#3 Gene Family Evolution

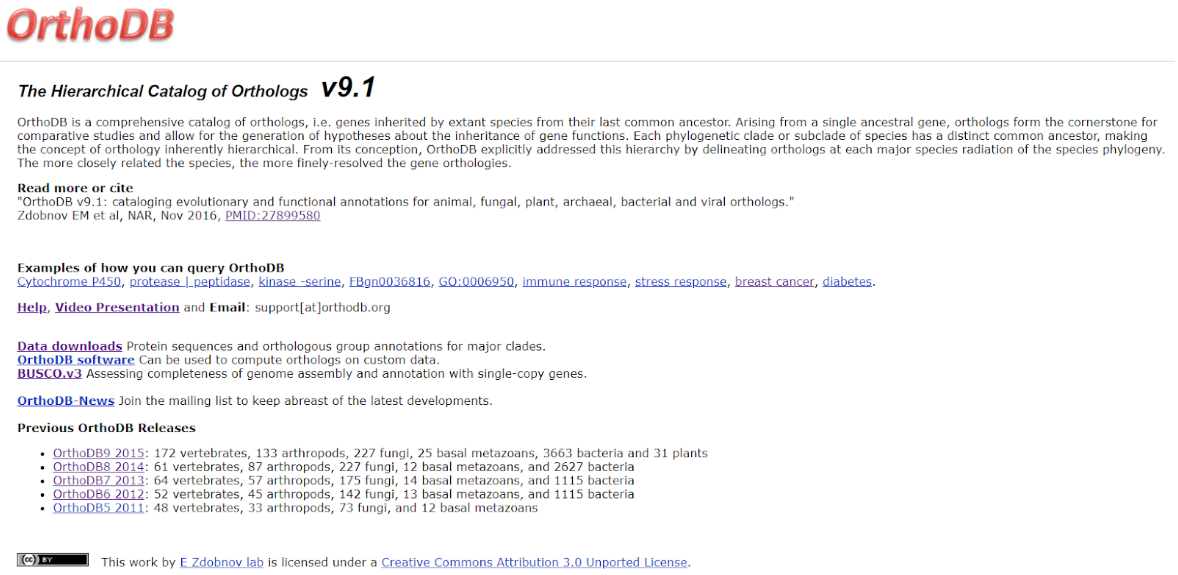
For this exercise we will gather orthologues from all of the 10 insects that we have been working with, and then using the species phylogeny that we built we will perform an analysis of gene family evolution to try to identify the branches with the most rapidly and dramatically changing gene families across our set of species

By the end of this first exercise you should:

[1] have some knowledge of the orthology data available from OrthoDB and how to extract what you need
[2] have an understanding of how to perform gene family evolution analyses with the Computational Analysis of gene Family Evolution (CAFE) package
[3] have learnt how it is important to consider both technical and biological reasons that could explain your results from large-scale comparative genomics 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



OrthoDB

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!."*

[A1] Obtaining orthology data from OrthoDB

We will start by exploring a few options for retrieving orthology data from OrthoDB

[1] Online browsing of OrthoDB
* First go to www.orthodb.org
* Then use the species selector on the right to select the 10 insect species of interest

Aedes aegypti
Anopheles albimanus
Anopheles gambiae
Cimex lectularius
Culex quinquefasciatus
Glossina morsitans

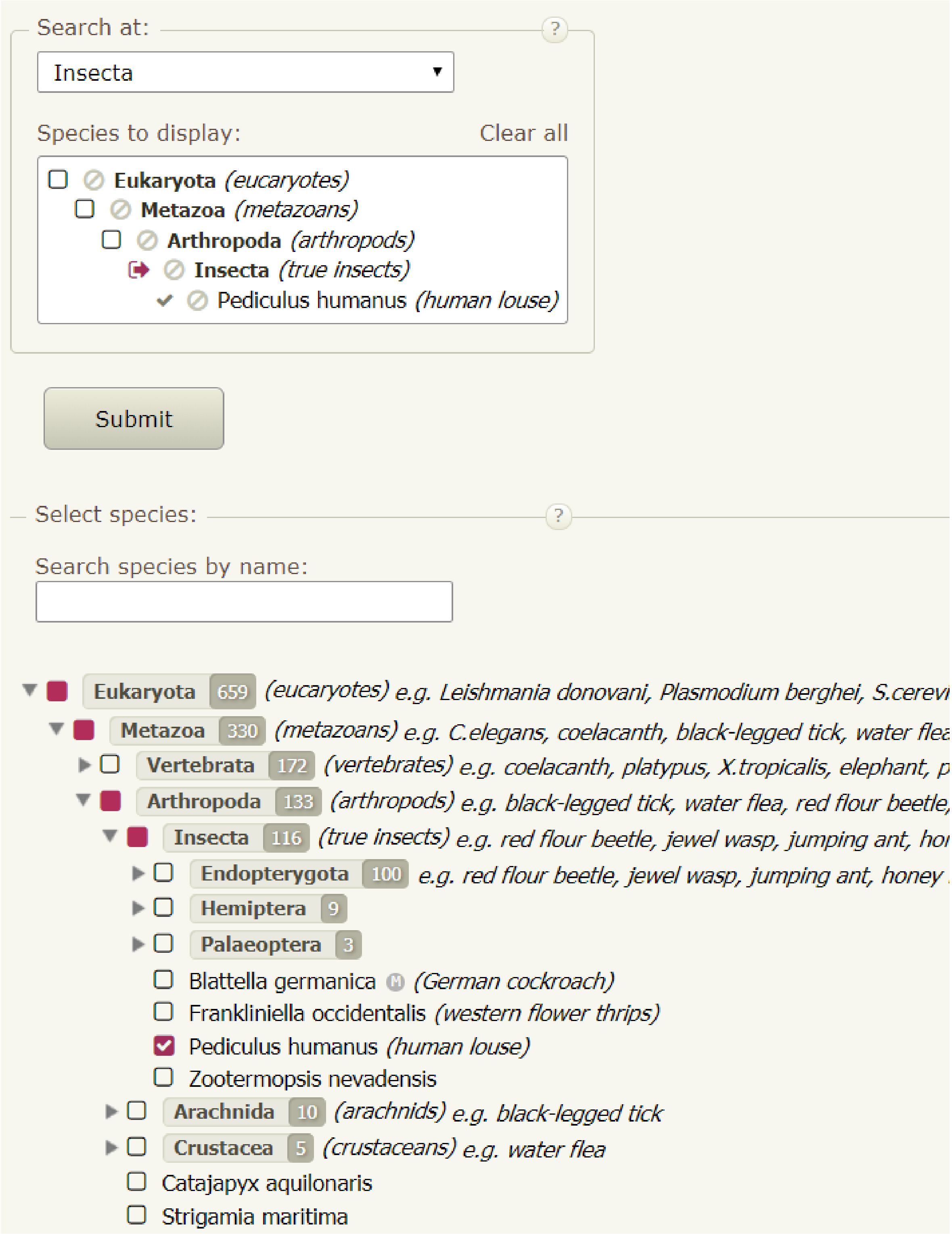
Lutzomyia longipalpis
Musca domestica
Pediculus humanus
Rhodnius prolixus

