# Bioconductor 简介及其在生物信息学中的应用

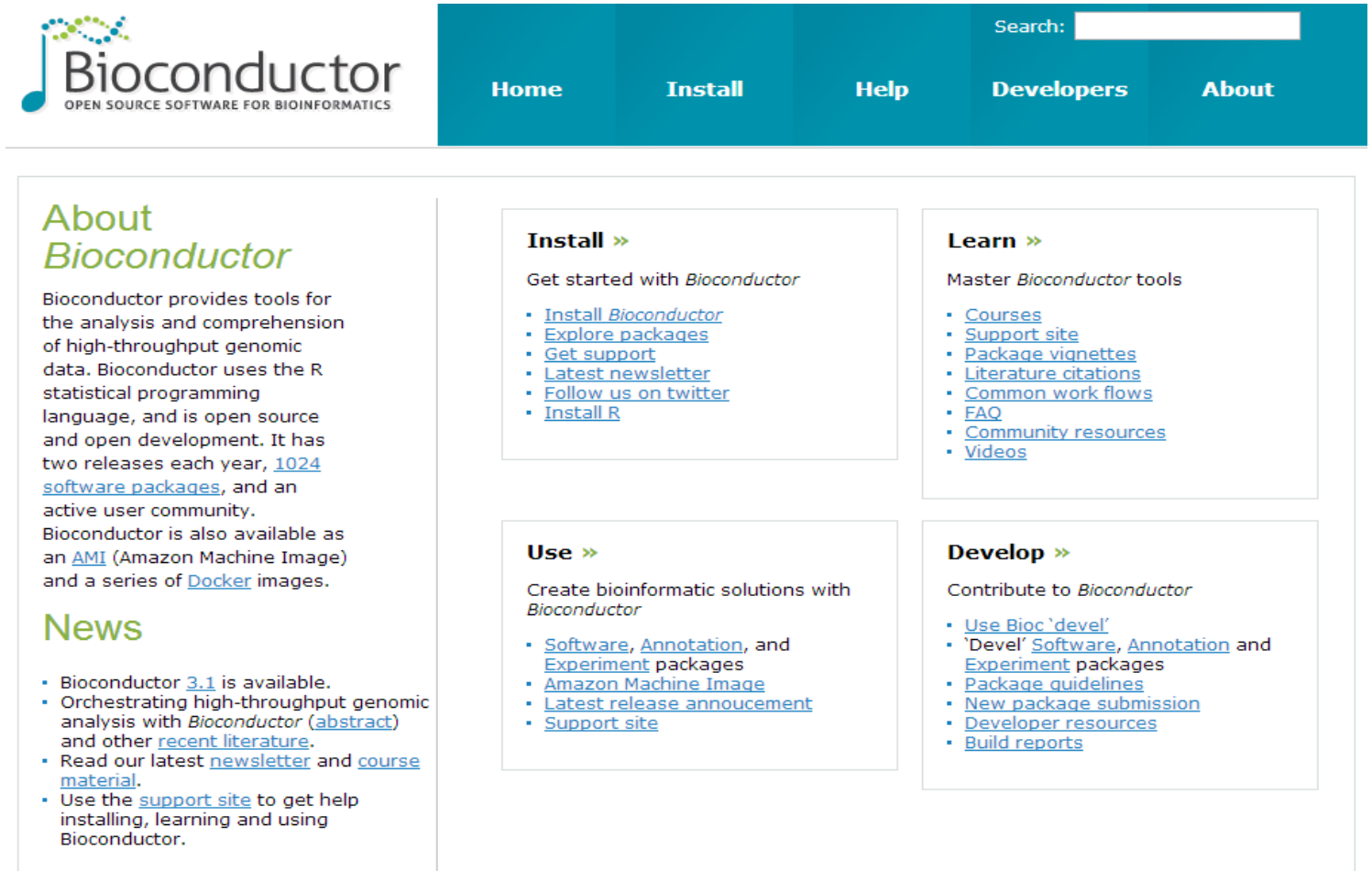
郑广勇
上海生命科学研究院

# 主要内容

- **Bioconductor 软件介绍**
- **Bioconductor 软件应用**
  - 基因芯片分析中的应用

# Bioconductor

Bioconductor 是一个基于R语言的生物信息软件包，主要用于生物数据的注释、分析、统计、以及可视化（http://www.bioconductor.org）



# Bioconductor

软件包的安装

Install the latest release of R, then get the latest version of Bioconductor by starting R and entering the commands

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

Install specific packages, e.g., "GenomicFeatures" and "AnnotationDbi", with

```
biocLite(c("GenomicFeatures", "AnnotationDbi"))
```

# Bioconductor

## (1) Experiment Data ; (2) Software ; (3) Annotation Data

### BioC 3.1: Multiple platform build/check report

*This page was generated on 2015-08-12 09:34:35 -0700 (Wed, 12 Aug 2015).*

***svn info***

Snapshot Date: **2015-08-11 17:20:10 -0700 (Tue, 11 Aug 2015)**
URL: **https://hedgehog.fhcrc.org/bioconductor/branches/RELEASE_3_1/madman/Rpacks**
Last Changed Rev: **107334** / Revision: **107334**
Last Changed Date: **2015-08-11 16:59:13 -0700 (Tue, 11 Aug 2015)**

| *Hostname* | *OS* | *Arch (*)* | *Platform label (**)* | *R version* | *Installed pkgs* |
|---|---|---|---|---|---|
| *zin2* | Linux (Ubuntu 14.04.2 LTS) | x86_64 | x86_64-linux-gnu | 3.2.1 (2015-06-18) -- "World-Famous Astronaut" | 1328 |
| *moscato2* | Windows Server 2008 R2 Enterprise SP1 (64-bit) | x64 | mingw32 / x86_64-w64-mingw32 | 3.2.1 (2015-06-18) -- "World-Famous Astronaut" | 1310 |
| *petty* | Mac OS X Snow Leopard (10.6.8) | x86_64 | i686-apple-darwin10 | 3.2.1 (2015-06-18) -- "World-Famous Astronaut" | 1315 |
| *morelia* | Mac OS X Mavericks (10.9.5) | x86_64 | x86_64-apple-darwin13.4.0 | 3.2.1 (2015-06-18) -- "World-Famous Astronaut" | 1274 |

*Click on any hostname to see more info about the system (e.g. compilers)* *(*) as reported by 'uname -p', except on Windows and Mac OS X Mavericks* *(**) as reported by 'gcc -v'*

