Verified assessment: Genome-scale data, Q1-4

Check whether you understand how to import, create and use *ExpressionSet* and *SummarizedExperiment* objects by investigating datasets related to cancer. These exercises cover:

- basics of the *ExpressionSet* and *SummarizedExperiment* classes,
- accessing data from GEO using `GEOquery`,
- importing NGS data from a BAM file,
- analyzing *ExpressionSets* and *SummarizedExperiments*.

This assessment, like all other end-of-section assessments, is available to verified users only.

Report all numerical answers to at least 3 significant digits.

These exercises require the following packages:

```
library(GEOquery)
library(NGScopyData)
library(Rsamtools)
library(GenomicAlignments)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
library(AnnotationDbi)
```

Q1: Functions to access ExpressionSet and SummarizedExperiment data

2.0/2.0 points (graded)

For each of the following tasks, choose the function that will return the desired data. Choices include `assay(x)`, `colData(x)`, `exprs(x)`, `fData(x)`, `pData(x)` and `rowData(x)`. Each function will be used once.

Given an *ExpressionSet* `x`, access experimental data.

exprs(x) ✔

Given an *ExpressionSet* `x`, access sample information.

pData(x) ✔

Given an *ExpressionSet* `x`, access feature information.

fData(x) ✔

Given a *SummarizedExperiment* `x`, access experimental data.

assay(x) ✔

Given a *SummarizedExperiment* `x`, access sample information.

colData(x) ✔

Given a *SummarizedExperiment* `x`, access feature information.

rowData(x) ✔

Submit You have used 2 of 5 attempts

Use the `getGEO()` function from the **GEOquery** package to download the data associated with the paper "Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9" by Krivtsov et al. (PMID: 16862118). This paper uses Affymetrix microarrays to study how macrophage precursors may become cancer cells. The data from this paper was uploaded to GEO with accession number GSE3725.

Note that `getGEO()` initially imports the data as a list. The experimental data are stored in the first element of the list. Extract the first element of the list and save the resulting *ExpressionSet* as `lstem`.

Q2: Number of samples and features

2/2 points (graded)

How many samples are in `lstem` (the *ExpressionSet* obtained above)?

28 ✔

28

How many features are in `lstem`?

22690 ✔

22690

Submit You have used 3 of 10 attempts

✔ Correct (2/2 points)

Q3: Sample characteristics

1/1 point (graded)

When working with data from GEO, sample metadata may often be missing or, more often, available but difficult to parse. The lack of a standard format for sample annotations means that extracting important sample characteristics can require additional work by the data analyst. (However, this is a small price for the convenience of obtaining entire datasets with a single call to `getGEO()`.)

In this case, the sample characteristic of interest is the type of cell on which expression measures were taken. There are five different cell types in the dataset, identified by text in parentheses in the `title` field of the sample data:

- `HSC`: hematopoetic stem cells,
- `GMP`: granulocyte macrophage progenitors,
- `CMP`: common myeloid precursors,
- `MEP`: megakaryocyte erythroid progenitors,

- `L-GMP` : GMP-like leukemic cells.

Before moving on, we will remove 6 samples corresponding to experimental controls from `lstem` .

```
lstem <- lstem[, !grepl("^GMP expressing", pData(lstem)$title)]
```

We can use a regular expression to extract the cell type from the `title` column of the sample data and add this information as a new column, `cell_type` . (A deep dive on regular expressions is beyond the scope of this exercise.)

```
titles <- as.character(pData(lstem)$title)
cell_type <- gsub(".*\\((.*?)( enriched)?\\).*", "\\1", titles)

# add cell_type column to pData
pData(lstem)$cell_type <- factor(cell_type)
```

How many samples are from the cell type L-GMP?

6 ✔

6

Submit You have used 1 of 10 attempts

✔ Correct (1/1 point)

Q4: Expression of a feature

3/3 points (graded)

Inspect the feature data. What gene symbol (`Gene Symbol` field) is associated with probe `1421579_at`?

Hoxa9 ✔

What is the mean expression level of `1421579_at` across all samples?
(Note that these data have not been normalized. We will learn more about normalization later.)

356.7458 ✔

356.7458

What is the mean expression level of `1421579_at` in samples with cell type L-GMP?

534.8235 ✔

534.8235

Submit You have used 1 of 10 attempts

✔ Correct (3/3 points)

Continue the verified assessment on management of genome-scale data on the following page.

Discussion

Hide Discussion

Topic: Section 3 / Genome-scale data, Q1-4

Add a Post

Show all posts ▼ by recent activity ▼

There are no posts in this topic yet.

✖