An introduction to Bioconductor

Martin Morgan
Fred Hutchinson Cancer Research Center
Seatle, WA, USA

March 31, 2007

Objectives

Provide a (superficial) appreciation for. . .

1. *Breadth* of computational opportunities already available in Bioconductor
2. *Strengths* of R and Bioconductor as analytic tools
3. *How to harness* Bioconductor to maximize effectiveness

Core technological focus

- High throughput 'expression' arrays
  - Affymetrix and other single-channel arrays, two-channel arrays, ...
  - Also: tiling arrays, exon arrays, SNPs, ...
- Analysis
  - Preporcessing (background correction, normalization, ...)
  - Visualization and interrogation
  - Statistical models (linear models, classification, ...)

Input and pre-processing

- Example packages: *affy* (Affymetrix), *marray* (two-channel), *vsn* (pre-processing), *affyQCReport* (quality control)
- Example: read `.cel` files

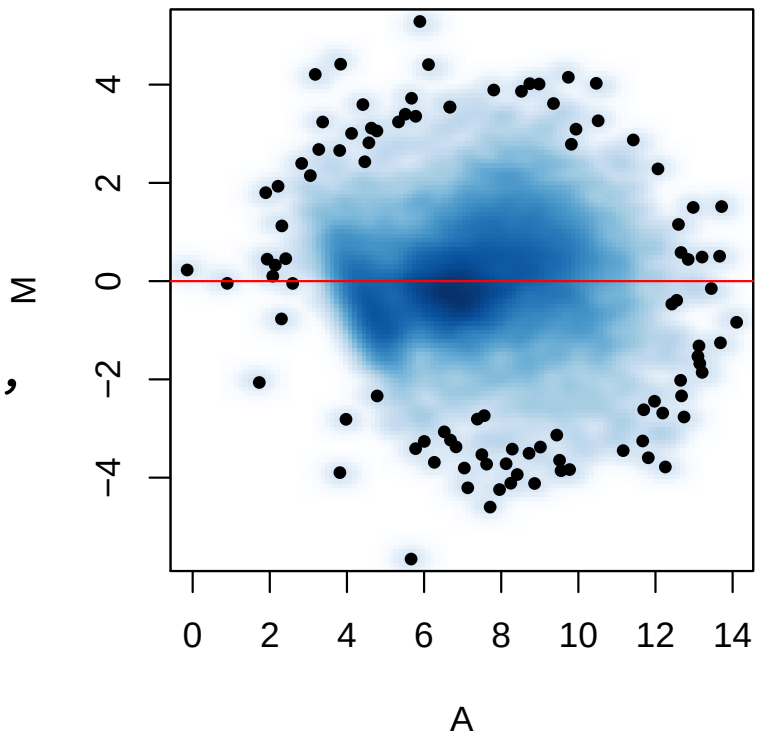
```
> library(affy)
> library(affydata)
> celPath <- system.file("celfiles",
+     package = "affydata")
> affyBatch <- ReadAffy(celfile.path = celPath)
```

- Example: background correction, between-array normalization, transformation

```
> library(vsn)
> data(lymphoma)
> vsnData <- justvsn(lymphoma)
```

Visualization

```
> library(geneplotter)
> exprVals <- exprs(vsnData)
> green <- exprVals[, 7]
> red <- exprVals[, 8]
> M <- green - red
> A <- (green + red)/2
> smoothScatter(A, M, pch = 20,
+     xlab = "A", ylab = "M")
> abline(h = 0, col = "red")
```



Interogation: probes ⇒ genes ⇒ pathways

```
> library(annotate)
> library(hgu95av2)
> library(GO)
> data(sample.ExpressionSet)
> obj <-
+     sample.ExpressionSet
> annotation(obj)
[1] "hgu95av2"
> print(feature <-
+     featureNames(obj)[2])
[1] "AFFX-MurIL10_at"
```

```
> ontologies <-
+     hgu95av2GO[[ feature ]]
> length(ontologies)
[1] 1
> ontologies[[1]]
[1] NA
```