* Note that you can type the species names into the 'Search species by name:' box, or you can expand the tree to find each species and select them
* Notice how the 'Species to display:' box above the selector section changes as you add more species (and if you have visited OrthoDB before then you might have to first clear the species to display box before selecting your 10 insects)
* Note also that the 'Search at:' box will automatically choose the phylogenetic level of the last common ancestor of all your selected species

* Once you have selected the 10 species, click the 'Submit' button (above the species selection panel) without specifying any text search or any phyletic profile filters, i.e. a blank search to return all groups at the Insecta level

* To familiarise yourself with OrthoDB orthologous groups you can expand the top result, e.g. EOG090W001L, by clicking on the double chevron '>>'
* You will see summary 'Functional descriptions' including Gene Ontology terms and InterPro domains
* You will see summary 'Evolutionary descriptions' including phyletic profile, evolutionary rate, and gene architecture
* Below that is the actual list of orthologs by organism - note only orthologues from the 10 selected species are currently displayed
* You can use the 'Selected species only' tick-box to show or hide all the other species with orthologues in this orthologous group
* Below the orthologs by organism you will see 'Sibling groups', these are other Insecta-level orthologous groups that share all or some InterPro domains with this group

Species selection



Search at: ?

Insecta ▼

Species to display: Clear all

□ ⊘ **Eukaryota** *(eucaryotes)*
□ ⊘ **Metazoa** *(metazoans)*
□ ⊘ **Arthropoda** *(arthropods)*
➔ ⊘ **Insecta** *(true insects)*
✓ ⊘ Pediculus humanus *(human louse)*

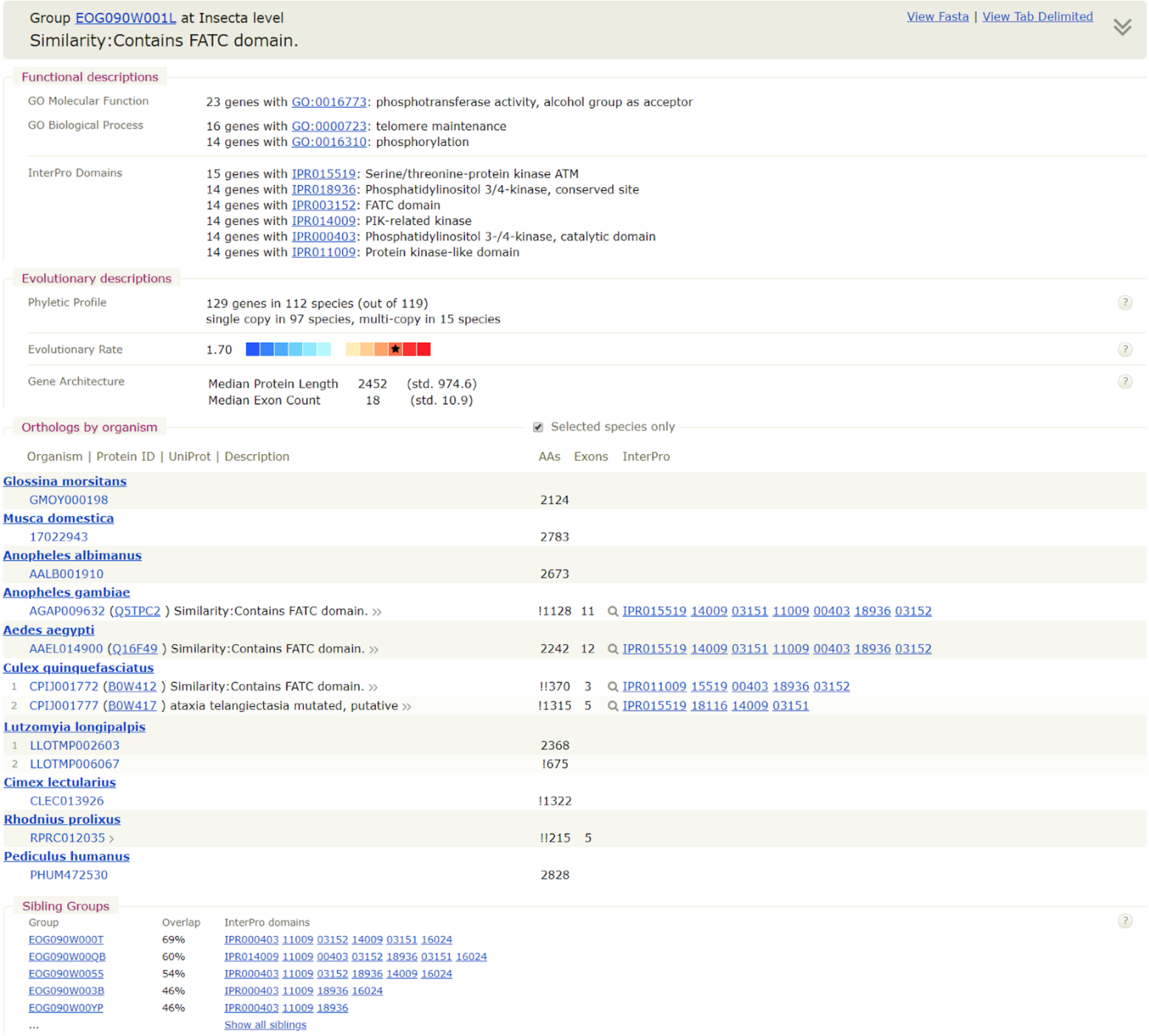
Submit

Select species: ?

