#1 BUSCO Genomes

For this first exercise we will run an assessment of BUSCO completeness on an insect genome

For the purposes of today's comparative genomics analysis we will in fact need the results from running BUSCO assessments for 10 insect genomes, as this takes some time the pre-computed results will be provided.

Here we will first attempt to run an assessment of BUSCO completeness on a minimised example insect genome

By the end of this first exercise you should:

[1] have successfully performed an assessment of BUSCO completeness on the example genome
[2] have become familiar with the output produced by running such an assessment

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

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] Getting the genome data

[1] First we need to create a directory in which we will perform this exercise
* From your HOME directory in the terminal

```
$ mkdir rmw1
$ cd rmw1
```

[2] Then we need to download the genome data that we are going to assess
* From the Moodle site, find the folder under 'Day 2 Rob Waterhouse' called 'BUSCO_genome_data', inside you should see the gzipped file called 'example_genome_subset.fa.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/6271/mod_folder/content/0/example_genome_subset.fa.gz
```

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

```
$ gunzip example_genome_subset.fa.gz
```

* This is a FASTA file of a subset of a genome, find out how many scaffolds there are in this file

```
$ grep '>' example_genome_subset.fa
```

* You should see two lines indicating two scaffolds:

```
>subscaf1
>subscaf2
```

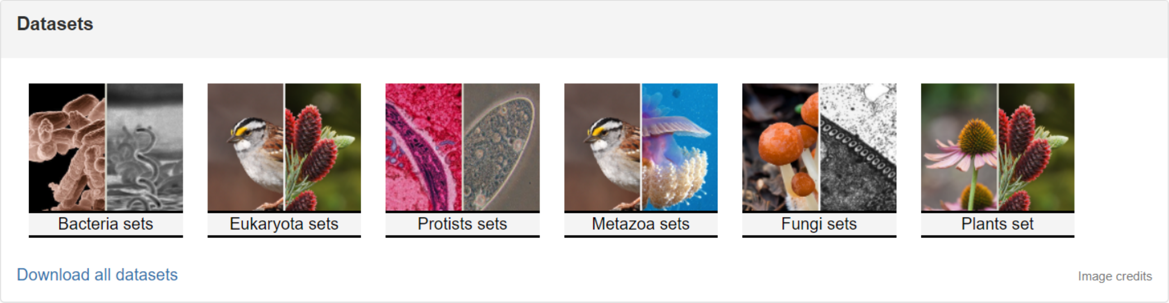
[3] BUSCO comes with various lineage-specific datasets with which to perform the assessments, so we will also need to fetch an appropriate dataset from the BUSCO website: https://busco.ezlab.org/

* Go to the BUSCO website and browse the datasets to find the Arthropoda lineage dataset (hint, arthropods are metazoans)
* Right click the image to get the full URL of the arthropoda_odb9 file (Copy Link Location) and then wget it to your VM
$ wget https://busco.ezlab.org/datasets/arthropoda_odb9.tar.gz
* unpack the tarball
$ tar -xf arthropoda_odb9.tar.gz
* list the contents of arthropoda_odb9
$ ls -l arthropoda_odb9
* You should see the following files and folders:
ancestral
ancestral_variants
dataset.cfg
hmms
info
lengths_cutoff
prfl
scores_cutoff
* You can download the BUSCO userguide from the BUSCO website (https://busco.ezlab.org/)
* Page 14 of the userguide explains the contents of BUSCO lineage datasets

On the Moodle site (find this folder)

BUSCO_genome_data

On the BUSCO site (find Arthropoda)



Datasets

2. **By exploring the contents of the arthropoda_odb9 dataset, and with the help of the userguide, how many BUSCOs are in this lineage and from how many species?** *

Mark only one oval.

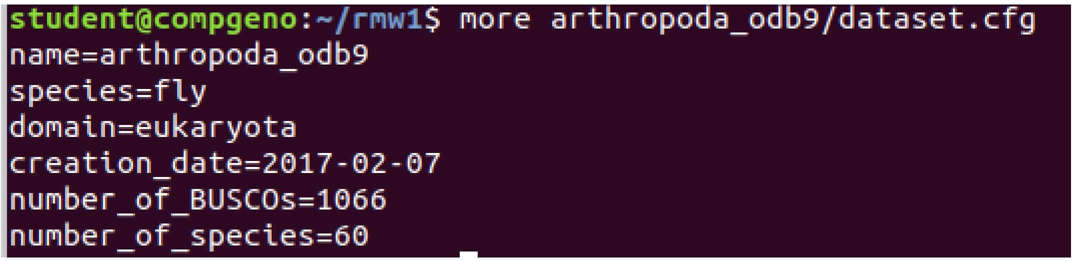
○ 1066 BUSCOs from 50 species *Skip to question 3.*

○ 1066 BUSCOs from 60 species *Skip to question 4.*

○ 6010 BUSCOs from 60 species *Skip to question 3.*

Are you sure?

Check the dataset.cfg file (cfg stands for configuration)



