BgeeDB: an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests

Julien Roux, Andrea Komljenovic,
Marc Robinson-Rechavi, Frédéric Bastian

UNIL | Université de Lausanne

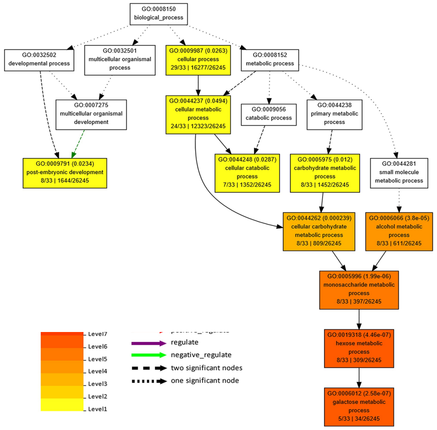
SIB
Swiss Institute of Bioinformatics

@_julien_roux

ENSMUSG00000023051, ENSMUSG00000040629, ENSMUSG00000058398, ENSMUSG00000025235, ENSMUSG00000048118,
ENSMUSG00000026567, ENSMUSG00000047014, ENSMUSG00000005506, ENSMUSG00000016758, ENSMUSG00000050799,
ENSMUSG00000026790, ENSMUSG00000062300, ENSMUSG00000001157, ENSMUSG00000048003, ENSMUSG00000040850,
ENSMUSG00000028614, ENSMUSG00000047003, ENSMUSG00000029707, ENSMUSG00000036478, ENSMUSG00000028962,
ENSMUSG00000060499, ENSMUSG00000063889, ENSMUSG00000062438, ENSMUSG00000040841, ENSMUSG00000053729,
ENSMUSG00000045179, ENSMUSG00000003549, ENSMUSG00000007907, ENSMUSG00000051306, ENSMUSG00000049470,
ENSMUSG00000026650, ENSMUSG00000024352, ENSMUSG00000024116, ENSMUSG00000063415, ENSMUSG00000072479,
ENSMUSG00000036211, ENSMUSG00000038994, ENSMUSG00000016626, ENSMUSG00000035246, ENSMUSG00000026360,
ENSMUSG00000029516, ENSMUSG00000060794, ENSMUSG00000028427, ENSMUSG00000028426, ENSMUSG00000068037,
ENSMUSG00000072663, ENSMUSG00000017767, ENSMUSG00000032921, ENSMUSG00000037017, ENSMUSG00000051965,
ENSMUSG00000038227, ENSMUSG00000005672, ENSMUSG00000003131, ENSMUSG00000028410, ENSMUSG00000028894,
ENSMUSG00000006527, ENSMUSG00000072770, ENSMUSG00000024176, ENSMUSG00000026234, ENSMUSG00000049539,
ENSMUSG00000051617, ENSMUSG00000040891, ENSMUSG00000096769, ENSMUSG00000037001, ENSMUSG00000039781,
ENSMUSG00000038210, ENSMUSG00000051977, ENSMUSG00000019834, ENSMUSG00000023070, ENSMUSG00000027794,
ENSMUSG00000026463, ENSMUSG00000040407, ENSMUSG00000027793, ENSMUSG00000028760, ENSMUSG00000002015,
ENSMUSG00000027433, ENSMUSG00000071470, ENSMUSG00000005883, ENSMUSG00000006731, ENSMUSG00000071359,
ENSMUSG00000030968, ENSMUSG00000031931, ENSMUSG00000005893, ENSMUSG00000002384, ENSMUSG00000000085,
ENSMUSG00000027660, ENSMUSG00000024392, ENSMUSG00000025482, ENSMUSG00000063972, ENSMUSG00000029848,
ENSMUSG00000090083, ENSMUSG00000075706, ENSMUSG00000096620, ENSMUSG00000014361, ENSMUSG00000038797,
ENSMUSG00000031922, ENSMUSG00000011349, ENSMUSG00000036529, ENSMUSG00000056131, ENSMUSG00000038709,
ENSMUSG00000020063, ENSMUSG00000020064, ENSMUSG00000032280, ENSMUSG00000049721, ENSMUSG00000081218,
ENSMUSG00000048516, ENSMUSG00000021038, ENSMUSG00000027938, ENSMUSG00000050957, ENSMUSG00000024426,
ENSMUSG00000068117, ENSMUSG00000047654, ENSMUSG00000069565, ENSMUSG00000027939, ENSMUSG00000035431,
ENSMUSG00000092118, ENSMUSG00000043050, ENSMUSG00000034579, ENSMUSG00000033487, ENSMUSG00000033486,
ENSMUSG00000031065, ENSMUSG00000021264, ENSMUSG00000083628, ENSMUSG00000020059, ENSMUSG00000024778,
ENSMUSG00000043289, ENSMUSG00000002768, ENSMUSG00000001558, ENSMUSG00000058328, ENSMUSG00000038932,
ENSMUSG00000037716, ENSMUSG00000056155, ENSMUSG00000021499, ENSMUSG00000074704, ENSMUSG00000025977,
ENSMUSG00000010592, ENSMUSG00000032498, ENSMUSG00000020390, ENSMUSG00000020150, ENSMUSG00000024990,
ENSMUSG00000071788, ENSMUSG00000021007, ENSMUSG00000046532, ENSMUSG00000000567, ENSMUSG00000050623,
ENSMUSG00000040828, ENSMUSG00000040829, ENSMUSG00000056215, ENSMUSG00000023010, ENSMUSG00000002799,
ENSMUSG00000001225, ENSMUSG00000041912, ENSMUSG00000023015, ENSMUSG00000027855, ENSMUSG00000024107,
ENSMUSG00000056223, ENSMUSG00000032076, ENSMUSG00000059970, ENSMUSG00000023000, ENSMUSG00000002324,
ENSMUSG00000020096, ENSMUSG00000020097, ENSMUSG00000079681, ENSMUSG00000049932, ENSMUSG00000027722,
ENSMUSG00000028938, ENSMUSG00000036551, ENSMUSG00000070999, ENSMUSG00000059625, ENSMUSG00000032187,
ENSMUSG00000033031, ENSMUSG00000022021, ENSMUSG00000048731, ENSMUSG00000079470, ENSMUSG00000044288,
ENSMUSG00000024207, ENSMUSG00000045378, ENSMUSG00000027719, ENSMUSG00000037992, ENSMUSG00000036545,
ENSMUSG00000013787, ENSMUSG00000035578, ENSMUSG00000037514, ENSMUSG00000020193, ENSMUSG00000021040,
ENSMUSG00000000365, ENSMUSG00000082639, ENSMUSG00000024430, ENSMUSG00000003873, ENSMUSG00000060985,
ENSMUSG00000025407, ENSMUSG00000014767, ENSMUSG00000071748, ENSMUSG00000037625, ENSMUSG00000094727,
ENSMUSG00000029155, ENSMUSG00000028063, ...