Interogation: *biomaRt*

```
> library(biomaRt)
> ensembl <-
+     useMart("ensembl", dataset="hsapiens_gene_ensembl")
> feature <- featureNames(obj)[100]
> gene <- getGene(id=feature,
+                 type="affy_hg_u95av2", mart=ensembl)
> names(gene)[1:4]

[1] "affy_hg_u95av2"  "hgnc_symbol"
[3] "description"     "chromosome_name"

> strwrap(gene$description, width=40)

[1] "protease inhibitor 15 preproprotein"
[2] "[Source:RefSeq_peptide;Acc:NP_056970]"
```

Analysis: linear models

- `M`: *ExpressionSet*; `exptlDesign`: *data.frame*

```
> library(limma)
> X <- model.matrix(~SFN * HGF, exptlDesign)
> fit <- lmFit(M, X)
> moderated <- eBayes(fit)

> summary(decideTests(moderated))
```

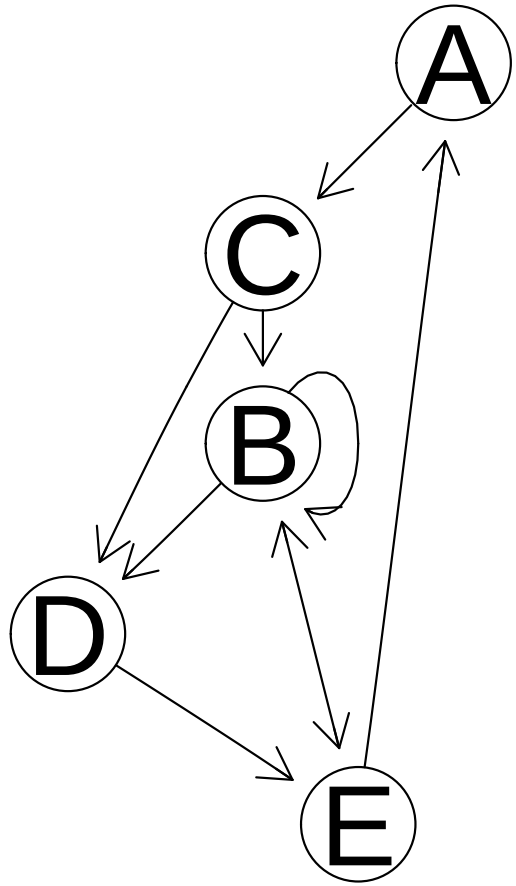
	(Intercept)	SFNHigh	HGFHigh	SFNHigh:HGFHigh
-1	2004	1	0	0
0	2778	7385	7386	7386
1	2604	0	0	0

Analysis: *many* other avenues

- Probe-level descriptions, e.g., *affyPLM*
- Differential expression, multiple comparison, experimental design
- Clustering and classification, e.g., *MLInterfaces*
- Pathways, gene ontologies, e.g., *GOstats*
- Gene set enrichment
- ...

Sophisticated resources

```
> library(RBGL)
> library(Rgraphviz)
> filePath <-
+   system.file("XML/dijkex.gxl",
+               package = "RBGL")
> dijk <- fromGXL(file(filePath))
> plot(dijk)
```



All of R

- $>$ 200 Bioconductor & $>$ 1000 R packages.
- Classical and cutting-edge statistical analysis
- Visualization
- Performance and interoperatbility
  - C and Fortran interface
  - Established non-R libraries (e.g., BOOST, curl, . . .)
- Packages to organize functionality
- Literate programming (e.g., *Sweave* and *weaver*)

Objectives revisited I

1. Very broad range of analyses:
   - Data import and pre-processing
   - Visualization; coordinated meta-data
   - Familiar analyses applied to new data
   - Novel analytic methods

Objectives revisited II

2. Strengths of Bioconductor include:
   - Documentation through help pages and vignettes.
   - Flexibility: user selection of appropriate analyses, with R scripts representing a natural way to coordinate analyses.
   - Access to diverse resources, e.g, *biomaRt* (internet data bases), *RBGL* (BOOST C++ library).
   - Research statistics: new packages added very frequently, representing state-of-the-art analytic methods produced by domain experts

Objectives revisited III

3. Harnessing Bioconductor:
   - *Packages* as the basic unit of software integration.
   - The *R programming langauge* for effectively accessing existing statistical and graphical faciltiies.
   - *Foriegn language interface* for fast computaton and access to existing solutions.
   - *Objects* providing for standardized communication between packages.