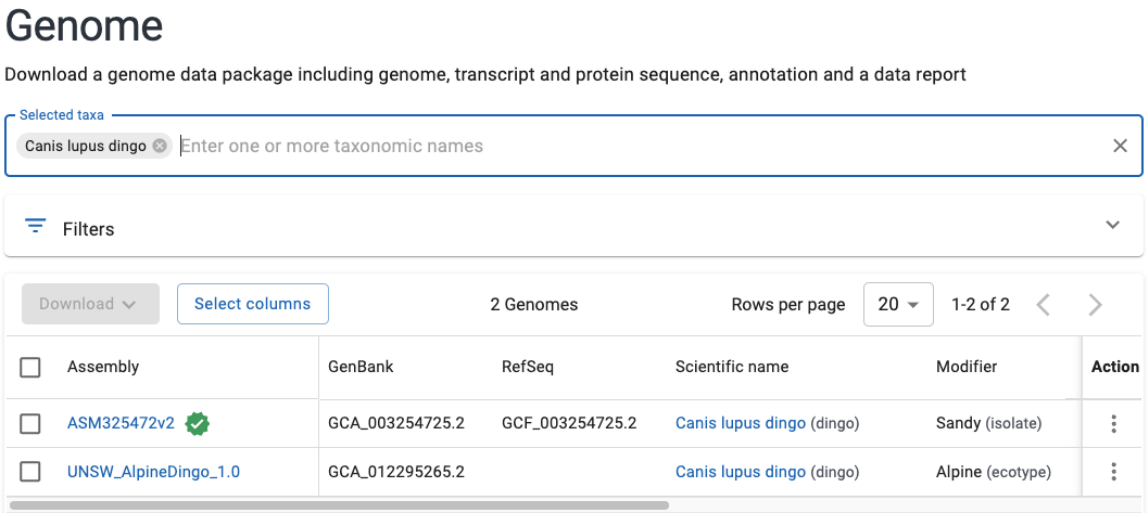
We start by going to https://www.ncbi.nlm.nih.gov/ and searching for our first species



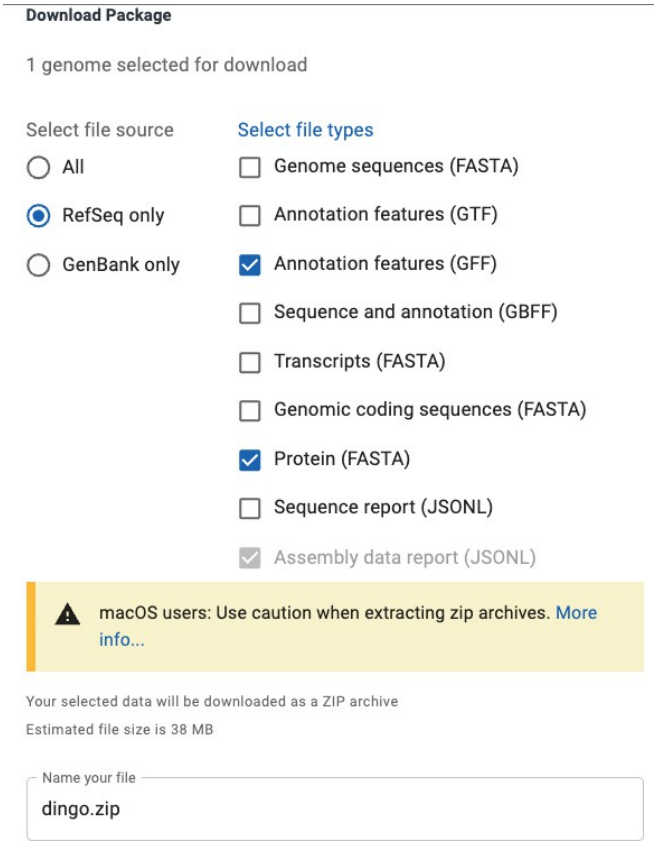
We want to find the genome of this species, so scroll down to the ‘Genomes’ section, and click on ‘Assembly / Genome’



There might be several genomes listed on this page. We are going to click on the one with the green tick, which shows us the reference genome.



We can then click the blue ‘Download’ button, and click the boxes to select the curated ‘RefSeq only’ annotation, and that we want the GFF and protein FASTA files. We can also give our download a useful name (such as the name of the species).



We can then click the ‘Download’ button, which downloads a zip file.

Repeat this step for all of the species that you want to use. For some species in our example list, like the Kiwi, there is no RefSeq genome with protein fasta available. For these species, it is recommended to use Ensembl.