How to characterize gene lists?

- Functional categories enriched among these genes
  - Gene Ontology enrichment test
  - GSEA
  - Pathways analysis
  - …



Gene Ontology enrichment test

- For each functional category:

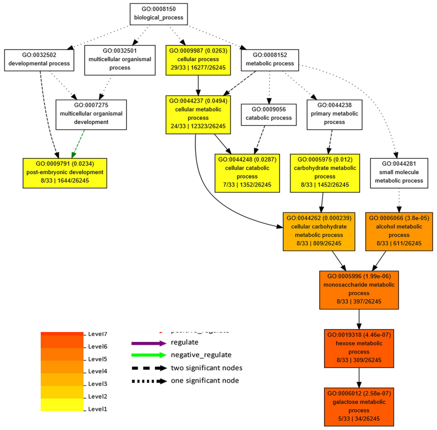
	Gene list	Other genes
Annotated	n_1	n_3
Not annotated	n_2	n_4

- Fisher / Hypergeometric test
- Bioconductor: topGO, GOstats, goseq,…

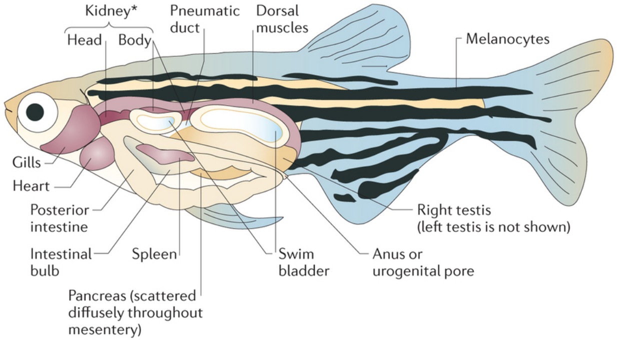
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

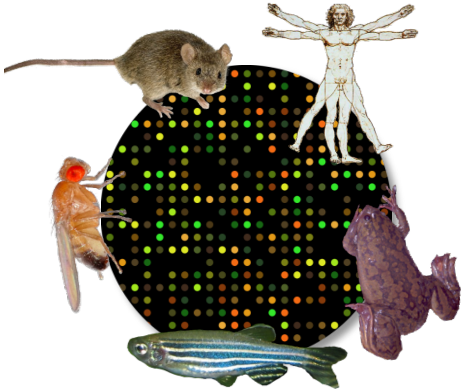
How to characterize gene lists?