```
student@compgeno:~/rmw1$ more arthropoda_odb9/dataset.cfg
name=arthropoda_odb9
species=fly
domain=eukaryota
creation_date=2017-02-07
number_of_BUSCOs=1066
number_of_species=60
```

3. **How many BUSCOs are in this lineage and from how many species?** *

Mark only one oval.

○ 1066 BUSCOs from 50 species *Skip to question 3.*

○ 1066 BUSCOs from 60 species *Skip to question 4.*

○ 6010 BUSCOs from 60 species *Skip to question 3.*

[B] Running a BUSCO assessment

Now we have a 'genome' and a BUSCO lineage dataset, we can run an assessment

[1] First check that BUSCO is installed and findable, e.g. by asking for the BUSCO version
* BUSCO has been installed for you in the home directory under software/
$ python3 ~/software/busco/scripts/run_BUSCO.py -v
* the '~' indicates that the path starts from your home directory
* the '-v' is short for '--version'
* BUSCO should print out its version for you:
BUSCO 3.0.2

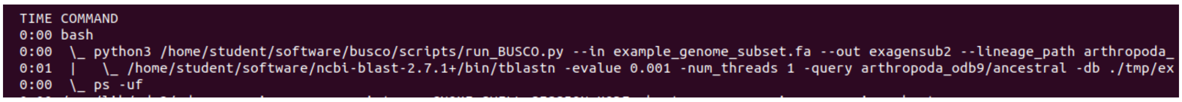
[2] Launching an assessment requires 4 mandatory arguments (see userguide page 8 for more details)
* -i (or --in) the name of the file with the data you want to assess (in this case our example_genome_subset.fa)
* -o (or --out) the label you wish to give this analysis, e.g. exagensub
* -l (or --lineage_path) the path to the lineage dataset you want to use (in this case our downloaded arthropoda_odb9)
* -m (or --mode) the assessment mode geno or genome, tran or transcriptome, prot or proteins (in this case genome)
* So your full command to launch BUSCO might look something like the following:
$ python3 ~/software/busco/scripts/run_BUSCO.py --in example_genome_subset.fa --out exagensub --lineage_path arthropoda_odb9 --mode genome >& exagensub.log.txt &

* NB: if you set up your VM with more than one CPU then you could specify in the command to use more than one CPU, to do this you would add the argument, e.g. for 2 CPUs: -c 2 (or --cpu 2)

* NB: the command above ends with two features that some of you might not be familiar with: '>& exagensub.log.txt' simply redirects the terminal output to a log file instead, and the final '&' means the job will run in the background (otherwise you would have to wait for it to finish before being able to enter further commands)

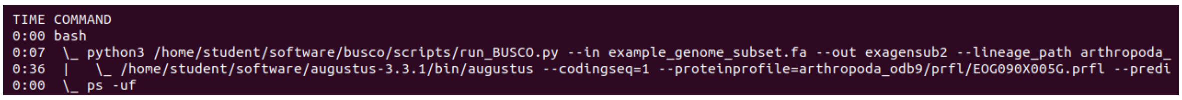
* NB: you can check that your command is actually running using the unix 'ps' command, e.g.
$ ps -uf

E.g. the first step is to run a BLAST tblastn search, so here you can see that the python3 call to run_BUSCO.py is running, and below that is the child process tblastn search



