思博奥科生物信息科技(北京)有限公司

R/Bioconductor在生物多维组学数据整合中的应用

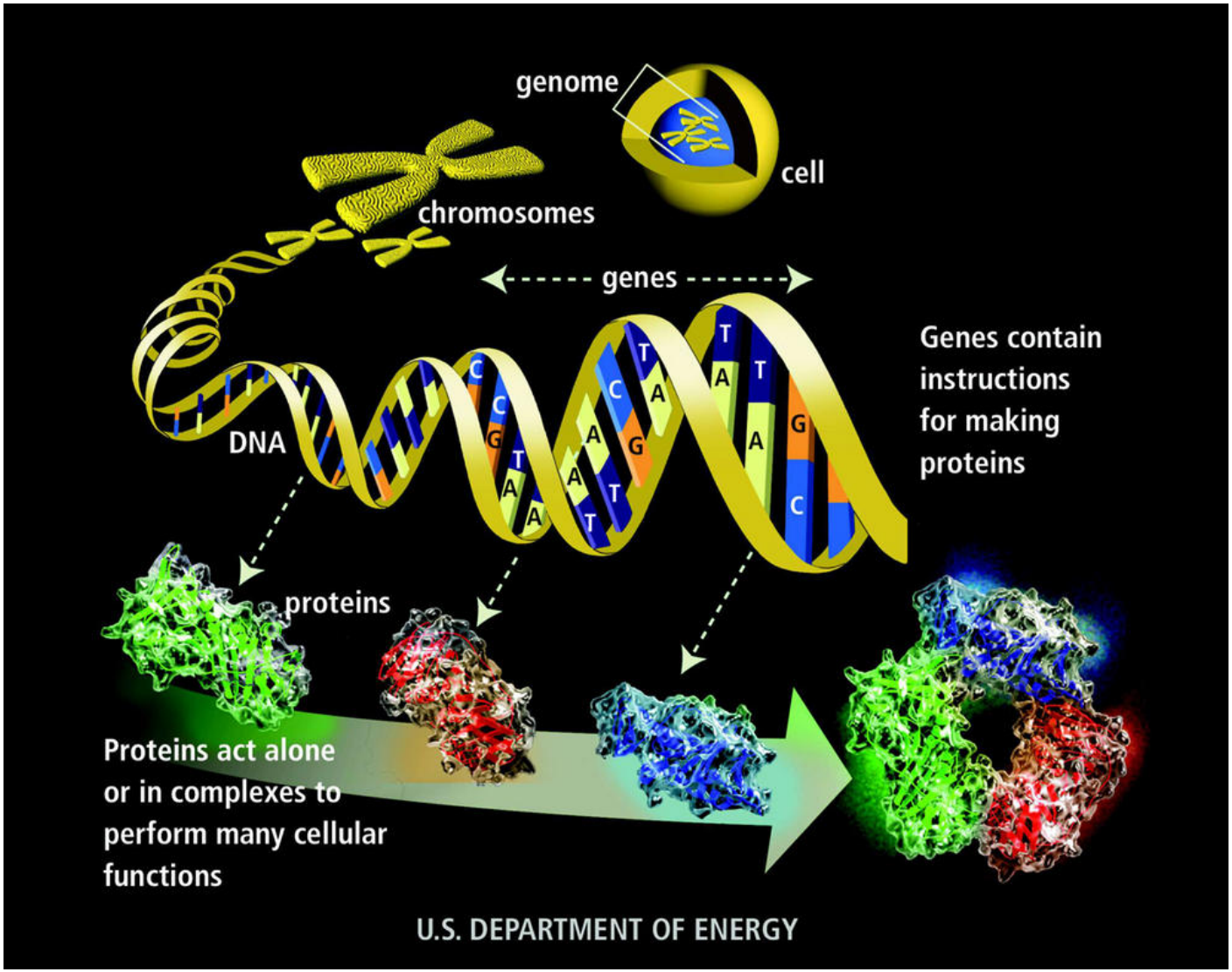
第五届中国R语言会议（上海）

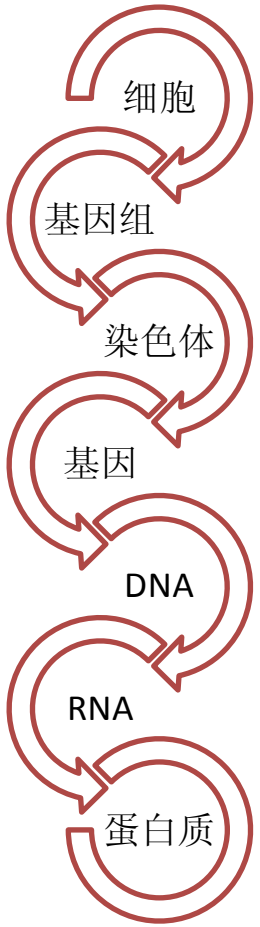
2012年11月03日

概要

- 涉及的基本生物学概念
  - 中心法则和组学(Omics)
- 组学数据整合的哲学基础和应用意义
- R/Bioconductor在组学数据整合中的案例
- 挑战和展望