- Functional categories enriched among these genes?
  - Gene Ontology enrichment test
  - GSEA
  - Pathways analyses
  - …



- Tissues enriched for expression of these genes?
  - Gene expression atlases
  - TopAnat





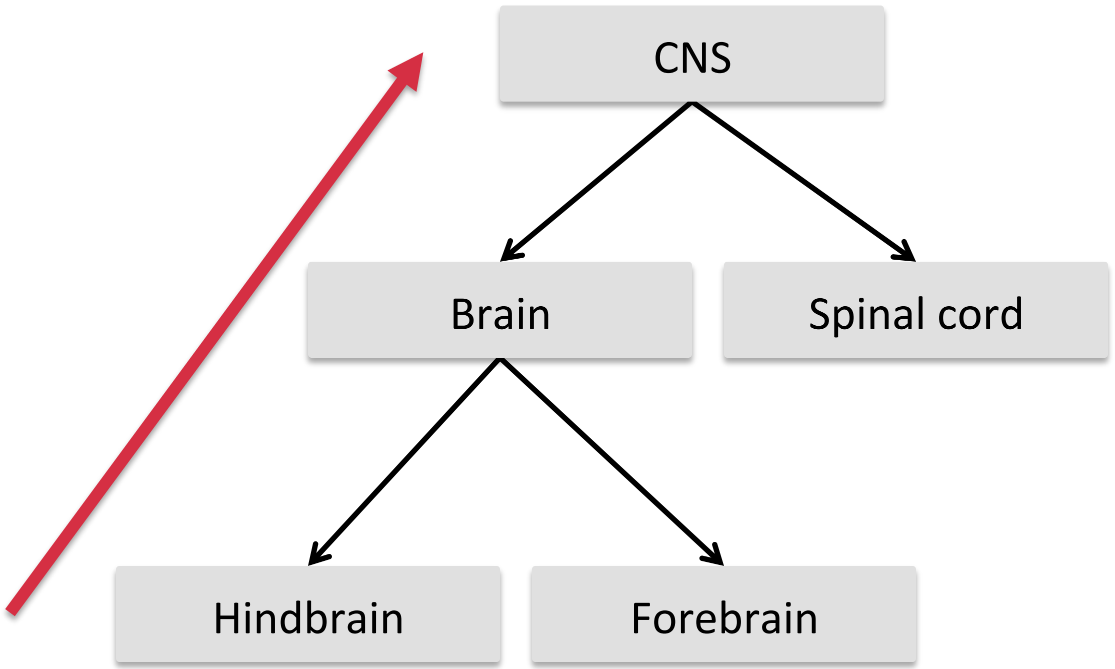
Bgee

http://bgee.org

Quick reminder:

- Only “normal” samples: no tumors, no mutants, no treatments
- RNA-seq, microarray, EST, in situ hybridization data from 17 animal species
- Manual mapping to Uberon ontology of anatomy and development

Uberon anatomical ontology



Bgee

http://bgee.org

Quick reminder:

- Only “normal” samples: no tumors, no mutants, no treatments
- RNA-seq, microarray, EST, in situ hybridization data from 17 animal species
- Manual mapping to Uberon ontology of anatomy and development
- Data reprocessed as presence/absence calls

Gene Ontology enrichment test

- For each functional category:

	Gene list	Other genes
Annotated	n_1	n_3
Not annotated	n_2	n_4

- Fisher / Hypergeometric test

TopAnat test

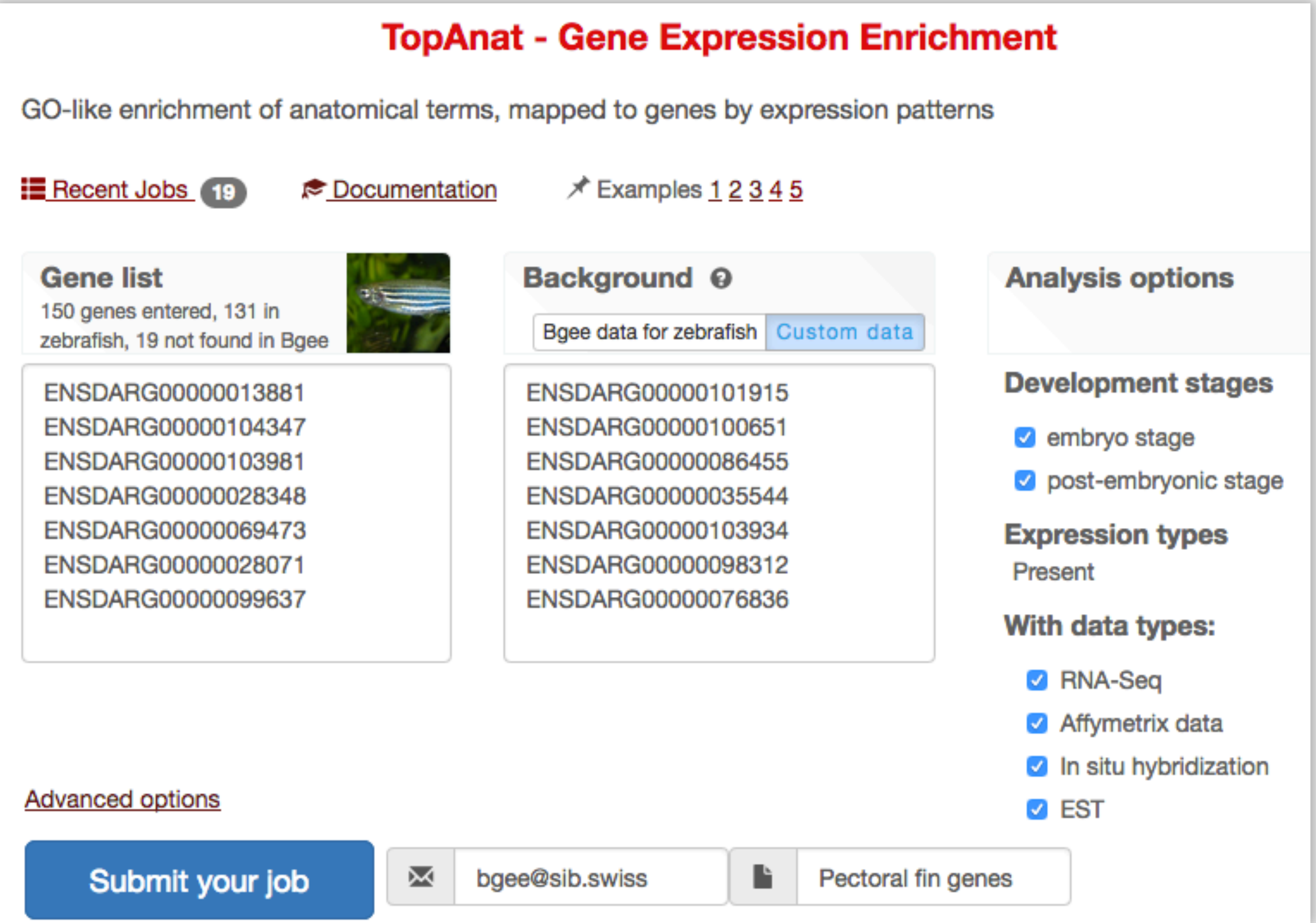
- For each anatomical structure:

	Gene list	Other genes
Expressed	n_1	n_3
Not expressed	n_2	n_4

- Fisher / Hypergeometric test

Implementation

- Based on topGO package
- Extension of *topGOdata* class
  - Accommodate Uberon Ontology
  - Use custom gene mapping



TopAnat - Gene Expression Enrichment

GO-like enrichment of anatomical terms, mapped to genes by expression patterns

Gene list

150 genes entered, 131 in zebrafish, 19 not found in Bgee

```
ENSDARG00000013881
ENSDARG00000104347
ENSDARG00000103981
ENSDARG00000028348
ENSDARG00000069473
ENSDARG00000028071
ENSDARG00000099637
```

Background

```
ENSDARG00000101915
ENSDARG00000100651
ENSDARG00000086455
ENSDARG00000035544
ENSDARG00000103934
ENSDARG00000098312
ENSDARG00000076836
```

Analysis options

Development stages

☑ embryo stage

☑ post-embryonic stage

Expression types

Present

With data types:

☑ RNA-Seq

☑ Affymetrix data

☑ In situ hybridization

☑ EST

✉ bgee@sib.swiss

Pectoral fin genes

http://bgee.org/?page=top_anat

BgeeDB

- http://www.bioconductor.org/packages/BgeeDB/

- Komljenovic*, Roux*, Robinson-Rechavi and Bastian (2016) BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. *F1000Research*, 5:2748