```
TIME COMMAND
0:00 bash
0:00  \_ python3 /home/student/software/busco/scripts/run_BUSCO.py --in example_genome_subset.fa --out exagensub2 --lineage_path arthropoda_
0:01  |   \_ /home/student/software/ncbi-blast-2.7.1+/bin/tblastn -evalue 0.001 -num_threads 1 -query arthropoda_odb9/ancestral -db ./tmp/ex
0:00  \_ ps -uf
```

E.g. once the tblastn search is complete, then BUSCO will launch Augustus commands which attempt to predict gene models in each of the identified regions in your genome. Here you can see that the python3 call to run_BUSCO.py is still running, and below that is the child process augustus prediction



```
TIME COMMAND
0:00 bash
0:07  \_ python3 /home/student/software/busco/scripts/run_BUSCO.py --in example_genome_subset.fa --out exagensub2 --lineage_path arthropoda_
0:36  |   \_ /home/student/software/augustus-3.3.1/bin/augustus --codingseq=1 --proteinprofile=arthropoda_odb9/prfl/EOG090X005G.prfl --predi
0:00  \_ ps -uf
```

4. **This could take some time (~10 minutes) so if you can see that the assessment is running then you can continue to the next step, or if not then ask for help.** *

Mark only one oval.

○ Continue *Skip to question 5.*

○ Help! *Skip to "We're here to help!."*

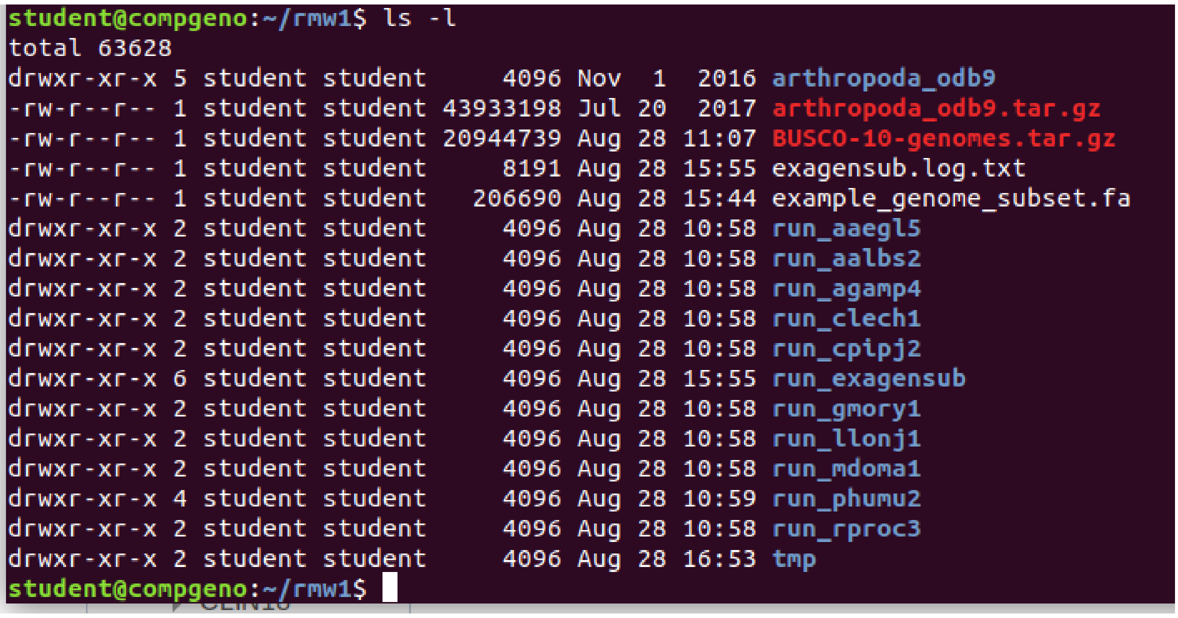
[C] Examining results from 10 genomes

We have pre-computed BUSCO genome assessments for you for 10 insect species, so while your example assessment finishes running you can fetch these pre-computed results.

[1] We need to download the genome assessment results that we are going to examine
* From the Moodle site, find the folder under 'Day 2 Rob Waterhouse' called 'BUSCO_genome_data', inside you should see the gzipped tarball file called 'BUSCO-10-genomes.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/6271/mod_folder/content/0/BUSCO-10-genomes.tar.gz
* untarzip the file using the tar command
$ tar -xzf BUSCO-10-genomes.tar.gz

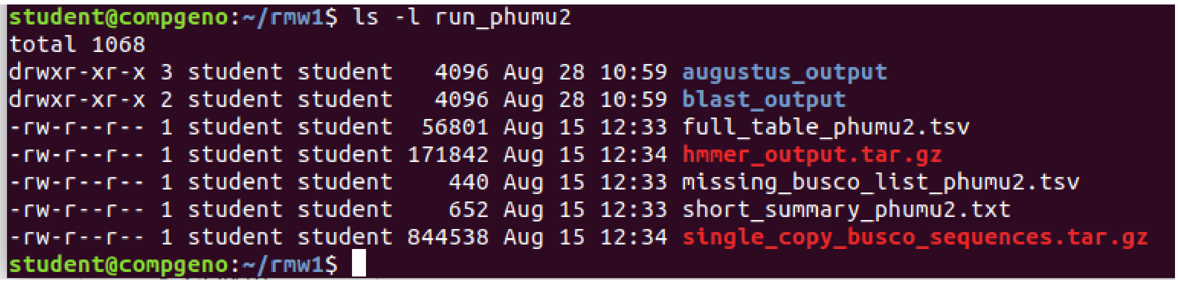
[2] Results files
* Your results should already be being written into a directory that starts with 'run_' and ends with the label you gave it, e.g. it could be 'run_exagensub' if you followed the previous step exactly
* The results files that you have just downloaded also start with 'run_'
* They each end with a species code and version number, e.g. aaegl5 refers to the mosquito Aedes aegypti assembly version L5
* Take a look at the contents of 'run_phumu2' (Pediculus humanus assembly U2)
$ ls -l run_phumu2
* Explore what's in these files and subdirectories, take a look at page 10 of the userguide for more details about what each results file or folder contains
* View the contents of the 'short_summary' file for a summary of the results obtained for phumu2
$ more run_phumu2/short_summary_phumu2.txt
* If your own assessment is now finished then you can also take a look at those results
$ ls -l run_exagensub
$ more run_exagensub/short_summary_exagensub.txt

The contents of your directory should look something like this. Where 'run_exagensub' should be filling up with your own results (or already filled up if your assessment has finished by now), and where the other 'run_xxxxxx' directories contain pre-computed results from 10 insect genomes



