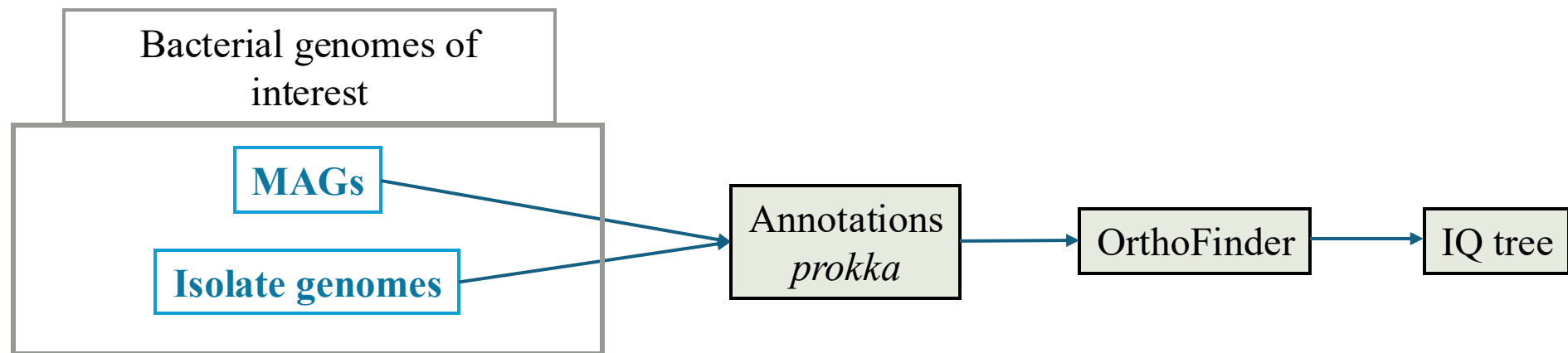


Phylogenetic tree reconstruction

Scripts of the Vonaesch Lab for the curnagl cluster

Simon Yersin

With the help of Garance Sarton-Lohéac



01: Gene annotations with Prokka

Prokka is a software tool to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files.

- To load Prokka correctly, we need to set up the databases, if necessary, follow the instruction on:

`install_prokka.sh`

- Annotate your genome using *Prokka*:

 `01_prokka.sh`





- Prokka produce a set of files per MAGs:

Extension	Description
.gff	This is the master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV.
.gbk	This is a standard Genbank file derived from the master .gff. If the input to prokka was a multi-FASTA, then this will be a multi-Genbank, with one record for each sequence.
.fna	Nucleotide FASTA file of the input contig sequences.
.faa	Protein FASTA file of the translated CDS sequences.
.ffn	Nucleotide FASTA file of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)
.sqn	An ASN1 format "Sequin" file for submission to Genbank. It needs to be edited to set the correct taxonomy, authors, related publication etc.
.fsa	Nucleotide FASTA file of the input contig sequences, used by "tbl2asn" to create the .sqn file. It is mostly the same as the .fna file, but with extra Sequin tags in the sequence description lines.
.tbl	Feature Table file, used by "tbl2asn" to create the .sqn file.
.err	Unacceptable annotations - the NCBI discrepancy report.
.log	Contains all the output that Prokka produced during its run. This is a record of what settings you used, even if the --quiet option was enabled.
.txt	Statistics relating to the annotated features found.
.tsv	Tab-separated file of all features: locus_tag,ftype,len_bp,gene,EC_number,COG,product

OrthoFinder: phylogenetic orthology inference for comparative genomics

- **OrthoFinder** is a fast, accurate and comprehensive platform for **comparative genomics**. It finds orthogroups and orthologs, infers rooted gene trees for all orthogroups and identifies all of the gene duplication events in those gene trees. It also infers rooted species tree for the species being analysed and maps the gene duplication events from the gene trees to branches in the species tree. OrthoFinder also provides comprehensive statistics for comparative genomic analyses.