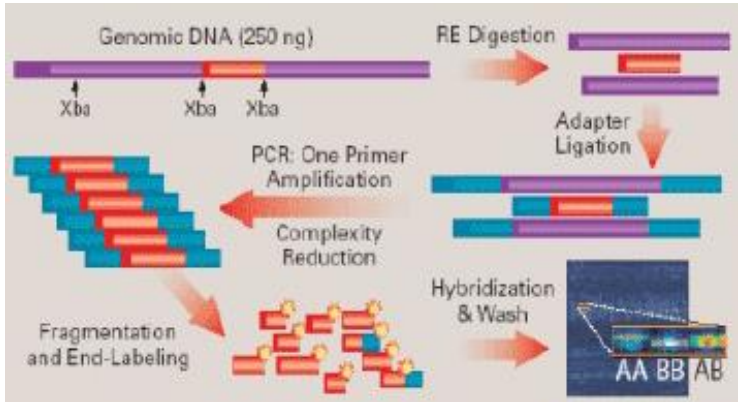
***Package STATUS** - Package status is indicated by one of the following glyphs:*
- TIMEOUT *INSTALL, BUILD, CHECK* or *BUILD BIN* of package took more than 40 minutes
- ERROR *INSTALL, BUILD, CHECK* or *BUILD BIN* of package returned an error
- WARNINGS *CHECK* of package produced warnings
- OK *INSTALL, BUILD, CHECK* or *BUILD BIN* of package was OK
- NotNeeded *INSTALL* of package was not needed (click on glyph to see why)
- **skipped** *CHECK* or *BUILD BIN* of package was skipped because the *BUILD* step failed (or because something bad happened with the Build System itself)

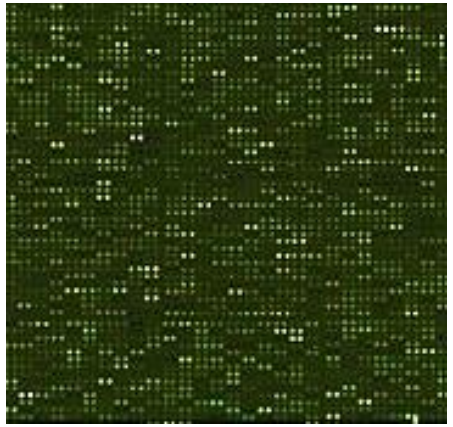
*Click on any glyph in the report below to see the status details (command output).*

***Package propagation STATUS** - Package propagation status is indicated by one of the following glyphs:*
- YES: Package was propagated because it didn't previously exist or version was bumped
- NO: Package was not propagated because of a problem (impossible dependencies, or version lower than what is already propagated)
- UNNEEDED: Package was not propagated because it is already in the repository with this version. A version bump is required in order to propagate it

*Use the check boxes to show only packages with the selected status types:* ☑ TIMEOUT ☑ ERROR ☑ WARNINGS ☑ OK