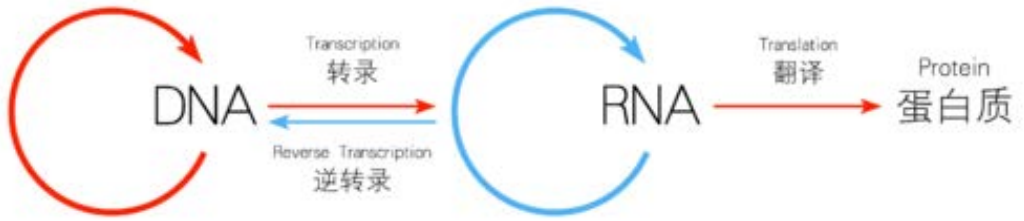
细胞中分子信息链


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY


细胞
基因组
染色体
基因
DNA
RNA
蛋白质

分子生物学“第一定律”

- 1953年James Watson & Francis Crick发现DNA双螺旋结构；
- 1958年Francis Crick提出中心法则“Central dogma”，并于1970年在Nature杂志上发表；
- 中心法则的多层次扩展产生了多种组学(Omics)数据；

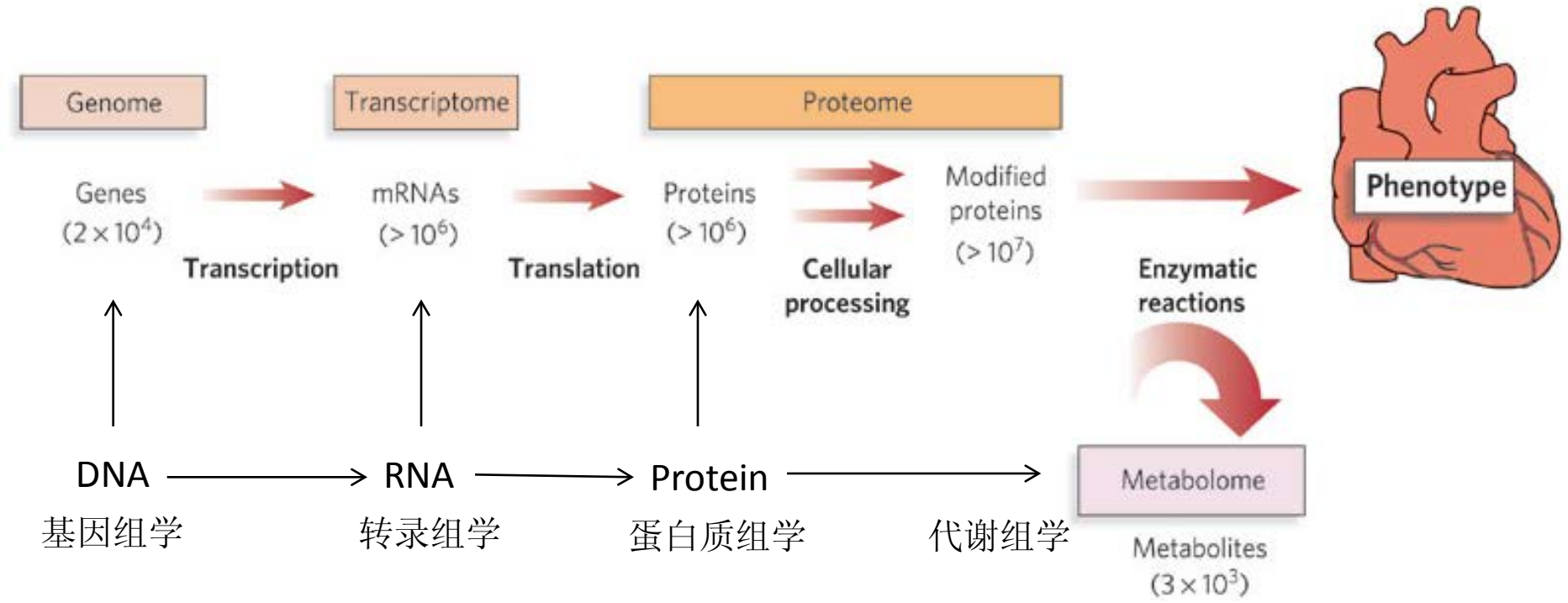


Watson J.D. and Crick F.H.C. A Structure for Deoxyribose Nucleic Acid. *Nature* 171, **1953.** 737-738.

Crick F., Central Dogma of Molecular Biology. *Nature* 227, **1970**. 561-563.