Search species by name:

▼ ■ **Eukaryota** 659 *(eucaryotes) e.g. Leishmania donovani, Plasmodium berghei, S.cerevi*
▼ ■ **Metazoa** 330 *(metazoans) e.g. C.elegans, coelacanth, black-legged tick, water flea*
▶ □ **Vertebrata** 172 *(vertebrates) e.g. coelacanth, platypus, X.tropicalis, elephant, p*
▼ ■ **Arthropoda** 133 *(arthropods) e.g. black-legged tick, water flea, red flour beetle,*
▼ ■ **Insecta** 116 *(true insects) e.g. red flour beetle, jewel wasp, jumping ant, ho*
▶ □ **Endopterygota** 100 *e.g. red flour beetle, jewel wasp, jumping ant, honey*
▶ □ **Hemiptera** 9
▶ □ **Palaeoptera** 3
□ Blattella germanica Ⓜ *(German cockroach)*
□ Frankliniella occidentalis *(western flower thrips)*
☑ Pediculus humanus *(human louse)*
□ Zootermopsis nevadensis
▶ □ **Arachnida** 10 *(arachnids) e.g. black-legged tick*
▶ □ **Crustacea** 5 *(crustaceans) e.g. water flea*
□ Catajapyx aquilonaris
□ Strigamia maritima

Example orthologous group



Group EOG090W001L at Insecta level
Similarity:Contains FATC domain.

Functional descriptions

GO Molecular Function	23 genes with GO:0016773: phosphotransferase activity, alcohol group as acceptor
GO Biological Process	16 genes with GO:0000723: telomere maintenance 14 genes with GO:0016310: phosphorylation
InterPro Domains	15 genes with IPR015519: Serine/threonine-protein kinase ATM 14 genes with IPR018936: Phosphatidylinositol 3/4-kinase, conserved site 14 genes with IPR003152: FATC domain 14 genes with IPR014009: PIK-related kinase 14 genes with IPR000403: Phosphatidylinositol 3-/4-kinase, catalytic domain 14 genes with IPR011009: Protein kinase-like domain

Evolutionary descriptions

Phyletic Profile	129 genes in 112 species (out of 119) single copy in 97 species, multi-copy in 15 species
Evolutionary Rate	1.70
Gene Architecture	Median Protein Length 2452 (std. 974.6) Median Exon Count 18 (std. 10.9)

Orthologs by organism ☑ Selected species only

Organism | Protein ID | UniProt | Description	AAs	Exons	InterPro
Glossina morsitans			
GMOY000198	2124		
Musca domestica			
17022943	2783		
Anopheles albimanus			
AALB001910	2673		
Anopheles gambiae			
AGAP009632 (Q5TPC2) Similarity:Contains FATC domain. »	!1128	11	IPR015519 14009 03151 11009 00403 18936 03152
Aedes aegypti			
AAEL014900 (Q16F49) Similarity:Contains FATC domain. »	2242	12	IPR015519 14009 03151 11009 00403 18936 03152
Culex quinquefasciatus			
1 CPIJ001772 (B0W412) Similarity:Contains FATC domain. »	!!370	3	IPR011009 15519 00403 18936 03152
2 CPIJ001777 (B0W417) ataxia telangiectasia mutated, putative »	!1315	5	IPR015519 18116 14009 03151
Lutzomyia longipalpis			
1 LLOTMP002603	2368		
2 LLOTMP006067	!675		
Cimex lectularius			
CLEC013926	!1322		
Rhodnius prolixus			
RPRC012035 ›	!!215	5	
Pediculus humanus			
PHUM472530	2828		

Sibling Groups

Group	Overlap	InterPro domains
EOG090W000T	69%	IPR000403 11009 03152 14009 03151 16024
EOG090W00QB	60%	IPR014009 11009 00403 03152 18936 03151 16024
EOG090W0055	54%	IPR000403 11009 03152 18936 14009 16024
EOG090W003B	46%	IPR000403 11009 18936 16024
EOG090W00YP	46%	IPR000403 11009 18936
...		Show all siblings

2. **How many groups in total are returned from the 'blank' search of Insecta?** *

Mark only one oval.

○ 5714 *Skip to question 3.*

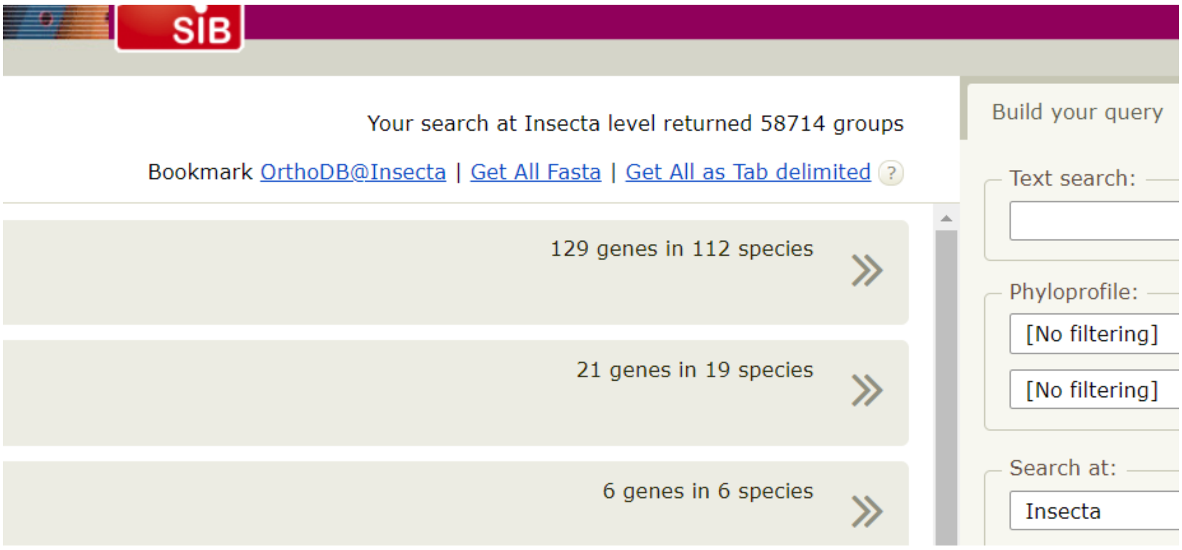
○ 58714 *Skip to question 4.*

○ 51792 *Skip to question 3.*

Are you sure?