# 基因芯片实验流程



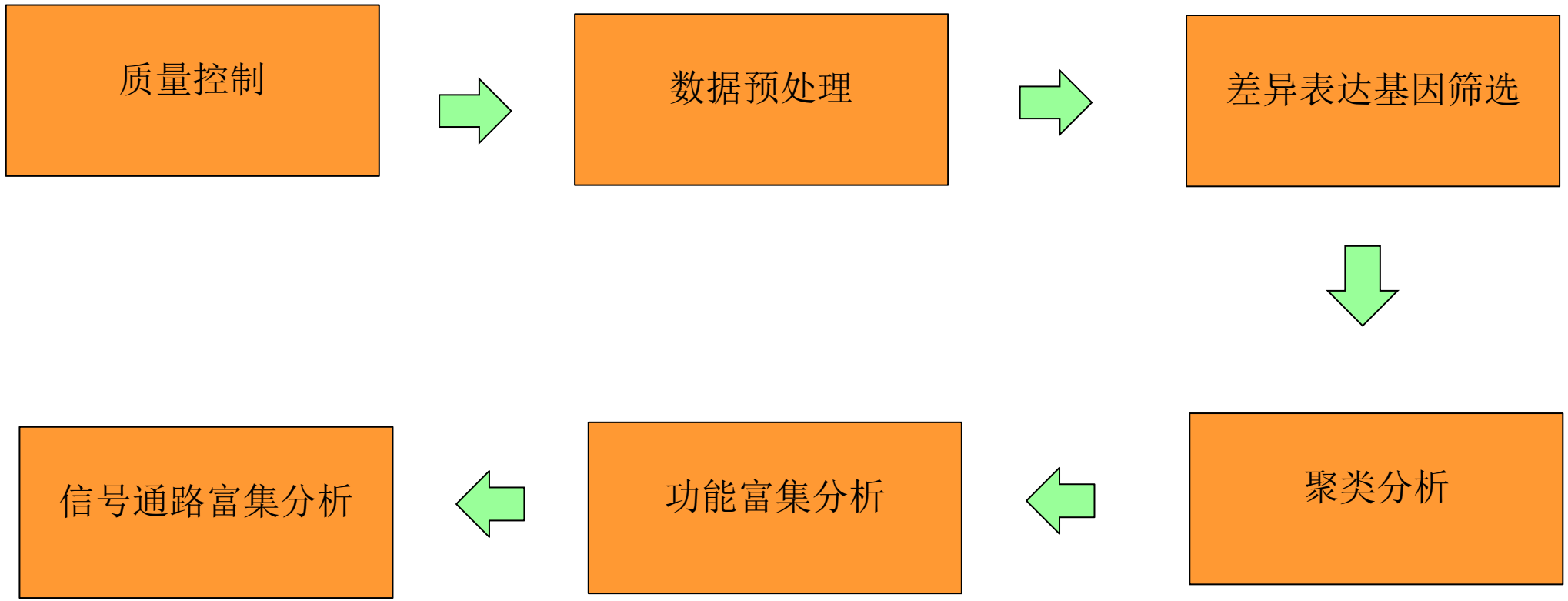



Gene-chip experiment

Picture scan

Raw picture

# 芯片数据分析流程



# 数据预处理

通过数据预处理，过滤掉低质量数据获取表达值数据，主要包括以下几个方面：

- 数据背景处理
- 数据标准化
- 综合表达量计算

# 差异表达基因分析

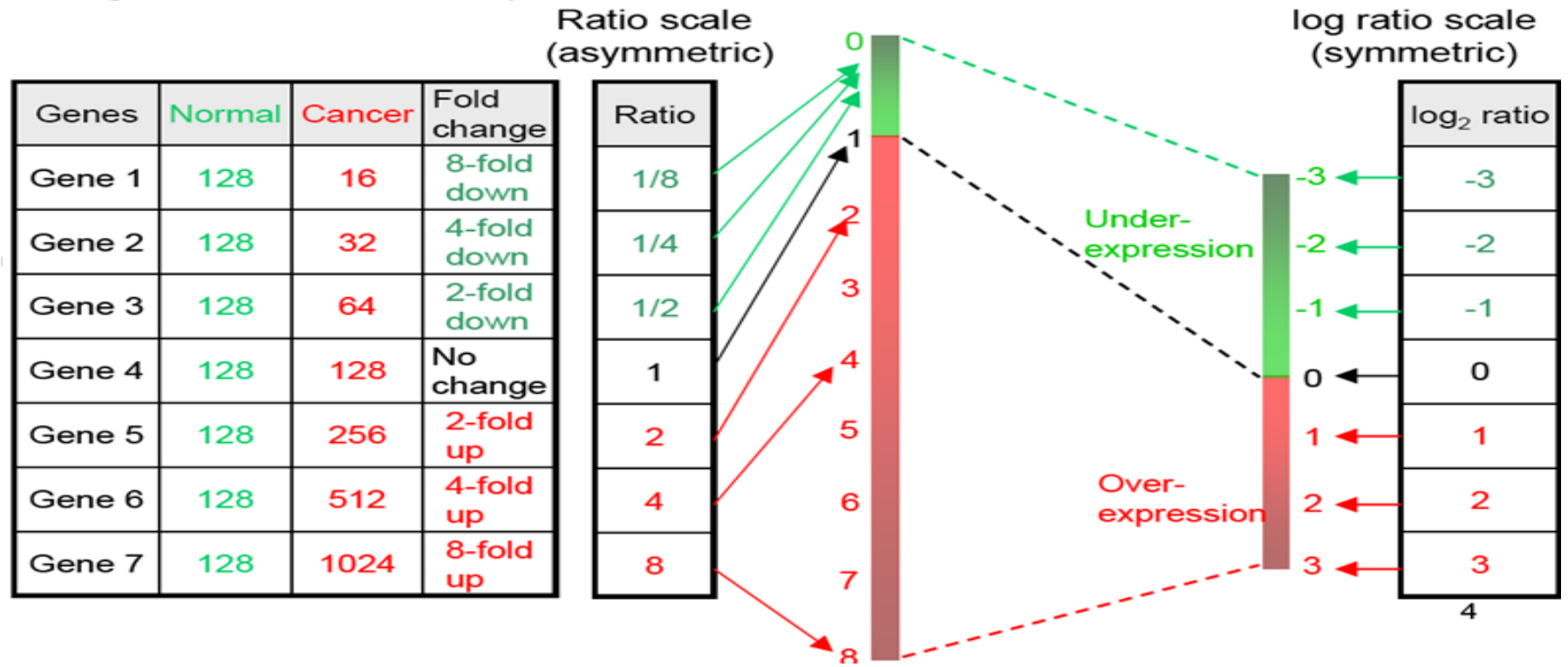
- Fold-change值
- T检验
- 经验贝叶斯（Empirical Bayes）
- Wilcoxon秩和检验
- 回归模型方法

# 差异表达基因筛选方法

- Fold-change

  最简单的判断差异基因的方法，在没有重复试验的条件下很常用。

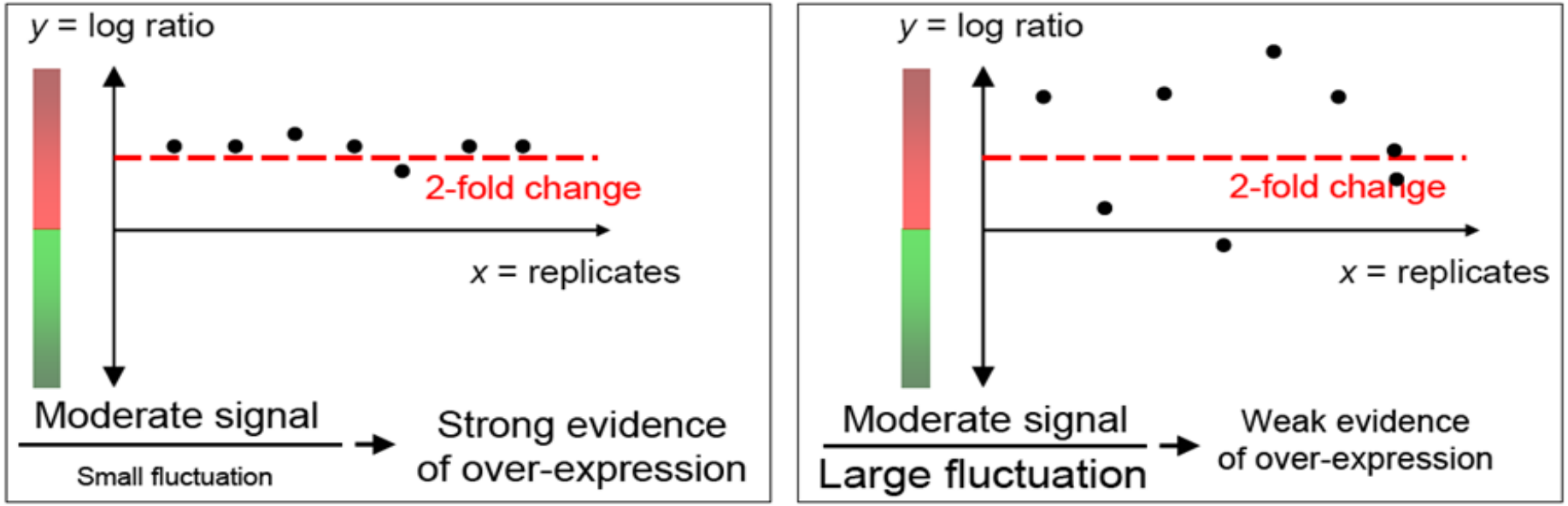
| Genes | Normal | Cancer | Fold change |
|---|---|---|---|
| Gene 1 | 128 | 16 | 8-fold down |
| Gene 2 | 128 | 32 | 4-fold down |
| Gene 3 | 128 | 64 | 2-fold down |
| Gene 4 | 128 | 128 | No change |
| Gene 5 | 128 | 256 | 2-fold up |
| Gene 6 | 128 | 512 | 4-fold up |
| Gene 7 | 128 | 1024 | 8-fold up |



# 差异表达基因筛选方法

- T检验

较常用的统计方法,用于判断某一基因在两个样本中其表达是否有显著性差异,不要求等方差，要求有重复试验



# 差异表达基因筛选方法

- 经验贝叶斯（Empirical Bayes)