中心法则与多层次组学

- 什么是组学(Omics)?
- 随着中心法则展开，生物的信息复杂度逐步增大；
- 基因型和表型密切相关，从而为疾病研究提供思路。

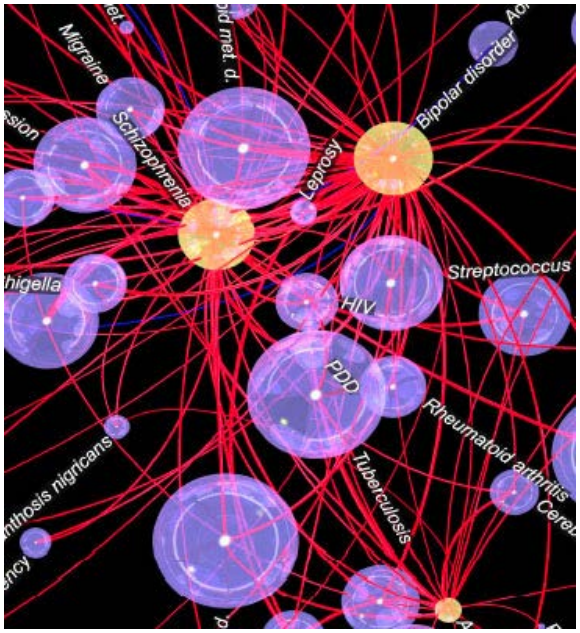


Robert E. Gerszten *et.al. Nature* .2008,451, 949-952.

数据整合的哲学基础

组学数据整合的医学意义

- 疾病是复杂系统，如何解决“Puzzle of Complex Disease”?
- 多维组学数据整合可更全面真实地模拟疾病自然机制；
- 多维组学数据整合可以有效提高生物信号的“信噪比”；
- 系统生物学理论指引的数据整合是转化医学和个体化医疗发展的基石。



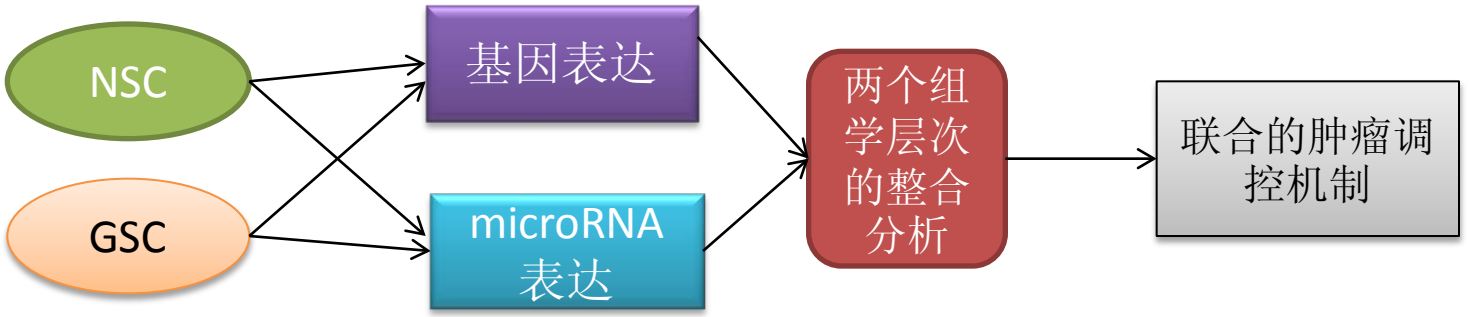
多维组学数据整合的方式

整合类型定义和平台建设：

- 第一类，组学数据与先验知识的横向整合
  - 文献、本体、生物医学数据库等。
  - 自有数据 + 同类的公共数据/已测基因组 。
- 第二类，不同层次组学数据之间的纵向整合
  - **SNPs / CNVs / DNA methylation / Gene Expression / microRNAs/Proteins**
    - DNA methylation/microRNAs/lncRNA profiling/CNVs + gene expression,
    - SNPs + CNVs, SNPs + gene expression,
    - Gene expression +Proteins, microRNA + Proteins.
- 第三类，组学数据与表型数据的关联性整合
  - 组学数据 + 临床资料(门诊、影像、生化和病理等)；
  - 组学数据 + 治疗数据(药理、药效和预后等)。
- 第四类，基于公共大数据的专题挖掘性整合

案例分析：神经胶质瘤干细胞 vs 正常神经干细胞的Gene和microRNA表达数据整合分析

- 神经胶质瘤(Glioma)是起源于神经胶质细胞的最常见的颅内肿瘤，约占所有颅内肿瘤的45%左右；
- 肿瘤干细胞学说：
  - 肿瘤干细胞和非肿瘤干细胞：神经胶质瘤干细胞(GSC)和正常神经干细胞(NSC)；
  - 基因调控(表达异常和突变等)可使NSC获得过度增殖能力，就具备了肿瘤细胞的特征；
  - 肿瘤发生、复发和转移的根源。