Make sure you are searching the Insecta level, without specifying any text search or any phyletic profile filters

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



Your search at Insecta level returned 58714 groups

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

129 genes in 112 species

21 genes in 19 species

6 genes in 6 species

Build your query

Text search:

Phyloprofile: [No filtering] [No filtering]

Search at: Insecta

3. **How many groups are returned?** *

Mark only one oval.

○ 5714 *Skip to question 3.*

○ 58714 *Skip to question 4.*

○ 51792 *Skip to question 3.*

[A2] Obtaining orthology data from OrthoDB

* If you try to download all the data for the full set of search results OrthoDB will politely suggest that you instead download the full datafiles from the downloads page - this is because download options from the search results pages are designed mainly for smaller, more focused, searches, e.g. all orthologous groups with proteins that have a specific InterPro domain, rather than the full set of all orthologous groups!

* Thus we will instead download the data from the downloads page and extract what we need from those datafiles.

* From the OrthoDB homepage, navigate through the 'Data downloads' link to the 'Flat files' page to see what is available for download from OrthoDB

* All the Insecta-level orthologous groups have already been downloaded and extracted for you
* FIRSTLY - from your HOME directory

```
$ mkdir rmw3
$ cd rmw3
```

* Now, from the Moodle site find the folder under 'Day 2 Rob Waterhouse' called 'OrthoDB_gene_families', inside you should see the gzipped file called 'ODB9_Insecta_OGs.txt.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/ODB9_Insecta_OGs.txt.gz
```

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

```
$ gunzip ODB9_Insecta_OGs.txt.gz
```

* Also from the Moodle site find the folder under 'Day 2 Rob Waterhouse' called 'OrthoDB_gene_families', where you should see the Perl script called 'get_ODB9_gene_counts_per_species.pl'
* 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/get_ODB9_gene_counts_per_species.pl
```

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

* Feel free to open up the script to get an idea of what it is trying to do, essentially the aim is to filter the groups to select only those with >7 of our 10 species, and that include our outgroup species, Pediculus humanus, and count the number of genes per species per orthologous group

* Look at the starting data

```
$ more ODB9_Insecta_OGs.txt
```

* It contains orthologous group IDs and gene IDs (gene IDs are composed of

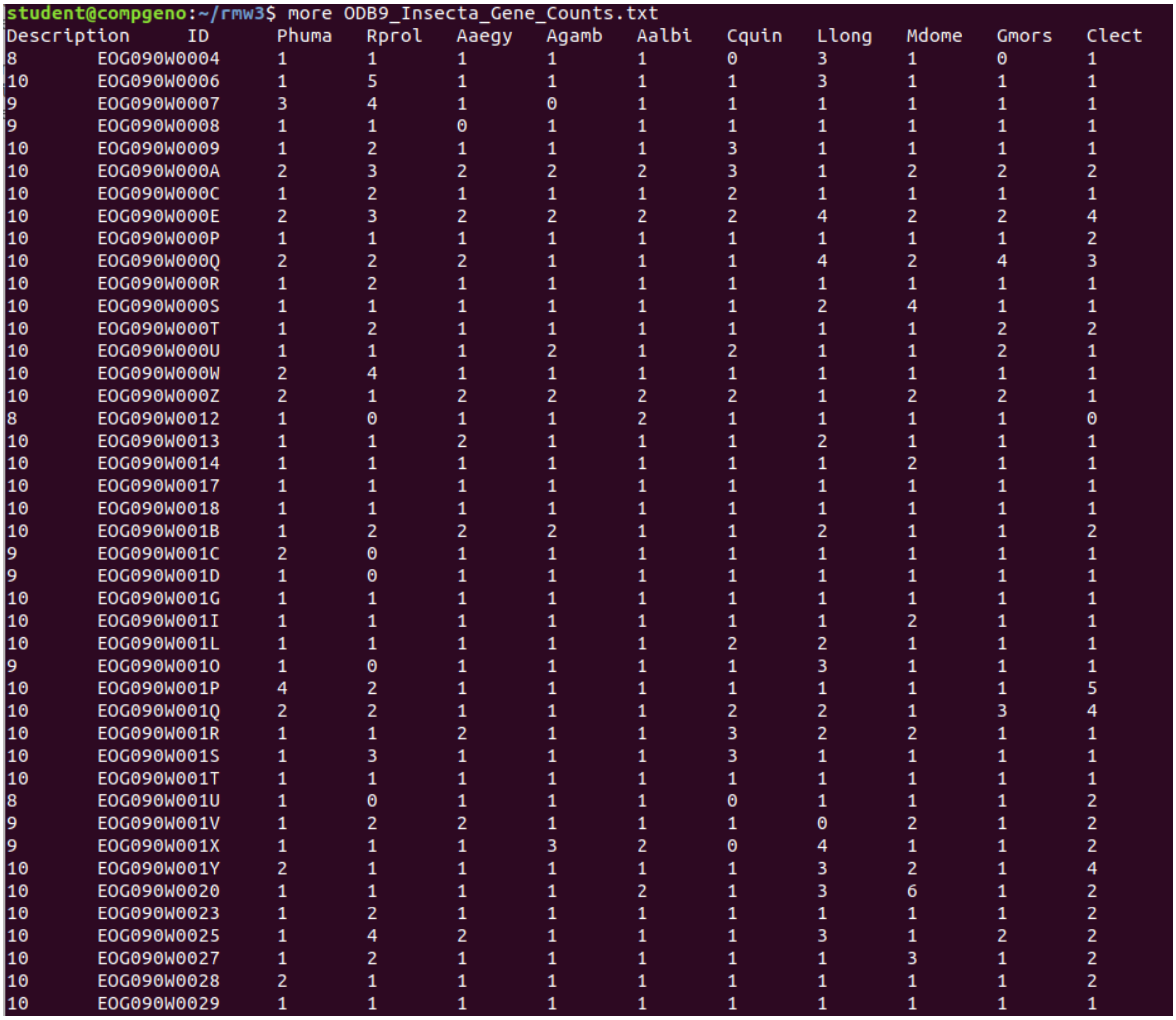
NCBItaxonomyID:Unique6characterID)
EOG090W0000	103372:0012a8
EOG090W0000	103372:0012a9
EOG090W0000	104421:001a20
EOG090W0000	1049336:004382
EOG090W0000	1049336:00465a
EOG090W0000	1049336:0046fc
EOG090W0000	1049336:0046fd