T-检验的一种改进方法，将标准差及信号强度的关系使用线性模型进一步强化，提高了准确率，目前比较常用的一种方法

- Wilcoxon秩和检验

是一种非参数的检验方法,该方法要比T-检验更加稳健, 更适合非正态分布的数据

- 线性回归模型

通过线性模型模拟不同实验条件下的基因表达情况， 其给出的回归方程不仅包括筛选差异表达基因部分, 还包括数据的预处理部分

# Bioconductor芯片分析包

- **affy**

  对数据进行表达值计算，质量控制，标准化等

- **simpleaffy**

  对表达数据进行质量控制，T检验，筛选出差异表达基因；

- **affyPLM**

  对芯片数据进行读取，质量控制，标准化；

- **gcRMA**

  对芯片数据进行读取，质量控制，标准化；

- **limma**

  采用回归模型方法进行差异表达基因筛选，读取数据，数据质量控制，标准化，用回归模型的方法筛选差异表达基因等，针对双通道数据比较全面的一套处理步骤；

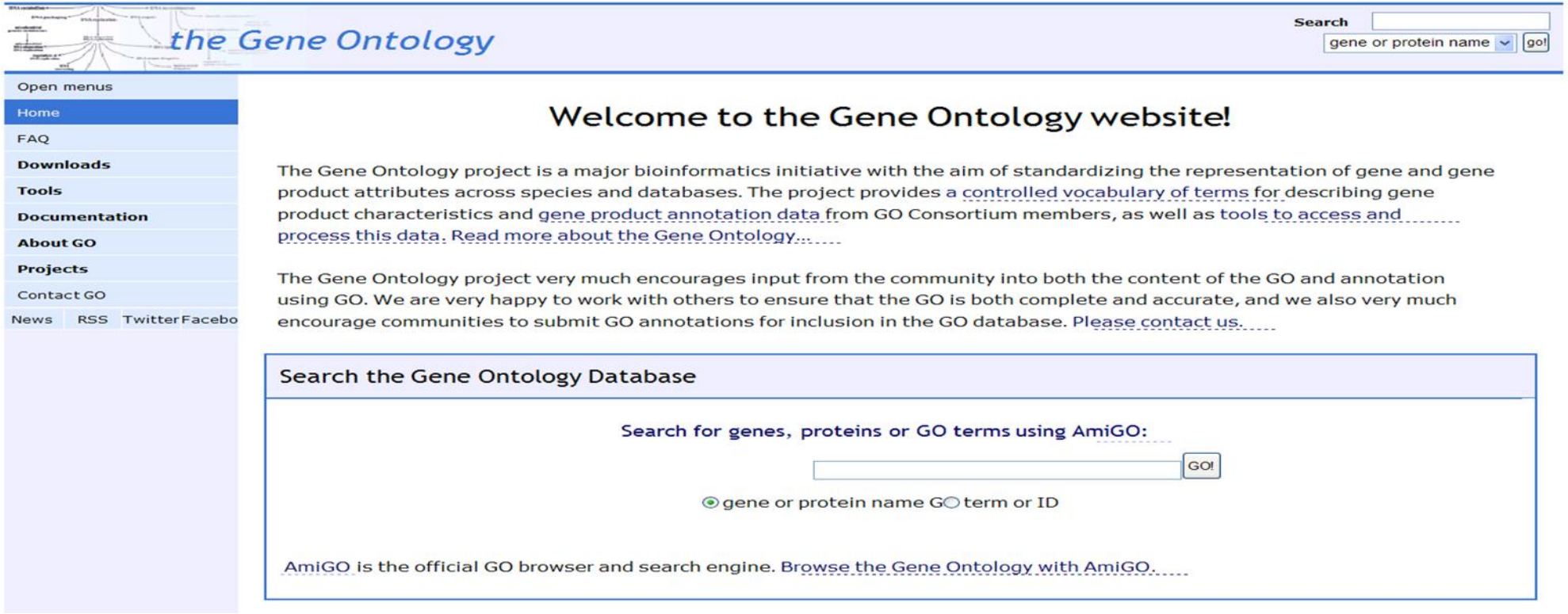
# 表达谱数据聚类分析

在基因表达数据分析中, 根据处理对象与目标的不同，将聚类方法分为三类:

- 基于基因的聚类(Gene-based clustering)
- 基于样本的聚类(Sample-based clustering)
- 双向聚类(Biclustering)

# 基因本体数据库

基因本体数据库（http://www.geneontology.org）是GO组织构建的一个结构化的标准生物学模型，旨在建立基因及其产物知识的标准词汇体系，涵盖了基因的细胞组分（cellular component）、分子功能（molecular function）、生物学过程（biological process）。



# 功能富集分析

- 分类注释

基因及其产物的功能，根据GO数据库的记录数据一一进行标注。并在细胞组分，分子功能，生物学过程中进行分类。

- 富集分析

一组差异表达基因与背景基因库，根据GO功能的注释结果进行对照比较，使用超几何分布等统计学方法，计算出两者差异的显著性，从而找到这组差异表达基因中富集的功能类别条目，从而揭示这组差异表达基因的整体功能特征。

# KEGG 信号通路数据库

KEGG 信号通路数据库（http://www.kegg.jp/kegg/pathway.html ）是京都基因组数据仓库下的一个子数据库，该数据库主要提供了分子相互作用及网络的通路信息，可以帮助人们理解重要的生理生化过程的分子机制。

# 通路富集分析

- 分类注释

  差异表达基因位于哪些通路上或者分类中，这些通路主要涉及那些生理生化过程。

- 富集分析

  通过统计分析，找出差异表达基因集富集在哪些生物学通路中，这些通路主要涉及那些生理生化过程，从而发现这组差异表达基因的在生物体内所有参与的通路信息。

# 芯片数据分析实例

选择HG-U133A平台的6个样本作为差异表达基因筛选实例

数据来源预 GEO 数据库 （id ： GSE21363）
样品:HITC6 细胞系（血管平滑肌细胞）
实验:血清缺乏诱导
目标:细胞形态影响

**R 软件下载**

http://www.r-project.org/

**Bioconductor工具**

```
source(“http://bioconductor.org/biocLite.R”)
options(BioC_mirror="http://mirrors.ustc.edu.cn/bioc/")
biocLite(“GEOquery”)
biocLite(c("affy","simpleaffy","affyPLM","gcRMA","limma","annotate"))
biocLite("hgu133a.db")
```

下载原始数据

```
library(GEOquery)
setwd(“E:/mywork/R/test”)
getGEOSuppFiles(GEO=“GSE21363”,baseDir=getwd())
untar("GSE21363/GSE21363_RAW.tar",exdir="data")
cels <- list.files("data/",pattern="[gz]")
sapply(paste("data",cels,sep="/"),gunzip)
celpath <- paste(getwd(),"data",sep="/")
setwd(celpath)
```

# 芯片数据分析步骤

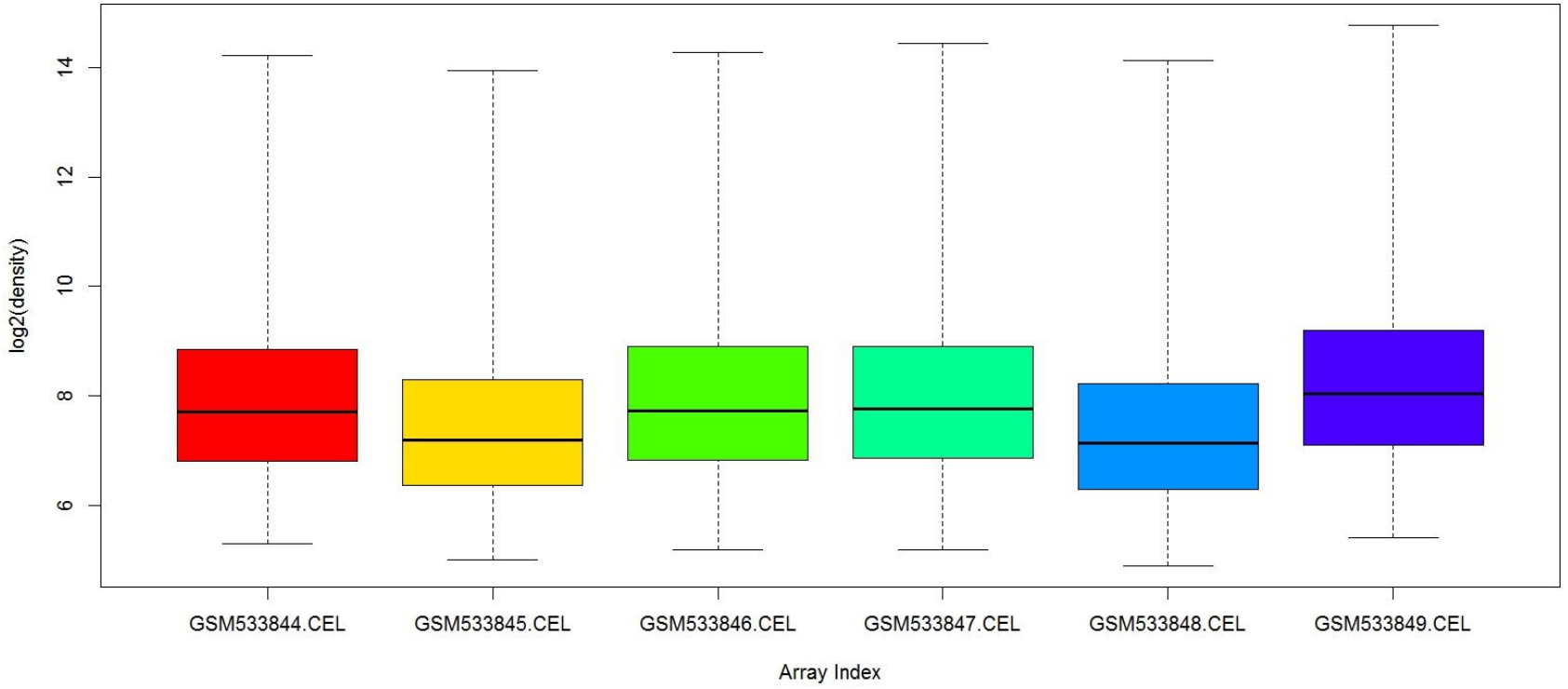
读入数据

```
library(affy)

celfiles <-
    c("GSM533844.CEL","GSM533845.CEL","GSM533846.CEL","
    GSM533847.CEL","GSM533848.CEL","GSM533849.CEL")

raw.data <- ReadAffy(filenames=celfiles)

pData(raw.data)$Treatment <- rep(c("Day0","Day8"),each=3)
```