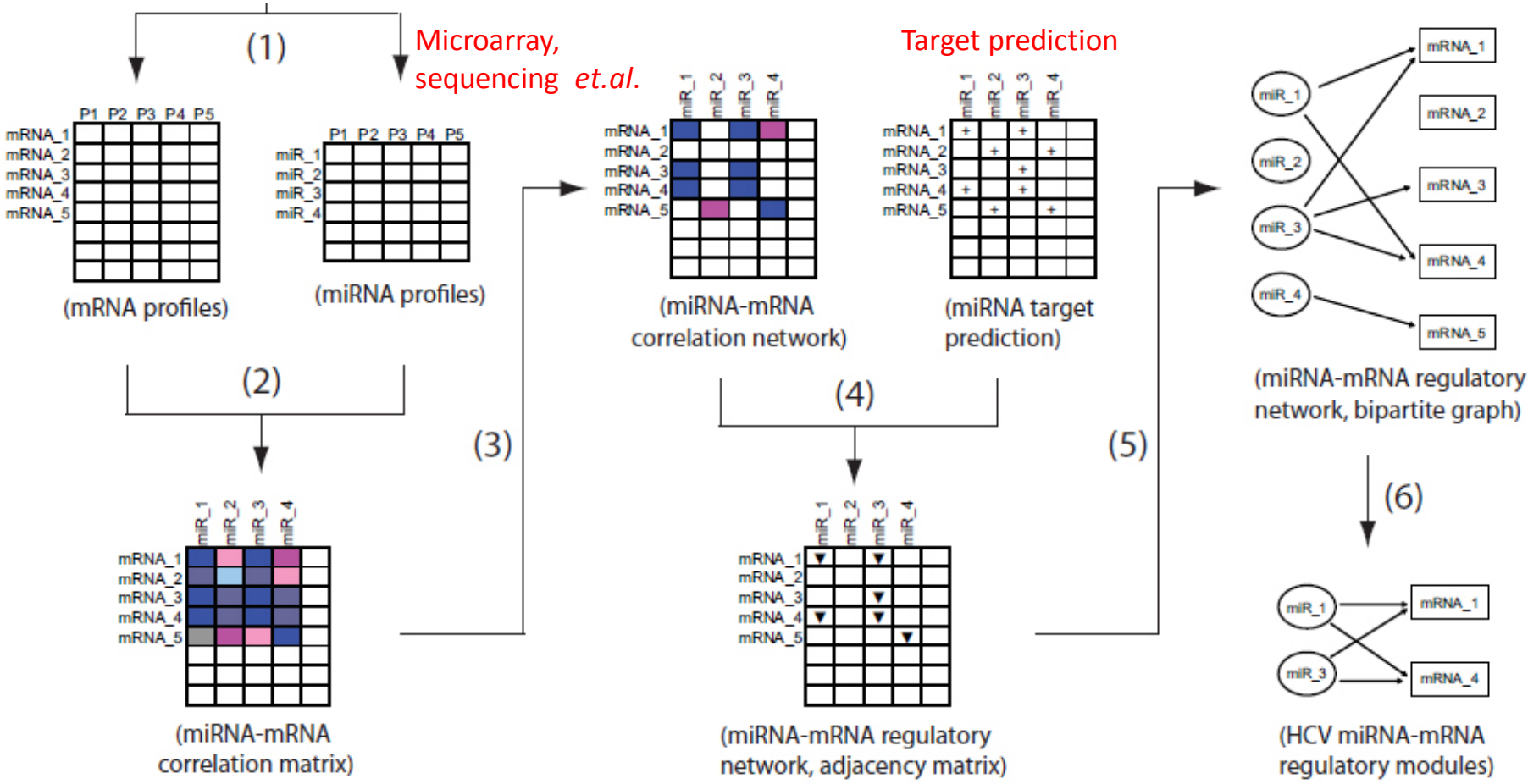
microRNA的基因调控

- microRNA是一类由内源基因编码的长度约为22 个核苷酸的非编码单链RNA 分子；
- microRNA可在转录后抑制基因的表达；
- microRNA倾向于和靶标基因表达负相关；
- 靶标预测是microRNA调控研究的难点；
- microRNA调控广泛存在于神经胶质瘤(Glioma)。

miRNA-mRNA整合分析流程

- **miRNA和mRNA表达谱数据矩阵化**
- 利用**miRNA-mRNA**的负相关关系筛选相关性
- **miRNA**的靶标预测及其功能通路分析

- 整合**miRNA-mRNA**相关性矩阵和靶标矩阵
- 利用整合相关性构建**miRNA-mRNA**调控关系

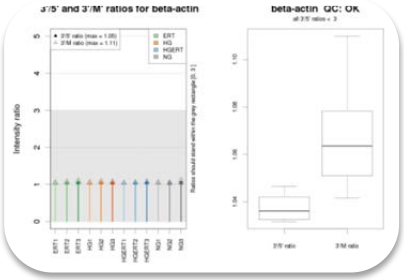


流程中的R和Bioconductor(芯片数据为例)

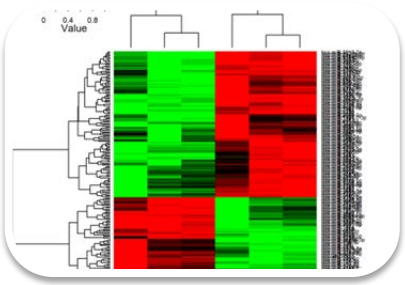
simpleaffy　　affy
affycoretools　　limma

heatmap.2()
plotPCA()
hclust()