Now, we need to deal with the issue of multiple transcripts per gene. For our first species, the Dingo, there are about 20,000 genes. However, if we look at the ‘protein.faa’ file that we have downloaded, it has about 75,000 sequences.

We can use the script ```ncbi_primary_transcript.py``` to extract the longest transcript per gene.

Place the script in the folder that has the zip folders, and run the following line of code in the command line

```python ncbi_primary_transcripts.py```

This will then make a folder named ‘primary_transcripts’, which is ready to run Orthofinder on.
```

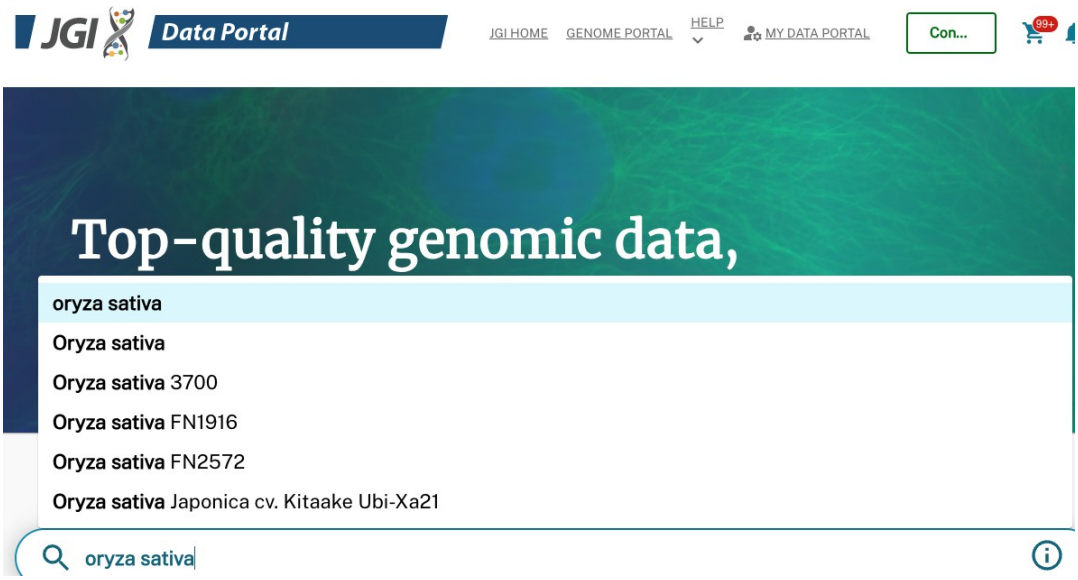
If you want to download data in bulk, you can also use the NCBI Datasets tool, which can run on the command line. For more info, see here https://www.ncbi.nlm.nih.gov/datasets/

Getting data from Phytozome

First, go to the Phytozome data portal website

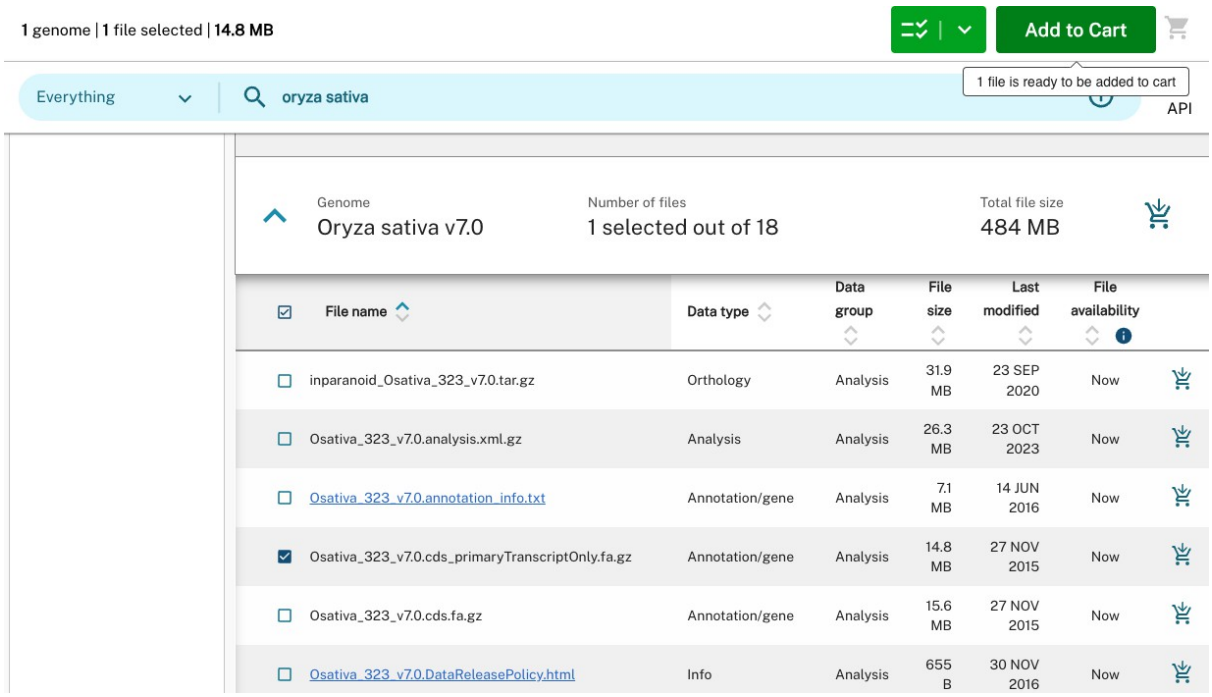
https://data.jgi.doe.gov/ Here, you can search for a species, such