| data | treatment |
| --- | --- |
| GSM533844.CEL | Day0 |
| GSM533845.CEL | Day0 |
| GSM533846.CEL | Day0 |
| GSM533847.CEL | Day8 |
| GSM533848.CEL | Day8 |
| GSM533849.CEL | Day8 |

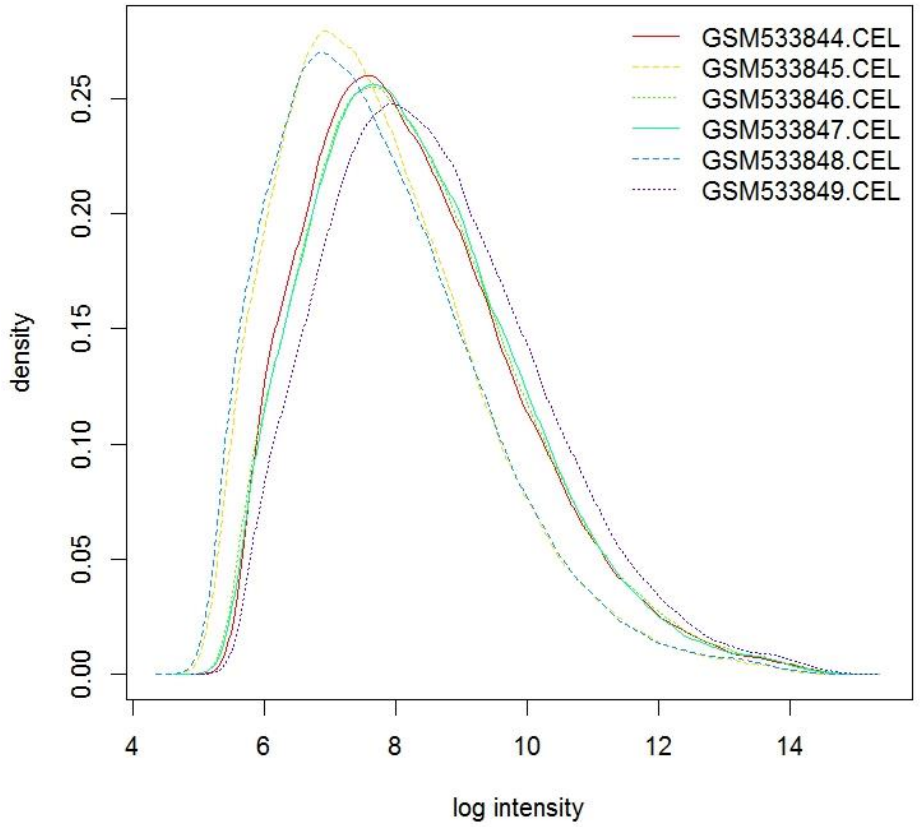
# 质量控制(1)

```
n.cel  <- length(celfiles)
cols <- rainbow(n.cel *1.2)
boxplot(raw.data, col=cols, xlab="Array Index", ylab="log2(density)")
```



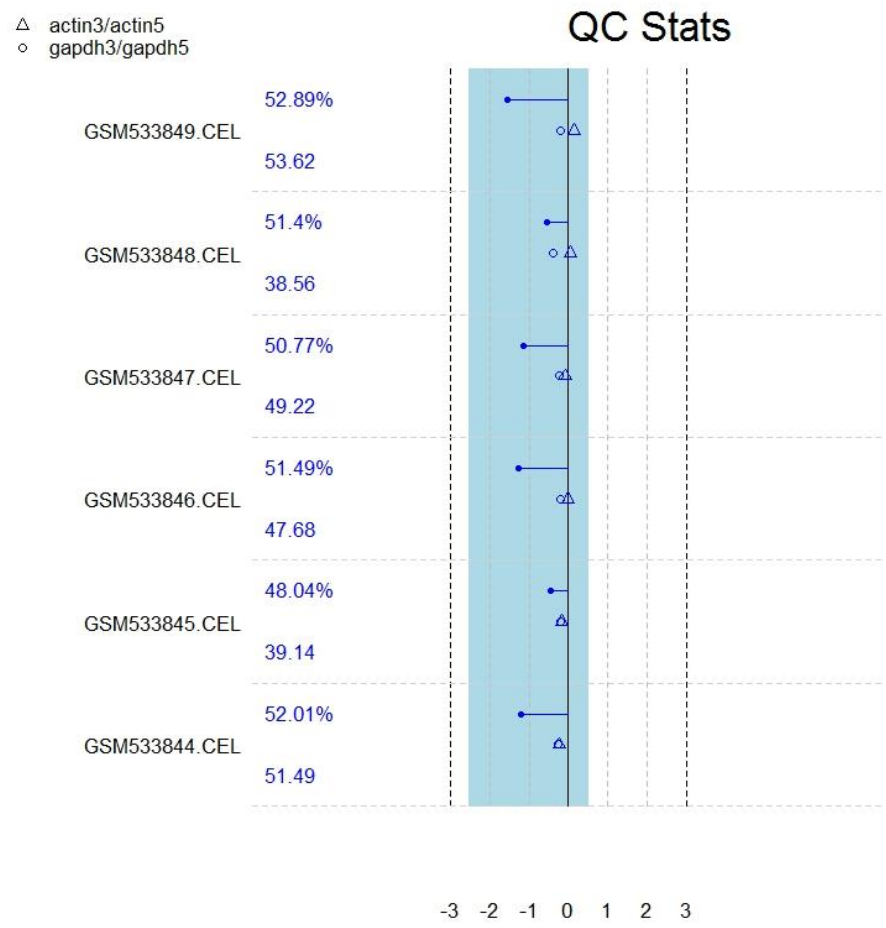
# 质量控制(2)

```
hist(raw.data, lty=1:3, col=cols)

legend("topright",legend=sampleNames(raw
.data),lty=1:3,col=cols,box.col="transparent"
,xpd=T)
```