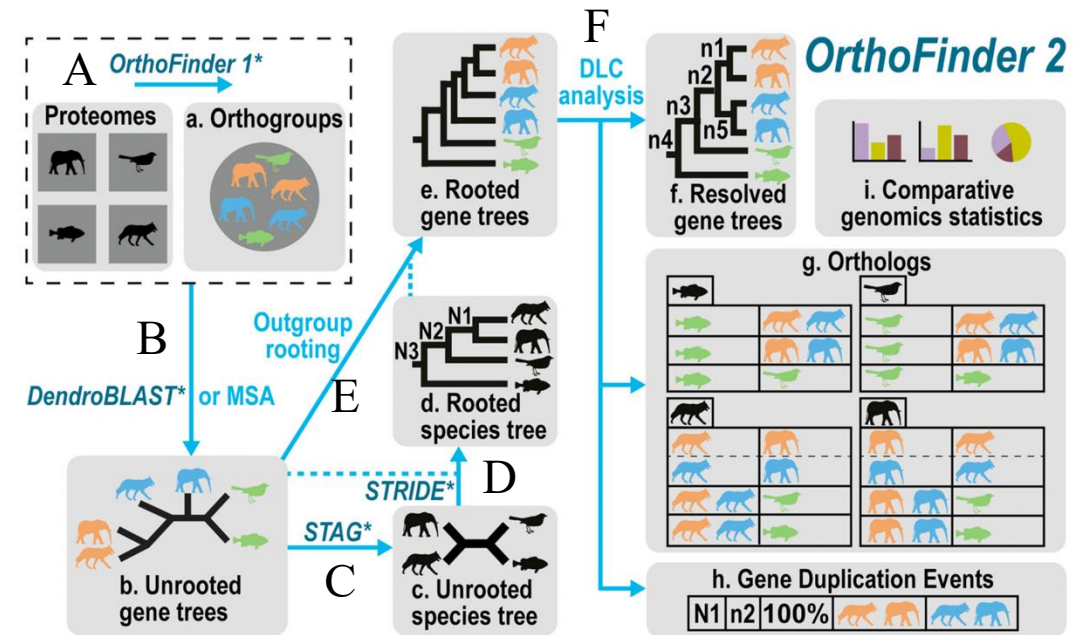
02:Phylogeny

- Move the .faa files produced by prokka in a dedicated directory genomesFAA (input for orthofinder)
- Run *OrthoFinder* using:
 [02_2_orthofinder.sh](#)
OrthoFinder produce a lot of different folder, the one that interest us is MultipleSequenceAlignment which include the SpeciesTreeAlignment.fa to produce a phylogenetic tree with IQ-Tree
- Produce a phylogenetic tree using *IQ-Tree*:
 [02_3_iq_tree.sh](#) (change date within the script for the data variable)
- Once *IQ-Tree* run is completed, download the files:
.treefiles OR .nex
- Tree visualization:
 **FigTree** (<http://tree.bio.ed.ac.uk/software/figtree/>)
 **RStudio**, package ggtree

OrthoFinder: Phylogeny orthology inference for comparative genomics

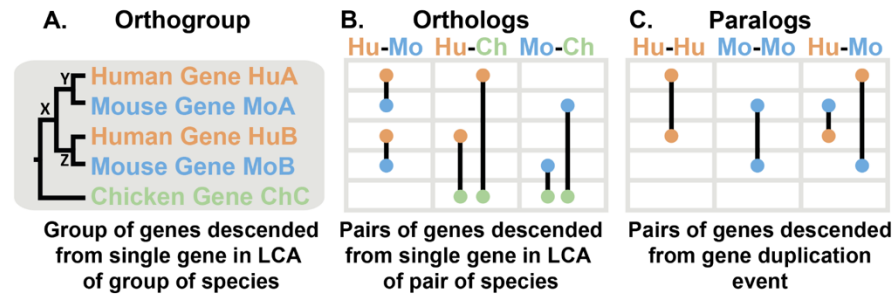
OrthoFinder is a fast, accurate and comprehensive platform for comparative genomics. It finds **orthogroups** and **orthologs**, infers **rooted gene trees** for all orthogroups and identifies all of the **gene duplication events** in those gene trees. It also infers a **rooted species tree** for the species being analysed and maps the gene duplication events from the gene trees to branches in the species tree. 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.

- A. Orthogroup inference
- B. Gene tree inference
- C. Species tree inference
- D. Species tree rooting
- E. Gene trees rooting
- F. DLC analysis for inference of orthologs and gene duplication events



The OrthoFinder workflow, each step is shown by an arrow with the algorithm used in italic with an asterisk, dotted line are additional data used to carry on to the next step.

