#4 Gene Family Evolution II

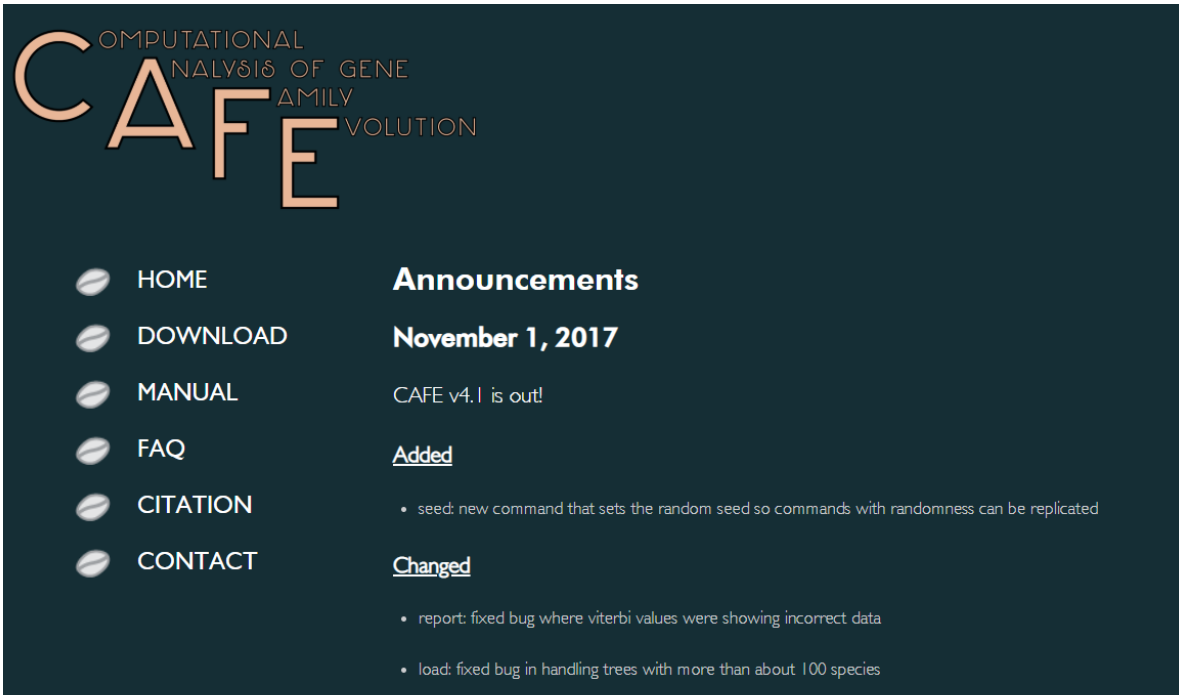
For this exercise we will explore in more detail the results of the CAFE analysis

By the end of this first exercise you should:

[1] have gained some more experience and knowledge of the orthology and sequence data available from OrthoDB and how to extract what you need
[2] have an understanding of how to build gene trees for rapidly evolving multi-copy gene families and some of the important data-accuracy issues that can impact on such analyses

NB: on the following pages, lines starting with a '*' are instructions or information, while lines starting with a '$' are commands to be typed into the terminal and executed

*Required



Announcements

November 1, 2017

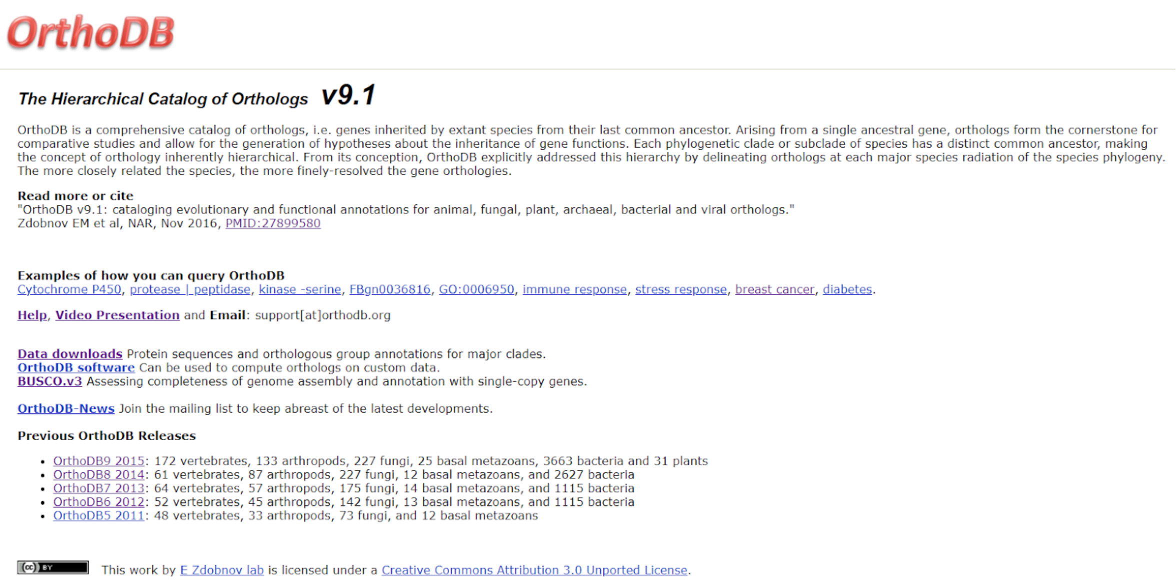
CAFE v4.1 is out!

Added

- seed: new command that sets the random seed so commands with randomness can be replicated

Changed

- report: fixed bug where viterbi values were showing incorrect data
- load: fixed bug in handling trees with more than about 100 species



The Hierarchical Catalog of Orthologs **v9.1**

OrthoDB is a comprehensive catalog of orthologs, i.e. genes inherited by extant species from their last common ancestor. Arising from a single ancestral gene, orthologs form the cornerstone for comparative studies and allow for the generation of hypotheses about the inheritance of gene functions. Each phylogenetic clade or subclade of species has a distinct common ancestor, making the concept of orthology inherently hierarchical. From its conception, OrthoDB explicitly addressed this hierarchy by delineating orthologs at each major species radiation of the species phylogeny. The more closely related the species, the more finely-resolved the gene orthologies.

Read more or cite
"OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs."
Zdobnov EM et al, NAR, Nov 2016, PMID:27899580

Examples of how you can query OrthoDB
Cytochrome P450, protease | peptidase, kinase -serine, FBgn0036816, GO:0006950, immune response, stress response, breast cancer, diabetes.