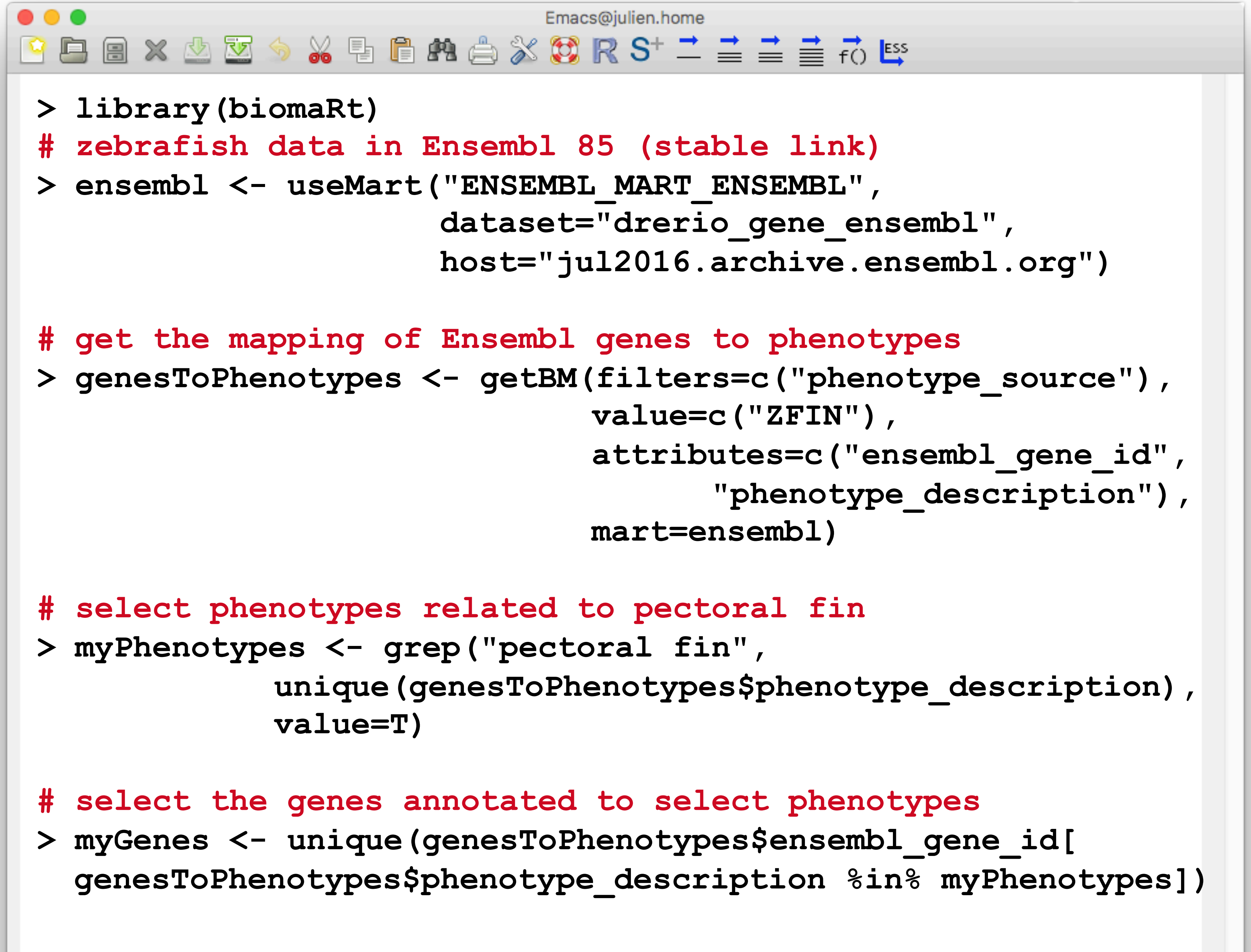
Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