* Check how many orthologous groups there are - this should be the same as our online search revealed previously
$ grep EOG ODB9_Insecta_OGs.txt | cut -f1 | uniq | wc
* grep, cut, uniq, wc are all very useful unix commands - just ask or Google them to find out more

* Now you can run the script
$ perl get_ODB9_gene_counts_per_species.pl &
* And have a look at the output - this is formatted for use with the Computational Analysis of gene Family Evolution (CAFE) tool that we will be using later
$ more ODB9_Insecta_Gene_Counts.txt
* NB: If this does not work you can obtain the ODB9_Insecta_Gene_Counts.txt.gz file from the Moodle site
* NB: the first column, 'Description', is required by CAFE, here we just enumerate the number of species in the group

=====
As an aside and for those that are interested, there are also options to extract data from OrthoDB using the application programming interface (API), and there are some example scripts available at this GitLab project: https://gitlab.com/rmwaterhouse/OrthoDB-API-Scripting
=====

ODB9_Insecta_Gene_Counts.txt



```
student@compgeno:~/rmw3$ more ODB9_Insecta_Gene_Counts.txt
Description     ID      Phuma   Rprol   Aaegy   Agamb   Aalbi   Cquin   Llong   Mdome   Gmors   Clect
8       EOG090W0004     1       1       1       1       1       0       3       1       0       1
10      EOG090W0006     1       5       1       1       1       1       3       1       1       1
9       EOG090W0007     3       4       1       0       1       1       1       1       1       1
9       EOG090W0008     1       1       0       1       1       1       1       1       1       1
10      EOG090W0009     1       2       1       1       1       3       1       1       1       1
10      EOG090W000A     2       3       2       2       2       3       1       2       2       2
10      EOG090W000C     1       2       1       1       1       2       1       1       1       1
10      EOG090W000E     2       3       2       2       2       2       4       2       2       4
10      EOG090W000P     1       1       1       1       1       1       1       1       1       2
10      EOG090W000Q     2       2       2       1       1       1       4       2       4       3
10      EOG090W000R     1       2       1       1       1       1       1       1       1       1
10      EOG090W000S     1       1       1       1       1       1       2       4       1       1
10      EOG090W000T     1       2       1       1       1       1       1       1       2       2
10      EOG090W000U     1       1       1       2       1       2       1       1       2       1
10      EOG090W000W     2       4       1       1       1       1       1       1       1       1
10      EOG090W000Z     2       1       2       2       2       2       1       2       2       1
8       EOG090W0012     1       0       1       1       2       1       1       1       1       0
10      EOG090W0013     1       1       2       1       1       1       2       1       1       1
10      EOG090W0014     1       1       1       1       1       1       1       2       1       1
10      EOG090W0017     1       1       1       1       1       1       1       1       1       1
10      EOG090W0018     1       1       1       1       1       1       1       1       1       1
10      EOG090W001B     1       2       2       2       1       1       2       1       1       2
9       EOG090W001C     2       0       1       1       1       1       1       1       1       1
9       EOG090W001D     1       0       1       1       1       1       1       1       1       1
10      EOG090W001G     1       1       1       1       1       1       1       1       1       1
10      EOG090W001I     1       1       1       1       1       1       1       2       1       1
10      EOG090W001L     1       1       1       1       1       2       2       1       1       1
9       EOG090W001O     1       0       1       1       1       1       3       1       1       1
10      EOG090W001P     4       2       1       1       1       1       1       1       1       5
10      EOG090W001Q     2       2       1       1       1       2       2       1       3       4
10      EOG090W001R     1       1       2       1       1       3       2       2       1       1
10      EOG090W001S     1       3       1       1       1       3       1       1       1       1
10      EOG090W001T     1       1       1       1       1       1       1       1       1       1
8       EOG090W001U     1       0       1       1       1       0       1       1       1       2
9       EOG090W001V     1       2       2       1       1       1       0       2       1       2
9       EOG090W001X     1       1       1       3       2       0       4       1       1       2
10      EOG090W001Y     2       1       1       1       1       1       3       2       1       4
10      EOG090W0020     1       1       1       1       2       1       3       6       1       2
10      EOG090W0023     1       2       1       1       1       1       1       1       1       2
10      EOG090W0025     1       4       2       1       1       1       3       1       2       2
10      EOG090W0027     1       2       1       1       1       1       1       3       1       2
10      EOG090W0028     2       1       1       1       1       1       1       1       1       2
10      EOG090W0029     1       1       1       1       1       1       1       1       1       1
```

4. **How many groups remain after our filtering?** *

Mark only one oval.

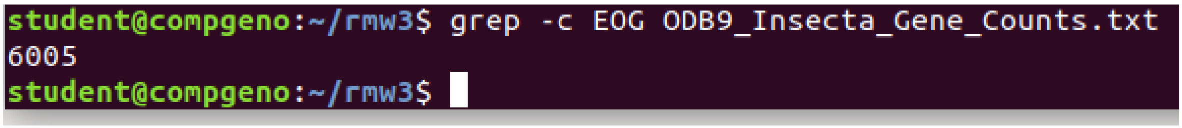
○ 6006 *Skip to question 5.*

○ 5999 *Skip to question 5.*

○ 6005 *Skip to question 6.*

Are you sure?

Try:
$ grep -c EOG ODB9_Insecta_Gene_Counts.txt



```
student@compgeno:~/rmw3$ grep -c EOG ODB9_Insecta_Gene_Counts.txt
6005
student@compgeno:~/rmw3$ █
```

5. **How many groups remain after our filtering?** *

Mark only one oval.