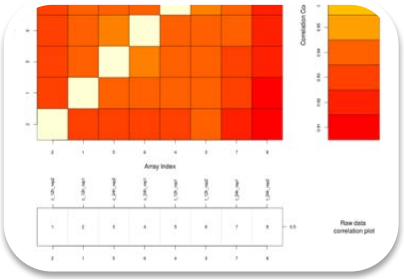
Hmisc.rcorr(), cor()



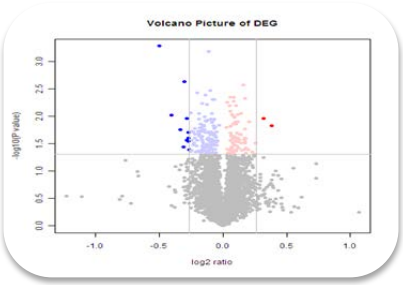
质控/预处理



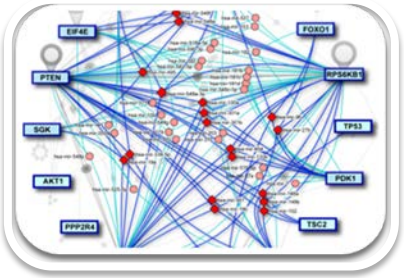
样本分析



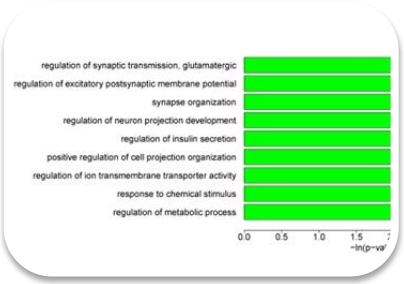
相关性计算



差异基因筛选



调控网络