as Oryza sativa



You can then find the genome that want to download, and click on it

The file that we want ends in 'primaryTranscriptOnly.fa.gz'. Luckily for us, Phytozome already deals with the problem of multiple transcripts. We can select that file, and click 'Add to Cart'.



We can then repeat this step for all of the species that we want to download. When we are ready, we then click the shopping cart icon, and press ‘Download’ on the next page. You will have to sign-in or register to download data.

These files can be placed in a folder, where they are now ready to run OrthoFinder.

A complete guide to OrthoFinder results files

Definitions for some terms that are used in these files:

- **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.
- **Unassigned gene**: A gene that has not been put into an orthogroup with any other genes.

Comparative_Genomics_Statistics

1. **Duplications_per_Orthogroup.tsv** is a tab separated text file that gives the number of duplications identified in each orthogroup.
2.
3. **Duplications_per_Species_Tree_Node.tsv** is a tab separated text file that gives the number of duplications identified as occurring along each branch of the species tree.
4.
5. **Orthogroups_SpeciesOverlaps.tsv** is a tab separated text file that contains the number of orthogroups shared between each species-pair as a square matrix.
6.
7. **OrthologuesStats_*.tsv** files are tab separated text files containing matrices giving the numbers of orthologues in one-to-one, one-to-many and many-to-many relationships between each pair of species.
    - **OrthologuesStats_one-to-one.tsv** is the number of one-to-one orthologues between each species pair.
    - **OrthologuesStats_many-to-many.tsv** contains the number of orthologues in a many-to-many relationship for each species pair (due to gene duplication events in both lineages post-speciation).
        i. Entry (i,j) is the number of genes in species i that are in a many-to-many orthology relationship with genes in species j.
    - **OrthologuesStats_one-to-many.tsv**: entry (i,j) gives the number of genes in species i that are in a one-to-many orthology relationship with genes from species j.
    - **OrthologuesStats_many-to-one.tsv**: entry (i,j) gives the number of genes in species i that are in a many-to-one orthology relationship with a gene from species j.
        i. There is a walk-through of an example results file here: #259.
    - **OrthologuesStats_Total.tsv** contains the totals for each species pair of orthologues of whatever multiplicity.
        i. Entry (i,j) is the total number of genes in species i that have orthologues in species j.

    There is a walk-through of an example results file here: #259.

8. **Statistics_Overall.tsv** is a tab separated text file that contains general statistics about orthogroup sizes and proportion of genes assigned to orthogroups.
9.
10. **Statistics_PerSpecies.tsv** is a tab separated text file that contains the same information as the Statistics_Overall.csv file but for each individual species.
11.

Gene_Duplication_Events

1. **Duplications.tsv** is a tab separated text file that lists all the gene duplication events identified by examining each node of each orthogroup gene tree. The columns are;
    a. "Orthogroup"
    b. "Species Tree node" (see Species_Tree/SpeciesTree_rooted_node_labels.txt)
    c. "Gene tree node" (see corresponding orthogroup tree in Resolved_Gene_Trees/)
    d. "Support" (proportion of expected species for which both copies of the duplicated gene are present)
    e. "Type"
        i. "Terminal": duplication on a terminal branch of the species tree
        ii. "Non-Terminal": duplication on an internal branch of the species tree & therefore shared by more than one species
        iii. "Non-Terminal: STRIDE": Non-Terminal duplication that also passes the very stringent STRIDE checks for what the topology of the gene tree should be post-duplication)
    f. "Genes 1" (the list of genes descended from one of the copies of the duplicate gene)
    g. "Genes 2" (the list of genes descended from the other copy of the duplicate gene.
2.
3. **SpeciesTree_Gene_Duplications_0.5_Support.txt** provides a summation of the above duplications over the branches of the species tree. The numbers after each node or species name are the number of gene duplication events with at least 50% support that occurred on the branch leading to the node/species. The branch lengths are as given in Species_Tree/SpeciesTree_rooted.txt. It is a text file, in newick format.

