ExpressionSet assessment

Integrity assurances with ExpressionSet

1/1 point (graded)

Let's use the genefu package again to work with breast cancer expression data.

```
library(Biobase)
library(genefu)
data(nkis)
dim(demo.nkis)
head(demo.nkis)[,1:8]
```

Try the following:

```
nkes = ExpressionSet(data.nkis, phenoData=AnnotatedDataFrame(demo.nkis),
  featureData=AnnotatedDataFrame(annot.nkis))
```

How many errors were generated with this command?

6 ✔

6

Submit You have used 2 of 5 attempts

✔ Correct (1/1 point)

Fixing the mismatched elements

1/1 point (graded)

What must be done to correct the errors in the previous attempt at making an ExpressionSet?

- transpose the sample data in demo.nkis
- don't supply the feature metadata in this step
- transpose the expression data matrix in data.nkis
- there was no error

✔

You have used 1 of 2 attempts

Submit

✔ Correct (1/1 point)

Working with GEO

1/1 point (graded)

Acquire the dataset associated with the paper "Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9", PMID 16862118, by Krivtsov and colleagues. This paper uses Affymetrix microarrays to study how macrophage precursors may become cancer cells.

```
# setup Bioconductor
library(GEOquery)
# retrieve the LSC data from GEO
lstem = getGEO("GSE3725")
```

What is the class of lstem after this operation completes?

- ExpressionSet
- list
- matrix
- indeterminate

Submit You have used 1 of 5 attempts

✔ Correct (1/1 point)

Experiment characteristics

1/1 point (graded)
Use

```
lstem = lstem[[1]]
```

after the command above to obtain a simple reference to the ExpressionSet instance of interest.

How many samples have been hybridized?
You can use ncol().

28 ✔

28

Submit You have used 1 of 5 attempts

✔ Correct (1/1 point)

Platform content

1/1 point (graded)

How many features are present on the array used in this experiment?

22690 ✔

22690

Submit You have used 1 of 5 attempts

✔ Correct (1/1 point)

Details of samples

1/1 point (graded)

One common difficulty of working with GEO is that the characteristics of different samples are not always easily determined. Sometimes there is no annotation, and sometimes the annotation is present in an unusual field. In this case, the sample characteristic of interest is the type of cell on which expression measures were taken. This can be found using the 'title' field of the pData(lstem). In other words,

```
pData(lstem)$title
```

generates a listing of the cell type descriptions. Let's ignore the first 6 samples:

```
lstem = lstem[, -c(1:6)] # note position of comma!
```

There are five different cell types present, identified by text in parentheses:

- HSC: hematopoetic stem cells
- GMP: granulocyte macrophage progenitors
- CMP: common myeloid precursors
- MEP: megakaryocyte erythroid progenitors
- L-GMP: GMP-like leukemic cells

How many samples are of type L-GMP?

6 ✔

6

Submit You have used 1 of 5 attempts

✔ Correct (1/1 point)

Improving sample and feature labeling for a heatmap

1/1 point (graded)

We'll conclude this problem set by producing a heatmap that compares the cell types in a useful way. The data in GEO are apparently not normalized. We will use a very crude approach to achieve constant median on the log scale, after recoding (rare) negative values to zero and then adding 1 to all values.

```
## perform an elementary normalization
ee = exprs(lstem)
ee[ee<0] = 0
eee = log(ee+1)
## boxplot(data.frame(eee))
meds = apply(eee,2,median)
tt = t(t(eee)-meds)
## boxplot(data.frame(tt))
## assign the normalized values to ExpressionSet
exprs(lstem) = tt
```

Now we will modify the feature names to be gene symbols instead of array probe names.

```
# simplify downstream labeling with gene symbol
featureNames(lstem) = make.names(fData(lstem)$"Gene Symbol", unique=TRUE)
```

The following code is somewhat complex, but it simplifies labeling of cell types by stripping away details of marker configurations.

```
  # reformat the naming of cell types
ct = pData(lstem)[,1]
ct = as.character(ct)
cct = gsub(".*(\\(.*\\)).*", "\\1", ct)
cct = make.unique(cct)
cct = gsub(" enriched", "", cct)
# use the cell types as sample names
sampleNames(lstem) = cct
```

Four genes identified in the stemness signature are given in a vector below. We will use these for a small-scale heatmap.

```
  # select some members of the stem cell signature
inds = which(fData(lstem)$"Gene Symbol" %in% c("Stat1", "Col4a1", "Hoxa9", "Itgb5"))
```

Finally we can produce the heatmap.

```
# obtain a simple heatmap
heatmap(exprs(lstem[inds,]), Colv=NA)
```

What's the total number of probes interrogating the four genes of interest?
You can add an optional tip or note related to the prompt like this.

11 ✔ **Answer:** 11

11

Submit You have used 1 of 5 attempts

Answers are displayed within the problem