Understanding orthology



The figure above shows a gene tree for three species: human, mouse and chicken. Orthologs are pairs of genes that descended from a single gene in the last common ancestor (LCA) of two species (Fig. B). They can be thought of as 'equivalent genes' between two species. An orthogroup is the extension of this concept 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 A). Genes within an orthogroup may be orthologs of one another or they may be paralogs, as explained next:

The tree 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).

To tell which genes are orthologs and which genes are paralogs we need to identify the gene duplication events in the tree. Orthologs are gene that diverged at a speciation event (e.g. HuA & MoA) while paralogs diverged at a gene duplication event (e.g. HuA & MoB, and others: Fig 2C). Because orthologs only diverged at the point when the species diverged, they are as closely related as any gene can be between the two species. 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 happens after the genes diverge, 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

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.

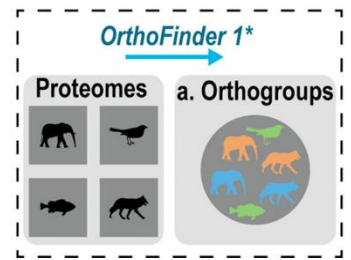
Orthogroups allow you to define the unit of comparison

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

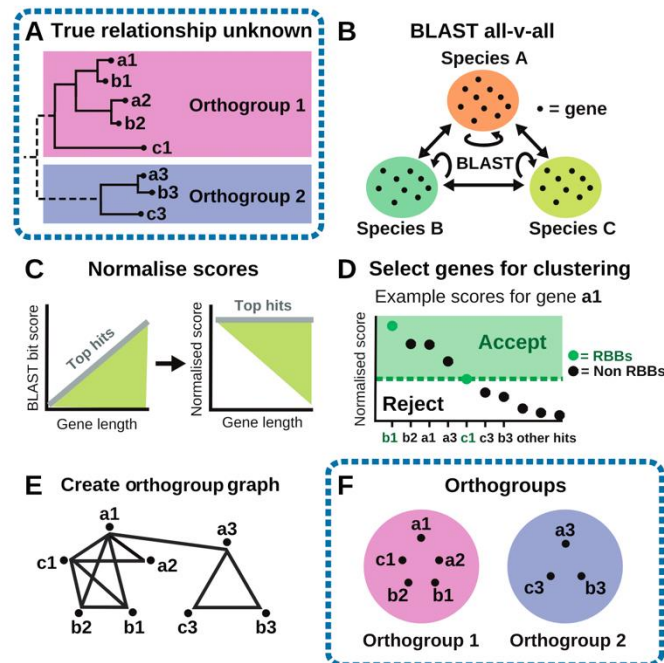
Orthology is defined by phylogeny. It is not definable by amino acid content, codon bias, GC content or other measures of sequence similarity. Methods that use such scores to define orthologs in the absence of phylogeny can only provide guesses. The only way to be sure that the orthology assignment is 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.

Default method:
DIAMOND



A. Orthogroup inference

- Orthologs are the set of genes in a species pair descended from a single gene in the last common ancestor of those two species.
- An **orthogroup** is the set of genes from multiple species descended from a single gene in the last common ancestor (LCA) of that set of species.
 - Orthogroup is the extension of orthology to multiple species
- The first step of the OrthoFinder workflow is the inference of the orthogroups using the original OrthoFinder algorithm (<https://doi.org/10.1186/s13059-015-0721-2>):

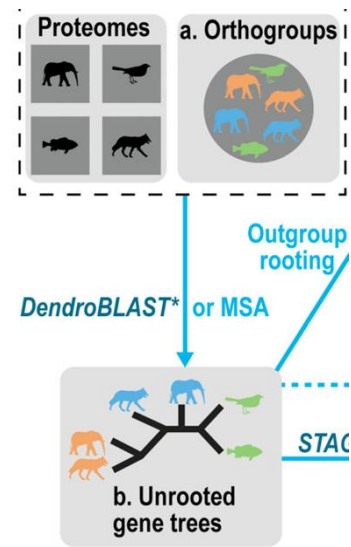


Overview of the steps in the OrthoFinder algorithm for two example orthogroups of genes from three species. **A** The unknown orthogroups that the algorithm must recover, shown as a gene tree. **B** BLAST search of all genes against all genes. **C** Gene length and phylogenetic distance normalisation of BLAST bit scores to give the scores to be used for orthogroup inference. **D** Selection of putative cognate gene-pairs from normalised BLAST scores. **E** Construction of orthogroup graph, genes are nodes in the graph and pairs of genes are connected by an edge with edge weights given by the normalised bit score. **F** Clustering of genes into discrete orthogroups using MCL.