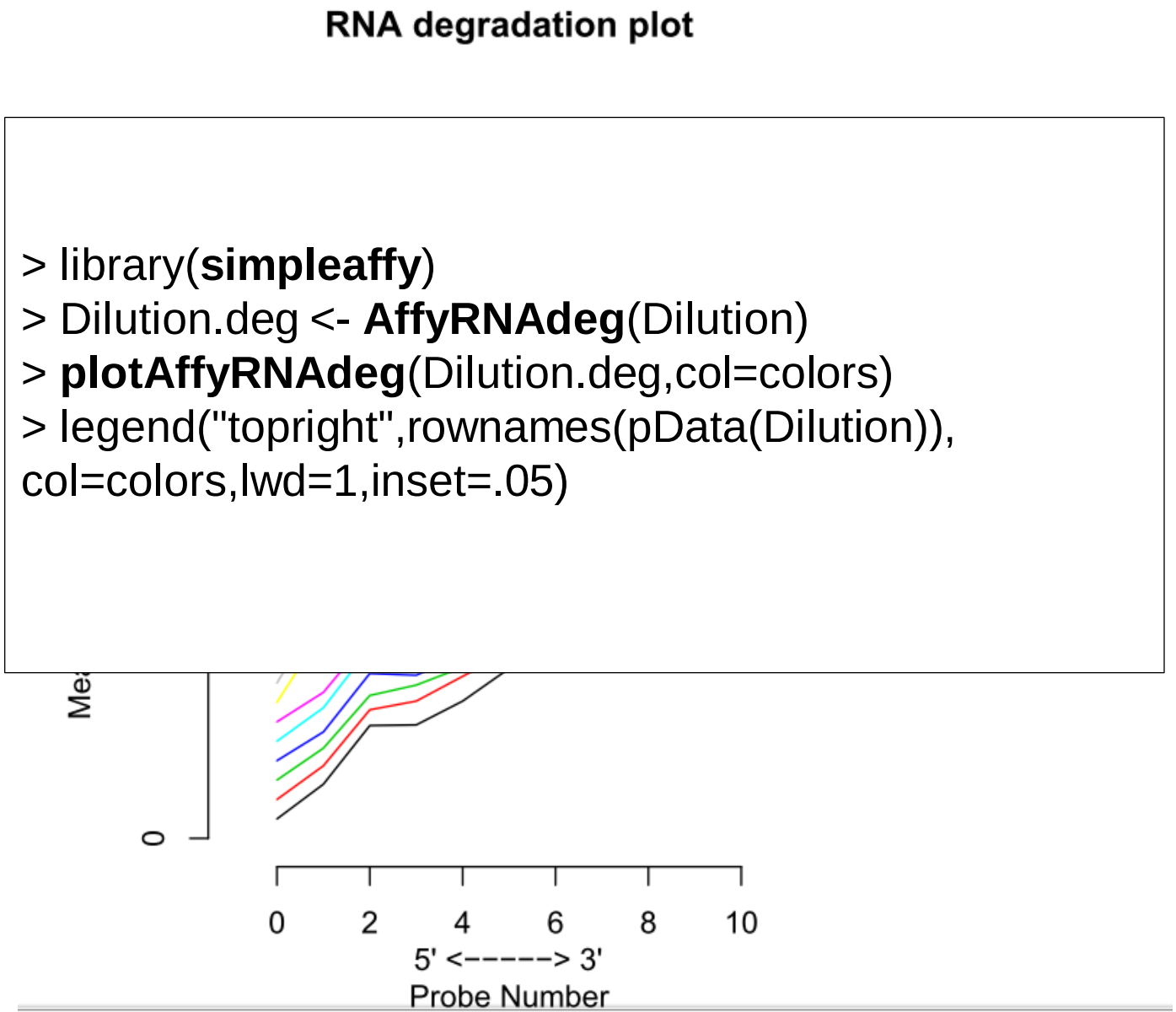
功能分析

t.test()　　genefilter
samr　　TANOVA
limma　　siggenes

igraph
mirDIP

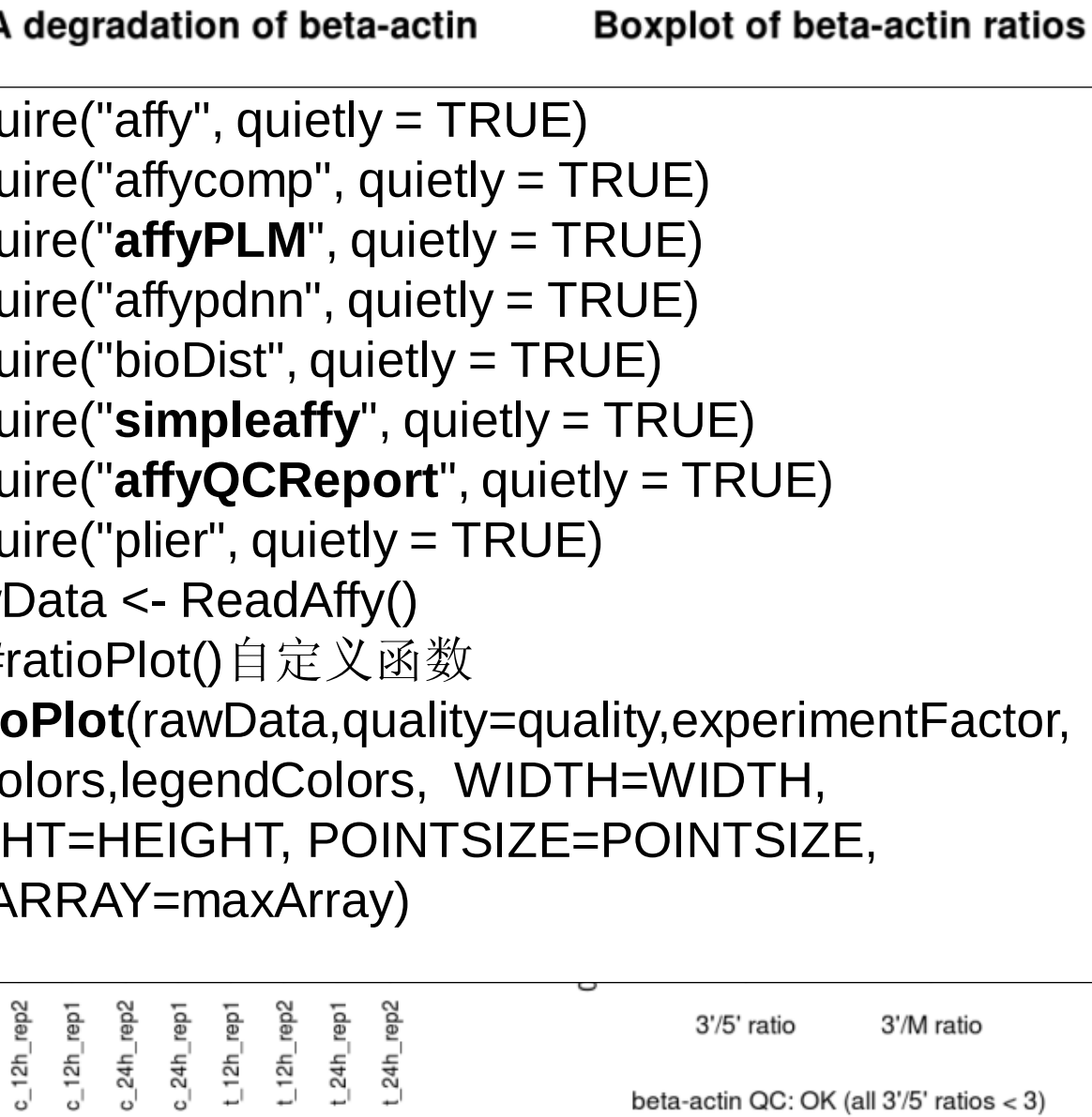
GO.db　　**AnnotationDbi**　　**barplot()**
org.Hs.eg.db　　**Bsgenome**　　**ggplot2**
org.Mm.eg.db　　**phyper()**

