

# Goals

- Get a chance to process dataset to explore genome evolutionary questions
- Practice creating graphical plots of data in R using existing templates of scripts
- Get a feel for some of the datatypes and datasets you might work with

# Get the Tutorial

```
$ git clone https://github.com/hyphaltip/Tutorial\_GenomeEvolution  
$ cd data  
$ bash download.sh # will download datasets
```

# Install some R packages

- May need to install extra libraries, in R:
  - `install.packages("ggplot2", "gridExtra", "dplyr", "RColorBrewer", "pheatmap")`
  - ```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("AnnotationDbi", version = "3.8")
BiocManager::install("tximport", version = "3.8")
```
  - see <https://bioconductor.org/packages/release/bioc/html/AnnotationDbi.html>  
<http://bioconductor.org/packages/release/bioc/html/tximport.html>
- open up the code in an editor or RStudio and take a look

# Explore datasets

- data folder has genome, protein, and GFF annotation in it  
or it will when you run the `download.sh` script
- How many annotate genes (gene features ) are in the GFF file for each species? How many transcripts (mRNA features)
  - Can solve this many ways, unix commands `grep`, `awk`, `wc` will suffice
  - Can also do this in R

# Explore datasets

- analysis/ortho\_set1/Results
- Has OrthoFinder pre-computed results
- Explore these result files

# Tasks/Questions: OrthoFinder data

- Make a heat map of gene family sizes - pick a cutoff like gene family total size  $> 25$  but experiment with this
- `scripts/plot_heatmap_family.R` shows you how to work with this. Run it with  
`Rscript scripts/plot_heatmap_family.R`  
on UNIX
- I recommend Rstudio for interactive session (comment out the `pdf()` line if you do this so you can see the plots in your session
- You may need to install `pheatmap` in your R installation  
Do `install.packages("pheatmap")` in your R console.

# Other challenges/questions

- How many singletons - genes that have no homologs so aren't in a orthogroup - are there per species
  - Make a table with these numbers
- How many single copy gene families are there across this data (eg 1:1:1 ... orthologs)

# Explore genome statistics

- Run the Rscript/plot\_chroms\_1.R
- Examine some genome wide statistics - like intron density and exon size and chromosome-wide plots of genes or introns/per gene etc.
- Explore and experiment with plotting different things like number of genes histograms across chromosomes

# Some reference links

**TxDB and Genomic Features**

**<https://kasperdanielhansen.github.io/genbioconductor/html/GenomicFeatures.html>**

**pheatmap**

**<https://davetang.org/muse/2018/05/15/making-a-heatmap-in-r-with-the-pheatmap-package/>**