Workshop 3- Comparative Genomics

Yocelyn T. Gutierrez Guerrero

(ygutierrezgro@berkeley.edu (mailto:ygutierrezgro@berkeley.edu))

Programs that we are going require

-Seqkit (conda install -c bioconda seqkit)

-Mafft (conda install -c bioconda mafft)

-backtranseq (conda install -c bioconda emboss)

-raxmlHPC-PTHREADS

-PAML

-codeml

Phylogenomics

First section:

1- How many single copy genes are in the file “Table_Rodents_FamilyGene.txt”?

Choose only one of the folders: Mus_scg or Cricetidae_scg

2- How many species and single copy genes are in the folder?

Genome assembly and/or annotation quality can impact the completeness of the genes annotated, an important step is evaluate the length of the sequences

3- Choose some multifasta files (for example: “5137_scg.faa”), and calculate if there are differences in the length of the sequences between species

```
##Some options

awk '/^>/ {if (seqlen) print seqlen;print;seqlen=0;next} {seqlen+=length($0)}END{prin
t seqlen}' your_file.faa

## or install seqkit and run:

seqkit fx2tab -lg your_file.faa
```

4- Concatenate all the multifasta files in one file. Then, concatenate each sequence for each species

Tip:To generate only one sequence for each species, you can use the command seqkit concat

5- What is the length of the sequence for the complex *Mus* sp.?

6- Generate an alignment using the multifasta file created in the step 5, but only including the *Mus* species

Mus complex: *M. musculus, M. caroli, M. pahari, M. castaneus, M. spicilegus, M. spretus*)

Cricetidae clade: (*M. ochrogaster ,P. m. bairdii, P. leucopus, O. cuniculus, C. griseus, M. fortis, M. oregoni, M. auratus, M. glareolus, A. amphibius*)

```
mafft --anysymbol --parttree --retree 1 --thread 4 multifasta_concatenate.faa > align
_multifasta.faa
```

Estimate the best-fit model of protein evolution can be performed with tools as: Prottest3 (https://doi.org/10.1093/bioinformatics/btr088)

7- Evaluate the missing data (visualizing the alignment)

8- Generate the phylogenetic tree. There are different tools to perform the phylogenetic tree: ASTRAL III (https://doi.org/10.1186/s12859-018-2129-y) highly recommended. Here, we will use RAxML (you should increase the bootstraping, this is only an example of how to run RAxML)

```
raxmlHPC-PTHREADS -f a -p 12345 -m PROTCATJTT -s align_multifasta.faa -n rodent.tree
-# 100 -T 10 -x 10
```

9- Visualizing the phylogenetic tree using FigTree or R libraries (for example: *ggtree* and *ape*)

dN/dS selection test

Second section

For the second part of the workshop, we'll use the scripts that are located into the folder dn_ds_Scripts

To test if specific branches in the phylogenetic tree have experienced signals of positive selection, we will compare two hypothesis using a neutral (**codeml.ctl**) and an alternative model (**codeml_alt.ctl**) of evolution

Generate a random list of single copy genes sequences (maximum 100) and save the list with the name: *list_scg.txt* For example:

10276

10295

10298

10429

10436

10458

10460

10501

10567

10569

10586

10587

Run the script **align_trees.sh** in the same directory where the scg multifasta files and the list of genes are storage

```
nohup ./align_tree.sh &
```

10- What type of files were generated?

Time to run the selection test

Assign internal branch labels to the phylogenetic tree Indicate which species might have an alternative model of evolution

```
nohup ./script_codeml_alt.sh &
nohup ./script_codeml.sh &
```

11- How many genes are statistically significant (file: seq_qval.txt)?

```
./script_calculateLRT.sh
```

12- Explore the function of the gene under selection **Tip**= blast the sequence to obtain the gene

13- Which site or sites are under positive selection? (check files in the folder Paml_alt)

```
awk '/Bayes Empirical Bayes/,/The grid/' gene_alt.ctl
```

Enrichment analysis with using the **topGO** R library

```
library("topGO")
library("ggplot2")

##annotation_universe:GO background

geneID2GO <- readMappings(file = "annotation_universe")
geneUniverse <- names(geneID2GO)

##gene.list: gene of interest
genesOfInterest <- read.table("gene.list", header=FALSE)


genesOfInterest <- as.character(genesOfInterest$V1)
geneList <- factor(as.integer(geneUniverse %in% genesOfInterest))
names(geneList) <- geneUniverse
myGOdataMF <- new("topGOdata", description="My project", ontology="MF", allGenes=gene
List,  annot = annFUN.gene2GO, gene2GO = geneID2GO)
sg <- sigGenes(myGOdataMF)
str(sg)
resultFisher <- runTest(myGOdataMF, algorithm="weight01", statistic="fisher")
allRes <- GenTable(myGOdataMF, classicFisher = resultFisher, orderBy = "resultFisher"
, ranksOf = "classicFisher", topNodes = 20)
allRes$FDR <- p.adjust(allRes$classicFisher, method = "fdr")
write.table(allRes, "Splicing_GOs_MF.txt", sep="\t", quote = FALSE, row.names=FALSE)

myGOdataCC <- new("topGOdata", description="My project", ontology="CC", allGenes=gene
List,  annot = annFUN.gene2GO, gene2GO = geneID2GO)
sg <- sigGenes(myGOdataCC)
str(sg)
resultFisher <- runTest(myGOdataCC, algorithm="weight01", statistic="fisher")
allRes <- GenTable(myGOdataCC, classicFisher = resultFisher, orderBy = "resultFisher"
, ranksOf = "classicFisher", topNodes = 20)
allRes$FDR <- p.adjust(allRes$classicFisher, method = "fdr")
write.table(allRes, "Splicing_GOs_CC.txt", sep="\t", quote = FALSE, row.names=FALSE)

myGOdataBP <- new("topGOdata", description="My project", ontology="BP", allGenes=gene
List,  annot = annFUN.gene2GO, gene2GO = geneID2GO)
sg <- sigGenes(myGOdataBP)
str(sg)
resultFisher <- runTest(myGOdataBP, algorithm="weight01", statistic="fisher")
allRes <- GenTable(myGOdataBP, classicFisher = resultFisher, orderBy = "resultFisher"
, ranksOf = "classicFisher", topNodes = 20)
allRes$FDR <- p.adjust(allRes$classicFisher, method = "fdr")
write.table(allRes, "Splicing_GOs_BP.txt", sep="\t", quote = FALSE, row.names=FALSE)
```

14- Filter *p-values* using the column of false discovery rate

15.Which are the functions with a significant enrichment??