Microarray Analysis with R/Bioconductor

Jiangwen Zhang, Ph.D.
Outline

- Overview of R and Bioconductor
  - Installation, updating and self learning resources
- Basic commands in R
- Using Bioconductor for various steps of microarray analysis (both 1 & 2 channels)
A quick overview

- **R & Bioconductor**
  - [http://www.r-project.org](http://www.r-project.org)
  - A statistical environment which implements a dialect of the S language that was developed at AT&T Bell lab
  - Open source, cross platform
  - Mainly command-line driven
  - Current release: 2.9
  - [http://www.bioconductor.org](http://www.bioconductor.org)
  - Open source software packages written in R for bioinformatics application
  - Mainly for microarray analysis at the moment
  - Current release: 2.4
Advantages

- Cross platform
  - Linux, windows and MacOS
- Comprehensive and centralized
  - Analyzes both Affymetrix and two color spotted microarrays, and covers various stages of data analysis in a single environment
- Cutting edge analysis methods
  - New methods/functions can easily be incorporated and implemented
- Quality check of data analysis methods
  - Algorithms and methods have undergone evaluation by statisticians and computer scientists before launch. And in many cases there are also literature references
- Good documentations
  - Comprehensive manuals, documentations, course materials, course notes and discussion group are available
- A good chance to learn statistics and programming
Limitations

- Not easy to learn
  - Require a substantial effort to learn statistics and programming skills before one can do a meaningful data analysis
- Not intuitive
  - Mainly command-line based analysis
  - There are limited wrapper functions and GUIs for certain basic functions
How to download, install and update the basic packages

R
- R website -> download (CRAN) -> select mirror -> select system

Install and update packages by R-GUI
- Menu bar -> Packages
- The installation of a standard Bioconductor suite:
  >source("http://www.bioconductor.org/getBioC.R")
download, install R and bioconductor packages

Visit:

http://cran.r-project.org/
click windows ->base to install R-2.7.1-win32.exe
launch R
copy and paste after prompt  ">

source("http://www.bioconductor.org/getBioC.R")
getBioC("limma")
getBioC("affy")
getBioC("hgu95av2")
getBioC("estrogen")
getBioC("hgu95av2cdf")
getBioC("simpleaffy")
getBioC("annotate")

getBioC("XML")
library(affy)
library(limma)
library(simpleaffy)
An overview of 5 important types of learning resources

1. **Web materials**
   - R website -> documentation
   - Bioconductor website -> documentation

2. **Books**
   - The best one to start with:
     - Peter Dalgaard “Introductory Statistics with R”

3. **Mailing lists**
   - R website -> Project -> Mailing lists
   - Bioconductor website -> Project -> Mailing lists
   - Searchable Bioconductor mailing lists
     - Bioconductor website -> Project -> searching mail archives

4. **Help pages**
   - Start browser help page
     - > help.start()
   - Command line
     - > help(“ls”)
     - > ?ls
Create a separate sub-directory, say work, to hold data files on which you will use R for this problem. This will be the working directory whenever you use R for this particular problem.

To start

Click shortcut of R for window system
Unix: bash$ R to start

> getwd()

Create one subdirectory “estrogen” in that root directory
File-> Change Dir...

Load library, affy
> library(affy)

to quit
R> q()
Getting Help

0 Details about a specific command whose name you know (input arguments, options, algorithm):

> ? t.test
> help(t.test)
> help.start() to launch html web page, then use search engine link
• Simple manipulations numbers and vectors
• Factors, Arrays and matrices
• Lists and data frames
• Reading data from files
• Probability distributions
• Loops and conditional execution
• Writing your own functions
• Statistical models in R
• Graphics
• Packages
R as a Calculator

> log2(32)
[1] 5

> print(sqrt(2))
[1] 1.414214

> pi
[1] 3.141593

> seq(0, 5, length=6)
[1] 0 1 2 3 4 5

> 1+1:10
[1] 2 3 4 5 6 7 8 9 10 11
R as a Graphics Tool

\[
> \text{plot}(\sin(\text{seq}(0, 2*\pi, \text{length}=100)))
\]
Variables

```r
> a <- 49
> sqrt(a)
[1] 7
> b <- "The dog ate my homework"
> sub("dog","cat",b)
[1] "The cat ate my homework"
> c <- (1+1==3)
> c
[1] FALSE
> as.character(b)
[1] "FALSE"
```
Vectors and assignment

**vector:** an ordered collection of data of the same type

```r
> a <- c(1, 2, 3) # In most contexts the = operator can be used as an alternative for <-
> a*2
[1]  2  4  6
```

```r
>x<-c(2, 4, 3, 67, 98, 100, 45, 23, 87, 29, 50)
>sum((x-mean(x))^2)/(length(x)-1)
```

In R, a single number is the special case of a vector with 1 element.

Other vector types: character strings, logical
Matrices and Arrays

matrix: rectangular table of data of the same type

Example: the expression values for 10000 genes for 30 tissue biopsies is a numeric matrix with 10000 rows and 30 columns.

array:
Suppose, for example, z is a vector of 1500 elements. The assignment
> dim(z) <- c(3,5,100)

matrices or more generally arrays are multi-dimensional generalizations of vectors.
Lists

list: ordered collection of data of arbitrary types. lists are a general form of vector in which the various elements need not be of the same type, and are often themselves vectors or lists.

Example:

```r
> doe <- list(name="john",age=28,married=F)
> doe$name
[1] "john"
> doe$age
[1] 28
> doe[[3]]
[1] FALSE
```

Typically, vector elements are accessed by their index (an integer) and list elements by $name (a character string). But both types support both access methods. Slots are accessed by @name.
Data Frames

data frame: rectangular table with rows and columns; data within each column has the same type (e.g. number, text, logical), but different columns may have different types. Represents the typical data table that researchers come up with – like a spreadsheet.

Example:

```r
> a <- data.frame
  (localization, tumorsize, progress, row.names=patients)

> a
   localization  tumorsize progress
XX348     proximal       6.3    FALSE
XX234       distal       8.0     TRUE
XX987     proximal      10.0    FALSE
```
Reading data from files

Large data objects will usually be read as values from external files rather than entered during an R session at the keyboard.

```
> HousePrice <- read.table("houses.data", header=TRUE)
```

The data frame may then be read as
```
> HousePrice <- read.table("houses.data", header=TRUE)
```
where the `header=TRUE` option specifies that the first line is a line of headings, and hence,
Importing and Exporting Data

There are many ways to get data in and out.