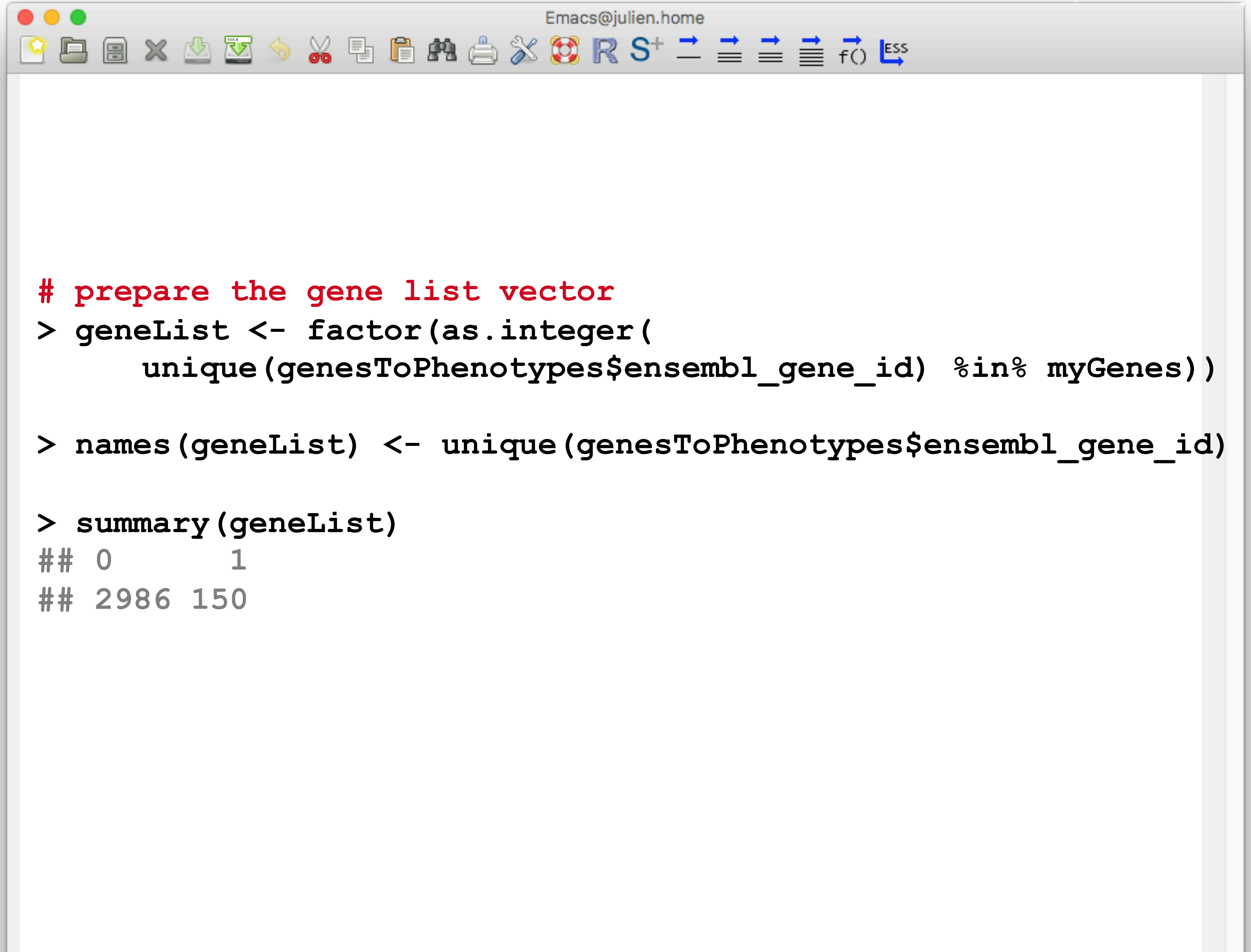
BgeeDB use case

TopAnat test:

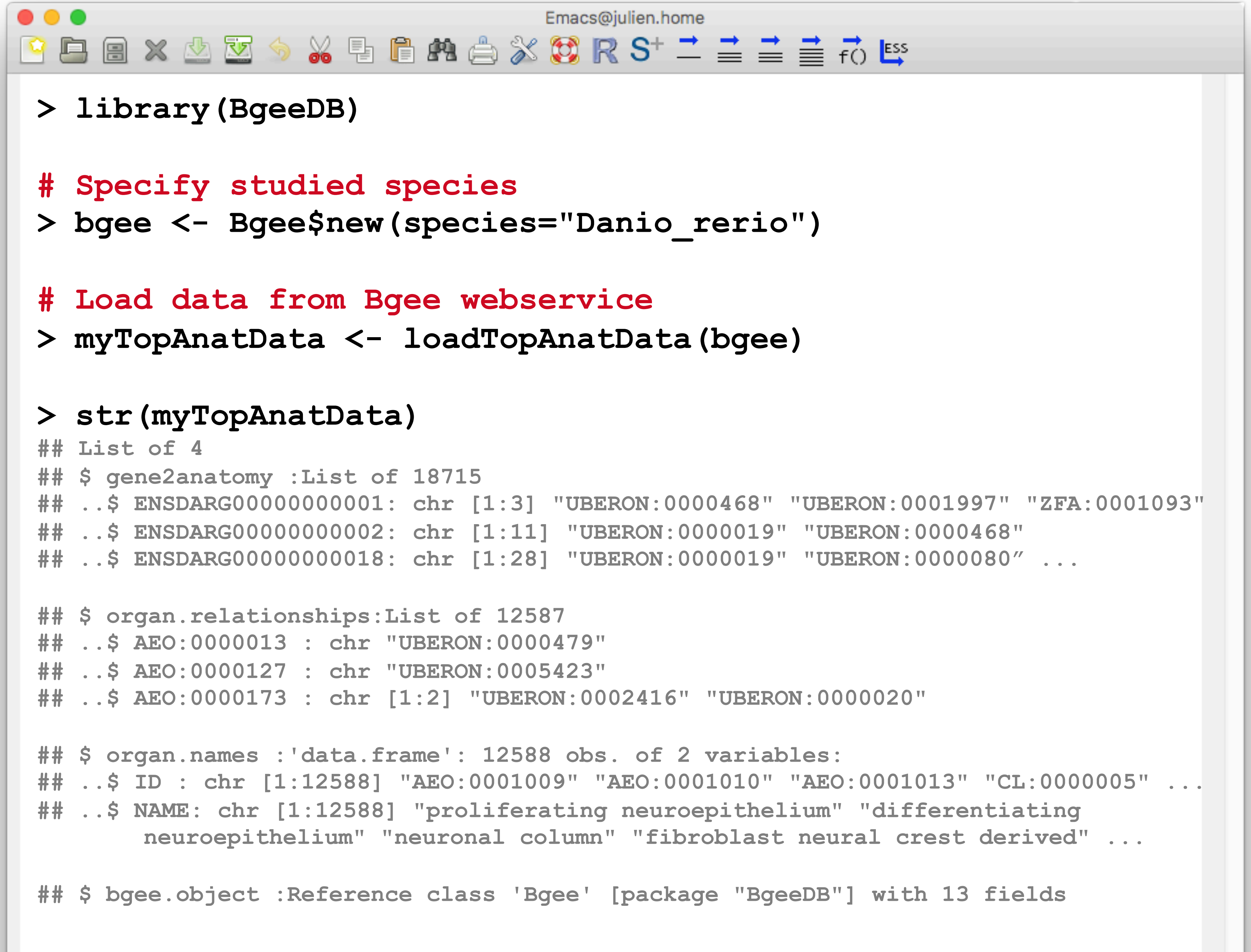
- **Foreground**: 150 Ensembl genes with *phenotype related to pectoral fin*, retrieved from ZFIN database
- **Background**: 3,136 Ensembl genes with an annotated phenotype in ZFIN