Help, **Video Presentation** and **Email**: support[at]orthodb.org

Data downloads Protein sequences and orthologous group annotations for major clades.
OrthoDB software Can be used to compute orthologs on custom data.
BUSCO.v3 Assessing completeness of genome assembly and annotation with single-copy genes.

OrthoDB-News Join the mailing list to keep abreast of the latest developments.

Previous OrthoDB Releases

- OrthoDB9 2015: 172 vertebrates, 133 arthropods, 227 fungi, 25 basal metazoans, 3663 bacteria and 31 plants
- OrthoDB8 2014: 61 vertebrates, 87 arthropods, 227 fungi, 12 basal metazoans, and 2627 bacteria
- OrthoDB7 2013: 64 vertebrates, 57 arthropods, 175 fungi, 14 basal metazoans, and 1115 bacteria
- OrthoDB6 2012: 52 vertebrates, 45 arthropods, 142 fungi, 13 basal metazoans, and 1115 bacteria
- OrthoDB5 2011: 48 vertebrates, 33 arthropods, 73 fungi, and 12 basal metazoans

1. **My VM is up and running and I'm ready to proceed.** *

Mark only one oval.

○ Yes *Skip to question 2.*

○ No *Skip to "We're here to help!."*

[A] Identifying the most 'interesting' families from the CAFE results

We will start by exploring the CAFE results from the previous exercise

[1] Get the CAFE results
* FIRSTLY - from your HOME directory

```
$ mkdir rmw4
$ cd rmw4
```

* To make sure we are all working with exactly the same results, from the Moodle site find the folder under 'Day 2 Rob Waterhouse' called 'OrthoDB_gene_families', inside you should see control file called 'report_run1.cafe.gz'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM

```
$ wget https://edu.sib.swiss/pluginfile.php/6280/mod_folder/content/0/report_run1.cafe.gz
```

* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* unzip the results

```
$ gunzip report_run1.cafe.gz
```

[2] Exploring the CAFE results
* Note that the results per orthoologous group take the following format:
'ID' 'Newick' 'Family-wide P-value' 'Viterbi P-values' 'cut P-value' 'Likelihood Ratio'
* For example:

```
EOG090W0004	((((((Aaegy_1:49,Cquin_0:49)_1:37,(Aalbi_1:43,Agamb_1:43)_1:43)_1:83,Llong_3:169)_1:16,(Gmors_0:71,Mdome_1:71)_1:114)_1:98,(Clect_1:143,Rprol_1:143)_1:140)_1:83,Phuma_1:366)_1	0.09	((-,-),(-,-),(-,-),(-,-),(-,-),(-,-),(-,-),(-,-),(-,-))
```

* We applied a p-value cut-off of '0.01', and as this orthologous group obtained a p-value of only 0.09 it was not analysed further by CAFE
* So we can focus on those orthologous groups that obtained a family-wide p-value below the cut-off

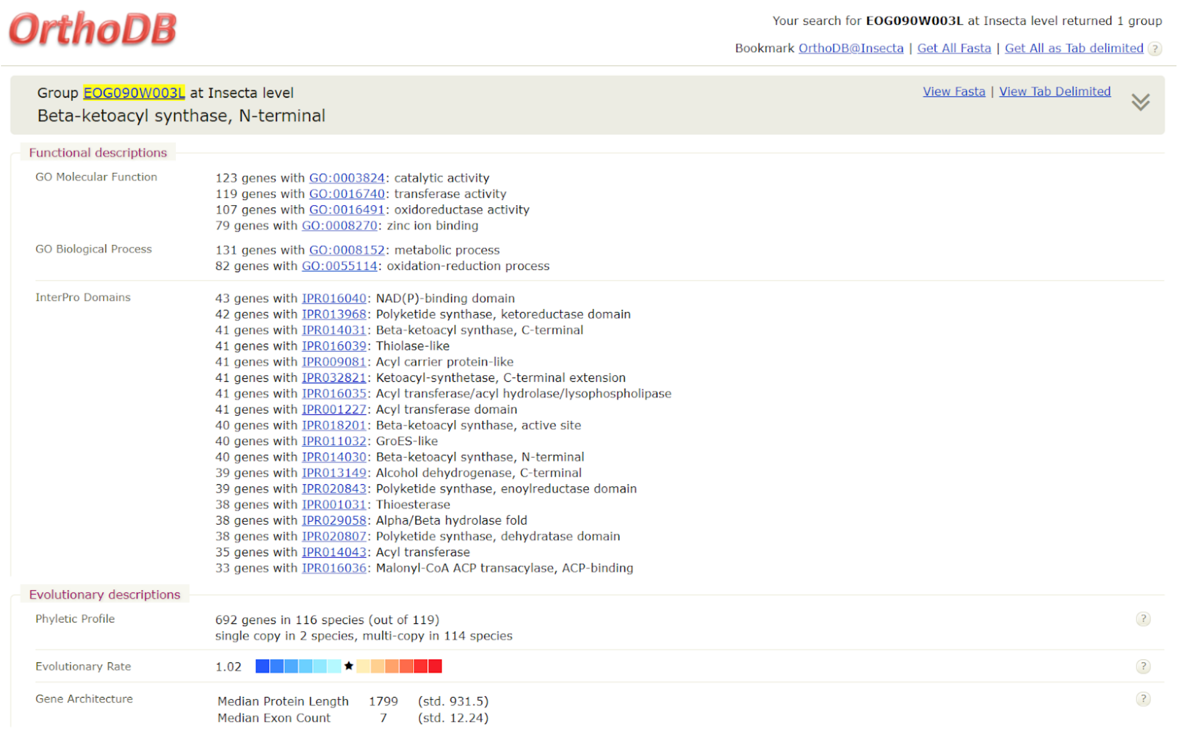
```
$ grep 'EOG' report_run1.cafe | sort -nrk3
```

* This simply sorts all the family result lines by the p-value
* So now that last line after sorting should be:

```
EOG090W003L	((((((Aaegy_7:49,Cquin_4:49)_4:37,(Aalbi_3:43,Agamb_4:43)_4:43)_4:83,Llong_1:169)_4:16,(Gmors_7:71,Mdome_2:71)_4:114)_4:98,(Clect_5:143,Rprol_4:143)_4:140)_4:83,Phuma_3:366)_4	0	((0.000557668,0.620016),(0.595206,0.607931),(0.16375,0.607931),(0.678048,0.00522876),(0.545031,0.71859),(0.0015282,0.00872628),(0.698918,0.745781),(0.187113,0.748588),(0.678048,0.58917))
```

* This orthologous group has a p-value of zero
* Note the gene counts per species appended to each species label e.g. 'Aaegy_7' means 7 genes in Aedes aegypti
* Note also the inferred gene counts at the internal nodes of the tree, e.g. Aalbi has 3 and Agamb has 4 and their inferred ancestor has 4 genes

[3] Going back to OrthoDB
* For any of these groups with significant p-values indicating dynamic changes across the phylogeny, we can now look them up at OrthoDB to learn about their putative functions and extract their protein sequence data
* Go to www.orthodb.org and search for EOG090W003L
* If OrthoDB has not remembered your species selection (10 species) from earlier you can select those species again
* Or if you are impatient then this URL will look up EOG090W003L at the Insecta level with your 10 species already selected for viewing:
https://www.orthodb.org/?query=EOG090W003L&level=50557&species=7167%2C7165%2C7159%2C7176%2C7200%2C7394%2C7370%2C79782%2C13249%2C121225



OrthoDB

Your search for **EOG090W003L** at Insecta level returned 1 group

Bookmark OrthoDB@Insecta | Get All Fasta | Get All as Tab delimited ?