Most programs (e.g. Excel), as well as humans, know how to deal with rectangular tables in the form of tab-delimited text files.

```r
> x <- read.delim("filename.txt")

Also: read.table, read.csv, scan

> write.table(x, file="x.txt", sep="\t")

Also: write.matrix, write
```
## Subsetting

Individual elements of a vector, matrix, array or data frame are accessed with “[ ]” by specifying their index, or their name.

```r
> a

    localization tumorsize progress
XX348        proximal  6.3        0
XX234        distal    8.0        1
XX987        proximal 10.0        0

> a[3, 2]
[1] 10

> a["XX987", "tumorsize"]
[1] 10

> a["XX987",]

    localization tumorsize progress
XX987        proximal  10        0
```
Loops

When the same or similar tasks need to be performed multiple times; for all elements of a list; for all columns of an array; etc.

```r
for(i in 1:10) {
  print(i*i)
}
```

```r
i<-1
while(i<=10) {
  print(i*i)
  i<-i+sqrt(i)
}
```

Also: repeat, break, next
Functions do things with data
“Input”: function arguments (0,1,2,...)
“Output”: function result (exactly one)

Example:
add <- function(a,b) {
  result <- a+b
  return(result)
}
## Statistical functions

<table>
<thead>
<tr>
<th>Function(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>rnorm, dnorm, pnorm, qnorm</td>
<td>Normal distribution random sample, density, cdf and quantiles</td>
</tr>
<tr>
<td>lm, glm, anova</td>
<td>Model fitting</td>
</tr>
<tr>
<td>loess, lowess</td>
<td>Smooth curve fitting</td>
</tr>
<tr>
<td>sample</td>
<td>Resampling (bootstrap, permutation)</td>
</tr>
<tr>
<td>.Random.seed</td>
<td>Random number generation</td>
</tr>
<tr>
<td>mean, median</td>
<td>Location statistics</td>
</tr>
<tr>
<td>var, cor, cov, mad, range</td>
<td>Scale statistics</td>
</tr>
<tr>
<td>svd, qr, chol, eigen</td>
<td>Linear algebra</td>
</tr>
<tr>
<td>Distribution</td>
<td>R name</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------</td>
</tr>
<tr>
<td>beta</td>
<td>beta</td>
</tr>
<tr>
<td>binomial</td>
<td>binom</td>
</tr>
<tr>
<td>Cauchy</td>
<td>cauchy</td>
</tr>
<tr>
<td>chi-squared</td>
<td>chisq</td>
</tr>
<tr>
<td>exponential</td>
<td>exp</td>
</tr>
<tr>
<td>F</td>
<td>f</td>
</tr>
<tr>
<td>gamma</td>
<td>gamma</td>
</tr>
<tr>
<td>geometric</td>
<td>geom</td>
</tr>
<tr>
<td>hypergeometric</td>
<td>hyper</td>
</tr>
<tr>
<td>log-normal</td>
<td>Inorm</td>
</tr>
<tr>
<td>logistic</td>
<td>logis</td>
</tr>
<tr>
<td>negative binomial</td>
<td>nbinom</td>
</tr>
<tr>
<td>normal</td>
<td>norm</td>
</tr>
<tr>
<td>Poisson</td>
<td>pois</td>
</tr>
<tr>
<td>Student's t</td>
<td>t</td>
</tr>
<tr>
<td>uniform</td>
<td>unif</td>
</tr>
<tr>
<td>Weibull</td>
<td>weibull</td>
</tr>
<tr>
<td>Wilcoxon</td>
<td>wilcox</td>
</tr>
</tbody>
</table>
Hypergeometric distribution:

<table>
<thead>
<tr>
<th>drawn</th>
<th>not drawn</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>In_pathway</td>
<td>k</td>
<td>m - k</td>
</tr>
<tr>
<td>out_pathway</td>
<td>n - k</td>
<td>N + k - n</td>
</tr>
<tr>
<td>Total</td>
<td>n</td>
<td>N - n</td>
</tr>
</tbody>
</table>

\[ P(X = k) = \binom{m}{k} \binom{N-m}{n-k} \binom{N}{n} \]

m <- 10;
N <-17;
b<-N-m
n<- 8
x <- 0:n

rbind(phyper(x, m, b, n), dhyper(x, m, b, n))
Bioconductor & Microarray

**Affymetrix**: exon, expression, chip-chip, SNP

**Agilent**:  
**Illumina**: beadarray  
**Nimblegen**:  
**Solexa**: sequencing based chip-seq

- microRNA detection  
- Comparative Genomic Hybridization (CGH)  
  - detects deletions or amplifications of genomic sequence  
- ChIP on chip  
  - chromatin immunoprecipitation  
- Single Nucleotide Polymorphism screening (SNP)  
  - measures an individual’s genotype at known sites of variance  

**Resequencing**: chip-seq  
**Cell Arrays**  
**Protein Arrays**  
**Tissue Arrays**

- Flow-cytometry  
- Mass data
Challenges in Genomics

• Diverse biological data types: Genotype, Copy Number, Transcription, Methylation,…
• Diverse technologies to measure the above: Affymetrix, Nimblegen,…
• Integration of multiple and diverse data structures
• Large datasets

Functional genomics, gene regulation network, signaling pathway, motif identification
Aims of Bioconductor

- Provide access to powerful statistical and graphical methods for the analysis of genomic data.
- Facilitate the integration of biological metadata (GenBank, GO, LocusLink, PubMed) in the analysis of experimental data.
- Allow the rapid development of extensible, interoperable, and scalable software.
- Promote high-quality documentation and reproducible research.
- Provide training in computational and statistical methods.
Install, start and update bioconductor

- Find getBioC()
  - Bioc website -> How To -> getBioC
  > *source("http://www.bioconductor.org/getBioC.R")*
  > getBioC.R
  - to install extra packages not in standard suite
    - Menu -> packages -> install packages from bioconductor

- Start Bioconductor
  > library(Biobase)

- Update Bioconductor packages
  - Menu -> packages -> update packages from bioconductor
Microarray Data analysis workflow