B. Gene tree inference

- An orthogroup is the smallest set of genes such as that, for all genes it contains, the orthologs of these genes are also in the same set. Since gene tree inference scales super-linearly with the number of genes, partitioning genes into the smallest possible sets is the most efficient way of constructing a set of gene trees that encompasses all orthology relationships.
- For this step, an unrooted gene tree is inferred for each orthogroup.
- By default, DendroBLAST is the selected method to infer the gene trees. The option MSA can be used to further infer maximum likelihood trees from multiple sequence alignment (MSA). By default, MAFFT is used for the alignment and FastTree for the tree inference.

Default method:
DendroBLAST



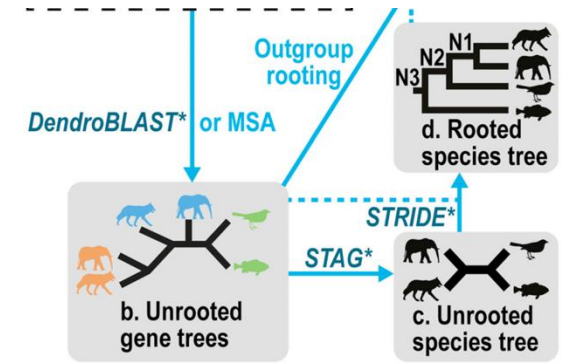
C. Species tree inference

- To obtain the rooted species tree (step E), an unrooted species tree is first required. The algorithm, Species Tree from All Gene (STAG), allows species tree inference even for species sets with few or no complete sets of one-to-one orthologs present in all species.
- STAG infers the species tree using the most closely related genes within single-copy or multi-copy orthogroups.



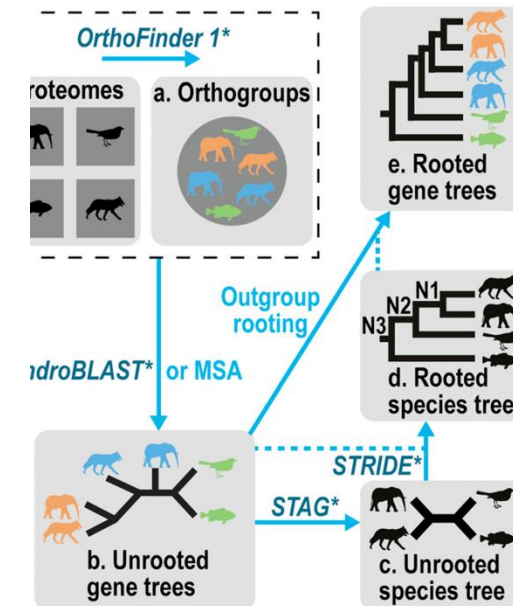
D. Species tree rooting

- The Species Tree Root Inference from Duplication Events (STRIDE) algorithm is used to root the species tree. It uses only information available in the set of gene trees to root the species tree.
- STRIDE identifies the set of well-supported in-group gene duplication events in the complete set of unrooted orthogroups tree and uses these events to infer a probability distribution over an unrooted STAG species tree for the location of its root.