```
student@compgeno:~/rmw1$ ls -l
total 63628
drwxr-xr-x 5 student student     4096 Nov  1  2016 arthropoda_odb9
-rw-r--r-- 1 student student 43933198 Jul 20  2017 arthropoda_odb9.tar.gz
-rw-r--r-- 1 student student 20944739 Aug 28 11:07 BUSCO-10-genomes.tar.gz
-rw-r--r-- 1 student student     8191 Aug 28 15:55 exagensub.log.txt
-rw-r--r-- 1 student student   206690 Aug 28 15:44 example_genome_subset.fa
drwxr-xr-x 2 student student     4096 Aug 28 10:58 run_aaegl5
drwxr-xr-x 2 student student     4096 Aug 28 10:58 run_aalbs2
drwxr-xr-x 2 student student     4096 Aug 28 10:58 run_agamp4
drwxr-xr-x 2 student student     4096 Aug 28 10:58 run_clech1
drwxr-xr-x 2 student student     4096 Aug 28 10:58 run_cpipj2
drwxr-xr-x 6 student student     4096 Aug 28 15:55 run_exagensub
drwxr-xr-x 2 student student     4096 Aug 28 10:58 run_gmory1
drwxr-xr-x 2 student student     4096 Aug 28 10:58 run_llonj1
drwxr-xr-x 2 student student     4096 Aug 28 10:58 run_mdoma1
drwxr-xr-x 4 student student     4096 Aug 28 10:59 run_phumu2
drwxr-xr-x 2 student student     4096 Aug 28 10:58 run_rproc3
drwxr-xr-x 2 student student     4096 Aug 28 16:53 tmp
student@compgeno:~/rmw1$ 
```

The contents of run_phumu2 (see page 10 of the userguide for details)