- Image analysis
- Raw data
- Diagnostic plots
- Normalization
- Filtering
- Estimate missing values
- Differential gene inference
  - Linear modeling and factorial experiment
- Clustering
- Classification

- Annotation
- GO analysis
- Pathway analysis
Microarray data analysis

Pre-processing

CEL, CDF

- affy
- vsn

.gpr, .spot,

- marray
- limma
- vsn

exprSet

Differential expression

- siggenes
- genefilter
- limma
- multtest

Graphs & networks

- graph
- RBGL
- Rgraphviz

Cluster analysis

- CRAN
- class
- cluster
- MASS
- mva

Prediction

- CRAN
- class
- e1071
- ipred
- LogitBoost
- MASS
- nnet
- randomForest
- rpart

Annotation

- annotate
- annaffy
- + metadata packages

Graphics

- geneplotter
- hexbin
- + CRAN
## Useful R/BioC Packages

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marray, limma</td>
<td>Spotted cDNA array analysis</td>
</tr>
<tr>
<td>affy</td>
<td>Affymetrix array analysis</td>
</tr>
<tr>
<td>vsn</td>
<td>Variance stabilization</td>
</tr>
<tr>
<td>annotate</td>
<td>Link microarray data to metadata on the web</td>
</tr>
<tr>
<td>ctest</td>
<td>Statistical tests</td>
</tr>
<tr>
<td>genefilter, limma,</td>
<td>Gene filtering (e.g.: differential expression)</td>
</tr>
<tr>
<td>multtest, siggenes</td>
<td></td>
</tr>
<tr>
<td>mva, cluster, clust</td>
<td>Clustering</td>
</tr>
<tr>
<td>class, rpart, nnet</td>
<td>Classification</td>
</tr>
</tbody>
</table>
What’s Your Question?

What are the targets genes for my knock-out gene?
Gene discovery, differential expression

Is a specified group of genes (genes from a pathway) all up-regulated in a specified condition?
Gene set enrichment analysis

Can I use the expression profile of cancer patients to predict chemotherapy outcome?
Class prediction, classification

Pathways/network affected?
Kegg, Biocarta
Considering Pathway/network Topology
- Pre-processing microarray data
diagnostic, normalization
- Differential Gene Expression
identification of up and down regulated genes
- Annotation and metadata
get the DE genes’ id, pathway involvement, GO
- Distances, Prediction, and Cluster Analysis
sample similarity calculation and visualization by heatmap
- Class prediction
provide expression profile of type-known samples to computer, train it, and
let computer to classify type-unknown samples
Pre-processing packages

- **marray, limma**: Spotted DNA microarrays.
- **affy**: Affymetrix oligonucleotide chips.
- **vsn**: Variance stabilization for both types of arrays.

- Reading in intensity data, diagnostic plots, normalization, computation of expression measures.

- The packages start with very different data structures, but produce similar objects of class **exprSet**.

- One can then use other Bioconductor and CRAN packages, e.g., **mva, genefilter, geneplotter**.
LIMMA and LIMMA GUI

- LIMMA is another library to perform basic 2 channels analysis, linear modeling for both single channel and 2 channel
- There is a nice GUI for LIMMA
- library(limma)
- library(limmaGUI)
Affymetrix chips

- **DAT** file: Image file, \(~10^7\) pixels, \(~50\) MB.
- **CEL** file: Cell intensity file, probe level PM and MM values.
- **CDF** (Chip Description File): Describes which probes belong to which probe-pair set and the location of the probes.
GeneChip® Eukaryotic Target Labeling Assays for Expression Analysis

One-Cycle Target Labeling
(for 1-15 μg total RNA or 0.2-2 μg mRNA)

- Total RNA Sample: 5' AAAAA 3'
- Poly-A RNA Control addition
- 1st strand cDNA synthesis: 5' TTTTTT 3'
- 2nd strand cDNA synthesis: 5' AAAAA 3'

Two-Cycle Target Labeling
(for 10-100 ng total RNA)

- Total RNA Sample: 5' AAAAA 3'
- Poly-A RNA Control addition
- 1st cycle, 1st strand cDNA synthesis: 5' T7-Oligo(dT) Primer 3' TTTTT 5'
- 1st cycle, 2nd strand cDNA synthesis: 5' AAAAA 3' TTTTT 5'
- 1st cycle, In vitro transcription 3' Un-labeled Ribonucleotides 5'
- 1st cycle, Cleanup of antisense RNA (cRNA): 5' UUUUU 3' 1.5 hours
- 2nd cycle, 1st strand cDNA synthesis: 5' AAAAA 3' UUUUU 5'
- 2nd cycle, 2nd strand cDNA synthesis: 5' T7-Oligo(dT) Primer 3' TTTTT 5' 2.5 hours

Approximate Experiment Time:
- 1.5 hours
- 2 hours
- Overnight
- 0.5 hours
- 1.5 hours
- 2.5 hours

Products Used:
- Poly-A RNA Control Kit
- Two-Cycle cDNA Synthesis Kit
- One-Cycle cDNA Synthesis Kit
- Sample Cleanup Module

FAS-IT RESEARCH COMPUTING GROUP
Harvard University
affy package

CEL and CDF files \rightarrow Class AffyBatch

\[
\begin{align*}
\text{rma or mas5} \\
\text{expresso} \\
\text{express}
\end{align*}
\]

\rightarrow Class exprSet

Save data to file using \texttt{write.exprs} or continue analysis using other Bioconductor and CRAN packages.
**affy and simpleaffy package**

- Class definitions for probe-level data: **AffyBatch, ProbSet, Cdf, Cel.**
- Basic methods for manipulating microarray objects: printing, plotting, subsetting.
- Functions and widgets for data input from **CEL** and **CDF** files, and automatic generation of microarray data objects.
- Diagnostic plots: 2D spatial images, density plots, boxplots, MA-plots.
  - image: 2D spatial color images of log intensities (AffyBatch, Cel).
  - boxplot: boxplots of log intensities (AffyBatch).
  - mva.pairs: scatter-plots with fitted curves (apply exprs, pm, or mm to AffyBatch object).
  - hist: density plots of log intensities (AffyBatch).
- Check RNA degradation and couple control metrics defined by affymetrix company

```r
library(simpleaffy)
Data.qc<-qc(data)
avbg(Data.qc)
sfs(Data.qc)
percent.present(Data.qc)
ratios(Data.qc[,1:2])
```
Quality control (2)