Orthogroup Sequences

1. A FASTA file for each orthogroup giving the amino acid sequences for each gene in the orthogroup.

Orthogroups Directory

Orthogroups.tsv is a tab separated text file that contains the genes in each orthogroup, with columns for each species.

Orthogroups.txt is a text file that contains the genes in each orthogroup (one orthogroup per line).

Orthogroups_UnassignedGenes.tsv is a tab separated text file that contains all of the genes that were not assigned to any orthogroup.

Orthogroups.GeneCount.tsv is a tab separated text file that contains counts of the number of genes for each species in each orthogroup.

Orthogroups_SingleCopyOrthologues.txt is a list of orthogroups that contain exactly one gene per species i.e. they contain one-to-one orthologues.

Orthologues Directory

One sub-directory for each species that in turn contains a file for each pairwise species comparison, listing the orthologs between that species pair.

Orthologues can be one-to-one, one-to-many or many-to-many depending on the gene duplication events since the orthologs diverged.

Each row in a file contains the gene(s) in one species that are orthologues of the gene(s) in the other species and each row is cross-referenced to the orthogroup that contains those genes.

Phylogenetic Hierarchical Orthogroups Directory

OrthoFinder infers HOGs, orthogroups at each hierarchical level (i.e. at each node in the species tree) by analysing the rooted gene trees.

1. **N1.tsv** 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. Additional columns give the HOG (Hierarchical Orthogroup) ID and the node in the gene tree from which the HOG was determined
2. **N2.tsv**, **N3.tsv**, ...: Orthogroups inferred from the gene trees corresponding to the clades of species in the species tree N2, N3, etc.

Phylogenetically_Misplaced_Genes

Genes in "Phylogenetically_Misplaced_Genes/" are those that appear to be out of place in the gene tree, and would otherwise negatively affect orthology analysis if not identified.

Putative_Xenologs

Xenologs are sets of genes descended from a common ancestor, but where there has been horizontal transfer on the evolutionary path to the gene copies in extant species, rather than just speciation and duplication. OrthoFinder tries to identify xenologs, but we call them ‘putative’, since many arise from contamination during sequencing. Each species has a file in this folder, listing genes and their putative xenologs from all other species

Resolved Gene Trees Directory

A rooted phylogenetic tree inferred for each orthogroup with 4 or more sequences and resolved using the OrthoFinder hybrid species-overlap/duplication-loss coalescent model.

2.

Single Copy Orthologue Sequences

1. The same files as the "Orthogroup Sequences" directory but restricted to only those orthogroups which contain exactly one gene per species.

Species Tree Directory

SpeciesTree_rooted.txt: A STAG species tree inferred from all orthogroups, containing STAG support values at internal nodes and rooted using STRIDE.

SpeciesTree_rooted_node_labels.txt: The same tree as above but with the nodes given labels (instead of support values) to allow other results files to cross-reference branches/nodes in the species tree (e.g. location of gene duplication events).

WorkingDirectory

This contains all the files necessary for orthofinder to run

Understanding Orthology

Orthogroups, Orthologs & Paralogs

Orthogroup = the group of genes descended from a single gene in the last common ancestor of a group of species

Orthologs = pairs of genes that descended from a single gene in the last common ancestor of two species

Paralogs = pairs of genes descended from a gene duplication event

Orthologs can be thought of as 'equivalent genes' between two species, as they descended from a single gene in the last common ancestor of that species. For example, the last common ancestor of humans and mice is a small mammal which lived alongside the dinosaurs. Individual genes present in that ancestor still exist in some form through their descendants in both humans and mice, and those genes are orthologs.

Orthologs describe relationships between pairs of species, but we can extend this idea to larger groups of species. Humans, mice, and chickens share a common ancestor from a few hundred million years ago, before the dinosaurs had even emerged. We can describe a group of genes across all three species that were descended from a single gene in this ancestor - these genes form an orthogroup.

Look at the figure below, which shows data for three species: human, mouse and chicken.