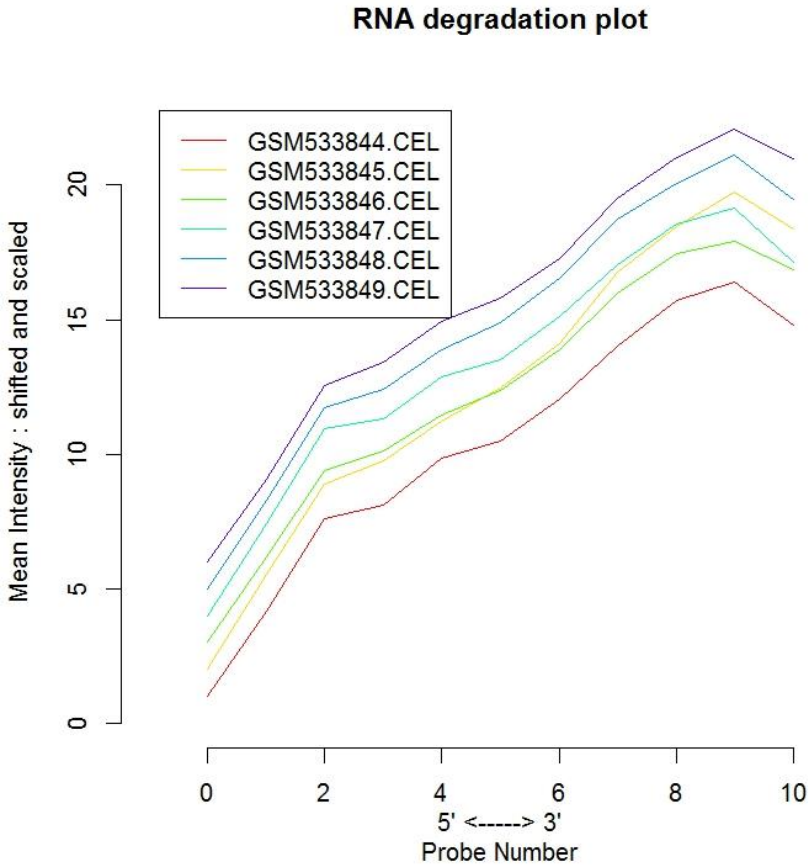
# 质量控制 (3)

```
library(simpleaffy)

data.qc <- qc(raw.data)

plot(data.qc)
```



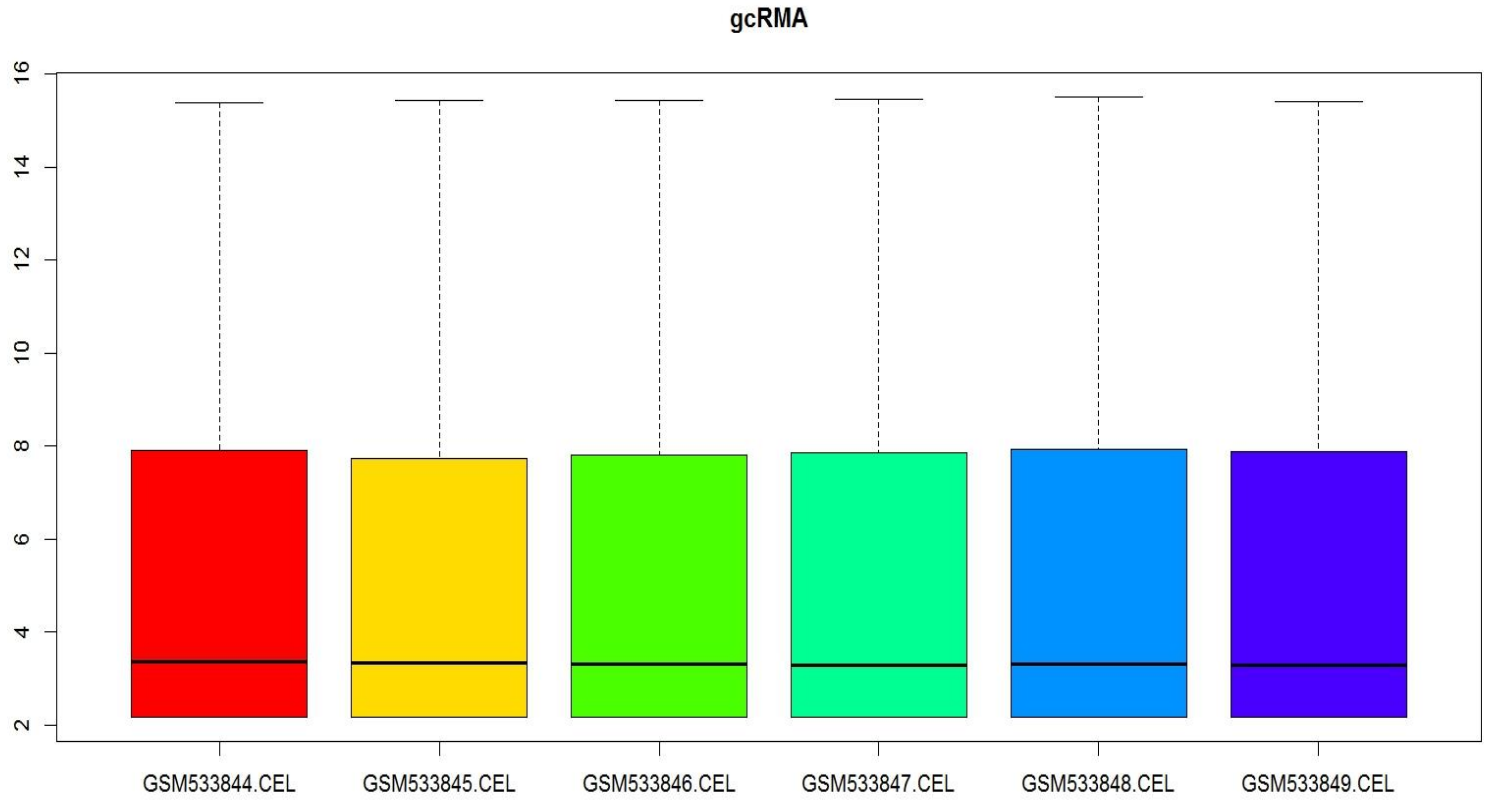
# 质量控制（4）

```
data.deg <- AffyRNAdeg(raw.data)

plotAffyRNAdeg(data.deg,col=cols)

legend("topleft",rownames(pData(raw.
data)),col=cols,lwd=1,inset=0.05)
```