QC metrics

- 1. Average background
- 2. Scale factor
- 3. Number of genes called present
- 4. 3′ to 5′ ratios of actin and GAPDH
- 5. Uses ordered probes in all probeset to detect possible RNA degradation.
Lab

>library(affy)
>library("simpleaffy")
data <- ReadAffy()
>qc(data) -> data.qc
>avbg(data.qc) # comparable bg expected
>sfs(data.qc) # within 3folds of each other expected
>percent.present(data.qc) # extremely low value is a problem
>ratios(data.qc)
> AffyRNAdeg(data) -> RNAdeg
>plotAffyRNAded(RNAdeg)
>summaryAffyRNAded(RNAdeg)
QC Stats

Scale factor for this chip falls outside the 3-fold region therefore all scale factor lines are coloured red.

High beta-actin 3':5' ratio: coloured red because it is greater than 3.

High GAPDH 3':5' ratio: coloured red because it is greater than 1.25.

Blue region represents the spread where all scale factors fall within 3 fold of the mean scale factor fall all chips.

Chip dividers

Acceptable beta-actin 3':5' ratio: coloured blue because it is less than 3.

Acceptable GAPDH 3':5' ratio: coloured blue because it is less than 1.25.
Affymetrix spike-in controls, ploy-A control and hybridization control

- The ploy-A controls AFFX-r2-Bs-Dap, AFFX-r2-Bs-Thr, AFFX-r2-Bs-Phe and AFFX-r2-Bs-Lys (morespikes slot) are modified B. subtilis genes and should be called present at a decreasing intensity, to verify that there was no bias during the retro-transcription between highly expressed genes and low expressed genes. Note that the linearity for lys, phe and thr (dap is present at a much higher concentration) is affected by a double amplification.

- hybridization control BioB, BioC, BioD in increasing concentration.
Normalization by affy package

- Background estimation.
- Probe-level normalization: quantile and curve-fitting normalization (Bolstad et al., 2003).
- Expression measures: MAS 4.0 AvDiff, MAS 5.0 Signal, MBEI (Li & Wong, 2001), RMA (Irizarry et al., 2003).
- Main functions: **ReadAffy, rma, mas5, expresso, express**.

Pre-processing oligonucleotide chip data:
- diagnostic plots,
- background correction,
- probe-level normalization,
- computation of expression measures.
Low level analysis

Normalization

- The main goal is to remove the systematic bias in the data as completely as possible, while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription.

- A basic assumption of most normalization procedures is that the average gene expression level does not change in an experiment.

- Normalization is different in spotted/two-color compared with high-density-oligonucleotides (Affy) technology.
Low level analysis

- Bioconductor affy R package

<table>
<thead>
<tr>
<th>BG correction</th>
<th>Normalization</th>
<th>PM correction</th>
<th>Expression index</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>None</td>
<td>none</td>
<td>Avg Diff (MAS 4)</td>
</tr>
<tr>
<td>MAS</td>
<td>Linear constant</td>
<td>Subtract MM (MAS 4)</td>
<td>Signal (MAS 5)</td>
</tr>
<tr>
<td>RMA</td>
<td>Contrasts</td>
<td>Subtract IM (MAS 5)</td>
<td>Li.Wong model</td>
</tr>
<tr>
<td>RMA2</td>
<td>Invariant set</td>
<td>PM only</td>
<td>RMA (median polish)</td>
</tr>
<tr>
<td>gcrma</td>
<td>Loess</td>
<td></td>
<td>GCRMA</td>
</tr>
<tr>
<td></td>
<td>Cyclic loess</td>
<td></td>
<td>Playerout</td>
</tr>
<tr>
<td></td>
<td>Cubic spline</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quantiles (robust)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VSN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Error model of Rosetta reso
RMA, MAS5, dCHIP

Folds change consistency comparison using 1.25 µg of RNA vs using 20 µg of RNA
Code for affy analysis

- For Affymetrix data
- Data loading
  > library(affy)
  > ReadAffy() -> affybatch
- Some diagnostic plots
  > hist(affybatch)
  > image(affybatch)
  > boxplot(affybatch)
  > RNAdeg <- AffyRNAdeg(affybatch)
  > plotAffyRNAdeg(RNAdeg(affybatch))
  > summary(RNAdeg)
  > mva.pairs(affybatch)

The sequence of affy probe level analysis
-> background correction -> normalization -> pm correction -> expression summarization
-> rma(affybatch)
-> Mas(affybatch)
hist(Dilution, col=1:4, type="l")
> AffyRNADEg(AffyBatch)
> plotAffyRNADEg()
Differential Gene Expression
Typical questions

- Detect genes are differentially expressed between two or more samples. [t-test, F-test and many others]
- Identification of groups of genes with characterizing a particular class of tumors. [Discrimination]
- Discover at molecular level, potential sub-classes of tumors / disease. [Clustering]
- Detection of gene regulatory mechanisms. [Network and meta data]
Significance Inference

1. Calculation of a statistic based on replicate array data for ranking genes according to their possibilities of differential expression

2. Selection of a cut-off value for rejecting the null-hypothesis that the gene is not differentially expressed

- Underlying distribution
- Sample size
- Parametric test
- Nonparametric test
- Permutation test
- Bayesian mixture model
Significance Inference

- Student’s T-test \( t_i = (M_i/SE_i) \)
  where \( SE_i \) (standard error of \( M_i \)) = \( s_i/\sqrt{n} \)

- Introduces some conservative protection against outlier M-values and poor quality spots

Limitations

- **large t-statistic can be driven by an unrealistically small value for s**
- Suffers from multiple hypothesis testing problem
- Suffers from inflated type I error

- Moderated t-statistic from limma package

- Uses empirical Bayes to estimate a posterior variance of the gene with the information borrow from all genes

Ebayes: borrow information from ensemble of genes, a good strategy for small sample project, superior to t-test.

has more functions than just linear modeling, it basically provides another ways for prepossessing, normalizing and plotting data as marray packages
Significance Inference

- Fold change as a threshold cut-off is inadequate, even if it is an average of replicates

- Limitations:
  - Fixed threshold does not account for statistical significance
  - Does not take into account of the variability of the expression levels for each gene
  - Genes with larger variances have a good chance of giving a large fold change even if they are not differentially expressed
  - Inflated type I & type II errors
Statistical filtering

- Ultimately, what matters is biological relevance.