○ 6006 *Skip to question 5.*

○ 5999 *Skip to question 5.*

○ 6005 *Skip to question 6.*

[B] Computational Analysis of gene Family Evolution (CAFE)

Computational Analysis of gene Family Evolution (CAFE) is a popular tool for statistical analyses of gene family dynamics across phylogenies
CAFE GitHub page: https://github.com/hahnlab/CAFE
CAFE website: https://hahnlab.github.io/CAFE/

[1] Usage (from GitHub):

The necessary inputs for CAFE v4.2 are:
- a data file containing gene family sizes for the taxa included in the phylogenetic tree
- a Newick formatted phylogenetic tree, including branch lengths

From the inputs above, CAFE v4.2 will compute:
- the maximum likelihood value of the birth & death parameter, λ (or of separate birth and death parameters (λ and μ, respectively), over the whole tree or for user-specified subsets of branches in the tree
- ancestral states of gene family sizes for each node in the phylogenetic tree
- p-values for each gene family describing the likelihood of the observed sizes given average rates of gain and loss
- average gene family expansion along each branch in the tree
- numbers of gene families with expansions, contractions, or no change along each branch in the tree

* We have already prepared:
[a] our gene family sizes input data
[b] our ultrametric species phylogeny

[2] Gene count data
* Check the top of our gene family sizes file
$ head ODB9_Insecta_Gene_Counts.txt
* You will see that the species names don't exactly match the assembly codes we have in our phylogeny, so first we must fix this so that they are the same
Sizes: Phuma Rprol Aaegy Agamb Aalbi Cquin Llong Mdome Gmors Clect
Phylogeny: ((((((aaegl5:49,cpipj2:49):37,(aalbs2:43,agamp4:43):43):83,llonj1:169):16,(gmory1:71,mdoma1:71):114):98,(clech1:143,rproc3:143):140):83,phumu2:366);
Edited phylogeny: ((((((Aaegy:49,Cquin:49):37,(Aalbi:43,Agamb:43):43):83,Llong:169):16,(Gmors:71,Mdome:71):114):98,(Clect:143,Rprol:143):140):83,Phuma:366)
* Also note that the semi-colon has been removed (CAFE requires no semi-colon at the end of the newick tree)

[3] Setting up to run CAFE

* So now you can check that CAFE is installed and accessible

```
$ cafe -v
```

* Create a directory where you want the CAFE results to be saved

```
$ mkdir reports
```

* CAFE runs with a control file where the options and starting parameters are defined
* For details see the GitHub and website, here briefly to explain the contents of the control file:

```
=====
#!shell    <- the header for a shell script
date        <- here we print the date and time just to record the start time

# load gene counts
load -i ODB9_Insecta_Gene_Counts.txt -p 0.01 -t 1 -l log.txt
# -i for 'input', i.e. the gene counts input file
# -p for 'p-value', i.e. the p-value cut-off for each orthologous group, only groups with count variation p-values lower than this cut-off will be analysed for expansions/contractions
# -t for 'threads', i.e. the number of threads to use when running, here just 1
# -l for 'log', i.e. a filename where CAFE should write the log details

# load ultrametric tree <- here we give our previously computed tree
tree ((((((Aaegy:49,Cquin:49):37,(Aalbi:43,Agamb:43):43):83,Llong:169):16,(Gmors:71,Mdome:71):114):98,(Clect:143,Rprol:143):140):83,Phuma:366)

# search for 1-parameter model
lambda -s -t ((((((1,1)1,(1,1)1)1,1)1,(1,1)1)1,(1,1)1)1,1)
# In CAFE terms, lambda represents the rate of change of evolution in a tree
# Here we use the '-s' option, so CAFE will search using an optimization algorithm to find the value(s) of lambda that maximize the log likelihood of the data for all families
# The '-t' allows users to investigate whether different parts of the tree are evolving at different rates, the user may specify which branches of the tree will take the same or different lambda values, here we set all to the same value (1), note that the lambda tree must have the same structure as the ultrametric tree

# Where should the results be written
report reports/report_runINSECTA
date        <- here we print the date and time just to record the end time
=====
```

[4] Running CAFE

* From the Moodle site find the folder under 'Day 2 Rob Waterhouse' called 'OrthoDB_gene_families', inside you should see the control file called 'run_cafe_params.txt'
* 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/run_cafe_params.txt
```

* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* Now all you have to do is call CAFE and point it to the control file

```
$ cafe run_cafe_params.txt >& cafe.log.txt &
```

* Remember you can check that is it actually running

```
$ ps -uf
```

[5] CAFE results

* If you don't want to wait for CAFE to finish running then you can download the pre-computed 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
```

* Have a look at the results

```
$ more report_run1.cafe
```

* Here is a description of the report contents from the CAFE website:

Tree: The current tree;

λ(s) and likelihood: The current λ values set by the lambda command; it can be either specified by the user (-t) or obtained by searching for the maximum likelihood value (-s). The likelihood of the data given the current λ value;

Average expansion: Mean number of genes gained or lost per family, where "minus" expansion is a net contraction;

Expansions and contractions: Total count of families that experienced expansions, contractions, or no change along each branch of the species tree;

List of family and description;

List of overall p-value for each family: The p-values are based on a Monte-Carlo re-sampling procedure. To determine the probability of a gene family with the observed sizes among taxa, CAFE will generate the expected distribution of family sizes under the stochastic birth-death model for the tree specified in the load command with the current λ value. Running the simulations uses the most machine resources and thus is the most time intensive step in CAFE. For each family in the data file, CAFE computes a probability (p-value) of observing the data given the average rate of gain and loss of genes. All else being equal, families with more variance in size are expected to have lower p-values.

List of branch-specific p-values for the significant families: The branch-specific p-values are obtained by the Viterbi method with the randomly generated likelihood distribution. This method calculates exact p-values for transitions between the parent and child family sizes for all branches of the phylogenetic tree. A low p-value indicates a rapidly evolving branch. This information is reported only for the families with an overall p-value less than the p-value cutoff set with the load command.

List of ancestral states for each family: Reports the maximum likelihood values of the ancestral number of genes at all inner nodes of all gene families.