# 预处理及标准化

```
library(gcrma)

data.gcrma <- gcrma(raw.data)

eset <- exprs(data.gcrma)

boxplot(eset,col=cols,main="gcRMA")
```



# 差异表达基因筛选 (1)

```
library(limma)

day_ <- factor(raw.data$Treatment)

design <- model.matrix(~-1+day_)


contrast.matrix <-
    makeContrasts(contrasts="day_Day8 -
    day_Day0", levels=design)
```

| sample_id | day_0 | day_8 |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 0 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |

# 差异表达基因筛选 (2)

```
fit <- lmFit(eset,design)

fit1 <- contrasts.fit(fit,contrast.matrix)

fit2 <- eBayes(fit1)

dif <- topTable(fit2, coef="day_Day8 - day_Day0", n=nrow(fit2), lfc=log2(1.5))

dif1 <- dif[dif[,"adj.P.Val"]<0.01,]
```

| | logFC | AveExpr | t | P.Value | adj.P.Val | B |
|---|---|---|---|---|---|---|
| 209496_at | 7.2094948 | 8.543996 | 106.61839 | 1.443313e-08 | 0.0002944602 | 11.112713 |
| 211737_x_at | -3.1632914 | 5.887529 | -92.70522 | 2.642914e-08 | 0.0002944602 | 10.546330 |
| 218471_s_at | 0.8324238 | 7.369664 | 64.67449 | 1.254324e-07 | 0.0009316700 | 8.940172 |
| 222156_x_at | 2.7442888 | 8.644903 | 52.11789 | 3.189226e-07 | 0.0017766380 | 7.903490 |
| 212558_at | 4.0550318 | 6.373138 | 48.41763 | 4.384334e-07 | 0.0019018261 | 7.540945 |
| 221511_x_at | 2.1564365 | 8.609168 | 46.70834 | 5.120925e-07 | 0.0019018261 | 7.362661 |

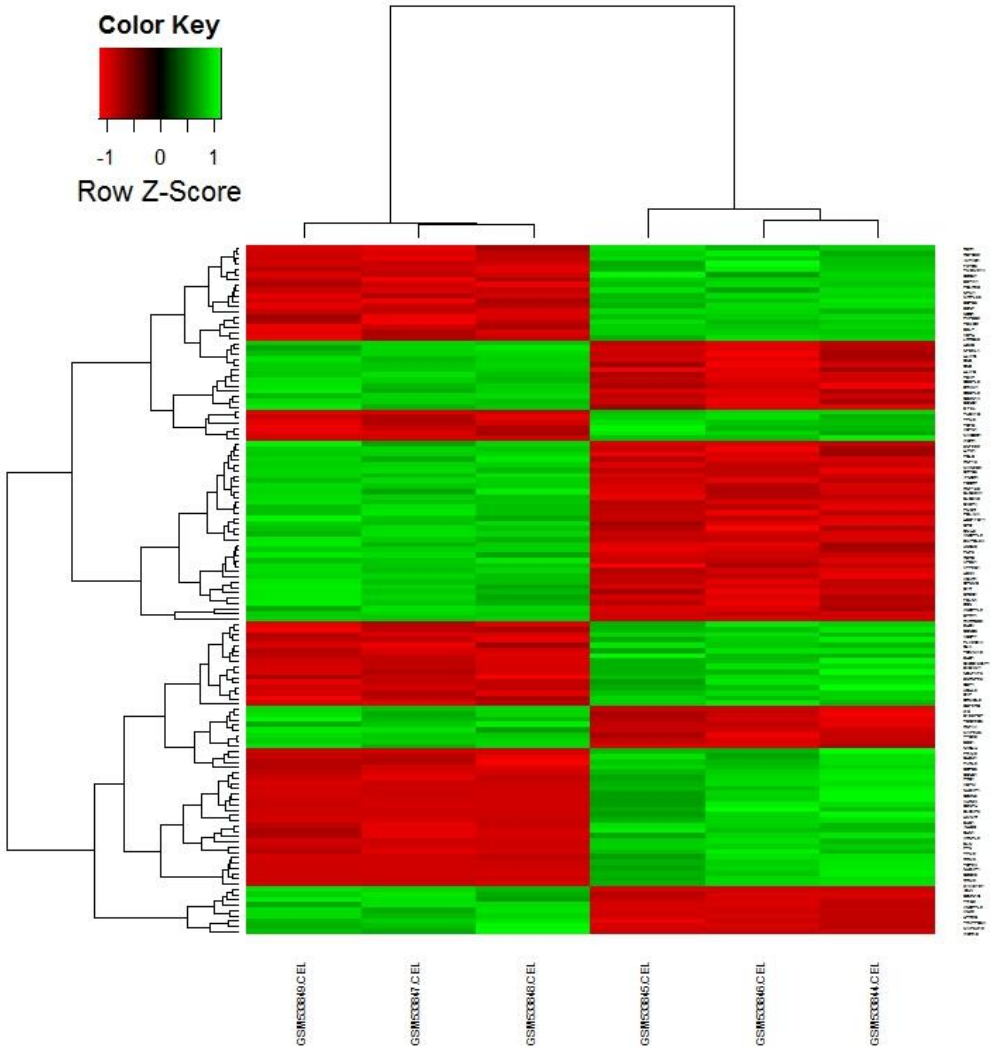
# 差异表达基因筛选 (3)

```
library(annotate)
affydb <- annPkgName(raw.data@annotation,type="db")
library(affydb,character.only=T)
dif1$symbols <- getSYMBOL(rownames(dif1),affydb)
dif1$EntrezID <- getEG(rownames(dif1),affydb)
dif2 <- dif1[(!is.na(dif1$symbols)),]
head(dif2)
```

| | logFC | AveExpr | t | P.Value | adj.P.Val | B | symbols | EntrezID |
|---|---|---|---|---|---|---|---|---|
| 209496_at | 7.2094948 | 8.543996 | 106.61839 | 1.443313e-08 | 0.0002944602 | 11.112713 | RARRES2 | 5919 |
| 211737_x_at | -3.1632914 | 5.887529 | -92.70522 | 2.642914e-08 | 0.0002944602 | 10.546330 | PTN | 5764 |
| 218471_s_at | 0.8324238 | 7.369664 | 64.67449 | 1.254324e-07 | 0.0009316700 | 8.940172 | BBS1 | 582 |
| 212558_at | 4.0550318 | 6.373138 | 48.41763 | 4.384334e-07 | 0.0019018261 | 7.540945 | SPRY1 | 10252 |
| 211072_x_at | -1.0317262 | 13.020444 | -37.73577 | 1.286618e-06 | 0.0031265394 | 6.289628 | TUBA1B | 10376 |
| 214703_s_at | 0.6270494 | 9.083220 | 36.31292 | 1.518793e-06 | 0.0031265394 | 6.094044 | MAN2B2 | 23324 |