Group EOG090W003L at Insecta level

Beta-ketoacyl synthase, N-terminal

View Fasta | View Tab Delimited

Functional descriptions

GO Molecular Function

123 genes with GO:0003824: catalytic activity
119 genes with GO:0016740: transferase activity
107 genes with GO:0016491: oxidoreductase activity
79 genes with GO:0008270: zinc ion binding

GO Biological Process

131 genes with GO:0008152: metabolic process
82 genes with GO:0055114: oxidation-reduction process

InterPro Domains

43 genes with IPR016040: NAD(P)-binding domain
42 genes with IPR013968: Polyketide synthase, ketoreductase domain
41 genes with IPR014031: Beta-ketoacyl synthase, C-terminal
41 genes with IPR016039: Thiolase-like
41 genes with IPR009081: Acyl carrier protein-like
41 genes with IPR032821: Ketoacyl-synthetase, C-terminal extension
41 genes with IPR016035: Acyl transferase/acyl hydrolase/lysophospholipase
41 genes with IPR001227: Acyl transferase domain
40 genes with IPR018201: Beta-ketoacyl synthase, active site
40 genes with IPR011032: GroES-like
40 genes with IPR014030: Beta-ketoacyl synthase, N-terminal
39 genes with IPR013149: Alcohol dehydrogenase, C-terminal
39 genes with IPR020843: Polyketide synthase, enoylreductase domain
38 genes with IPR001031: Thioesterase
38 genes with IPR029058: Alpha/Beta hydrolase fold
38 genes with IPR020807: Polyketide synthase, dehydratase domain
35 genes with IPR014043: Acyl transferase
33 genes with IPR016036: Malonyl-CoA ACP transacylase, ACP-binding

Evolutionary descriptions

Phyletic Profile

692 genes in 116 species (out of 119)
single copy in 2 species, multi-copy in 114 species

Evolutionary Rate

1.02 🟦🟦🟦🟦🟦🟦 ★ 🟨🟨🟧🟧🟥🟥

Gene Architecture

Median Protein Length	1799	(std. 931.5)
Median Exon Count	7	(std. 12.24)

2. **How many genes and species in TOTAL are found in this orthologous group?** *

Mark only one oval.

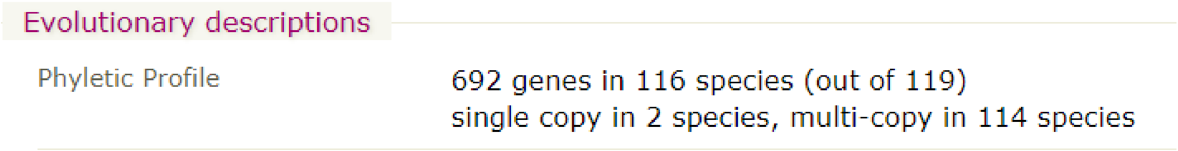
○ 116 genes & 692 species *Skip to question 3.*

○ 123 genes & 114 species *Skip to question 3.*

○ 692 genes & 116 species *Skip to question 4.*

Are you sure?

Gene and species counts are found under the 'Evolutionary descriptions' section



Evolutionary descriptions

Phyletic Profile

692 genes in 116 species (out of 119)
single copy in 2 species, multi-copy in 114 species

3. **How many genes and species in TOTAL are found in this orthologous group?** *

Mark only one oval.

○ 116 genes & 692 species *Skip to question 3.*

○ 123 genes & 114 species *Skip to question 3.*

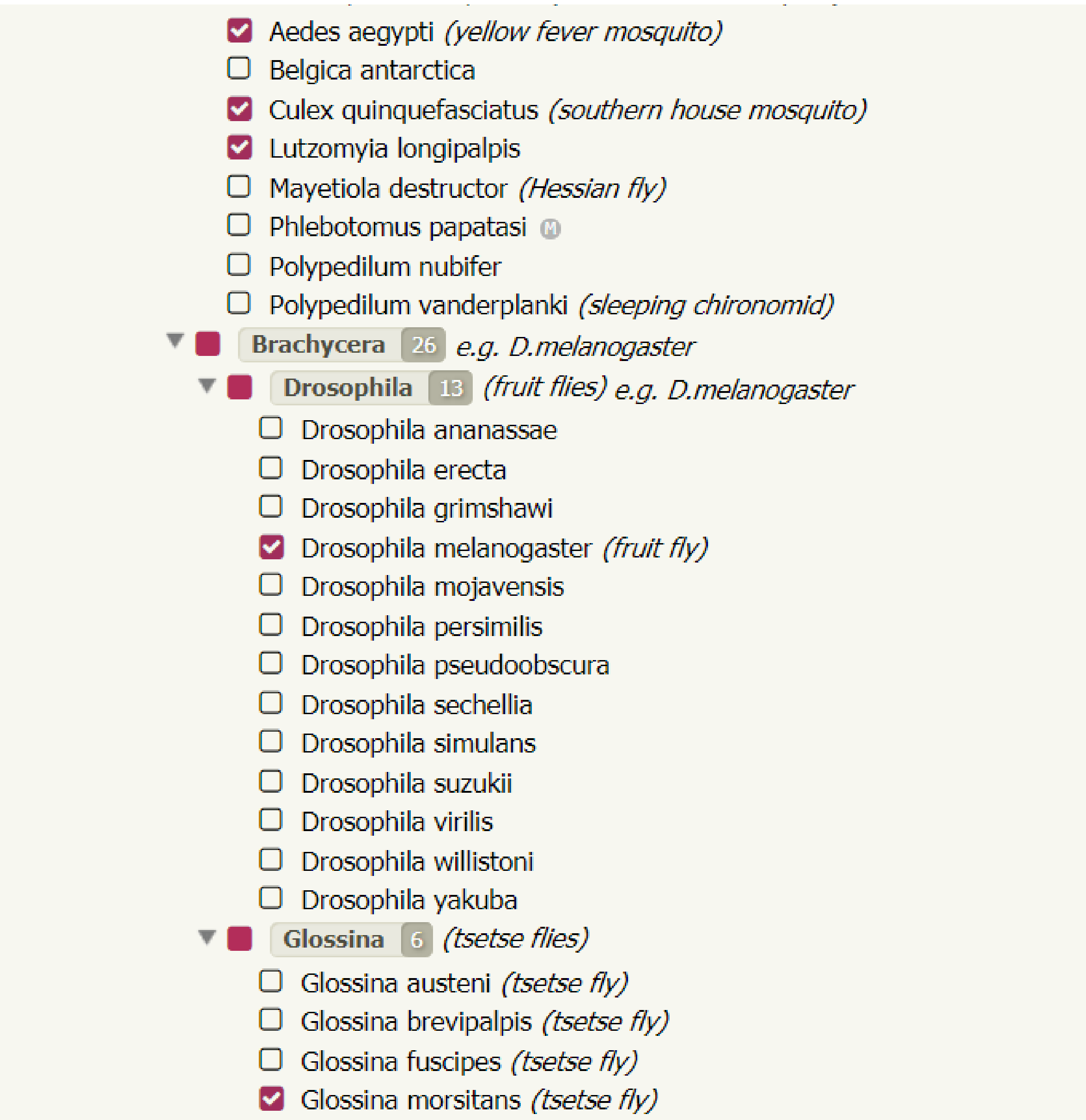
○ 692 genes & 116 species *Skip to question 4.*