- P-values should help you evaluate the strength of the evidence, rather than being used as an absolute yardstick of significance.

- Statistical significance is not necessarily the same as biological significance.
Filtering before DE study: library (genefilter)

Two filters: gene should be above “100” for 5 times and have a Cox-PH-model p-value < 0.01

\[ kF \leftarrow kOverA(5, 100) \]

Assemble them in a filtering function

\[ ff \leftarrow filterfun(kF, cF) \]

Apply the filter

\[ sel \leftarrow genefilter(exprs(DATA), ff) \]

Select the relevant subset of the data

\[ mySub \leftarrow DATA[sel,] \]
Annotation and metadata
Biological metadata

- Biological attributes that can be applied to the experimental data.
- E.g. for genes
  - chromosomal location;
  - gene annotation (LocusLink, GO);
  - relevant literature (PubMed).
- Biological metadata sets are large, evolving rapidly, and typically distributed via the WWW.
- Tools: **annotate**, **annaffy**, and **AnnBuilder** packages, and annotation data packages.
annotate, annafy, and AnnBuilder

Metadata package hgu95av2

- Assemble and process genomic annotation data from public repositories.
- Build annotation data packages or XML data documents.
- Associate experimental data in real time to biological metadata from web databases such as GenBank, GO, KEGG, LocusLink, and PubMed.
- Process and store query results: e.g., search PubMed abstracts.
- Generate HTML reports of analyses.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENENAME</td>
<td>zinc finger protein 261</td>
</tr>
<tr>
<td>LOCUSID</td>
<td>9203</td>
</tr>
<tr>
<td>ACCNUM</td>
<td>X95808</td>
</tr>
<tr>
<td>MAP</td>
<td>Xq13.1</td>
</tr>
<tr>
<td>SYMBOL</td>
<td>ZNF261</td>
</tr>
<tr>
<td>PMID</td>
<td>10486218</td>
</tr>
<tr>
<td></td>
<td>9205841</td>
</tr>
<tr>
<td></td>
<td>8817323</td>
</tr>
<tr>
<td>GO</td>
<td>GO:0003677</td>
</tr>
<tr>
<td></td>
<td>GO:0007275</td>
</tr>
<tr>
<td></td>
<td>GO:0016021</td>
</tr>
<tr>
<td></td>
<td>+ many other mappings</td>
</tr>
</tbody>
</table>

mappings between different gene identifiers for hgu95av2 chip.
Distances, Prediction, and Cluster Analysis
clustering
1a. Find the genes whose expression fits specific, predefined patterns.
1b. Find the genes whose expression follows the pattern of predefined gene or set of genes.
2. Carry out some kind of exploratory analysis to see what expression patterns emerge; cluster analysis (usually on genes).

tumor classification:
1. The identification of new/unknown tumor classes using gene expression profiles; cluster analysis (usually on samples).
2. The classification of malignancies into known classes; Discrimination.
3. The identification of “marker” genes that characterize the different tumor classes; variable selection.
Cluster analysis for microarray data

- **Hierarchical method**
  - Bottom-up method
    - Single linkage
    - Complete linkage
    - Average linkage
  - Top-down method
    - TSVQ clustering
    - Macnaughton-Smith clustering

- **Partitioning method**
  - k-means clustering
  - k-medoids clustering
  - Self-organizing maps (SOMs)
Hierarchical method

Hierarchical clustering methods produce a tree or dendrogram. The tree can be built in two distinct ways:
- **bottom-up**: agglomerative clustering;
- **top-down**: divisive clustering.

**Merging:**
- Computationally simple
- Precise at bottom of tree
- Good for many small clusters

**Divisive**
- More complex, but more precise at the top of the tree
- Good for looking at large and/or few clusters

**For Gene expression applications, divisive makes more sense.**

Applying Euclidean distance to categorical data is invalid.
Correlation metric applied to highly skewed data will give misleading results.
Partition methods

- Partition the data into a prespecified number $k$ of mutually exclusive and exhaustive groups.
- Iteratively reallocate the observations to clusters until some criterion is met, e.g. minimize within cluster sums of squares.

**Examples:**
- $k$-means, self-organizing maps (SOM), PAM, etc.;
- Fuzzy: needs stochastic model, e.g. Gaussian mixtures.
Partitioning vs. hierarchical

Partitioning:

Advantages
- Optimal for certain criteria.

Disadvantages
- Need initial \( k \);
- Often require long computation times.

Hierarchical

Advantages
- Faster computation.

Disadvantages
- Rigid;
- Cannot correct later for erroneous decisions made earlier.
Hierarchical clustering

Average linkage

Complete linkage

Cut tree at given height yields
7 clusters

15 clusters
Visualization

- Principal components analysis (PCA), an exploratory technique that reduces data dimensionality to 2 or 3 dimensional space.

- For a matrix of m genes x n samples, create a new covariance matrix of size n x n

- Thus transform some large number of variables into a smaller number of uncorrelated variables called principal components (PCs).
Visualization

PCA Mapped Data of spr s2 (83.7%)
Distances

- Microarray data analysis often involves
  - clustering genes and/or samples;
  - classifying genes and/or samples.
- Both types of analyses are based on a measure of distance (or similarity) between genes or samples.
- R has a number of functions for computing and plotting distance and similarity matrices.
Distances

Distance functions
- **dist** (*mva*): Euclidean, Manhattan, Canberra, binary;
- **daisy** (*cluster*).

Correlation functions
- **cor**, **cov.wt**.

