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