[B1] Exploring orthology data at OrthoDB

* Amongst insects Drosophila melanogaster generally has the best annotated genes both in terms of their structural annotations (exons, introns, untranslated regions etc.) and their functional annotations, so it is often useful when exploring orthologues to include Drosophila melanogaster in your searches

* Using the species selector on the bottom right, add Drosophila melanogaster either by typing in the 'Search species by name:' box, or by expanding the tree to find Drosophila (hint Drosophila are Brachycera like Glossina)

* After adding Drosophila melanogaster then click submit again to run the search again



☑ Aedes aegypti *(yellow fever mosquito)*
□ Belgica antarctica
☑ Culex quinquefasciatus *(southern house mosquito)*
☑ Lutzomyia longipalpis
□ Mayetiola destructor *(Hessian fly)*
□ Phlebotomus papatasi Ⓜ
□ Polypedilum nubifer
□ Polypedilum vanderplanki *(sleeping chironomid)*

▼ ■ Brachycera 26 *e.g. D.melanogaster*

▼ ■ Drosophila 13 *(fruit flies) e.g. D.melanogaster*

□ Drosophila ananassae
□ Drosophila erecta
□ Drosophila grimshawi
☑ Drosophila melanogaster *(fruit fly)*
□ Drosophila mojavensis
□ Drosophila persimilis
□ Drosophila pseudoobscura
□ Drosophila sechellia
□ Drosophila simulans
□ Drosophila suzukii
□ Drosophila virilis
□ Drosophila willistoni
□ Drosophila yakuba

▼ ■ Glossina 6 *(tsetse flies)*

□ Glossina austeni *(tsetse fly)*
□ Glossina brevipalpis *(tsetse fly)*
□ Glossina fuscipes *(tsetse fly)*
☑ Glossina morsitans *(tsetse fly)*

4. **How many Drosophila melanogaster genes are found in this group (EOG090W003L)?** *

Mark only one oval.

○ 3 *Skip to question 6.*

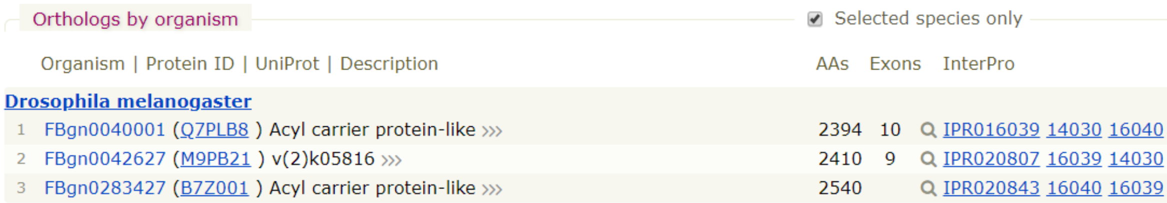
○ 2 *Skip to question 5.*

○ 1 *Skip to question 5.*

Are you sure?

URL now including Drosophila melanogaster

https://www.orthodb.org/?query=EOG090W003L&level=50557&species=7167%2C7165%2C7159%2C7176%2C7200%2C7227%2C7394%2C7370%2C79782%2C13249%2C121225



Orthologs by organism ☑ Selected species only

	Organism | Protein ID | UniProt | Description	AAs	Exons	InterPro
	Drosophila melanogaster			
1	FBgn0040001 (Q7PLB8) Acyl carrier protein-like ›››	2394	10	IPR016039 14030 16040
2	FBgn0042627 (M9PB21) v(2)k05816 ›››	2410	9	IPR020807 16039 14030
3	FBgn0283427 (B7Z001) Acyl carrier protein-like ›››	2540		IPR020843 16040 16039

5. **How many Drosophila melanogaster genes are found in this group (EOG090W003L)?** *

Mark only one oval.

○ 3 *Skip to question 6.*

○ 2 *Skip to question 5.*

○ 1 *Skip to question 5.*

[B2] Exploring orthology data at OrthoDB

Useful clues about data quality - very important when assessing gene family expansions/contractions, as wrongly annotated neighbouring genes can be split (looks like more gene copies) or fused (looks like fewer gene copies)

[1] Protein functions

* Two of the three Drosophila melanogaster proteins are described as 'Acyl carrier protein-like'
1 FBgn0040001 (Q7PLB8) Acyl carrier protein-like
2 FBgn0042627 (M9PB21) v(2)k05816
3 FBgn0283427 (B7Z001) Acyl carrier protein-like

* The most frequently occurring Gene Ontology terms and InterPro domains are shown as counted across all genes from this orthologous group (not just the selected species)
* These appear in line with the 'Acyl carrier protein-like' descriptions of the Drosophila melanogaster genes, i.e. the terms or domains are related to functions of acyl carrier proteins
* https://en.wikipedia.org/wiki/Acyl_carrier_protein "The acyl carrier protein (ACP) is an important component in both fatty acid and polyketide biosynthesis with the growing chain bound during synthesis as a thiol ester at the distal thiol of a 4'-phosphopantetheine moiety. The protein is expressed in the inactive apo form and the 4'-phosphopantetheine moiety must be post-translationally attached to a conserved serine residue on the ACP by the action of holo-acyl carrier protein synthase (ACPS), a 4'-phosphopantetheinyl transferase."

