BioC Introduction

Chao-Jen Wong

Fred Hutchinson Cancer Research Center

November 23, 2009

1. Getting Acquainted with Bioconductor

2. The ALL Dataset and ExpressionSet

3. BioC Introduction

4. Summary

5. Exercise

Outline

1. Getting Acquainted with Bioconductor
2. The ALL Dataset and ExpressionSet
3. BioC Introduction
4. Summary
5. Exercise

Preparation

- Get acquainted with the Bioconductor website
  - biocView packages
    `http://bioconductor.org/download`
  - Getting help: mailing list
    `http://www.bioconductor.org/docs/mailList.html`
  - Searchable mailing list
    `http://dir.gmane.org/gmane.science.biology.informatics.conductor`
- Easy approach to install packages

```
source("http://bioconductor.org/biocLite.R")
biocLite()
biocLite(pkgs)
```

Outline

1. Getting Acquainted with Bioconductor
2. The ALL Dataset and ExpressionSet
3. BioC Introduction
4. Summary
5. Exercise

The ALL ExpressionSet

Code

```
> 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 la
  st seen
    (21 total)
featureData
  featureNames: 1000_at, 1001_at, ..., A
  FFX-YEL024w/RIP1_at  (12625 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
  pubMedIds: 14684422 16243790
Annotation: hgu95av2
```

ExpressionSet

Structure for genomic data

- assayData: Expression data from microarray experiments.

```
> exprs(ALL)
```

- metadata: phenoData, featureData, annotation – A description of the samples and features in experiment.

```
> phenoData(ALL)
> sampleNames(ALL)
> featureData(ALL)
> head(featureNames(ALL))
> annotation(ALL)
```

- experimentData: A flexible structure to describe expeirment.

```
> experimentData(ALL)
> abstract(ALL)
```

- protocoldata: Equipment-generated variables describing sample phenotypes.

Some operations on ExpressionSet

Code

```
> class(ALL)

[1] "ExpressionSet"
attr(,"package")
[1] "Biobase"

> dim(ALL)

Features  Samples
   12625      128

> exprs(ALL)[1:3, 1:3]

             01005    01010    03002
1000_at   7.597323 7.479445 7.567593
1001_at   5.046194 4.932537 4.799294
1002_f_at 3.900466 4.208155 3.886169

> names(pData(ALL))
> varMetadata(ALL)[1:5,,drop=FALSE]
> colnames(exprs(ALL))
> table(ALL$BT) # dollar-sign returns phenodata selection
```

Some operations on ExpressionSet

Exercise

1. Get familiar with the generic functions to access the phenotypical data and meta-data associated with `ALL`.
2. Use pData, varLabels and VarMatedata to extract details of phenotype information of `ALL`.
3. Try to find covariates carrying the information of the molecular biology and cell types (B- and T-cells) of the `ALL` samples.

Data subsetting

Select samples originating from B-cell tumors (BT covariate) found to carry out BCR/ABL mutation and NEG with no cytogenetic abnormalities (mol.biol covariate).

Code: sebsetting

```
> bcell <- grep("^B", as.character(ALL$BT))
> types <- c("NEG", "BCR/ABL")
> moltyp <- which(as.character(ALL$mol.biol) %in% types)
> ALL_bcrneg <- ALL[, intersect(bcell, moltyp)]
```

Code: reshaping

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

Nonspecific filtering

`nsFilter` – filter out probe sets for a number of different criteria.

code: nsFilter

```
> library("genefilter")
> library("hgu95av2.db")
> #openVignette("genefilter")
> filt_bcrneg <- nsFilter(ALL_bcrneg,
+                      require.entrez=TRUE,
+                      require.GOBP=TRUE,
+                      remove.dupEntrez=TRUE,
+                      feature.exclude="^AFFX",
+                      var.cutoff=0.5)
> ALLfilt_bcrneg <- filt_bcrneg$eset
> dim(ALLfilt_bcrneg)

Features  Samples
    3842       79
```

Outline

1. Getting Acquainted with Bioconductor
2. The ALL Dataset and ExpressionSet
3. **BioC Introduction**
4. Summary
5. Exercise

Finding help in R

- ? foo gets the manual page of function `foo`.
- class ? foo gets manual page of class `foo`.
- `help.start(foo)` gets html manual page of object `foo`.
- openVignette() provides interface for opening vignettes. Note that this function is in namespace of package *Biobase*.
- apropos finds objects in the search path that partially match the given character string.
- sessionInfo() prints version information of R and loaded packages.
- search() gives a list of attached packages in current working R session.

Finding help in R

Exercise:

1. There are many number of different plotting functions available. Can you find them?
2. Try to find function to use to perform a MannWhitney test.
3. Open the PDF version of the vignette "Bioconductor Overview" which is part of the *Biobase* package. Use either `biocLite()` or `install.packages()`.
4. What is the output of function `sessionInfo()`?
5. Try to install the *xtable* packages.

Annotation mapping

hgu95av2.db: mappings between Affymetrix IDs and various forms of biological annotation.

```
> hgu95av2()
> ls("package:hgu95av2.db")
```

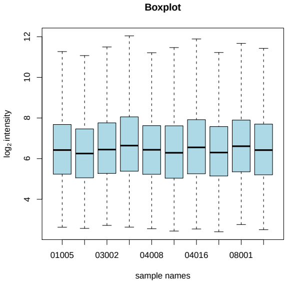
Code: mapping

```
> hgu95av2SYMBOL$"1001_at"

[1] "TIE1"

> mget("1001_at", hgu95av2SYMBOL)

$`1001_at`
[1] "TIE1"

> rmap <- revmap(hgu95av2SYMBOL) ## reverse mapping
> get("TIE1", rmap)

[1] "1001_at"
```

Graphics

Code: visualizing expression patterns

```
> apropos("plot")
> x <- exprs(ALLfilt_bcrneg)[, 1]
> y <- exprs(ALLfilt_bcrneg)[, 2]
> plot(x=x, y=y)
> smoothScatter(x=x, y=y)
> boxplot(exprs(ALLfilt_bcrneg)[, 1:10])
```



Outline

1. Getting Acquainted with Bioconductor
2. The ALL Dataset and ExpressionSet
3. BioC Introduction
4. **Summary**
5. Exercise

Summary

- Logistics of access.
  - Install packages using biocLite().
  - Load packages into the session using library().
- ExpressionSet objects.
  - Fundamental facilities: exprs(), $.
  - Others: phenoData(), featureData(), varLabels(), annotation().
- Annotation mapping and remapping.
  - Fundamental facilities: get(), mget(), and revmap().
  - Annotation packages for certain platform *platfrom*.db.
- Visualization: boxplot(), heatmap().
- Session information: sessionInfo().

Outline

1. Getting Acquainted with Bioconductor
2. The ALL Dataset and ExpressionSet
3. BioC Introduction
4. Summary
5. **Exercise**

Exercise

- hgu95avMAP *environment* contains the mappings between affymetrix identifiers and chromosome band locations.
- apply family of functions: apply(), sapply(), lapply(), and eapply().

eapply

1. Find the chromosome band to which the probe 1001_at maps.
2. Find all genes that map to the p arm of chromosome 17 (17p) using functions grep and eapply.