原始数据质量控制



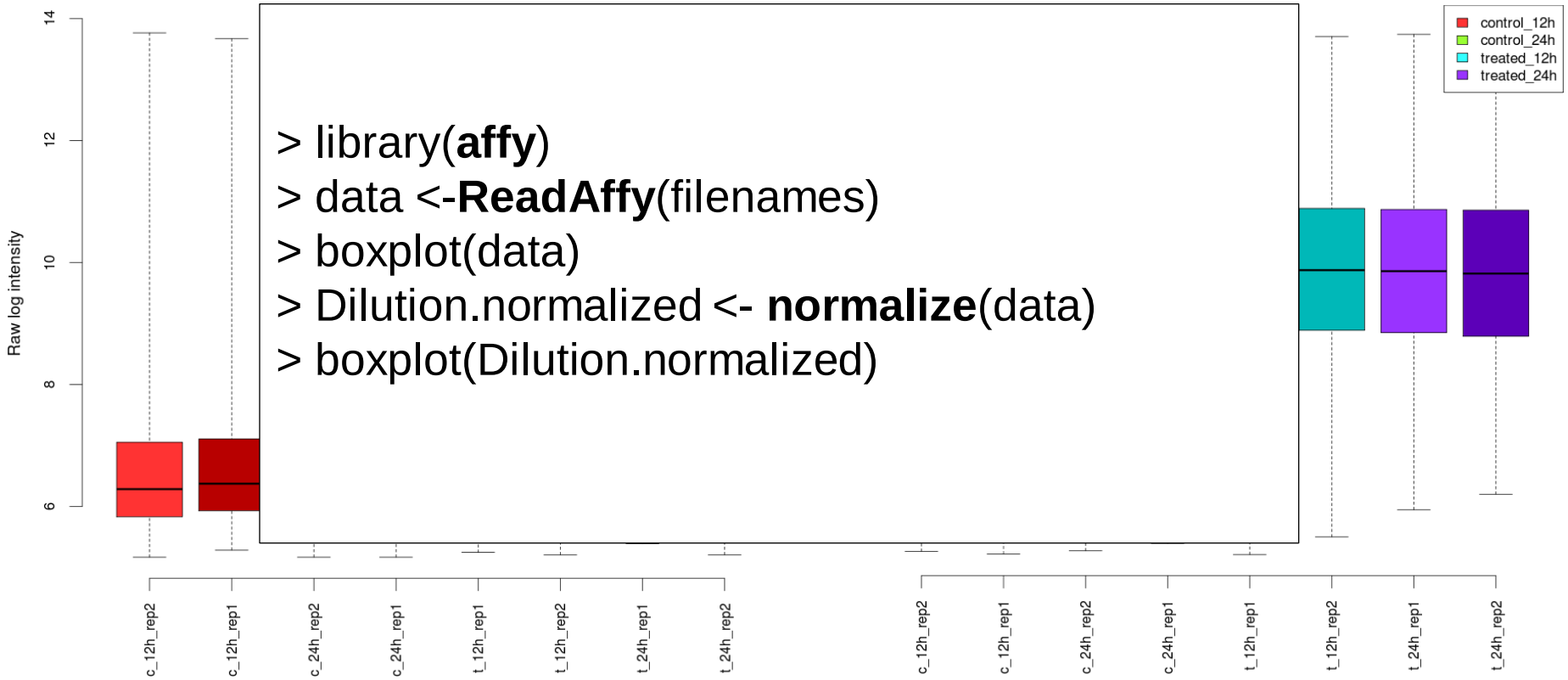

```
> library(simpleaffy)
> Dilution.deg <- AffyRNAdeg(Dilution)
> plotAffyRNAdeg(Dilution.deg,col=colors)
> legend("topright",rownames(pData(Dilution)),
col=colors,lwd=1,inset=.05)
```