* If there are chevrons '>' or '>>' or '>>>' next to the gene identifiers or short descriptions then you can click on these to expand the available information to find out more and to link to various external databases such as FlyBase

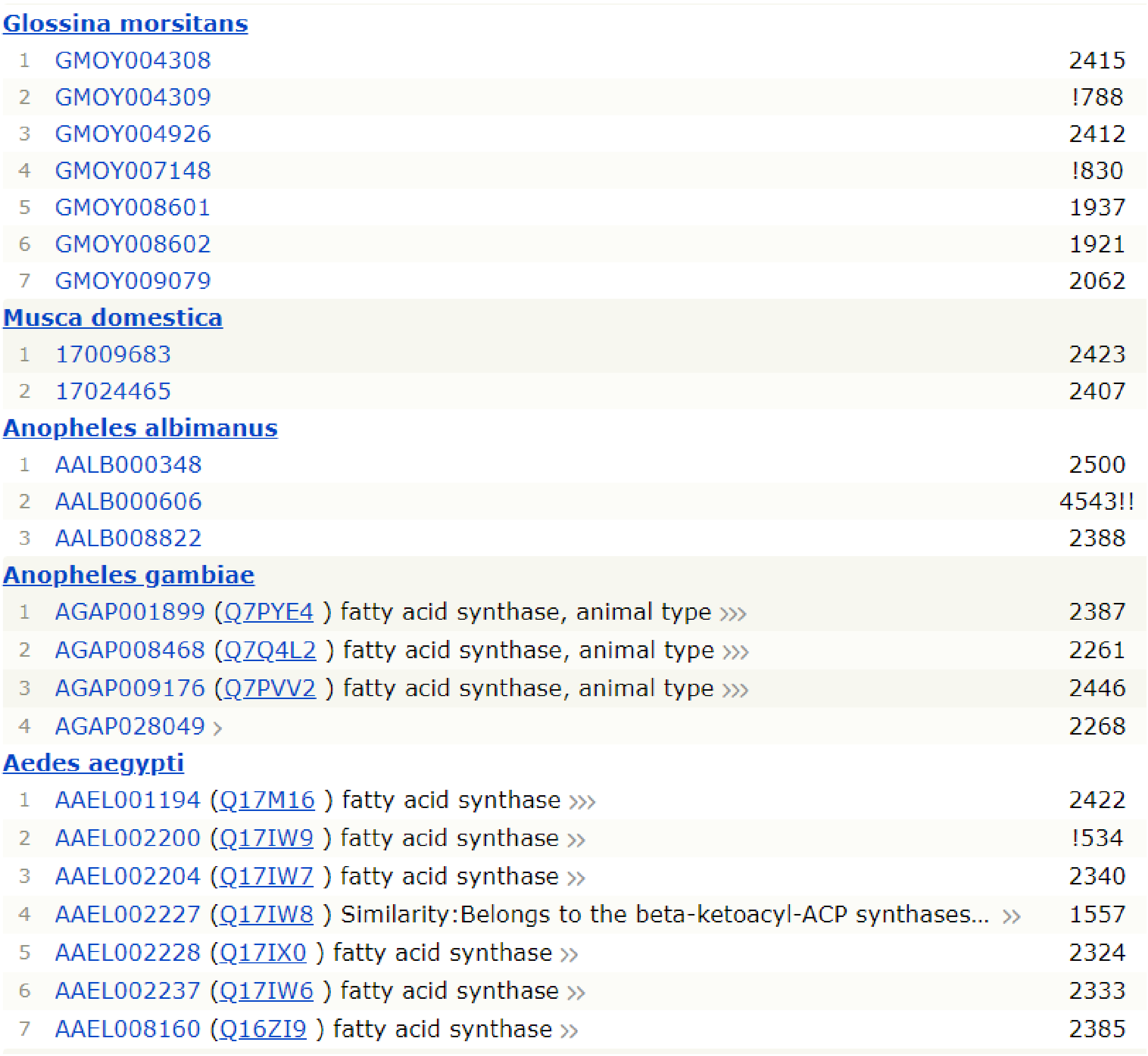
[2] Protein lengths, deviation from the median

* The three Drosophila melanogaster proteins have lengths (AAs) of 2394, 2410, and 2540
1 FBgn0040001 (Q7PLB8) Acyl carrier protein-like 2394
2 FBgn0042627 (M9PB21) v(2)k05816 2410
3 FBgn0283427 (B7Z001) Acyl carrier protein-like 2540
* Evolution happens, and lengths of orthologues can be different, but in general we would not expect very large differences unless there were truly dynamic structural changes occurring (rare)

* Under the 'Evolutionary descriptions' at the top of the page we can see
Gene Architecture: Median Protein Length 1799 (std. 931.5)
* So across all the genes in this group the Drosophila genes appear longer than the median, and if we trust the Drosophila ones then that means there are probably some short fragmented genes included in this orthologous group
* Looking at the protein lengths (AAs) of the genes from our selected species we can get a sense of what the 'normal' protein length should be, and those that are shorter or longer by one (!) or two (!!) standard deviations from the median are labelled with exclamation marks to highlight potential gene annotation issues

* So from amongst our selected species we can see most protein lengths seem to be more-or-less reasonable
* Apart from:
- Glossina morsitans has two short fragments
- Anopheles albimanus has one very long protein
- Aedes aegypti has one short fragment
* So when building a gene tree we might exclude the short fragments or trim the very long protein or attempt some manual curation of the gene models in order to produce better quality input data for the alignments

Protein lengths (AAs) can indicate gene fragments or fusions



Glossina morsitans		
1	GMOY004308	2415
2	GMOY004309	!788
3	GMOY004926	2412
4	GMOY007148	!830
5	GMOY008601	1937
6	GMOY008602	1921
7	GMOY009079	2062
Musca domestica		
1	17009683	2423
2	17024465	2407
Anopheles albimanus		
1	AALB000348	2500
2	AALB000606	4543!!
3	AALB008822	2388
Anopheles gambiae		
1	AGAP001899 (Q7PYE4) fatty acid synthase, animal type ›››	2387
2	AGAP008468 (Q7Q4L2) fatty acid synthase, animal type ›››	2261
3	AGAP009176 (Q7PVV2) fatty acid synthase, animal type ›››	2446
4	AGAP028049 ›	2268
Aedes aegypti		
1	AAEL001194 (Q17M16) fatty acid synthase ›››	2422
2	AAEL002200 (Q17IW9) fatty acid synthase ››	!534
3	AAEL002204 (Q17IW7) fatty acid synthase ››	2340
4	AAEL002227 (Q17IW8) Similarity:Belongs to the beta-ketoacyl-ACP synthases... ››	1557
5	AAEL002228 (Q17IX0) fatty acid synthase ››	2324
6	AAEL002237 (Q17IW6) fatty acid synthase ››	2333
7	AAEL008160 (Q16ZI9) fatty acid synthase ››	2385

6. **What is the length of the longest Anopheles gambiae orthologue?** *

Mark only one oval.