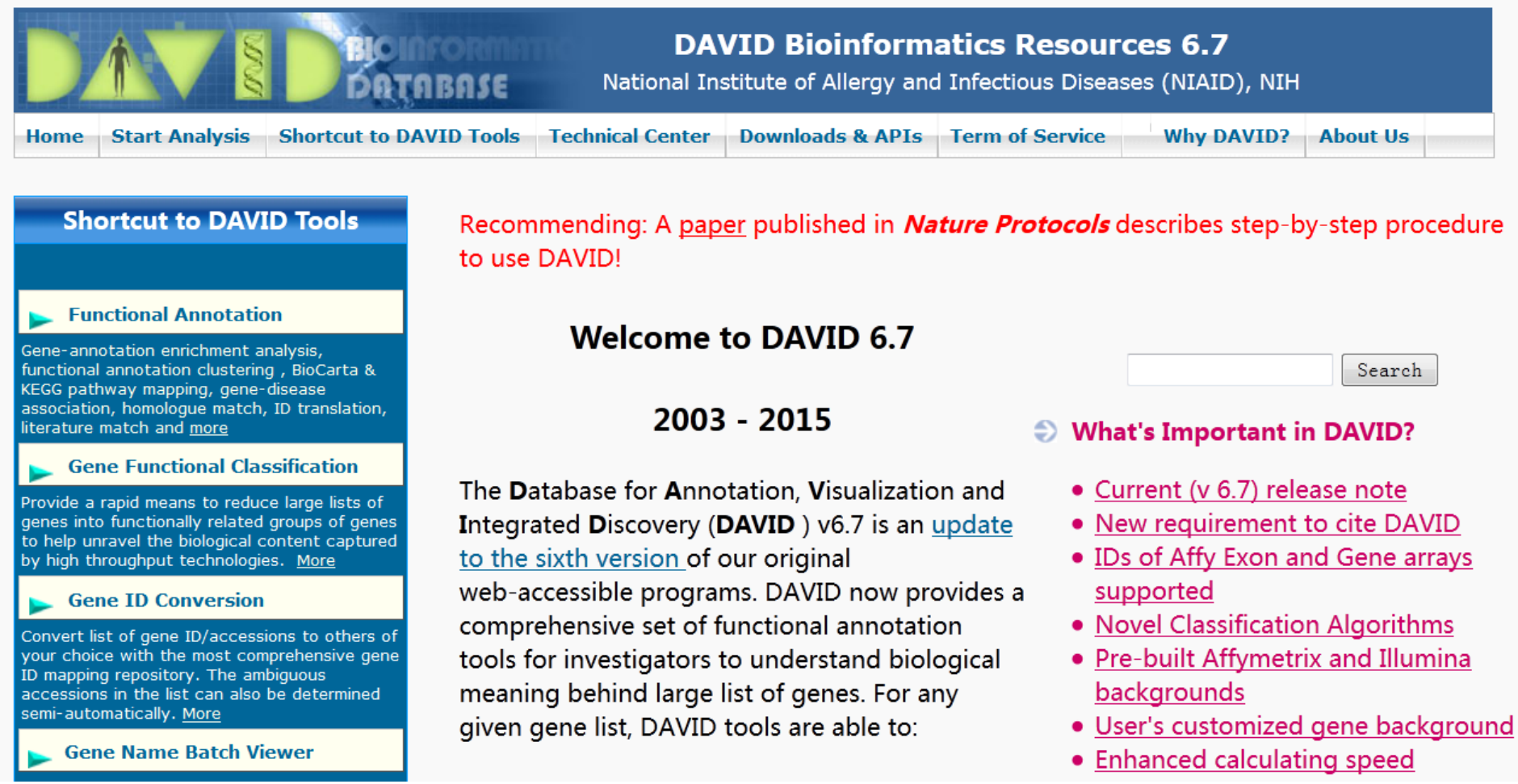
# 聚类分析

```
eset1 <- eset[(rownames(dif2)),]

row.names(eset1) <- dif2$symbols

Library(gplots)

heatmap.2(as.matrix(eset1),col=redgreen(75),
cexRow=0.2,cexCol=0.5,scale="row",trace="no
ne",key=T,keysize=1.2,density.info="none")
```



# 功能和信号通路分析（1）

**https://david.ncifcrf.gov/**



# 功能和信号通路分析（3）

## GO 富集分析结果 (top 10)

| Sublist | Category | Term | RT | Genes | Count | % | P-Value | Benjamini |
|---|---|---|---|---|---|---|---|---|
| | GOTERM_BP_FAT | mitotic cell cycle | RT | | 15 | 12.3 | 6.9E-7 | 7.5E-4 |
| | GOTERM_BP_FAT | regulation of cyclin-dependent protein kinase activity | RT | | 7 | 5.7 | 3.1E-6 | 1.7E-3 |
| | GOTERM_BP_FAT | cell cycle process | RT | | 17 | 13.9 | 4.7E-6 | 1.7E-3 |
| | GOTERM_BP_FAT | regulation of cell cycle | RT | | 13 | 10.7 | 7.1E-6 | 1.9E-3 |
| | GOTERM_BP_FAT | cell cycle phase | RT | | 14 | 11.5 | 1.4E-5 | 2.9E-3 |
| | GOTERM_BP_FAT | cell cycle | RT | | 19 | 15.6 | 1.8E-5 | 3.2E-3 |
| | GOTERM_CC_FAT | spindle | RT | | 9 | 7.4 | 1.8E-5 | 3.7E-3 |
| | GOTERM_CC_FAT | microtubule cytoskeleton | RT | | 15 | 12.3 | 8.1E-5 | 8.3E-3 |
| | GOTERM_CC_FAT | lytic vacuole | RT | | 9 | 7.4 | 2.3E-4 | 1.5E-2 |
| | GOTERM_CC_FAT | lysosome | RT | | 9 | 7.4 | 2.3E-4 | 1.5E-2 |

# 功能和信号通路分析（4）

## 信号通路富集分析结果

| Sublist | Category | Term | RT | Genes | Count | % | P-Value | Benjamini |
|---|---|---|---|---|---|---|---|---|
| | KEGG_PATHWAY | Glycosaminoglycan degradation | RT | | 5 | 4.1 | 3.4E-5 | 2.6E-3 |
| | KEGG_PATHWAY | p53 signaling pathway | RT | | 6 | 4.9 | 3.7E-4 | 1.4E-2 |
| | KEGG_PATHWAY | Lysosome | RT | | 7 | 5.7 | 6.5E-4 | 1.6E-2 |
| | KEGG_PATHWAY | Cell cycle | RT | | 7 | 5.7 | 9.2E-4 | 1.7E-2 |
| | KEGG_PATHWAY | Glycosphingolipid biosynthesis | RT | | 3 | 2.5 | 7.1E-3 | 1.0E-1 |
| | KEGG_PATHWAY | Other glycan degradation | RT | | 3 | 2.5 | 9.2E-3 | 1.1E-1 |
| | BIOCARTA | Cell Cycle | RT | | 3 | 2.5 | 2.9E-2 | 6.6E-1 |
| | BBID | 26.cyclin-CDK_complexes | RT | | 3 | 2.5 | 3.1E-2 | 2.9E-1 |
| | BIOCARTA | Cyclins and Cell Cycle Regulation | RT | | 3 | 2.5 | 3.4E-2 | 4.7E-1 |

# Thank You