Plotting functions
- **image**;
- **plotcorr** (*ellipse*);
- **plot.cor**, **plot.mat** (*sma*).
R cluster analysis packages

- **cclust**: convex clustering methods.
- **class**: self-organizing maps (SOM).
- **cluster**:
  - AGglomerative NESting (**agnes**),
  - Clustering LARe Applications (**clara**),
  - DIsisive ANalysis (**diana**),
  - Fuzzy Analysis (**fanny**),
  - MONothetic Analysis (**mona**),
  - Partitioning Around Medoids (**pam**).
- **e1071**:
  - fuzzy C-means clustering (**cmeans**),
  - bagged clustering (**bclust**).
- **flexmix**: flexible mixture modeling.
- **fpc**: fixed point clusters, clusterwise regression and discriminant plots.
- **GeneSOM**: self-organizing maps.
- **mclust, mclust98**: model-based cluster analysis.
- **mva**:
  - hierarchical clustering (**hclust**),
  - k-means (**kmeans**).

Specialized summary, plot, and print methods for clustering results.

Download from CRAN
Heatmaps

Golub et al. ALL AML dataset, random 50 genes

Heatmap function from mva package
Class prediction

- Old and extensive literature on class prediction, in statistics and machine learning.
- Examples of classifiers
  - nearest neighbor classifiers (k-NN);
  - discriminant analysis: linear, quadratic, logistic;
  - neural networks;
  - classification trees;
  - support vector machines.
- Aggregated classifiers: bagging and boosting
**R class prediction packages**

- **class**:  
  - k-nearest neighbor (**knn**),  
  - learning vector quantization (**lvq**).
- **classPP**: projection pursuit.  
- **e1071**: support vector machines (**svm**).  
- **ipred**: bagging, resampling based estimation of prediction error.  
- **knnTree**: k-nn classification with variable selection inside leaves of a tree.  
- **LogitBoost**: boosting for tree stumps.  
- **MASS**: linear and quadratic discriminant analysis (**lda**, **qda**).  
- **mlbench**: machine learning benchmark problems.  
- **nnet**: feed-forward neural networks and multinomial log-linear models.  
- **pamR**: prediction analysis for microarrays.  
- **randomForest**: random forests.  
- **rpart**: classification and regression trees.  
- **sma**: diagonal linear and quadratic discriminant analysis, naïve Bayes (**stat.diag.da**).
Lab

- Estrogen dataset at Bioconductor
  

  8 arrays (duplicate): ER+ breast cancer cells were exposed to estrogen 10 hours, 48 hours.

  E10, E10, C10, C10, E48, E48, C48, C48

- Create working directory
  - Move .cel files to the directory

- Create experiment description file
Brief introduction:

Data gives results from a 2x2 factorial experiment on MCF7 breast cancer cells using Affymetrix HGU95av2 arrays. The factors in this experiment were estrogen (present or absent) and length of exposure (10 or 48 hours). The aim of the study is to identify genes which respond to estrogen and to classify these into early and late responders. Genes which respond early are putative direct-target genes while those which respond late are probably downstream targets in the molecular pathway.

This experiment studied the effect of estrogen on the gene expression in estrogen receptor positive breast cancer cells over time. After serum starvation, samples were exposed to estrogen, and mRNA was harvested at two time points (10 or 48 hours). The control samples were not exposed to estrogen and were harvested at the same time points. Table 1 shows the experimental design, and corresponding samples names. The full data set (12,625 probes, 32 samples) and its analysis are discussed in Scholtens, et al. Analyzing Factorial Designed Microarray Experiments.
Visit:
http://cran.r-project.org/

click windows ->base to install R-2.7.1-win32.exe

launch R

copy and paste after prompt ">

source("http://www.bioconductor.org/getBioC.R")

getBioC("limma")

getBioC("marray")

getBioC("affy")

getBioC("hgu95av2")

getBioC("estrogen")

getBioC("hgu95av2cdf")

getBioC("simpleaffy")

library(affy)

library(limma)

library(simpleaffy)
Table 1: Experimental Conditions for .cel Files

<table>
<thead>
<tr>
<th>time</th>
<th>estrogen</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>absent</td>
<td>present</td>
</tr>
<tr>
<td>10 hours</td>
<td>et1</td>
<td>Et1</td>
</tr>
<tr>
<td></td>
<td>et2</td>
<td>Et2</td>
</tr>
<tr>
<td>48 hours</td>
<td>eT1</td>
<td>ET1</td>
</tr>
<tr>
<td></td>
<td>eT2</td>
<td>ET2</td>
</tr>
<tr>
<td>filename</td>
<td>estrogen</td>
<td>time.h</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>low10-1.cel</td>
<td>absent</td>
<td>10</td>
</tr>
<tr>
<td>low10-2.cel</td>
<td>absent</td>
<td>10</td>
</tr>
<tr>
<td>high10-1.cel</td>
<td>present</td>
<td>10</td>
</tr>
<tr>
<td>high10-2.cel</td>
<td>present</td>
<td>10</td>
</tr>
<tr>
<td>low48-1.cel</td>
<td>absent</td>
<td>48</td>
</tr>
<tr>
<td>low48-2.cel</td>
<td>absent</td>
<td>48</td>
</tr>
<tr>
<td>high48-1.cel</td>
<td>present</td>
<td>48</td>
</tr>
<tr>
<td>high48-2.cel</td>
<td>present</td>
<td>48</td>
</tr>
</tbody>
</table>
gene i in sample j (j = 1, ..., 8).

\[ y_{ij} = \mu_i + \beta_{ES} x_{ES} + \beta_{TIME} x_{TIME} + \beta_{ES:TIME} x_{ES:TIME} + \epsilon_{ij} \]

\[ y_{ij} = \mu_i + \beta_{TIME} x_{TIME} + \epsilon_i \]

```R
design <- model.matrix(~factor(estrogen)*factor(time.h), pdat)
design
```

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>present</th>
<th>48h</th>
<th>present:48h</th>
</tr>
</thead>
<tbody>
<tr>
<td>low10-1.cel</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>low10-2.cel</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>high10-1.cel</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>high10-2.cel</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>low48-1.cel</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>low48-2.cel</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>high48-1.cel</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>high48-2.cel</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

colnames(design) <- c("Intercept", "ES", "T48", "ES:T48")
fit <- lmFit(esEset, design)
fit$coefficients[1:3,]
```
## Project

### Launch necessary Bioconductor packages

```r
library(estrogen)
library(limma)
library(hgu95av2cdf)
```

### Define data directory

```r
datadir <- system.file(package="estrogen", "extdata") # OR
datadir <- file.path(.find.package("estrogen"), "extdata")
datadir
```
Project