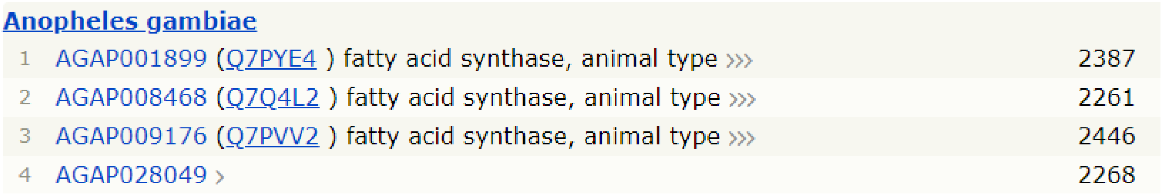
○ 2446 *Skip to question 8.*

○ 2387 *Skip to question 7.*

○ 2268 *Skip to question 7.*

Are you sure?

AGAP009176 is 2446 amino acids long



Anopheles gambiae		
1	AGAP001899 (Q7PYE4) fatty acid synthase, animal type ›››	2387
2	AGAP008468 (Q7Q4L2) fatty acid synthase, animal type ›››	2261
3	AGAP009176 (Q7PVV2) fatty acid synthase, animal type ›››	2446
4	AGAP028049 ›	2268

7. **What is the length of the longest Anopheles gambiae orthologue?** *

Mark only one oval.

○ 2446 *Skip to question 8.*

○ 2387 *Skip to question 7.*

○ 2268 *Skip to question 7.*

[C] OrthoDB API to download sequence data

We will now use the OrthoDB application programming interface to download the protein sequences of interest

[1] The OrthoDB API
* From the OrthoDB homepage, navigate through the 'Data downloads' link to the API information page
* If you cannot find it try here: https://www.orthodb.org/?page=api
* The API expects commands (defining the type of query) and arguments (refining the query)
* Scroll down to the 'fasta' section as this is where we will focus for now
* The fasta query is very simple as it takes only two arguments:
id => OrthoDB cluster id
species => list of NCBI species taxonomy id's
* So to obtain the protein sequences of the genes in a given orthologous groups from a set of species the API's URL construction would take the following form:
The type of query: https://www.orthodb.org/fasta?
The group ID: id= OrthoDB group identifier
And then the list of species: &species= comma-separated list of NCBI taxonomy identifiers

* So for the orthologous group we are interested in, and our 10 species, plus Drosophila melanogaster, the URL would be the following:

https://www.orthodb.org/fasta?id=EOG090W003L&species=7167,7165,7159,7176,7200,7227,7394,7370,79782,13249,121225

[2] Using the API from the command line
* Imagine you wanted to retrieve the protein sequence data for all the orthologous groups that showed significant variation with expansions and contractions from our CAFE analysis
* Instead of browsing each one online at OrthoDB and constructing API URLs to retrieve the data, you could write a script that collects everything that you need (sequence data, functional data, etc.) using the API from the command line
* Here we will retrieve just the fasta sequences for one group from our 11 species

```
$  wget "https://www.orthodb.org/fasta?id=EOG090W003L&species=7167,7165,7159,7176,7200,7227,7394,7370,79782,13249,121225" -O EOG090W003L.fas
```

* If you cannot make this work, then from the Moodle site find the folder under 'Day 2 Rob Waterhouse' called 'OrthoDB_gene_families', inside you should see the fasta file called 'EOG090W003L.fas'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM

```
$ wget https://edu.sib.swiss/pluginfile.php/6280/mod_folder/content/0/EOG090W003L.fas
```

* NB: if the URL you copied ends with '?forcedownload=1' then delete this part

8. **How many sequences were downloaded?** *

Mark only one oval.

○ 11 *Skip to question 9.*
○ 40 *Skip to question 9.*
○ 43 *Skip to question 10.*

Are you sure?

Try a grep on the '>' character that starts all FASTA header lines

```
$ grep -c '>' EOG090W003L.fas
```

9. **How many sequences were downloaded?** *

Mark only one oval.

○ 11 *Skip to question 9.*
○ 40 *Skip to question 9.*
○ 43 *Skip to question 10.*

[D] Align & trim

Similar to what we did earlier using the single-copy BUSCO sequences, now we will build phylogenetic trees for this multi-copy gene family (if you have chosen to work with a different orthologous group then simply substitute your orthologous group ID in the commands below)

[0] replace ':' with '_' in FASTA header
* Trimal will cut headers at the first colon, so we need to first replace them here in our input FASTA file, using sed
$ sed -i -e 's/:/_/g' EOG090W003L.fas

[1] Alignment
* Again we will use MAFFT
$ mafft EOG090W003L.fas > EOG090W003L.aln

[2] Trimming
* Again we will use TrimAl with the -strictplus option
$ trimal -in EOG090W003L.aln -out EOG090W003L.aln.sp.trm -strictplus
* We will also try the -automated1 option
$ trimal -in EOG090W003L.aln -out EOG090W003L.aln.at.trm -automated1
* TrimAl also has options for the 'Automated removal of spurious sequences' (see userguide), and as we know there are likely some fragment genes in our dataset this could be useful
$ trimal -in EOG090W003L.aln.sp.trm -out EOG090W003L.aln.spov.trm -resoverlap 0.7 -seqoverlap 70
$ trimal -in EOG090W003L.aln.at.trm -out EOG090W003L.aln.atov.trm -resoverlap 0.7 -seqoverlap 70

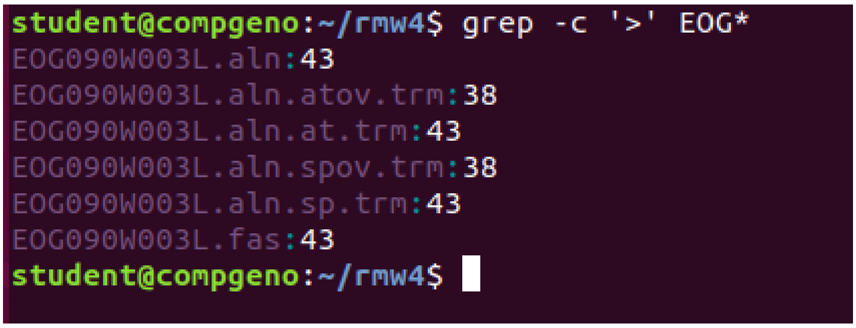
10. **How many genes/proteins were removed after trimming and removal of spurious sequences?** *

Mark only one oval.