[6] Expansions and contractions across the phylogeny
* Looking at the CAFE report (either open it, or use the given grep commands below):
* (a) can you find the input tree line?

```
$ grep 'Tree:' reports/report_runINS.cafe
```

* (b) can you find the node-labelled tree (i.e. the numbers that CAFE assigns each node on the tree) line?

```
$ grep '# IDs of nodes:' reports/report_runINS.cafe
```

* (c) can you find the line that shows the order (in terms of node pairs, i.e. branches) in which the results are shown?

```
$ grep '# Output format for:' reports/report_runINS.cafe
```

* (d) can you find the line that reports the numbers of expansions?

```
$ grep 'Expansion :' reports/report_runINS.cafe
```

* (e) can you find the line that reports the numbers of non-changing orthologous groups?

```
$ grep 'nRemain :' reports/report_runINS.cafe
```

* (f) can you find the line that reports the numbers of contractions?

```
$ grep 'nDecrease :' reports/report_runINS.cafe
```

* What we would really like is our beautiful ultrametric tree with all the branches labelled with the numbers of expanding and contracting gene families, so you can painstakingly copy and paste the correct values onto your newick-formatted tree and then import this into a tree visualisation tool like EvolView that we used previously.
* This is a bit boring so it has been done for you here:

```
((((((Aaegy<0>e761.c201:49,Cquin<2>e938.c204:49)<1>e303.c22:37,
(Aalbi<4>e121.c695:43,Agamb<6>e335.c151:43)<5>e24.c86:43)
<3>e217.c61:83,Llong<8>e521.c1381:169)<7>e24.c4:16,
(Gmors<10>e466.c682:71,Mdome<12>e647.c88:71)<11>e177.c181:114)<9>e136.c78:98,
(Clect<14>e562.c365:143,Rprol<16>e499.c544:143)<15>e159.c249:140)
<13>e11.c3:83,Phuma<18>e264.c249:366)<17>;
```

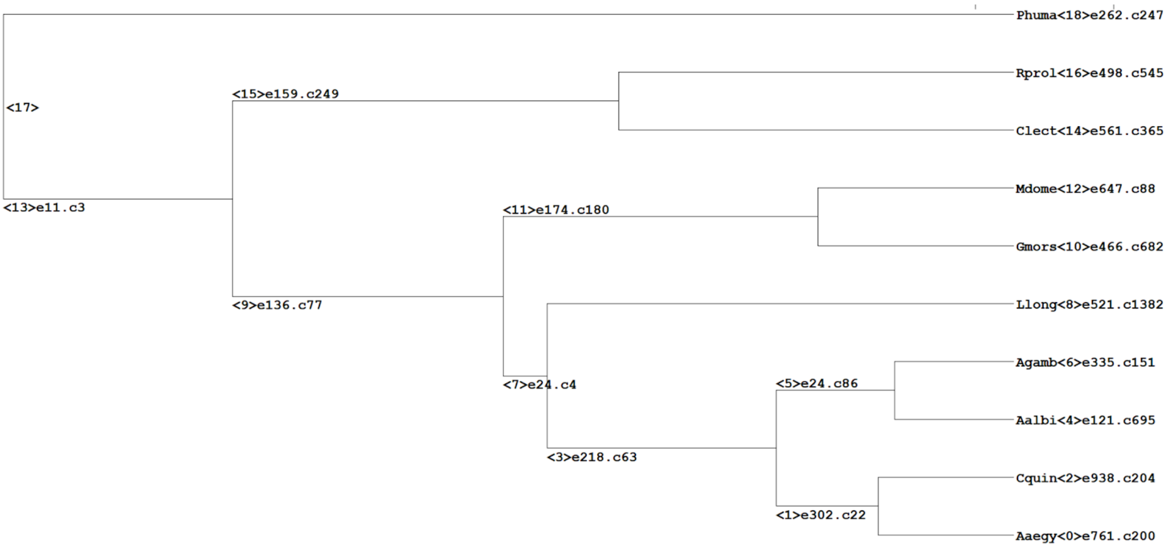
* Alternatively, you can download the script provided on the Moodle site called 'place_CAFE_counts_on_tree.pl' - this has not been extensive tested with lots of different CAFE runs, but it should at least work on the downloaded 'report_runINS.cafe' in your reports directory

```
$ perl place_CAFE_counts_on_tree.pl reports/report_runINS.cafe
```

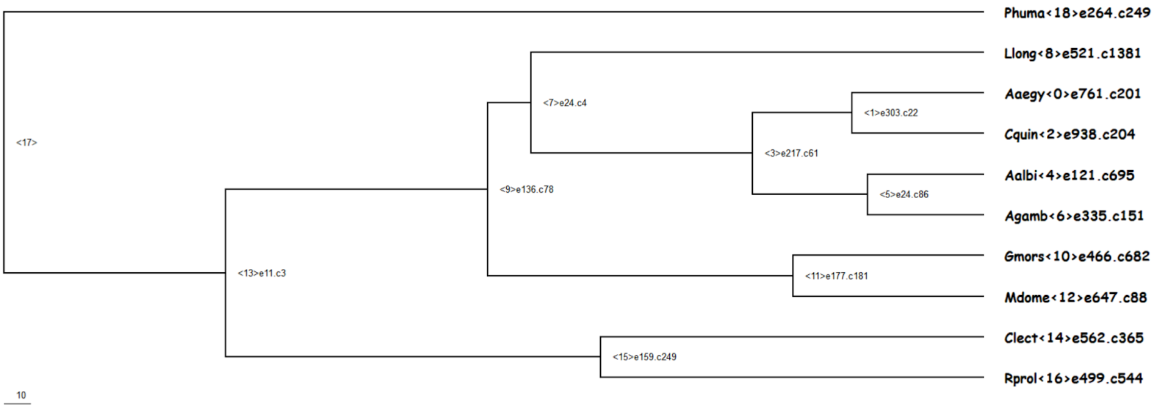
* NB: different tree viewing tools treat the formatting of internal branch, node, and bootstrap labels differently, so make sure you check what you want to be displayed where, and that it is rendered

correctly by the tool you use.
* E.g. NJplot: http://doua.prabi.fr/software/njplot