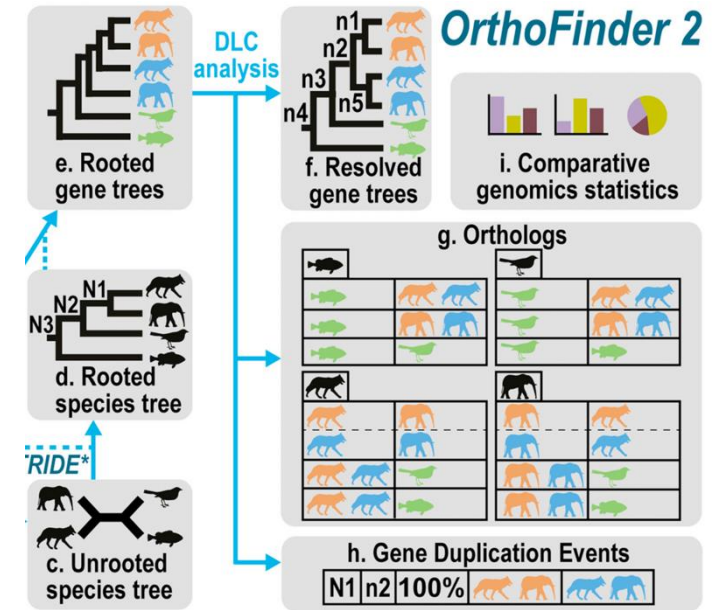
D. Gene tree rooting

- A gene tree must be correctly rooted in order for it to show the correct evolutionary history of the gene family and thus allow correct ortholog inference.
- In general, the rooted species tree can be used to root the orthogroup trees by identifying the outgroup clade in each orthogroup tree and placing the root on the branch separating this out-group from the remaining genes. As this approach might not work all the time for all trees, OrthoFinder uses a generalization of this outgroup rooting method to be able to root any potential gene tree.

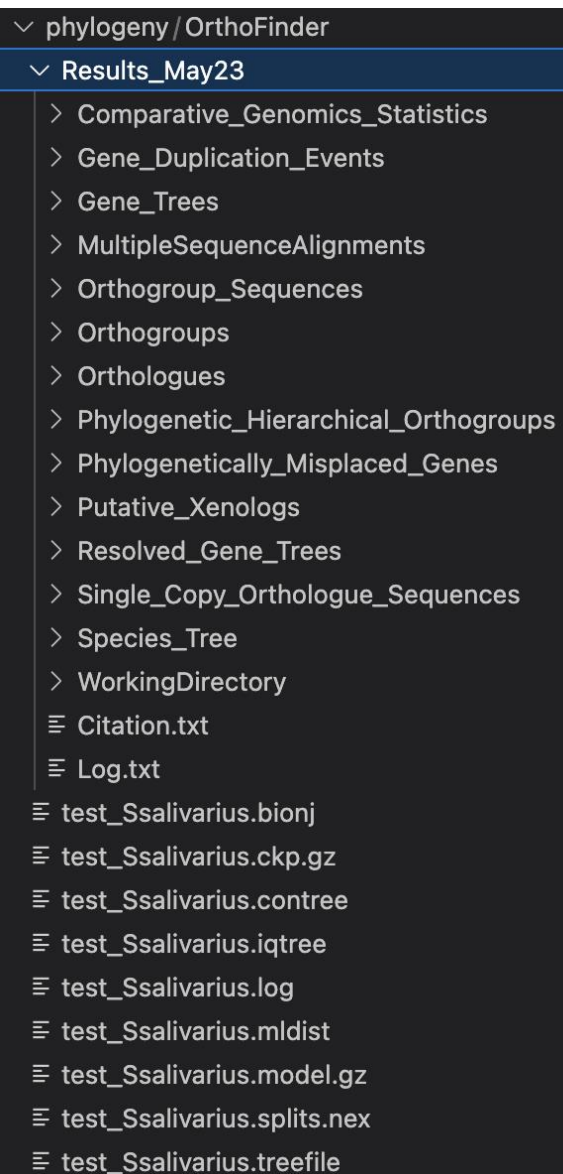


F, G, H. DLC analysis for inference of orthologs and gene duplication events

- Given a set of rooted orthogroup gene trees, the final major challenge in accurately dissecting phylogenetic relationships between genes is to account for incomplete lineage sorting and gene tree error.
- OrthoFinder use a hybrid duplication-loss coalescent (DLC) model to distinguish orthologs from paralogs in gene trees and determine the resolved gene tree. Orthologs and gene duplication events are then determined from the resolved gene tree according to the species overlap method.