○ 4 *Skip to question 11.*
○ 6 *Skip to question 11.*
○ 5 *Skip to question 12.*

Are you sure?

Try grep -c '>' on all your fasta files



```
student@compgeno:~/rmw4$ grep -c '>' EOG*
EOG090W003L.aln:43
EOG090W003L.aln.atov.trm:38
EOG090W003L.aln.at.trm:43
EOG090W003L.aln.spov.trm:38
EOG090W003L.aln.sp.trm:43
EOG090W003L.fas:43
student@compgeno:~/rmw4$ 
```

11. **How many genes/proteins were removed after trimming and removal of spurious sequences?** *

Mark only one oval.

○ 4 *Skip to question 11.*
○ 6 *Skip to question 11.*
○ 5 *Skip to question 12.*

[E] Gene family trees

[1] Now we can use RAxML again to build our gene trees

* You might ask RAxML to select the best substitution model, but here we will stick with PROTGAMMAJTT
* First the 'automated1' trimmed alignment

```
$ raxmlHPC-SSE3 -s EOG090W003L.aln.atov.trm -n EOG090W003L_atov -f a -N 3 -x 12345 -p 12345 -m PROTGAMMAJTT >& log-atov.txt &
```

* Then the 'strictplus' trimmed alignment

```
$ raxmlHPC-SSE3 -s EOG090W003L.aln.spov.trm -n EOG090W003L_spov -f a -N 3 -x 12345 -p 12345 -m PROTGAMMAJTT >& log-spov.txt &
```

* NB: you would normally do this with more than just 3 bootstrap samples
* If the alignment and trimming did not work for you and/or you don't want to wait for the trees to finish, then from the Moodle site find the folder under 'Day 2 Rob Waterhouse' called 'OrthoDB_gene_families', inside you should see the tarball called 'AlnTrmTre.tar.gz'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM

```
$ wget https://edu.sib.swiss/pluginfile.php/6280/mod_folder/content/0/AlnTrmTre.tar.gz
```

* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* Untarzip the tarball

```
$ tar -xf AlnTrmTre.tar.gz
```

* NB: these trees were run with 10 bootstrap samples

* Note you have several output files from running RAxML:

RAxML_bestTree.xxx
RAxML_bipartitionsBranchLabels.xxx
RAxML_bipartitions.xxx
RAxML_bootstrap.xxx
RAxML_info.xxx

* The bipartitions files will have the bootstrap support values on the trees.
* As we had NCBI taxonomy IDs as part of the gene IDs let's first convert them to species codes so we can more easily interpret the tree
* First make copies of the trees

```
$ cp RAxML_bipartitions.EOG090W003L_atov EOG090W003L_atov_tree.txt
$ cp RAxML_bipartitions.EOG090W003L_spov EOG090W003L_spov_tree.txt
```

* Then run the following sed commands to perform the replacements on the tree files:

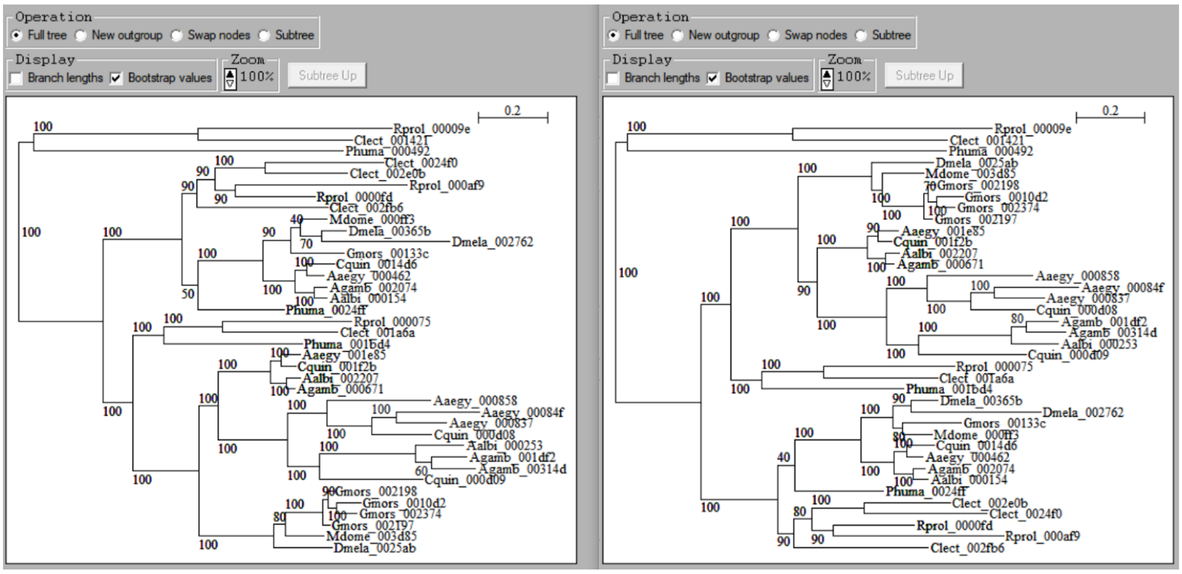
```
sed -i -e 's/121225_/Phuma_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/13249_/Rprol_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/79782_/Clect_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/7159_/Aaegy_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/7176_/Cquin_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/7165_/Agamb_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/7167_/Aalbi_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/7200_/Llong_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/7370_/Mdome_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/7394_/Gmors_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
sed -i -e 's/7227_/Dmela_/g' EOG090W003L_atov_tree.txt EOG090W003L_spov_tree.txt
```

* Now you can visualise the trees using your favourite tree viewer
* It can be hard to spot any differences by eye, so here newick utilities can be useful
* Take the bestTree files (no bootstrap values) and extract just their topologies, then order them, then compare

```
$ nw_topology RAxML_bestTree.EOG090W003L_atov | nw_order - > atov.txt
$ nw_topology RAxML_bestTree.EOG090W003L_spov | nw_order - > spov.txt
$ diff atov.txt spov.txt
```

* NB: http://phylo.io/ is a handy tool for comparing similar trees

EOG090W003L spov_tree (left) atov_tree (right)



12. **Can you find the (topology) difference?** *

Mark only one oval.