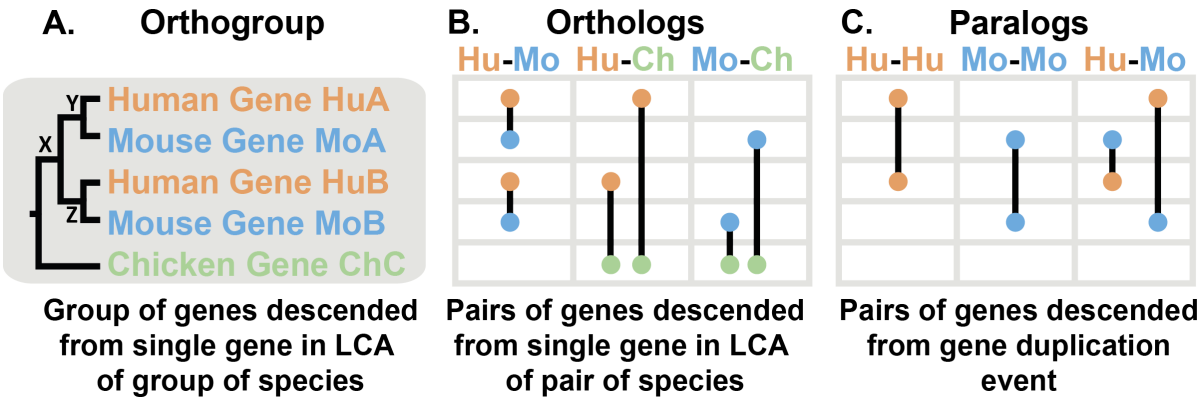


Figure 1: Orthologues, Orthogroups & Paralogues

The tree in Figure 1A shows the evolutionary history of a gene. First, there was a speciation event where the chicken lineage diverged from the human-mouse ancestor. In the human-mouse ancestor, there was a gene duplication event at X producing two copies of the gene in that ancestor, Y & Z. When human and mouse diverged they each inherited gene Y (becoming HuA & MoA) and gene Z (HuB & MoB). In general, we can identify a gene duplication event because it creates two copies of a gene in a species (e.g. HuA & HuB).

The mouse gene MoB is closer related to a human gene than it is to the other mouse gene MoA. This is because the gene duplication occurred in the ancestor, so each mouse gene is

more closely related to its human ortholog (e.g. HuA and MoA), as they were both descended from a single gene. By contrast, HuA and MoB diverged at the gene duplication event. They aren't descended from a single gene in the common ancestor of human and mice, so aren't orthologs. Instead, they are paralogs (Figure 1C). Paralogs are more distantly related, they diverged at a gene duplication event in a common ancestor. Such a gene duplication event must have occurred further back in time than when the species diverged and so paralogs between a pair of species are always less closely related than orthologs between that pair of species. Paralogs are also possible within a species (e.g. HuA & HuB).

The chicken gene diverged from the other genes when the lineage leading to chicken split from the lineage leading to human and mouse. Therefore, the chicken gene ChC is an ortholog of HuA & HuB in human and an ortholog of MoA & MoB in mouse. Depending on what happend after the genes diverged, orthologs can be in one-to-one relationships (HuA - MoA), many-to-one (HuA & HuB - ChC), or many-to-many (no examples in this tree, but would occur if there were a duplication in chicken). All of these relationships are identified by OrthoFinder.

Why Orthogroups?

There are several reasons why Orthogroups are the relevant way of analysing orthology relationships between species:

Orthogroups allow you to analyse all of your data
All of the genes in an orthogroup are descended from a single ancestral gene. Thus, all the genes in an orthogroup started out with the same sequence and function. As gene duplication and loss occur frequently in evolution, one-to-one orthologs are rare and limitation of analyses to on-to-one orthologs limits an analysis to a small fraction of the available data. By analysing orthogroups you can analyse all of your data.

Orthogroups allow you to define the unit of comparison
It is important to note that with orthogroups you choose where to define the limits of the unit of comparison. For example, if you just chose to analyse human and mouse in the above figure then you would have two orthogroups.

Orthogroups are the only way to identify orthologs
Orthology is defined by phylogeny. It is not definable by amino acid content, codon bias, GC content or other measures of sequence similarity. Methods that use such scores to define orthologs in the absence of phylogeny can only provide guesses. The only way to be sure that the orthology assignment is correct is by conducting a phylogenetic reconstruction of all genes descended from a single gene the last common ancestor of the species under consideration. This set of genes is an orthogroup. Thus, the only way to define orthology is by analysing orthogroups.

For a comprehensive overview of orthology and the OrthoFinder approach, you can watch David Emms' conference talk, from the 2020 symposium on Phylogenomics and Comparative genomics [https://www.youtube.com/watch?v=L6eXJAE5J7g]