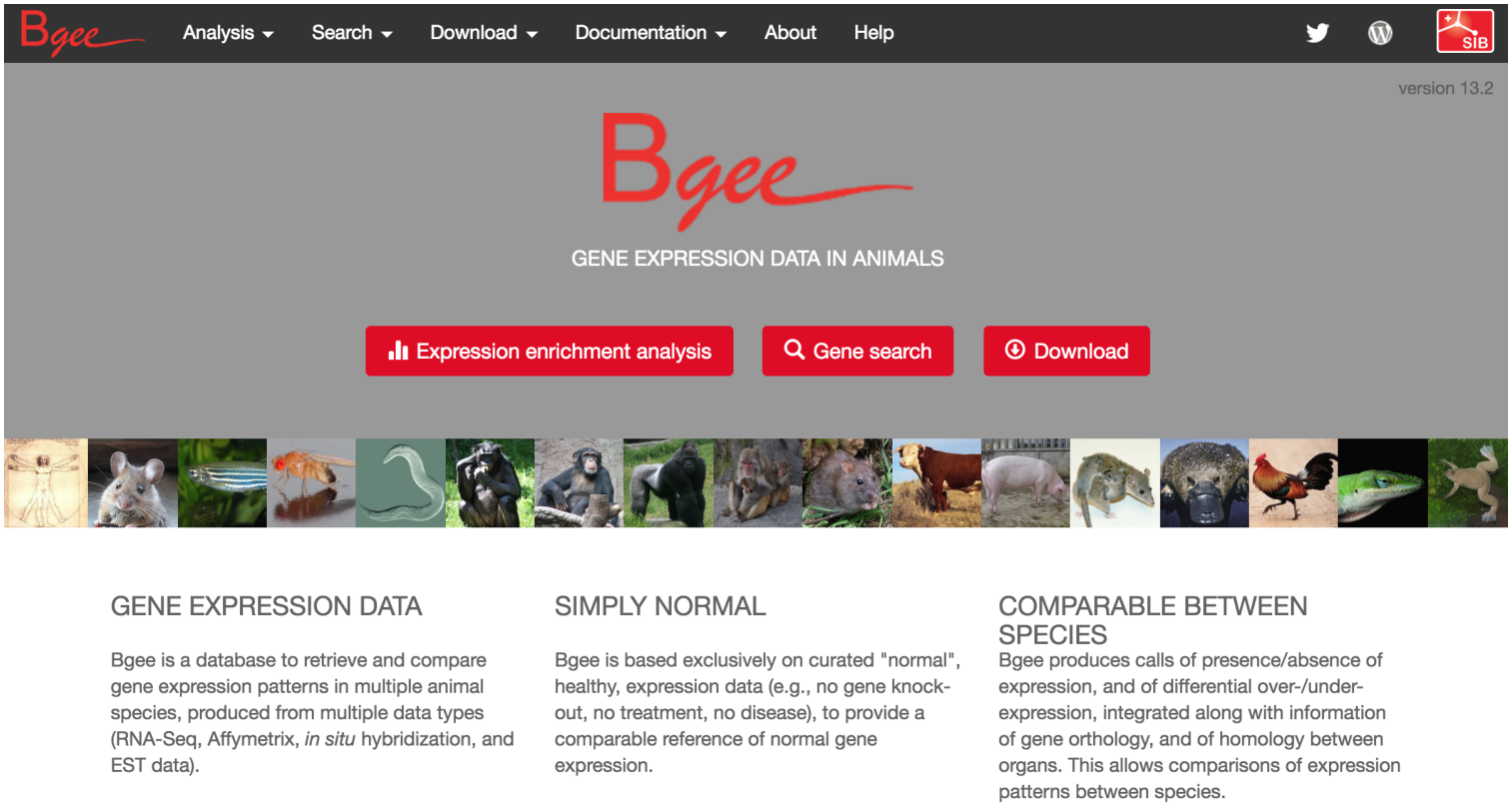
BgeeDB: an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests

Andrea Komljenovic*, Julien Roux*, Marc Robinson-Rechavi, Frederic B. Bastian

University of Lausanne, Switzerland

SIB Swiss Institute of Bioinformatics, Switzerland

European Bioconductor Developers' Meeting 2016
Basel, Switzerland



- database is accessible on: **bgee.org**
- 17 species
- RNA-Seq, Affymetrix microarrays, in situ hybridization and ESTs
- gene expression comparison across tissues, stages and species

Important features of **Bgee** database
that are easily accesible through **BgeeDB** package:

- **manually-curated** datasets
- exact **anatomical and stage mappings** to UBERON ontology

Manually-curated datasets

- Example: GSE1659 from GEO

Platforms (1)	GPL81 [MG_U74Av2] Affymetrix Murine Genome U74A Version 2 Array
Samples (12) Less...	GSM28550 Healthy 1 week C1
	GSM28551 Healthy 3 weeks C3
	GSM28552 Healthy 5 weeks C5
	GSM28553 Diabetic 1 week D1
	GSM28554 Diabetic 3 weeks D3
	GSM28555 Diabetic 5 weeks D5
	GSM28556 Trained diabetic 1 week DT1
	GSM28557 Trained diabetic 3 weeks DT3
	GSM28558 Trained diabetic 5 weeks DT5
	GSM28559 Trained 1 week T1
	GSM28560 Trained 3 weeks T3
	GSM28561 Trained 5 weeks T5

Manually-curated datasets

- GEOquery package keeps all 12 samples from GSE1659

```
## $GSE1659_series_matrix.txt.gz
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 12488 features, 12 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM28550 GSM28551 ... GSM28561 (12 total)
##   varLabels: title geo_accession ... data_row_count (26 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: 100001_at 100002_at ... AFFX-YEL024w/RIP1_at
##     (12488 total)
##   fvarLabels: ID GB_ACC ... Gene Ontology Molecular Function (16
##     total)
##   fvarMetadata: Column Description labelDescription
## experimentData: use 'experimentData(object)'
## Annotation: GPL81
```

Manually-curated datasets

- BgeeDB package includes only 3 healthy samples from GSE1659

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 9017 features, 3 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM28550 GSM28551 GSM28552
##   varLabels: Chip.ID Anatomical.entity.ID ... Stage.name (5 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: 100001_at 100002_at ...
##     AFFX-TransRecMur/X57349_M_at (9017 total)
##   fvarLabels: Probeset.ID Gene.ID
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

Anatomical and stage mapping to UBERON ontology

- Example: GSE1749 from GEO

Platforms (3)	GPL81	[MG_U74Av2] Affymetrix Murine Genome U74A Version 2 Array
	GPL339	[MOE430A] Affymetrix Mouse Expression 430A Array
	GPL340	[MOE430B] Affymetrix Mouse Expression 430B Array
Samples (57) Less...	GSM22541	Oocyte 1
	GSM22542	Oocyte 2
	GSM22543	Oocyte 3
	GSM22544	Oocyte 4
	GSM22545	1-Cell 1
	GSM22546	1-Cell 2
	GSM22547	1-Cell 3
	GSM22548	2-Cell 1
	GSM22549	2-Cell 2
	GSM22550	2-Cell 3
	GSM22551	8-Cell 1
	GSM22552	8-Cell 2
	GSM22553	8-Cell 3
	GSM22554	8-Cell 4
	GSM22555	Blastocyst 1
	GSM22556	Blastocyst 2
	GSM22557	Blastocyst 3

Anatomical and stage mapping to UBERON ontology

- GEOquery package keeps general mappings from GSE1749

```
##                  title type              source_name_ch1
## GSM22541     Oocyte 1  RNA preimplantation mouse embryo
## GSM22542     Oocyte 2  RNA preimplantation mouse embryo
## GSM22543     Oocyte 3  RNA preimplantation mouse embryo
## GSM22544     Oocyte 4  RNA preimplantation mouse embryo
## GSM22545     1-Cell 1  RNA preimplantation mouse embryo
## GSM22546     1-Cell 2  RNA preimplantation mouse embryo
## GSM22547     1-Cell 3  RNA preimplantation mouse embryo
## GSM22548     2-Cell 1  RNA preimplantation mouse embryo
## GSM22549     2-Cell 2  RNA preimplantation mouse embryo
## GSM22550     2-Cell 3  RNA preimplantation mouse embryo
## GSM22551     8-Cell 1  RNA preimplantation mouse embryo
## GSM22552     8-Cell 2  RNA preimplantation mouse embryo
## GSM22553     8-Cell 3  RNA preimplantation mouse embryo
## GSM22554     8-Cell 4  RNA preimplantation mouse embryo
## GSM22555 Blastocyst 1  RNA preimplantation mouse embryo
## GSM22556 Blastocyst 2  RNA preimplantation mouse embryo
## GSM22557 Blastocyst 3  RNA preimplantation mouse embryo
```

Anatomical and stage mapping to UBERON ontology

- BgeeDB package includes precise UBERON anatomical and stage mappings from GSE1749

```
##       Chip.ID Anatomical.entity.ID Anatomical.entity.name       Stage.ID               Stage.name
## 1561 GSM22541           CL:0000023                 oocyte UBERON:0000104                life cycle
## 1562 GSM22542           CL:0000023                 oocyte UBERON:0000104                life cycle
## 1563 GSM22543           CL:0000023                 oocyte UBERON:0000104                life cycle
## 1564 GSM22544           CL:0000023                 oocyte UBERON:0000104                life cycle
## 1565 GSM22555       UBERON:0000358             blastocyst UBERON:0000108            blastula stage
## 1566 GSM22556       UBERON:0000358             blastocyst UBERON:0000108            blastula stage
## 1567 GSM22557       UBERON:0000358             blastocyst UBERON:0000108            blastula stage
## 1568 GSM22545       UBERON:0000922                 embryo UBERON:0000106              zygote stage
## 1569 GSM22546       UBERON:0000922                 embryo UBERON:0000106              zygote stage
## 1570 GSM22547       UBERON:0000922                 embryo UBERON:0000106              zygote stage
## 1571 GSM22548       UBERON:0007010        cleaving embryo MmusDv:0000005 Theiler stage 02 (mouse)
## 1572 GSM22549       UBERON:0007010        cleaving embryo MmusDv:0000005 Theiler stage 02 (mouse)
## 1573 GSM22550       UBERON:0007010        cleaving embryo MmusDv:0000005 Theiler stage 02 (mouse)
## 1574 GSM22551       UBERON:0007010        cleaving embryo MmusDv:0000006 Theiler stage 03 (mouse)
## 1575 GSM22552       UBERON:0007010        cleaving embryo MmusDv:0000006 Theiler stage 03 (mouse)
## 1576 GSM22553       UBERON:0007010        cleaving embryo MmusDv:0000006 Theiler stage 03 (mouse)
## 1577 GSM22554       UBERON:0007010        cleaving embryo MmusDv:0000006 Theiler stage 03 (mouse)
```

The **BgeeDB** is a collection of functions to import data from the **Bgee** database directly into R.

- List annotation of RNA-seq and microarray
- Download the processed gene expression data
- Download the gene expression calls and use them to perform gene list expression localization enrichment tests analyses

Current release of the database

Checking for current release in **BgeeDB**:

```
> library(BgeeDB)
> listBgeeRelease()
```

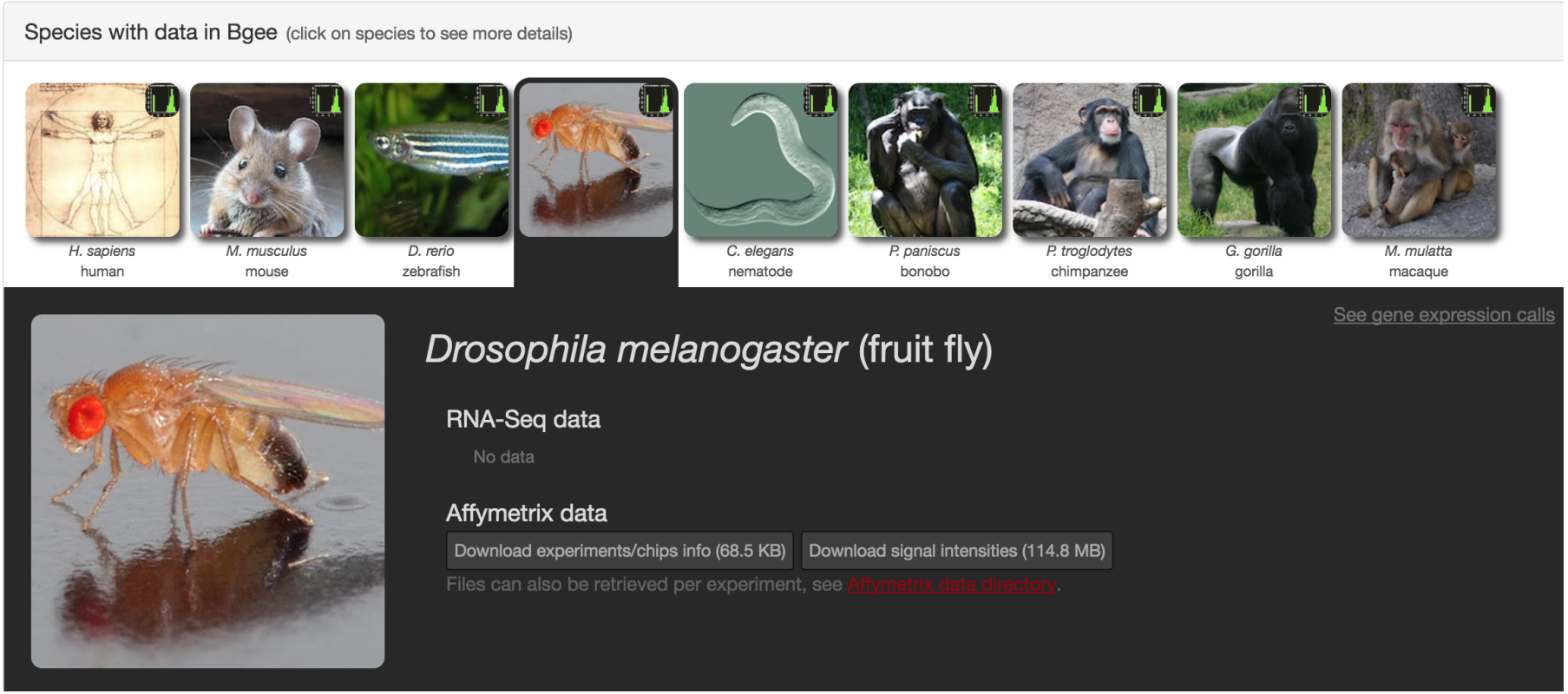
	Number of libraries	**Number of species**
Release 13	526 RNA-seq libraries	17 animal species
Release 14	5 746 RNA-seq libraries	29 animal species

Current release also offers 12 736 Affymetrix, 46 619 in situ hybridization and 3 185 EST libraries.

Availability of species and datatypes

Checking the species and their data types in **BgeeDB**:

```
> listBgeeSpecies()
```



i. Download part of package

- **getAnnotation()**
- **getData()**
- **formatData()**

ii. Enrichment part of package

The **getAnnotation()** function will output the list of RNA-seq experiments and libraries available in **Bgee** for mouse.

```
> bgee <- Bgee$new(species = "Mus_musculus",
+                  dataType = "rna_seq")
> annotation_bgee_mouse <- getAnnotation(bgee)
```

```
## $sample.annotation
##   Experiment.ID Library.ID Library.secondary.ID Anatomical.entity.ID
## 1      GSE30617  GSM759583            ERX012363       UBERON:0000948
## 2      GSE30617  GSM759584            ERX012348       UBERON:0000948
## 3      GSE30617  GSM759585            ERX012344       UBERON:0000948
## 4      GSE30617  GSM759586            ERX012362       UBERON:0000948
## 5      GSE30617  GSM759587            ERX012378       UBERON:0000948
## 6      GSE30617  GSM759588            ERX012374       UBERON:0000948
##   Anatomical.entity.name       Stage.ID      Stage.name
## 1                  heart MmusDv:0000052 8 weeks (mouse)
## 2                  heart MmusDv:0000052 8 weeks (mouse)
## 3                  heart MmusDv:0000052 8 weeks (mouse)
## 4                  heart MmusDv:0000052 8 weeks (mouse)
## 5                  heart MmusDv:0000052 8 weeks (mouse)
## 6                  heart MmusDv:0000052 8 weeks (mouse)
```

The **getData()** function will download processed RNA-seq data from all mouse experiments in Bgee as a list.

```
> data_bgee_mouse <- getData(bgee)
```

Name	Size	Date Modified
[parent directory]		
Mus_musculus_RNA-Seq_experiments_libraries.zip	9.1 kB	7/6/16, 11:54:00 AM
Mus_musculus_RNA-Seq_read_counts_RPKM_GSE30352.tsv.zip	4.8 MB	7/6/16, 11:54:00 AM
Mus_musculus_RNA-Seq_read_counts_RPKM_GSE30617.tsv.zip	10.2 MB	7/6/16, 11:54:00 AM
Mus_musculus_RNA-Seq_read_counts_RPKM_GSE36026.tsv.zip	3.5 MB	7/6/16, 11:54:00 AM
Mus_musculus_RNA-Seq_read_counts_RPKM_GSE41338.tsv.zip	1.7 MB	7/6/16, 11:54:00 AM
Mus_musculus_RNA-Seq_read_counts_RPKM_GSE41637.tsv.zip	7.9 MB	7/6/16, 11:54:00 AM
Mus_musculus_RNA-Seq_read_counts_RPKM_GSE43520.tsv.zip	2.6 MB	7/6/16, 11:54:00 AM
Mus_musculus_RNA-Seq_read_counts_RPKM_GSE43721.tsv.zip	898 kB	7/6/16, 11:54:00 AM
Mus_musculus_RNA-Seq_read_counts_RPKM.zip	31.3 MB	7/6/16, 11:54:00 AM

The **formatData()** function reformats the data into an ExpressionSet object including:

```
> mouse.counts <-
+           formatData(bgee,
+                      data_bgee_mouse,
+                      callType = "present", stats = "counts")
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 39179 features, 36 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM759583 GSM759584 ... GSM759618 (36 total)
##   varLabels: Library.ID Anatomical.entity.ID ... Stage.name (5
##     total)
##   varMetadata: labelDescription
## featureData
##   featureNames: ENSMUSG00000000001 ENSMUSG00000000003 ...
##     ENSMUSG00000099334 (39179 total)
##   fvarLabels: Gene.ID
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
```

The **BgeeDB** offers ExpressionSet object for downstream analysis:

```
> library(edgeR)
> # subset the dataset to brain and heart
> brain.heart <-
+   mouse.counts[,
+   pData(mouse.counts)$Anatomical.entity.name %in%
+   c("brain", "heart")]
> # filter out very lowly expressed genes
> brain.liver<-
+   brain.liver[rowSums(cpm(brain.liver) > 1) > 3,]
> # create edgeR DGEList object
> dge <- DGEList(counts=brain.liver.filtered,
+ group=pData(brain.liver.filtered)$Anatomical.entity.name)
> dge <- calcNormFactors(dge)
> dge <- estimateCommonDisp(dge)
> ...
```

i. Download part of package

- getAnnotation()
- getData()
- formatData()

ii. Enrichment part of package - Julien Roux

Acknowledgments

UNIL | Université de Lausanne

SIB Swiss Institute of Bioinformatics

- Bgee team
- Marc Robinson-Rechavi

Komljenovic A*, Roux J*, Robinson-Rechavi M and Bastian FB. BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests [version 1; referees: awaiting peer review]. F1000Research 2016, 5:2748