



*UniL SIB 3-5 September 2018*

© R.M. Waterhouse

# *Comparative Genomics Practicals*

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## ***#1 BUSCO Genomes***

*Assessment of BUSCO completeness of 10 insect genomes*

## ***#2 Phylogenetics to Phylogenomics***

*Use #1 results to build phylogenetic trees & species tree*

## ***#3 Gene Family Evolution***

*Identify dramatically changing gene families*

## ***#4 Gene Family Evolution II***

*Phylogenies of rapidly evolving multi-copy gene families*

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## ***BUSCO bonus***

*Use the BUSCO plotting tool to visualise the results from the first practical exercise*

## ***OrthoDB bonus***

*A 'quiz' that requires browsing and searching of OrthoDB online to find the answers*

## ***Orthology bonus***

*More programming-oriented (scripting) using the API to query OrthoDB and build an orthology landscape plot*

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## What's the link?

- ❑ The overarching aim - your boss asked you to identify the most dynamically evolving gene families in N insects
- ❑ So firstly you need a species tree (#1 & #2)
- ❑ Then you need the gene families to use the tree and the counts to estimate ancestral gene content changes (#3)
- ❑ Finally you need to investigate some examples in detail (#4) to convince and impress your boss

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## #1 BUSCO Genomes

### *Assessment of BUSCO completeness of 10 insect genomes*

#### #1 BUSCO Genomes

<https://goo.gl/forms/j0YzclZHCv8AnLm1>



#### #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

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## #1 BUSCO Genomes

<https://goo.gl/forms/j0Yzc1dZHCV8AnLm1>



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

☐ Yes

☐ No

NEXT

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## #1 BUSCO Genomes

<https://goo.gl/forms/j0Yzc1dZHCV8AnLm1>

### [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
```

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## #1 BUSCO Genomes

<https://goo.gl/forms/j0YzcldZHCV8AnLm1>

[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:

```
>subscf1
```

```
>subscf2
```




# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## #1 BUSCO Genomes


<https://goo.gl/forms/j0Yzc1dZHCV8AnLm1>


On the Moodle site (find this folder)

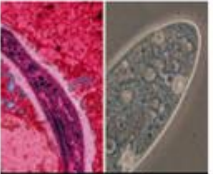
 BUSCO\_genome\_data


On the BUSCO site (find Arthropoda)


**Datasets**


  
Bacteria sets

  
Eukaryota sets

  
Protists sets

  
Metazoa sets

  
Fungi sets

  
Plants set

[Download all datasets](#)

Image credits

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## #1 BUSCO Genomes

<https://goo.gl/forms/j0Yzc1dZHCV8AnLm1>

[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](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

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## #1 BUSCO Genomes

<https://goo.gl/forms/j0Yzc1dZHCV8AnLm1>

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? \*

- ☒ 1066 BUSCOs from 50 species
- ☐ 1066 BUSCOs from 60 species
- ☐ 6010 BUSCOs from 60 species

BACK

NEXT

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## #1 BUSCO Genomes

<https://goo.gl/forms/j0Yzc1dZHCV8AnLm1>



# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

*Lutzomyia longipalpis* (photo Dr Ray Wilson)



BACK

SUBMIT

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## *#2 Phylogenetics to Phylogenomics*

*Use #1 results to build phylogenetic trees & species tree*

#2 Phylogenetics to Phylogenomics

<https://goo.gl/forms/dKr0ojmJDaYK8EBJ3>

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## ***#3 Gene Family Evolution***

***Identify dramatically changing gene families***

#3 Gene Family Evolution

<https://goo.gl/forms/11C59BWE44bf33E52>

# Comparative genomics with BUSCO & OrthoDB

© R.M. Waterhouse

## ***#4 Gene Family Evolution II***

***Phylogenies of rapidly evolving multi-copy gene families***

#4 Gene Family Evolution II

<https://goo.gl/forms/tpXX1M5Y9fB9SZA32>



## #1 BUSCO Genomes

<https://goo.gl/forms/j0YzcldZHCV8AnLm1>

## #2 Phylogenetics to Phylogenomics

<https://goo.gl/forms/dKr0ojmJDaYK8EBJ3>

## #3 Gene Family Evolution

<https://goo.gl/forms/l1C59BWE44bf33E52>

## #4 Gene Family Evolution II

<https://goo.gl/forms/tpXX1M5Y9fB9SZA32>

## OrthoDB bonus

<https://goo.gl/forms/Kfcbfnz4NinymV7y2>

## BUSCO bonus

<https://goo.gl/forms/dIaBwLehxY2qu6YE2>

## Orthology bonus

<https://goo.gl/forms/KAW3YHhUcO1Q6EKB2>