Exploring OrthoFinder's results



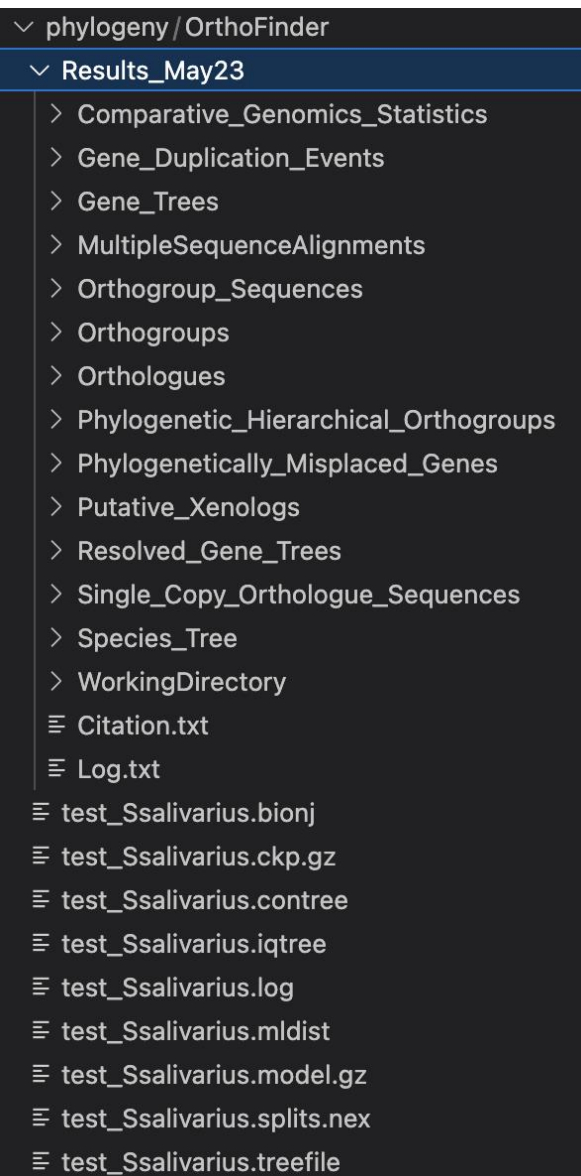
- In Comparative_Genomics_Statistics/Statistics_Overall.tsv, we can check how many genes were assigned to orthogroups.

```
Number of genes 7932
Number of genes in orthogroups 7561
Number of unassigned genes 371
Percentage of genes in orthogroups 95.3
Percentage of unassigned genes 4.7
```

At least 80% of the gene should be assigned to orthogroups, fewer means that we are probably missing orthology relationship that actually exist for some of the remaining genes, poor species sampling is the most likely cause.

- In the same folder, Statistics_PerSpecies.tsv, we obtain the same information for how many genes were assigned to orthogroup but for each of our species (samples).
- Species_Tree/SpeciesTree_rooted.txt contain the inferred rooted species tree by OrthoFinder using STAG and STRIDE after step D. It can be visualize using FigTree or any tree viewer.
- One of the most common reasons for running OrthoFinder is to find the orthologue of a gene we are interested in. In the Orthologues directory, there is a sub-directory for each species. Each species has then a table containing orthogroup of this species against another species. We can find
- The gene tree for a specific gene of interest can be then obtain from Gene_Trees/OG000XXX_tree.txt

Exploring OrthoFinder's results



- Gene duplication events can be explored with
 - Species_Tree/SpeciesTree_rooted_node_labels.txt
 - Gene_Duplication_Events/SpeciesTree_Gene_Duplication_0.5_Support.txt
 - Gene_Duplication_Events/Duplications.tsv
 - Resolved_Gene_Trees/OG000XXX_tree.txt
 - Comparative_Genomics_Statistics/Duplication_per_Orthogroup.tsv
 - Comparative_Genomics_Statistics/Duplication_per_Species_Tree_Node.tsv
- Orthogroups are in the file Orthogroups/Orthogroups.tsv and the sequences in Orthogroups_Sequences/
- Finally, the Multiple Sequences Alignment tree, in MultipleSequenceAlignments/SpeciesTreeAlignment.fa can be used for the next step and produce a phylogenetic tree with IQ-Tree.
- https://davidemms.github.io/orthofinder_tutorials/exploring-orthofinders-results.html
- <https://github.com/davidemms/OrthoFinder?tab=readme-ov-file#orthofinder-results-files>