```
student@compgeno:~/rmw1$ ls -l run_phumu2
total 1068
drwxr-xr-x 3 student student   4096 Aug 28 10:59 augustus_output
drwxr-xr-x 2 student student   4096 Aug 28 10:59 blast_output
-rw-r--r-- 1 student student  56801 Aug 15 12:33 full_table_phumu2.tsv
-rw-r--r-- 1 student student 171842 Aug 15 12:34 hmmer_output.tar.gz
-rw-r--r-- 1 student student    440 Aug 15 12:33 missing_busco_list_phumu2.tsv
-rw-r--r-- 1 student student    652 Aug 15 12:33 short_summary_phumu2.txt
-rw-r--r-- 1 student student 844538 Aug 15 12:34 single_copy_busco_sequences.tar.gz
student@compgeno:~/rmw1$ 
```

5. **How many Complete and single-copy BUSCOs were identified in phumu2?** *

Mark only one oval.

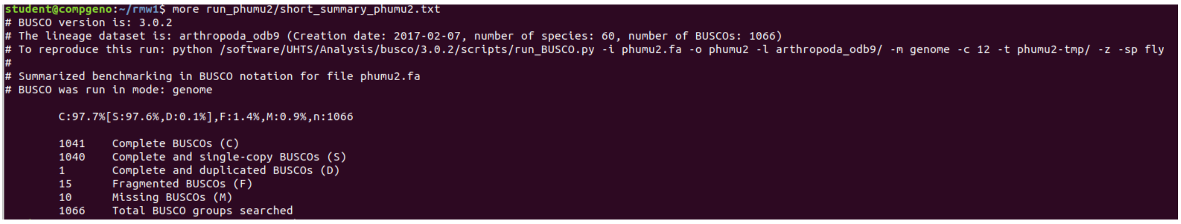
○ 1041 *Skip to question 6.*

○ 1066 *Skip to question 6.*

○ 1040 *Skip to question 7.*

Are you sure?

* View the contents of the 'short_summary' file for a summary of the results obtained for phumu2
$ more run_phumu2/short_summary_phumu2.txt



```
student@compgeno:~/rmw1$ more run_phumu2/short_summary_phumu2.txt
# BUSCO version is: 3.0.2
# The lineage dataset is: arthropoda_odb9 (Creation date: 2017-02-07, number of species: 60, number of BUSCOs: 1066)
# To reproduce this run: python /software/UHTS/Analysis/busco/3.0.2/scripts/run_BUSCO.py -i phumu2.fa -o phumu2 -l arthropoda_odb9/ -m genome -c 12 -t phumu2-tmp/ -z -sp fly
#
# Summarized benchmarking in BUSCO notation for file phumu2.fa
# BUSCO was run in mode: genome