```r
### read in experiment design information and .cel files

```r
targets <- readTargets("phenoData.txt",
path=datadir, sep="", row.names="filename")
targets

library(affy)
esAB <- ReadAffy(filenames=targets$filename, celfile.path=datadir)

### quality check by "simpleaffy" package

```r
library(simpleaffy)
qc(esAB)->qc
plot(qc)
```
## Project

### normalize###

```r
normalize.methods(esAB)
esEset<-rma(esAB)
pairs(exprs(esEset))
boxplot(esAB, col="red")
boxplot(data.frame(exprs(esEset)))# not code: boxplot(exprs(esEset)), this only generate one boxplot
```

### heatmap###

```r
###Select the 50 genes with the highest variation (standard deviation) across chips
rsd <- apply(exprs(esEset), 1, sd)
sel <- order(rsd, decreasing = TRUE)[1:50]
heatmap(exprs(esEset)[sel, ], col = gentlecol(256))
```
library(limma)

design <- model.matrix(~factor(estrogen) * factor(time.h), targets)
design
colnames(design) <- c("Intercept", "ES", "T48", "ES:T48")
fit <- lmFit(esEset, design)

#fit$coefficients[1:3,]
contM <- cbind(es10 = c(0,1,0,0), es48 = c(0,1,0,1))
fitC <- contrasts.fit(fit, contM)
fitC <- eBayes(fitC)

#glist <- topTable(fitC, n=20, coef=2, adjust="fdr")
glist <- topTable(fitC, n=20, coef=1, adjust="fdr")
library(marray)
table2html(glist, filename="estrogen_file.html", disp="file")
getwd()
edit(glist)
Project

```r
### annotation, query PubMed
getBioC("XML")
getBioC("annotate")
library(annotate)
library(XML)
annotation(esAB)# "hgu95av2"
library( "hgu95av2")
absts<-pm.getabst(glist$ID,"hgu95av2") ###pm.getabst query PubMed directly, be careful, too many queries can get you banned

### to look at second gene's pubmed abstract
titl<-sapply(absts[[2]], articleTitle)
sapply(absts[[2]], pubDate)
sapply(absts[[2]], pmid)
sapply(absts[[2]], journal)
sapply(absts[[2]], abstText)
strwrap(titl, simplify=F) #strwrap to format text to fit the page width
## search abstracts with key word
```
Project

sapply(absts[[2]], abstText)->try
grep("estrogen",try)->try1
pmAbst2HTML(absts[[2]][try1],filename="estrogen_pm.html")
#if(!interactive()) file.remove("estrogen_file.html")
ll<-getEG(glist$ID, "hgu95av2")
sym<-getSYMBOL(glist$ID, "hgu95av2")
##htmlpage(genelist, filename, title, othernames, table.head, table.center = TRUE, repository = list("en"), ...)
htmlpage(genelist=list(ll, glist$ID),filename="estrogen_ANN.html", title="estrogen effect", othernames=data.frame(sym, glist[-1]),table.head=c("EntrezID","AffyID", "Symbol", colnames(glist)[-1]), repos=list("en", "affy"))
Now we can start analysing our data for biological effects. We set up a linear model with main effects for the level of estrogen (estrogen) and the time (time.h). Both are factors with 2 levels.

