



We have shown how Bioconductor provides resources for studying the human genome sequence as well as SNPs. Bioconductor also provides resources that permit us to obtain information about genes. We will see how these databases can be quite complex. But before learning about these here we present some experimental data.

The driving force behind the formation of the Bioconductor project was the emergence of high throughput measurement of gene expression data. Unlike genome sequence and SNP data, gene expression data varies from cell to cell, from tissue to tissue and from individual to individual. Statistical techniques such as those implemented in R were natural tools to parse out variability and perform statistical inference. Furthermore, the ambitious use of newly invented technologies added issues of measurement error and bias to already difficult challenge.

Note that in the previous assessments we focused on *static* database information: genome sequence and SNPs. In previous courses we have seen the 'tissuesGeneExpression' data which is experimental data measured with microarrays. If you have not installed it you can do it like this:

```
library(devtools)
install_github("genomicsclass/tissuesGeneExpression")
```

You can load the data:

```
library(tissuesGeneExpression)
data(tissuesGeneExpression)
head(e[,1:5])
table(tissue)
```

The rows of the matrix `e` are the features, in this case representing genes, and the columns are the samples. The entries of the matrix are gene expression measurements (in log scale) obtained using a microarray technology.

### QUESTION 1.3.1 (1 point possible)

Look at the data for the feature with ID "209169\_at". You can index the rows of 'e' directly with this character string.

Which of the following best describes the data? (Hint: stratify data by tissue and create boxplots)

- ☐ This is human data and this gene has the same sequence across all tissues thus there is no difference in gene expression
- ☒ This gene is expressed in the brain but not the other tissues ✓
- ☐ This gene is differentially expressed between all tissues
- ☐ The individual to individual variability is much larger than the difference between tissues

#### EXPLANATION

We can stratify the data and make boxplots like this:

```
boxplot(e["209169_at",]~tissue,las=2)
```

[Hide Answer](#)

You have used 5 of 5 submissions

**QUESTION 1.3.2** (1 point possible)

Below is a vector of 6 IDs which index features of 'e':

IDs = c("201884\_at", "209169\_at", "206269\_at", "207437\_at", "219832\_s\_at", "212827\_at")

Which of the following ID(s) appear to represent a gene specific to placenta? Be careful when you are picking, to pick the correct name or names. Names often look similar. Also, if you get your guess wrong, you need to uncheck the ones you think are wrong to guess again.

[Help](#)

- ☐ "201884\_at"
- ☐ "209169\_at"
- ☒ "206269\_at" 
- ☐ "207437\_at"
- ☐ "219832\_s\_at"
- ☐ "212827\_at"

**EXPLANATION**

```
IDs = c("201884_at", "209169_at", "206269_at", "207437_at", "219832_s_at", "212827_at")
library(rafalib)
mypar2(3,2)
for(i in IDs){
  boxplot(e[i,]~tissue,las=2)
```


[Hide Answer](#)

You have used 5 of 5 submissions

**QUESTION 1.3.3** (1 point possible)

Note that there is much existing work on gene function and all we have here are identifiers provided by the manufacturer of the machine that makes the measurements. How would we go about finding more information about gene "206269\_at" for example? Does it have a known function? Where is it on the genome? What is its sequence? One of the strengths of Bioconductor is that it connects R, an existing comprehensive toolbox for data analysis, with the existing comprehensive databases annotating the genome. We will learn about these powerful resources in this class.

The microarray product used to make the measurements described here is the Affymetrix Human GeneChip HG133A. Search the Bioconductor website and determine which of the following packages provides a connection to gene information:

- ☐ Biobase
- ☐ simpleaffy
- ☐ hgu133a2cdf
- ☒ hgu133a.db 

☐ affy[Hide Answer](#)

You have used 5 of 5 submissions

**QUESTION 1.3.4** (1 point possible)

Another powerful aspect of Bioconductor is that it provides object classes specifically designed to keep high throughput data organized. Below we show an example of how the three tables that are needed to conduct data analysis are available from Bioconductor data objects. For example, for gene expression we can use the ExpressionSet object, which we will review in more detail in later weeks.

```
library(Biobase)
## the GSE5859 package can be installed like this:
## library(devtools)
## install_github("genomicsclass/GSE5859")
library(GSE5859)
data(GSE5859)
class(e)
```

These objects were originally designed for gene expression data so the methods to extract the high throughput measurements have related names:

```
dat = exprs(e)
dim(dat)
```

The information about samples is also stored in this object and the functions to create it try to guarantee that the columns of `exprs(e)` match the rows of the sample information table. `pData` is used as shorthand for phenotype data.

```
sampleInfo = pData(e)
dim(sampleInfo)
head(sampleInfo)
```

A final table, which we will cover in more detail later, is a table that describes the rows, in this case genes. Because each product will have a different table, these have already been created in Bioconductor.

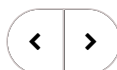
Because there are certain products that are widely used, Bioconductor makes databases available from which you can extract this information. This way, every object which is linked to this information does not have to carry around the table. Later we will learn to write code like this to extract specific information from these tables:

```
library(hgfocus.db)
annot = select(hgfocus.db,
              keys=featureNames(e),
              keytype="PROBEID",
              columns=c("CHR", "CHRLOC", "SYMBOL"))
## here we pick one column from the annotation
annot = annot[ match(featureNames(e),annot$PROBEID), ]
head(annot)
dim(annot)
```

For the `tissuesGeneExpression` data, which of the following would be the correct way to store the information, using the Bioconductor infrastructure?

☐ `e` should be in a data.frame with the columns given by tissue

- ☒ In an `eSet` object with the `e` in the `assayData` accessible with `exprs()` and `tissue` as one of the columns in the `phenoData` ✓
- ☐ In an sql database
- ☐ In `eSet` object with the `e` in the `featureData` and `tissue` one of the columns of `assayData`

[Hide Answer](#)*You have used 5 of 5 submissions*

edX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law, literature, math, medicine, music, philosophy, physics, science, statistics and more. edX is a non-profit online initiative created by founding partners Harvard and MIT.

© 2015 edX Inc.

EdX, Open edX, and the edX and Open edX logos are registered trademarks or trademarks of edX Inc.

[Terms of Service and Honor Code](#)

[Privacy Policy \(Revised 10/22/2014\)](#)



#### About edX

[About](#)

[News](#)

[Contact](#)

[FAQ](#)

[edX Blog](#)

[Donate to edX](#)

[Jobs at edX](#)

#### Follow Us

[Facebook](#)

[Twitter](#)

[LinkedIn](#)

[Google+](#)

[Tumblr](#)

[Meetup](#)

[Reddit](#)

[Youtube](#)