        C:97.7%[S:97.6%,D:0.1%],F:1.4%,M:0.9%,n:1066

        1041    Complete BUSCOs (C)
        1040    Complete and single-copy BUSCOs (S)
        1       Complete and duplicated BUSCOs (D)
        15      Fragmented BUSCOs (F)
        10      Missing BUSCOs (M)
        1066    Total BUSCO groups searched
```

6. **How many Complete and single-copy BUSCOs were identified in phumu2?** *

Mark only one oval.

○ 1041 *Skip to question 6.*

○ 1066 *Skip to question 6.*

○ 1040 *Skip to question 7.*

[D] And what about in the assessment you ran?

* It should be finished by now, you can check with the 'ps -uf' command
$ ps -uf
* Or you can check the end of your log file to see if it has finished, e.g.
$ tail exagensub.log.txt
* If the log file ends with something like the following then it should be done:
INFO Results written in /home/student/rmw1/run_exagensub/

7. **How many Complete and single-copy BUSCOs were identified in exagensub?** *

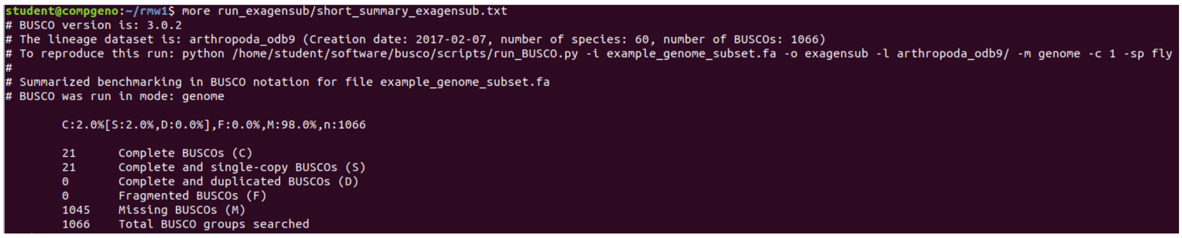
Mark only one oval.

○ 21 *Skip to question 9.*

○ 1045 *Skip to question 8.*

Are you sure?

* View the contents of the 'short_summary' file for a summary of the results obtained for exagensub
$ more run_exagensub/short_summary_exagensub.txt



```
student@compgeno:~/rmw1$ more run_exagensub/short_summary_exagensub.txt
# BUSCO version is: 3.0.2
# The lineage dataset is: arthropoda_odb9 (Creation date: 2017-02-07, number of species: 60, number of BUSCOs: 1066)
# To reproduce this run: python /home/student/software/busco/scripts/run_BUSCO.py -i example_genome_subset.fa -o exagensub -l arthropoda_odb9/ -m genome -c 1 -sp fly
#
# Summarized benchmarking in BUSCO notation for file example_genome_subset.fa
# BUSCO was run in mode: genome

        C:2.0%[S:2.0%,D:0.0%],F:0.0%,M:98.0%,n:1066

        21      Complete BUSCOs (C)
        21      Complete and single-copy BUSCOs (S)
        0       Complete and duplicated BUSCOs (D)
        0       Fragmented BUSCOs (F)
        1045    Missing BUSCOs (M)
        1066    Total BUSCO groups searched
```

8. **How many Complete and single-copy BUSCOs were identified in exagensub?** *

Mark only one oval.

○ 21 *Skip to question 9.*

○ 1045 *Skip to question 8.*

[E] What about the other species?

* To have a quick look at the results for all the runs in your directory you can use 'grep' to search for the summary line in each of the short_summary files by using a wildcard (*) in your search
$ grep 'C:' run_*/short_summary_*.txt

9. **Which species (file), other than your exagensub, has the most missing BUSCOs?** *

Mark only one oval.

○ aaegl5 Aedes aegypti mosquito *Skip to question 10.*

○ aalbs2 Anopheles albimanus mosquito *Skip to question 10.*

○ agamp4 Anopheles gambiae mosquito *Skip to question 10.*

○ clech1 Cimex lectularius bed bug *Skip to question 10.*

○ cpipj2 Culex quinquefasciatus mosquito *Skip to question 10.*

○ gmory1 Glossina morsitans tsetse fly *Skip to question 10.*

○ llonj1 Lutzomyia longipalpis sandfly *Skip to "Yep, that sandfly genome is not great."*

○ mdoma1 Musca domestica house fly *Skip to question 10.*

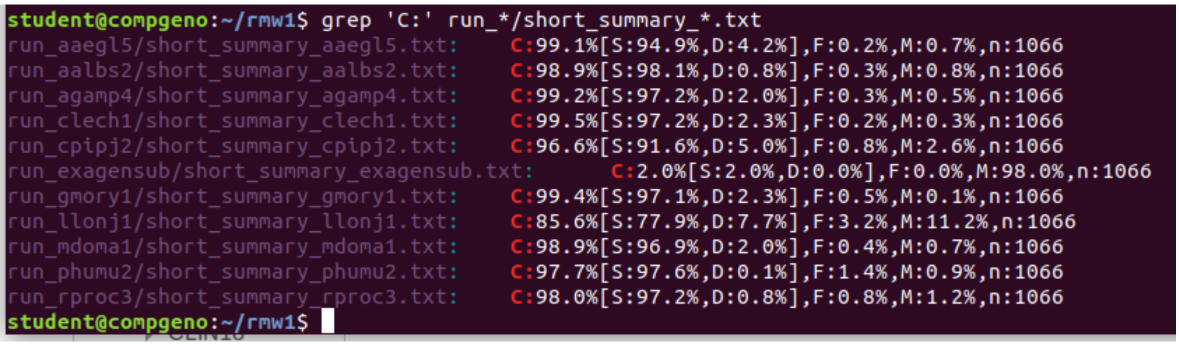
○ phumu2 Pediculus humanus body louse *Skip to question 10.*

○ rproc3 Rhodnius prolixus kissing bug *Skip to question 10.*

Are you sure?

* To have a quick look at the results for all the runs in your directory you can use 'grep' to search for the summary line in each of the short_summary files by using a wildcard (*) in your search
$ grep 'C:' run_*/short_summary_*.txt

The % missing is indicated by the 'M:'