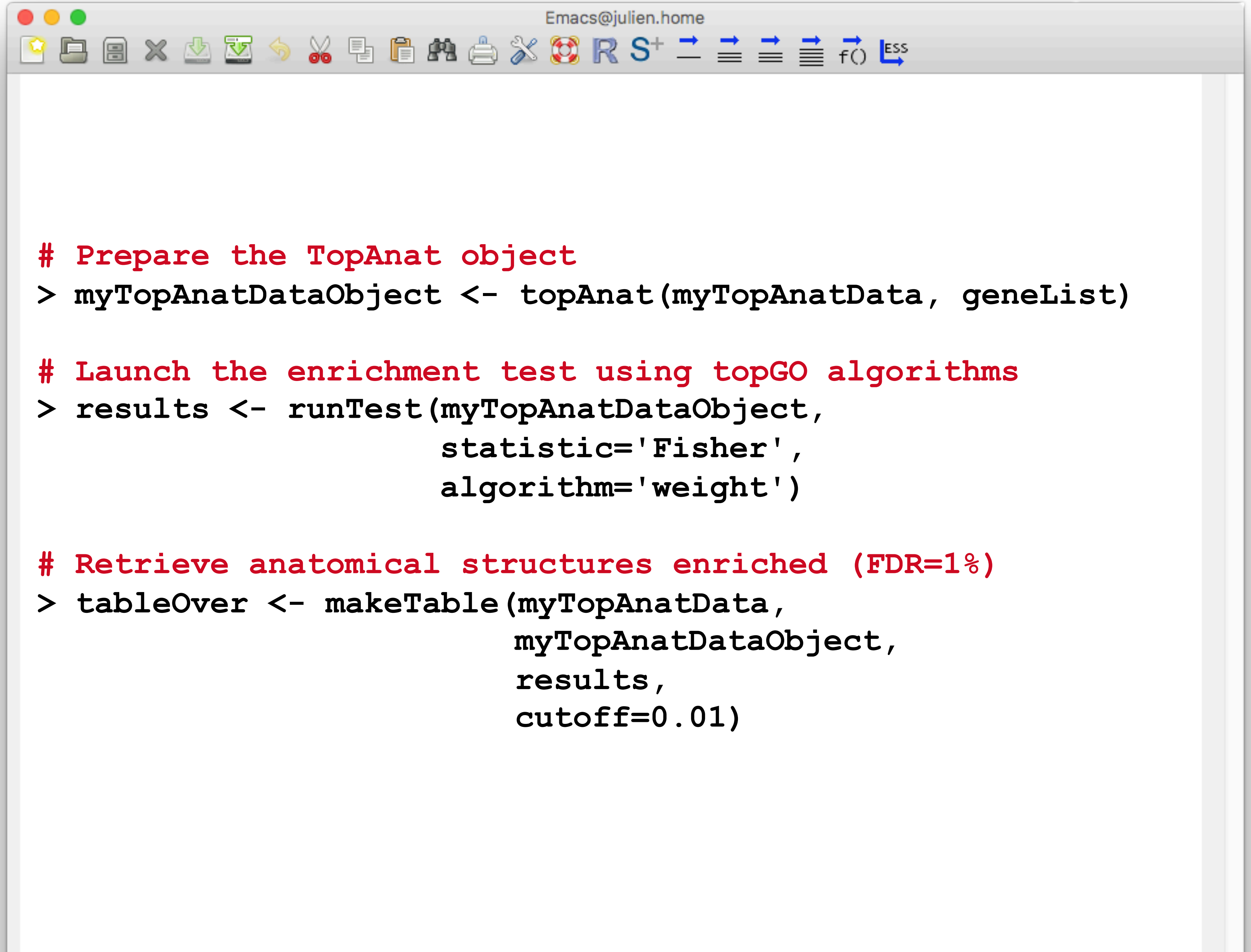
```
> library(biomaRt)
# zebrafish data in Ensembl 85 (stable link)
> ensembl <- useMart("ENSEMBL_MART_ENSEMBL",
                     dataset="drerio_gene_ensembl",
                     host="jul2016.archive.ensembl.org")

# get the mapping of Ensembl genes to phenotypes
> genesToPhenotypes <- getBM(filters=c("phenotype_source"),
                             value=c("ZFIN"),
                             attributes=c("ensembl_gene_id",
                                          "phenotype_description"),
                             mart=ensembl)

# select phenotypes related to pectoral fin
> myPhenotypes <- grep("pectoral fin",
                  unique(genesToPhenotypes$phenotype_description),
                  value=T)

# select the genes annotated to select phenotypes
> myGenes <- unique(genesToPhenotypes$ensembl_gene_id[
  genesToPhenotypes$phenotype_description %in% myPhenotypes])
```



```
# prepare the gene list vector
> geneList <- factor(as.integer(
      unique(genesToPhenotypes$ensembl_gene_id) %in% myGenes))

> names(geneList) <- unique(genesToPhenotypes$ensembl_gene_id)

> summary(geneList)
## 0       1
## 2986 150
```



```
> library(BgeeDB)

# Specify studied species
> bgee <- Bgee$new(species="Danio_rerio")

# Load data from Bgee webservice
> myTopAnatData <- loadTopAnatData(bgee)

> str(myTopAnatData)
## List of 4
## $ gene2anatomy :List of 18715
## ..$ ENSDARG00000000001: chr [1:3] "UBERON:0000468" "UBERON:0001997" "ZFA:0001093"
## ..$ ENSDARG00000000002: chr [1:11] "UBERON:0000019" "UBERON:0000468"
## ..$ ENSDARG00000000018: chr [1:28] "UBERON:0000019" "UBERON:0000080” ...

## $ organ.relationships:List of 12587
## ..$ AEO:0000013 : chr "UBERON:0000479"
## ..$ AEO:0000127 : chr "UBERON:0005423"
## ..$ AEO:0000173 : chr [1:2] "UBERON:0002416" "UBERON:0000020"

## $ organ.names :'data.frame': 12588 obs. of 2 variables:
## ..$ ID : chr [1:12588] "AEO:0001009" "AEO:0001010" "AEO:0001013" "CL:0000005" ...
## ..$ NAME: chr [1:12588] "proliferating neuroepithelium" "differentiating
        neuroepithelium" "neuronal column" "fibroblast neural crest derived" ...

## $ bgee.object :Reference class 'Bgee' [package "BgeeDB"] with 13 fields
```



```
# Prepare the TopAnat object
> myTopAnatDataObject <- topAnat(myTopAnatData, geneList)

# Launch the enrichment test using topGO algorithms
> results <- runTest(myTopAnatDataObject,
                     statistic='Fisher',
                     algorithm='weight')

# Retrieve anatomical structures enriched (FDR=1%)
> tableOver <- makeTable(myTopAnatData,
                         myTopAnatDataObject,
                         results,
                         cutoff=0.01)
```