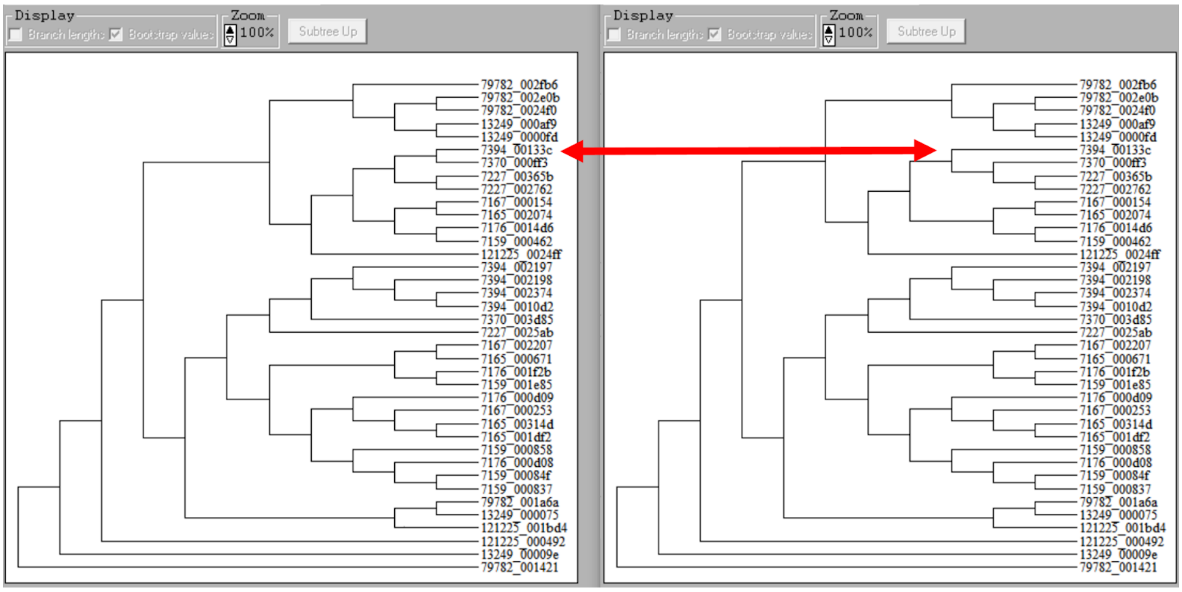
○ 13249_000075 *Skip to question 13.*

○ 7394_00133c *Skip to "Interpreting the tree."*

○ 79782_0024f0 *Skip to question 13.*

Are you sure?

Compare topologies: they are almost the same, except for one gene



13. **Can you find the (topology) difference?** *

Mark only one oval.

○ 13249_000075 *Skip to question 13.*

○ 7394_00133c *Skip to "Interpreting the tree."*

○ 79782_0024f0 *Skip to question 13.*

Interpreting the tree

Of course there are many parameters that one can explore at many stages of such analyses, from how the gene families (orthologous groups) were defined at the start, so how we selected the ones to analyse with CAFE, to which ones to then examine in more detail, to methods for alignment and filtering, and finally tree building and visualisation.

This exercise should hopefully have given you a flavour for how some of these steps can proceed, and here with EOG090W003L we can see (figure below) at least three potentially interesting expansions.

As you saw - we discarded some proteins, as they appeared to be incomplete fragments, so keep in mind that the tree below may not represent the whole evolutionary history of these genes in these species. To fully investigate this would require some manual curation and further sequence searches.

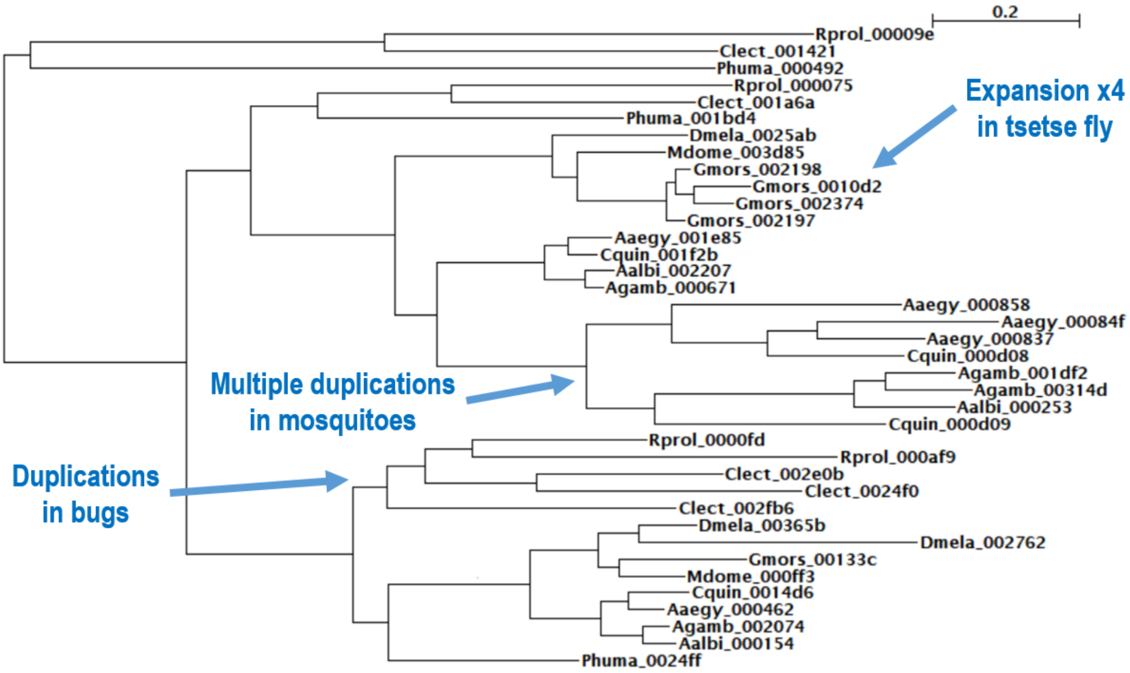
As well as the three expansions, there is also a duplication in Drosophila melanogaster (not highlighted on the figure but easy to find). This duplication, and the multiple duplications in mosquitoes are particularly interesting as they show rather long branch lengths which suggests perhaps sequence evolution under relaxed constraint and potentially adaptive selection - possibly with functional consequences for these proteins.

Hopefully by completing this exercise you:

[1] have gained some more experience and knowledge of the orthology and sequence data available from OrthoDB and how to extract what you need
[2] have an understanding of how to build gene trees for rapidly evolving multi-copy gene families and some of the important data-accuracy issues that can impact on such analyses

Click SUBMIT below to finish this exercise.

Several clear expansions are visible



Expansion x4 in tsetse fly

Multiple duplications in mosquitoes

Duplications in bugs

Stop filling out this form.

We're here to help!

If you're stuck, please raise your hand and hopefully one of us will be able to help you.

Hit BACK below to return to the practical.