```
student@compgeno:~/rmw1$ grep 'C:' run_*/short_summary_*.txt
run_aaegl5/short_summary_aaegl5.txt:     C:99.1%[S:94.9%,D:4.2%],F:0.2%,M:0.7%,n:1066
run_aalbs2/short_summary_aalbs2.txt:     C:98.9%[S:98.1%,D:0.8%],F:0.3%,M:0.8%,n:1066
run_agamp4/short_summary_agamp4.txt:     C:99.2%[S:97.2%,D:2.0%],F:0.3%,M:0.5%,n:1066
run_clech1/short_summary_clech1.txt:     C:99.5%[S:97.2%,D:2.3%],F:0.2%,M:0.3%,n:1066
run_cpipj2/short_summary_cpipj2.txt:     C:96.6%[S:91.6%,D:5.0%],F:0.8%,M:2.6%,n:1066
run_exagensub/short_summary_exagensub.txt:       C:2.0%[S:2.0%,D:0.0%],F:0.0%,M:98.0%,n:1066
run_gmory1/short_summary_gmory1.txt:     C:99.4%[S:97.1%,D:2.3%],F:0.5%,M:0.1%,n:1066
run_llonj1/short_summary_llonj1.txt:     C:85.6%[S:77.9%,D:7.7%],F:3.2%,M:11.2%,n:1066
run_mdoma1/short_summary_mdoma1.txt:     C:98.9%[S:96.9%,D:2.0%],F:0.4%,M:0.7%,n:1066
run_phumu2/short_summary_phumu2.txt:     C:97.7%[S:97.6%,D:0.1%],F:1.4%,M:0.9%,n:1066
run_rproc3/short_summary_rproc3.txt:     C:98.0%[S:97.2%,D:0.8%],F:0.8%,M:1.2%,n:1066
student@compgeno:~/rmw1$
```

10. **Which species (file), other than your exagensub, has the most missing BUSCOs?** *

Mark only one oval.

○ aaegl5 Aedes aegypti mosquito *Skip to question 10.*

○ aalbs2 Anopheles albimanus mosquito *Skip to question 10.*

○ agamp4 Anopheles gambiae mosquito *Skip to question 10.*

○ clech1 Cimex lectularius bed bug *Skip to question 10.*

○ cpipj2 Culex quinquefasciatus mosquito *Skip to question 10.*

○ gmory1 Glossina morsitans tsetse fly *Skip to question 10.*

○ llonj1 Lutzomyia longipalpis sandfly *Skip to "Yep, that sandfly genome is not great."*

○ mdoma1 Musca domestica house fly *Skip to question 10.*

○ phumu2 Pediculus humanus body louse *Skip to question 10.*

○ rproc3 Rhodnius prolixus kissing bug *Skip to question 10.*

Yep, that sandfly genome is not great

Hopefully by completing this exercise you:

[1] have successfully performed an assessment of BUSCO completeness on the example genome
[2] have become familiar with the output produced by running such an assessment

Click SUBMIT below to finish this exercise

Lutzomyia longipalpis (photo Dr Ray Wilson)

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.