Seminar III: R/Bioconductor

August-December 2009

Leonardo Collado Torres

Bachelor in Genomic Sciences (LCG),
UNAM, Cuernavaca, Mexico

lcollado@lcg.unam.mx

http://www.lcg.unam.mx/~lcollado/

August 27, 2009

Assistants: Alejandro Reyes areyes@lcg.unam.mx, José Reyes jreyes@lcg.unam.mx and Víctor Moreno jmoreno@lcg.unam.mx

Note: Questions through the forum please. Those who are not from the sixth LCG generation send us an email so we can register you on the forum.

Abstract

Expected solutions for the second set of exercises. Note that your template .Rnw file should just be a much simpler file than the one used to make this pdf file: answer_02_sweave.Rnw

1 Sweave

1. Create your own template Sweave document.
   - Title: course name, homework number
   - Author: name, email, include a link to your personal academic webpage if you have one. [1]

[1] You will probably make one this semester on the PHP course.

- Abstract: short description on the homework and any notes you might want to add
- A sample homework solution: meaning a short description and some code. For example, how to sum $2 + 3$.

```
> 2 + 3
```

```
[1] 5
```

2. For a proper template check: `http://www.lcg.unam.mx/~lcollado/B/quizzes/02_answer/answer_02_template.Rnw` which generates `http://www.lcg.unam.mx/~lcollado/B/quizzes/02_answer/answer_02_template.pdf`

2 ALL dataset

- You'll have to explore the ALL dataset[2] and create your first homework as a vignette document.
- Install the `ALL` package and explore the `ALL` object.

```
> library(ALL)
> data(ALL)
> ALL
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 12625 features, 128 samples
  element names: exprs
phenoData
  sampleNames: 01005, 01010, ..., LAL4  (128 total)
  varLabels and varMetadata description:
    cod:  Patient ID
    diagnosis:  Date of diagnosis
    ...: ...
    date last seen:  date patient was last seen
    (21 total)
featureData
  featureNames: 1000_at, 1001_at, ..., AFFX-YEL024w/RIP1_at  (12625 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
  pubMedIds: 14684422 16243790
Annotation: hgu95av2
```

[2] John Quackenbush mentioned it on Monday as the most studied dataset.

```
> varLabels(ALL)
```

```
 [1] "cod"            "diagnosis"      "sex"            "age"
 [5] "BT"             "remission"      "CR"             "date.cr"
 [9] "t(4;11)"        "t(9;22)"        "cyto.normal"    "citog"
[13] "mol.biol"       "fusion protein" "mdr"            "kinet"
[17] "ccr"            "relapse"        "transplant"     "f.u"
[21] "date last seen"
```

- Select the samples from the B-cell tumors.[3]

```
> bcell <- grep("^B", as.character(ALL$BT))
```

- Select those of molecular type `BCR/ABL` or `NEG`.[4]

```
> subset.mol.biol <- ALL$mol.biol %in% c("BCR/ABL", "NEG")
```

- Combine the previous two subsets and keep the *intersect*ion

```
> subset.pos <- intersect(bcell, which(subset.mol.biol == TRUE))
> ALL.work <- ALL[, subset.pos]
```

- Eliminate unused factor levels on your resulting subset.

```
> ALL.work$BT <- factor(ALL.work$BT)
> ALL.work$mol.biol <- factor(ALL.work$mol.biol)
```

- Use the `nsFilter` function from the `genefilter` package to keep those with *entrez* ID, *GOBP*, remove duplicate *entrez* and the following arguments:

```
> library(genefilter)
> library(hgu95av2.db)
> filtered <- nsFilter(ALL.work, var.fun = IQR, var.cutoff = 0.5,
+     feature.exclude = "^AFFX", require.entrez = TRUE, require.GOBP = TRUE,
+     remove.dupEntrez = TRUE)
```

- Meaning that we'll use the interquantile range with a variance cutoff of 0.5 to eliminate those with small variation and by excluding `AFFX` we'll take out the controls AFFY probes.

- How many:

[3] What's the name of the function to look for text in Unix? Well, a function with the same name is available in `R`. Use it

[4] A binary operator such as %in% is useful here

1. duplicates were removed?

   ```
   > filtered$filter.log$numDupsRemoved
   [1] 2653
   ```

2. control features were excluded?

   ```
   > filtered$filter.log$feature.exclude
   [1] 19
   ```

3. had low variance (small variation)?

   ```
   > filtered$filter.log$numLowVar
   [1] 3873
   ```

4. had no GO?

   ```
   > filtered$filter.log$numNoGO.BP
   [1] 1528
   ```

5. had no entrez ID?

   ```
   > filtered$filter.log$numRemoved.ENTREZID
   [1] 679
   ```