```r
lm.coef = function(y) lm(y ~ factor(targets$estrogen) * factor(targets$time.h))
$coefficients
eff = esApply(esEset, 1, lm.coef)
eff[2,order(abs(eff[2,]), decreasing=T)[1:20]]->top20
topTable(fitC, n=nrow(fitC), coef=1, adjust="fdr")->allLimma
allLimma[match(names(top20),allLimma$ID),]
```
## Project

<table>
<thead>
<tr>
<th>ID</th>
<th>logFC</th>
<th>AveExpr</th>
<th>t</th>
<th>P.Value</th>
<th>adj.P.Val</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>910_at</td>
<td>3.113733</td>
<td>9.660238</td>
<td>23.59225</td>
<td>4.96E-09</td>
<td>3.13E-05</td>
<td>9.942522</td>
</tr>
<tr>
<td>39642_at</td>
<td>2.939428</td>
<td>7.876515</td>
<td>23.71715</td>
<td>4.74E-09</td>
<td>3.13E-05</td>
<td>9.96681</td>
</tr>
<tr>
<td>38827_at</td>
<td>2.9322</td>
<td>6.829773</td>
<td>9.110154</td>
<td>1.15E-05</td>
<td>4.26E-03</td>
<td>3.928051</td>
</tr>
<tr>
<td>31798_at</td>
<td>2.800195</td>
<td>12.11578</td>
<td>16.38509</td>
<td>1.03E-07</td>
<td>3.51E-04</td>
<td>7.97729</td>
</tr>
<tr>
<td>1884_s_at</td>
<td>2.799396</td>
<td>9.034796</td>
<td>12.05054</td>
<td>1.26E-06</td>
<td>1.06E-03</td>
<td>5.949749</td>
</tr>
<tr>
<td>1536_at</td>
<td>2.662258</td>
<td>5.937222</td>
<td>13.26247</td>
<td>5.80E-07</td>
<td>7.32E-04</td>
<td>6.610486</td>
</tr>
<tr>
<td>40117_at</td>
<td>2.555282</td>
<td>9.676557</td>
<td>15.6807</td>
<td>1.47E-07</td>
<td>3.58E-04</td>
<td>7.705093</td>
</tr>
<tr>
<td>1854_at</td>
<td>2.507616</td>
<td>8.532099</td>
<td>15.15848</td>
<td>1.95E-07</td>
<td>3.58E-04</td>
<td>7.490766</td>
</tr>
</tbody>
</table>

910_at 39642_at 38827_at 31798_at 1884_s_at 1536_at 40117_at 1854_at
3.113733 2.939428 2.9322 2.800195 2.799396 2.662258 2.555282 2.507616
<table>
<thead>
<tr>
<th>EntrezID</th>
<th>AffyID</th>
<th>Symbol</th>
<th>logFC</th>
<th>AveExpr</th>
<th>t</th>
<th>P.Value</th>
<th>adj.P.Val</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>7083</td>
<td>910_at</td>
<td>TK1</td>
<td>3.86</td>
<td>9.66</td>
<td>29.21</td>
<td>0.00</td>
<td>0.00</td>
<td>11.61</td>
</tr>
<tr>
<td>7031</td>
<td>31798_at</td>
<td>TFF1</td>
<td>3.60</td>
<td>12.12</td>
<td>21.05</td>
<td>0.00</td>
<td>0.00</td>
<td>9.89</td>
</tr>
<tr>
<td>4605</td>
<td>1854_at</td>
<td>MYBL2</td>
<td>3.34</td>
<td>8.53</td>
<td>20.20</td>
<td>0.00</td>
<td>0.00</td>
<td>9.64</td>
</tr>
<tr>
<td>9768</td>
<td>38116_at</td>
<td>KIAA0101</td>
<td>3.76</td>
<td>9.51</td>
<td>16.86</td>
<td>0.00</td>
<td>0.00</td>
<td>8.48</td>
</tr>
<tr>
<td>3148</td>
<td>38065_at</td>
<td>HMGB2</td>
<td>2.99</td>
<td>9.10</td>
<td>16.21</td>
<td>0.00</td>
<td>0.00</td>
<td>8.21</td>
</tr>
<tr>
<td>7494</td>
<td>39755_at</td>
<td>XBP1</td>
<td>1.77</td>
<td>12.13</td>
<td>15.83</td>
<td>0.00</td>
<td>0.00</td>
<td>8.05</td>
</tr>
<tr>
<td>7153</td>
<td>1592_at</td>
<td>TOP2A</td>
<td>2.30</td>
<td>8.31</td>
<td>15.79</td>
<td>0.00</td>
<td>0.00</td>
<td>8.03</td>
</tr>
<tr>
<td>7083</td>
<td>41400_at</td>
<td>TK1</td>
<td>2.24</td>
<td>10.04</td>
<td>15.29</td>
<td>0.00</td>
<td>0.00</td>
<td>7.81</td>
</tr>
<tr>
<td>9052</td>
<td>33730_at</td>
<td>GPRC5A</td>
<td>-2.04</td>
<td>8.57</td>
<td>-15.14</td>
<td>0.00</td>
<td>0.00</td>
<td>7.74</td>
</tr>
<tr>
<td>11065</td>
<td>1651_at</td>
<td>UBE2C</td>
<td>2.97</td>
<td>10.50</td>
<td>14.78</td>
<td>0.00</td>
<td>0.00</td>
<td>7.57</td>
</tr>
<tr>
<td>991</td>
<td>38414_at</td>
<td>CDC20</td>
<td>2.02</td>
<td>9.46</td>
<td>14.59</td>
<td>0.00</td>
<td>0.00</td>
<td>7.48</td>
</tr>
<tr>
<td>890</td>
<td>1943_at</td>
<td>CCNA2</td>
<td>2.19</td>
<td>7.60</td>
<td>14.00</td>
<td>0.00</td>
<td>0.00</td>
<td>7.18</td>
</tr>
<tr>
<td>4175</td>
<td>40117_at</td>
<td>MCM6</td>
<td>2.28</td>
<td>9.68</td>
<td>13.97</td>
<td>0.00</td>
<td>0.00</td>
<td>7.17</td>
</tr>
<tr>
<td>332</td>
<td>40533_at</td>
<td>BIRC5</td>
<td>1.64</td>
<td>8.47</td>
<td>13.53</td>
<td>0.00</td>
<td>0.00</td>
<td>6.93</td>
</tr>
<tr>
<td>54898</td>
<td>39642_at</td>
<td>ELOVL2</td>
<td>1.61</td>
<td>7.88</td>
<td>13.03</td>
<td>0.00</td>
<td>0.00</td>
<td>6.65</td>
</tr>
<tr>
<td>6790</td>
<td>34851_at</td>
<td>AURKA</td>
<td>1.96</td>
<td>9.96</td>
<td>12.85</td>
<td>0.00</td>
<td>0.00</td>
<td>6.55</td>
</tr>
<tr>
<td>5111</td>
<td>1824_s_at</td>
<td>PCNA</td>
<td>1.64</td>
<td>9.24</td>
<td>12.76</td>
<td>0.00</td>
<td>0.00</td>
<td>6.50</td>
</tr>
<tr>
<td>11130</td>
<td>35995_at</td>
<td>ZWINT</td>
<td>2.76</td>
<td>8.87</td>
<td>12.68</td>
<td>0.00</td>
<td>0.00</td>
<td>6.46</td>
</tr>
<tr>
<td>27338</td>
<td>893_at</td>
<td>UBE2S</td>
<td>1.54</td>
<td>10.95</td>
<td>12.66</td>
<td>0.00</td>
<td>0.00</td>
<td>6.45</td>
</tr>
<tr>
<td>40079_at</td>
<td></td>
<td></td>
<td>-2.41</td>
<td>8.23</td>
<td>-12.63</td>
<td>0.00</td>
<td>0.00</td>
<td>6.42</td>
</tr>
</tbody>
</table>
limma package

- Fitting of gene-wise linear models to estimate log ratios between two or more target samples simultaneously:
- **ebayes**: moderated t-statistics based on empirical Bayes

For example, the log ratios in above figure are each the difference of the log abundance in two of the four samples; we may construct a design matrix which specifies how the log ratios in all ten experiments are derived from the log abundances in the four chips; then the best estimate of the log abundances is obtained by solving the least squares problem for this design matrix.

<table>
<thead>
<tr>
<th>Chip</th>
<th>WTT</th>
<th>WTC</th>
<th>KOT</th>
<th>KOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Matrix Multiplication

\[
\begin{pmatrix}
 y_1 \\
 y_2 \\
 y_3
\end{pmatrix} =
\begin{pmatrix}
 1 & 0 & 1 \\
 -1 & 0 & 1 \\
 1 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
 \beta_1 \\
 \beta_2 \\
 \beta_3
\end{pmatrix}
\]

\[
\beta = B - A
\]

Linear Model Estimates

Obtain a linear model for each gene \( g \)

\[
E(y_g) = X \beta_g \quad \text{var}(y_g) = W_g^{-1} \sigma^2_g
\]

Estimate model by robust regression, least squares or generalized least squares to get

- coefficients \( \hat{\beta}_{gj} \)
- standard deviations \( s_g \)
- standard errors \( \text{se}(\hat{\beta}_{gj})^2 = c_g s_g^2 \)