# **Hands on BgeeCall package**

**Duration of section 1h:30m**

RNA-Seq libraries available [here](https://bgee.org/ftp/tmp/bgeecall_exercice.tar.gz)

Vignette of BgeeCall available [here](https://bioconductor.org/packages/release/bioc/vignettes/BgeeCall/inst/doc/bgeecall-manual.html)

### Get transcriptome and annotation files

1. Extract transcriptome and annotation files for *Drosophila melanogaster* from the Ensembl database using the **download.file()** function (Note: use the following links to download annotation and cdna file:   
   <http://ftp.ensemblgenomes.org/pub/release-51/metazoa/gtf/drosophila_melanogaster/Drosophila_melanogaster.BDGP6.32.51.chr.gtf.gz> & <http://ftp.ensemblgenomes.org/pub/release-51/metazoa/fasta/drosophila_melanogaster/cdna/Drosophila_melanogaster.BDGP6.32.cdna.all.fa.gz>).

Code:

# install downloader from CRAN

install.packages("downloader")

library(downloader)

# Specify URL where file is stored

url <- "http://ftp.ensemblgenomes.org/pub/release-51/metazoa/gtf/drosophila\_melanogaster/Drosophila\_melanogaster.BDGP6.32.51.chr.gtf.gz"

url1 <- "http://ftp.ensemblgenomes.org/pub/release-51/metazoa/fasta/drosophila\_melanogaster/cdna/Drosophila\_melanogaster.BDGP6.32.cdna.all.fa.gz"

# Specify destination where file should be saved

destfile <- "C:/Users/Marion/Desktop/BgeeCall-assignment/Drosophila\_melanogaster.BDGP6.32.51.chr.gtf.gz"

destfile1 <- "C:/Users/Marion/Desktop/BgeeCall-assignment/Drosophila\_melanogaster.BDGP6.32.cdna.all.fa.gz"

# Download files

download.file(url, destfile)

download.file(url1, destfile1)

### Retrieve intergenic information

1. List all intergenic releases available in BgeeCall. How many exist?

5 intergenic releases exist.

# List of all intergenic releases

{

list\_intergenic\_release()

}

Truncated output:

Release Release Date FTPURL

1.0 2021-06-11 https://bgee.org/ftp/intergenic/1.0/

0.2 2019-02-07 https://bgee.org/ftp/intergenic/0.2/

0.1 2018-12-21 https://bgee.org/ftp/intergenic/0.1/

community 2019-07-22

custom 2019-07-22

1. Verify which species are available for the current Bgee intergenic release. How many exist?

There are 52 species in Bgee intergenic release 1.0

Code:

# Verify which species are available for the current Bgee intergenic release ()

{

bgee <- new("BgeeMetadata")

list\_bgee\_ref\_intergenic\_species(myBgeeMetadata = bgee)

list\_bgee\_ref\_intergenic\_species(release = '1.0')

}

1. Verify which species belong to the community. How many exist?

Two species belong to the community.

Code:

# Species that belong to the community.

{

list\_community\_ref\_intergenic\_species()

}

Truncate output:

Species ID:10036, 13686

### Use BgeeCall to download the pseudo-alignment software

1. Create an object of the KallistoMetadata class.
2. If you don’t have Kallisto software installed on your computer, specify the argument download\_kallisto = TRUE, otherwise leave download\_kallisto attribute by default FALSE.
   1. if you don’t know whether you have Kallisto installed just check that by typing the following command in the terminal: **kallisto version**

### Run analysis: Drosophila melanogaster 1 sample

1. Create a userMetadata object (note that you have to specify in the argument species\_id the Taxonomy ID, you can verify that in <https://bgee.org/> in the [See species information](https://bgee.org/search/species)).
2. What happens if the argument reads\_size is not specified by you when you create the new userMetadata object? What can be the impact in general? Reads size of RNA-Seq libraries can be found in SRA (e.g <https://www.ncbi.nlm.nih.gov/sra/?term=SRX109278>)

The default read size will be set at 51. This means that the kmer sizes will be lower than this value. If you set your own read size based on your libraries, you can then set kmers sizes higher than 50 kmers using read\_size\_kmer\_threshold().

1. Specify by using the following functions **setRNASeqLibPath()**, **setTranscriptomeFromFile()**, **setAnnotationFromFile(), setOutputDir() and setWorkingPath()** the path to your library **SRX109278**, transcriptome file, annotation file as well as the output and working directory.
2. Generate the present and absent calls for the library **SRX109278** by using **generate\_calls\_workflow()**. Which type of information is provided in the output files?
3. Plot the frequency of p-values for the correspondent library.

### Run analysis: multiple Drosophila melanogaster samples

1. Create a user input file describing all RNA-Seq libraries previously downloaded, see <https://github.com/BgeeDB/BgeeCall/blob/develop/inst/userMetadataTemplate.tsv> and the vignette of the package for more information
2. Run the generation of present and absent calls from the user file with default values for all .
3. Combine multiple libraries per species using the **merging\_libraries()** function. What is the proportion of genes present?
4. Modify the input file to combine libraries per species (**species\_id**) and developmental stage (**devStage**), see the structure of the file here: <https://github.com/BgeeDB/BgeeCall/blob/develop/inst/userMetadataTemplate_merging.tsv>   
   developmental stages of libraries :  
   - fully formed stage (ID : UBERON:0000066) for libraries SRX1720957 and SRX1720958  
   - day 4 of adulthood (ID : FBdv:00007079) for libraries SRX493950, SRX493999, SRX109278 and SRX109279
5. Generate the present and absent calls with a more restrictive p-value = 0.01
6. Get summary stats of all libraries by using **get\_summary\_stats()** function.
7. Plot the proportion of protein coding genes of all libraries for each p-value cut-off.

### Downstream analysis

The aim of this part is to show you that you can go from BgeeCall results to forward analysis.

1. Perform a differential expression analysis between different developmental stage conditions. (Note: that in the provided dataset we have 4 samples from FBdv:00007079 and 2 samples from UBERON:0000066, so you can select just 2 samples from FBdv:00007079 (like: SRX109278 and SRX109279) to balance the analysis. Note that statistically it is recommended to use at least 3 samples of each condition for differential expression analysis).
2. Filter the results by providing just genes with FDR < 0.01. Provide a visualization graphic as MA plot.
3. Make a GO analysis. Which type of information do you retrieve?