Organ name	Enrichment fold	P-value	FDR
pectoral appendage field	12.7	4.00E-10	7.14E-08
pectoral appendage cartilage tissue	10.7	2.41E-08	3.58E-06
ceratohyal cartilage	7.6	4.76E-08	6.06E-06
median fin fold	7.1	7.17E-12	2.13E-09
fin bone	6.5	4.29E-06	0.000478091
bone of free limb or fin	6.1	7.95E-05	0.004168941
irregular bone	6.0	8.17E-06	0.000659745
dorsal hyoid arch skeleton	5.9	0.000107699	0.004817668
paired limb/fin bud	5.7	1.62E-22	1.45E-19
endochondral bone	5.4	7.11E-06	0.000633841
dermal bone	4.8	8.89E-06	0.000659745
mouth	4.4	0.000101104	0.004817668
pharyngeal epithelium	4.0	4.89E-06	0.000483901
hypoblast (generic)	3.6	1.12E-05	0.000713807
pectoral fin	3.4	1.04E-18	4.62E-16
germ ring	3.2	3.95E-05	0.002344194
skin epidermis	3.0	6.22E-05	0.003463408
ear vesicle	2.7	3.14E-10	6.98E-08
endoderm	2.7	0.000108141	0.004817668
pharyngeal arch	2.4	1.08E-05	0.000713807
cranium	2.2	0.000129452	0.005492447
immature eye	1.8	0.000190297	0.007707016

Conclusions

- TopAnat is a new way to make biological sense of gene lists
- Gene annotation entirely experimental!
- BgeeDB is a versatile way of running TopAnat analyses

Add BgeeDB to your toolbox!

Thanks!

- http://www.bioconductor.org/packages/BgeeDB/
- Komljenovic*, Roux*, Robinson-Rechavi and Bastian (2016) BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. *F1000Research*, 5:2748

@_julien_roux / @antifreezeprot / @BgeeDB

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS