

gse21942

Guillermo Ayala

2025-03-11

Table of contents

Packages	1
Data	1
Downloading and preprocessing the data	1
Normalization	2
Multiple correspondences	3

Packages

```
pacman::p_load(GEOquery,affy,Biobase,hgu133plus2.db)
```

Data

The data set can be found at [GEO gse21942](#).

Downloading and preprocessing the data

```
GEOquery::getGEOSuppFiles("GSE21942")
setwd("GSE21942/")
system("tar xvf GSE21942_RAW.tar")
# Reading CEL files to a affybatch
gse21942raw = affy::ReadAffy()
setwd("../")
system("rm -fr GSE21942/")
```

Normalization

- Normalization using RMA

```
gse21942rma = rma(gse21942raw)
```

Let us modify the `rownames`.

```
rownames(pData(gse21942rma)) = sapply(rownames(pData(gse21942rma)),function(i)
  unlist(strsplit(i,split=".CEL."))[1])
```

We download the processed data set.

```
gse = getGEO("GSE21942")
```

The samples have the same position.

```
all(rownames(pData(gse[[1]])) == rownames(pData(gse21942rma)))
```

We take the phenotypic variables from the processed data.

```
pData(gse21942rma) = pData(gse[[1]])
```

We modify the name of the last phenotypic variable.

```
names(pData(gse21942rma))[ncol(pData(gse21942rma))] = "FactorValue..DISEASE.STATE."
```

The samples GSM545845 and GSM545846 are technical replications. We remove them from the ExpressionSet. First, we can see that they are the last two samples.

```
match(c("GSM545845", "GSM545846"), rownames(pData(gse21942rma)))
```

The new ExpressionSet would be

```
gse21942a = gse21942rma[,-c(28,29)]
```

Multiple correspondences

```
a = AnnotationDbi::select(hgu133plus2.db,
                           keys=featureNames(gse21942a),
                           columns=c("ENTREZID","ENSEMBL"),
                           keytype="PROBEID")

a = a[!is.na(a[, "ENTREZID"]),] ## Eliminamos sondas sin ENTREZID
c1 = match(unique(a[,1]),a[,1])
a1 = a[c1,]
c2 = match(unique(a1[,2]),a1[,2])
a2 = a1[c2,]
dim(a2)
gse21942 = gse21942a[match(a2[,1],featureNames(gse21942a)),]
fData(gse21942) = a2
all(featureNames(gse21942) == a2$PROBEID) ## Comprobamos la correspondencia

save(gse21942,file=paste0(dirTamiData,"gse21942.rda"))
```