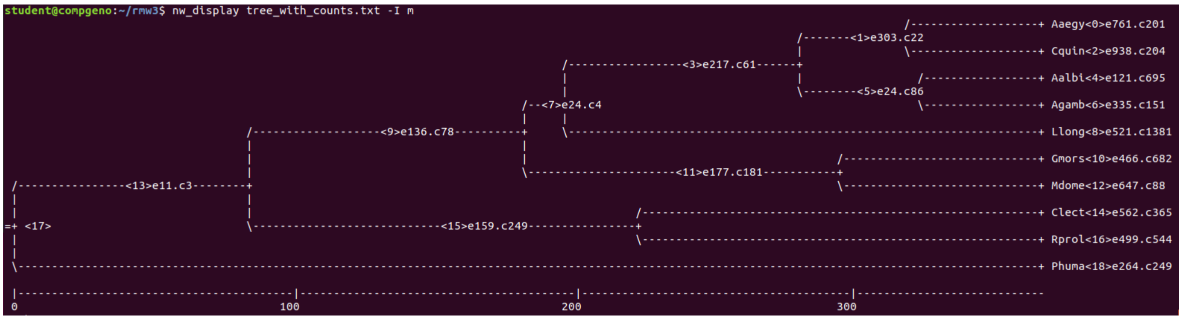
Tree viewed using NJplot with expansions (e) and contractions (c) labelled on all nodes



Tree viewed using TreeView with expansions (e) and contractions (c) labelled on all nodes



Tree viewed using newick utilities nw_display with expansions (e) and contractions (c) labelled on all nodes



```
student@compgeno:~/rmw3$ nw_display tree_with_counts.txt -I m
                                                                                                                                       /-------------------+ Aaegy<0>e761.c201
                                                                                                                       /-------<1>e303.c22
                                                                                                                       |               \-------------------+ Cquin<2>e938.c204
                                                                                    /-----------------<3>e217.c61------+
                                                                                    |                                  |                 /-----------------+ Aalbi<4>e121.c695
                                                                                    |                                  \--------<5>e24.c86
                                                                              /--<7>e24.c4                                               \-----------------+ Agamb<6>e335.c151
                                                                              |     |
                                     /------------------<9>e136.c78-----------+     \----------------------------------------------------------------------+ Llong<8>e521.c1381
                                     |                                        |
                                     |                                        |                                              /-----------------------------+ Gmors<10>e466.c682
                                     |                                        \----------------------<11>e177.c181-----------+
 /----------------<13>e11.c3---------+                                                                                       \-----------------------------+ Mdome<12>e647.c88
 |                                   |
 |                                   |                                                         /-----------------------------------------------------------+ Clect<14>e562.c365
=+ <17>                              \----------------------------<15>e159.c249----------------+
 |                                                                                             \-----------------------------------------------------------+ Rprol<16>e499.c544
 |
 \---------------------------------------------------------------------------------------------------------------------------------------------------------+ Phuma<18>e264.c249

 |-----------------------------------------|-----------------------------------------|-----------------------------------------|----------------------------
 0                                       100                                       200                                       300
```

6. **Which branch has experienced the most contractions?** *

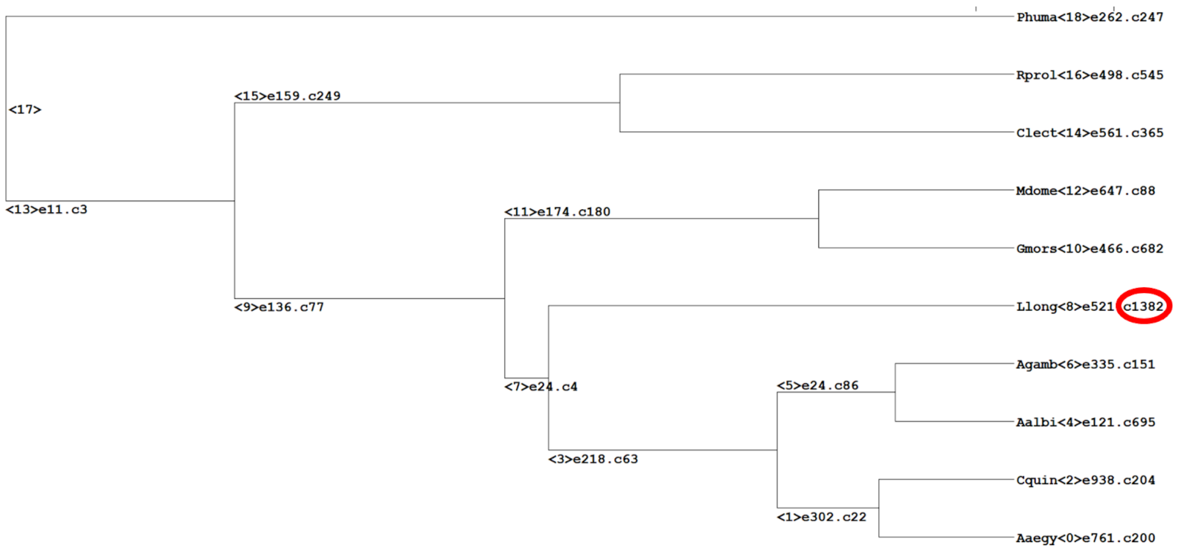
Mark only one oval.

○ <1> *Skip to question 7.*

○ <8> *Skip to "Yep, that sandfly - again!."*

○ <15> *Skip to question 7.*

Are you sure?



7. **Which branch has experienced the most contractions?** *

Mark only one oval.

○ <1> *Skip to question 7.*

○ <8> *Skip to "Yep, that sandfly - again!."*

○ <15> *Skip to question 7.*

Yep, that sandfly - again!

* However, considering your findings from the first exercise - do you really trust this result?
* I.e. given that the Llonj1 assembly had the highest number of missing BUSCOs, it is probably not completely and/or contiguously assembled, so gene annotation is likely to be incomplete, which could give the impression of many lost genes, just as we observe here with CAFE
* So, always remember to think about the technical aspects of your data before you jump to conclusions about putative biological explanations for what you observe!

Hopefully by completing this exercise you:

[1] have some knowledge of the orthology data available from OrthoDB and how to extract what you need
[2] have an understanding of how to perform gene family evolution analyses with the Computational Analysis of gene Family Evolution (CAFE) package
[3] have learnt how it is important to consider both technical and biological reasons that could explain your results from large-scale comparative genomics analyses

Click SUBMIT below to finish this exercise.

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.