原始数据质量控制

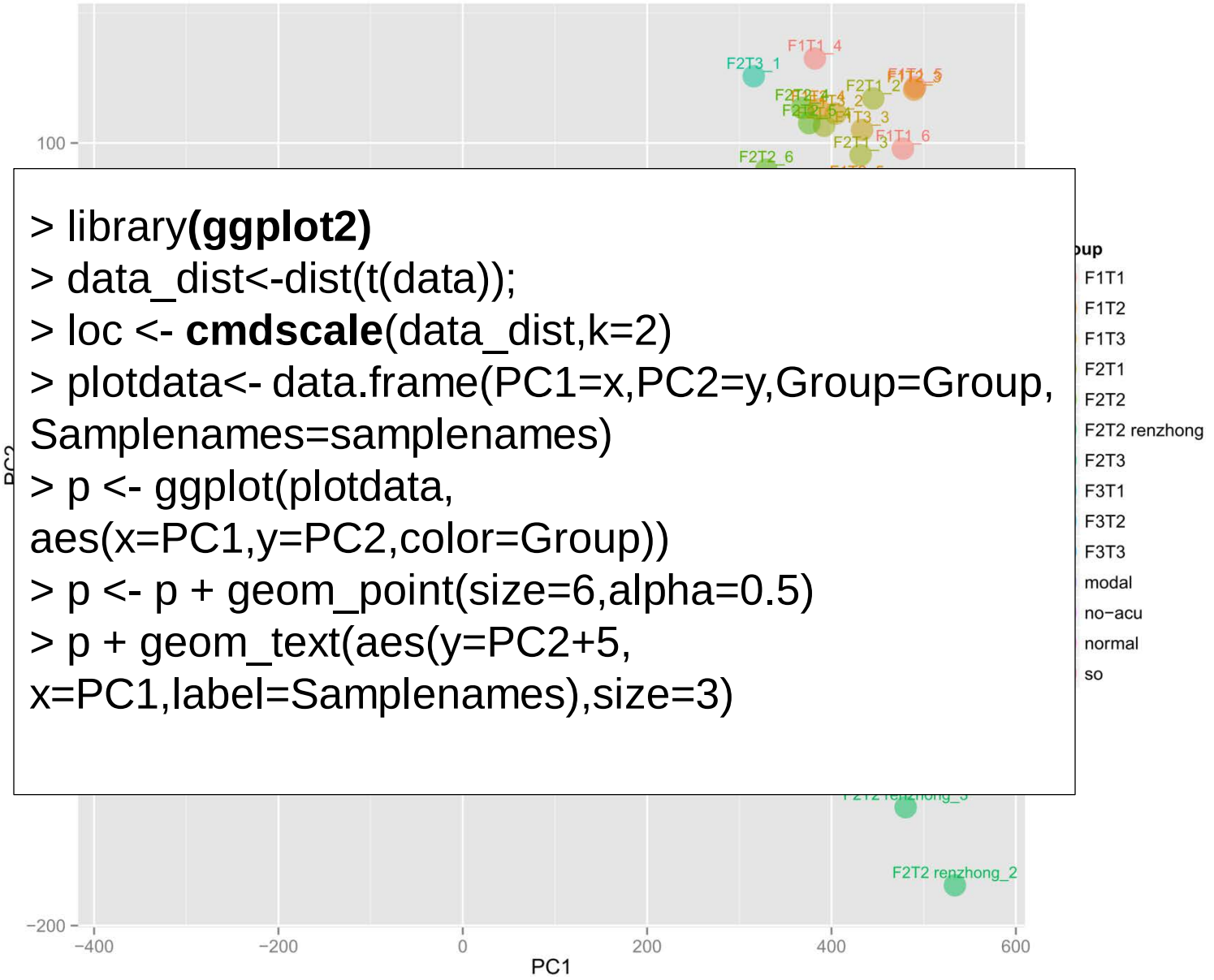



```
> require("affy", quietly = TRUE)
> require("affycomp", quietly = TRUE)
> require("affyPLM", quietly = TRUE)
> require("affypdnn", quietly = TRUE)
> require("bioDist", quietly = TRUE)
> require("simpleaffy", quietly = TRUE)
> require("affyQCReport", quietly = TRUE)
> require("plier", quietly = TRUE)
> rawData <- ReadAffy()
> ###ratioPlot()自定义函数
> ratioPlot(rawData,quality=quality,experimentFactor,
plotColors,legendColors,  WIDTH=WIDTH,
HEIGHT=HEIGHT, POINTSIZE=POINTSIZE,
MAXARRAY=maxArray)
```

数据标准化

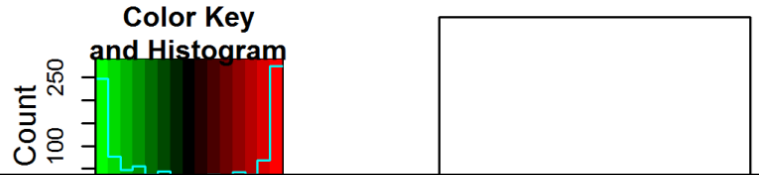

> library(**affy**)
> data <-**ReadAffy**(filenames)
> boxplot(data)
> Dilution.normalized <- **normalize**(data)
> boxplot(Dilution.normalized)
Raw log intensity
14
12
10
8
6
control_12h
control_24h
treated_12h
treated_24h
c_12h_rep2
c_12h_rep1
c_24h_rep2
c_24h_rep1
t_12h_rep1
t_12h_rep2
t_24h_rep1
t_24h_rep2
c_12h_rep2
c_12h_rep1
c_24h_rep2
c_24h_rep1
t_12h_rep1
t_12h_rep2
t_24h_rep1
t_24h_rep2

基于PCA分析的样本筛选

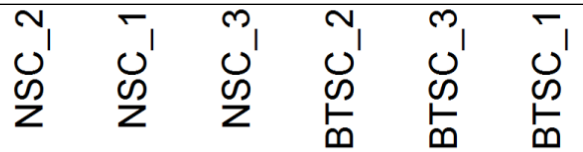



```
> library(ggplot2)
> data_dist<-dist(t(data));
> loc <- cmdscale(data_dist,k=2)
> plotdata<- data.frame(PC1=x,PC2=y,Group=Group,
Samplenames=samplenames)
> p <- ggplot(plotdata,
aes(x=PC1,y=PC2,color=Group))
> p <- p + geom_point(size=6,alpha=0.5)
> p + geom_text(aes(y=PC2+5,
x=PC1,label=Samplenames),size=3)
```

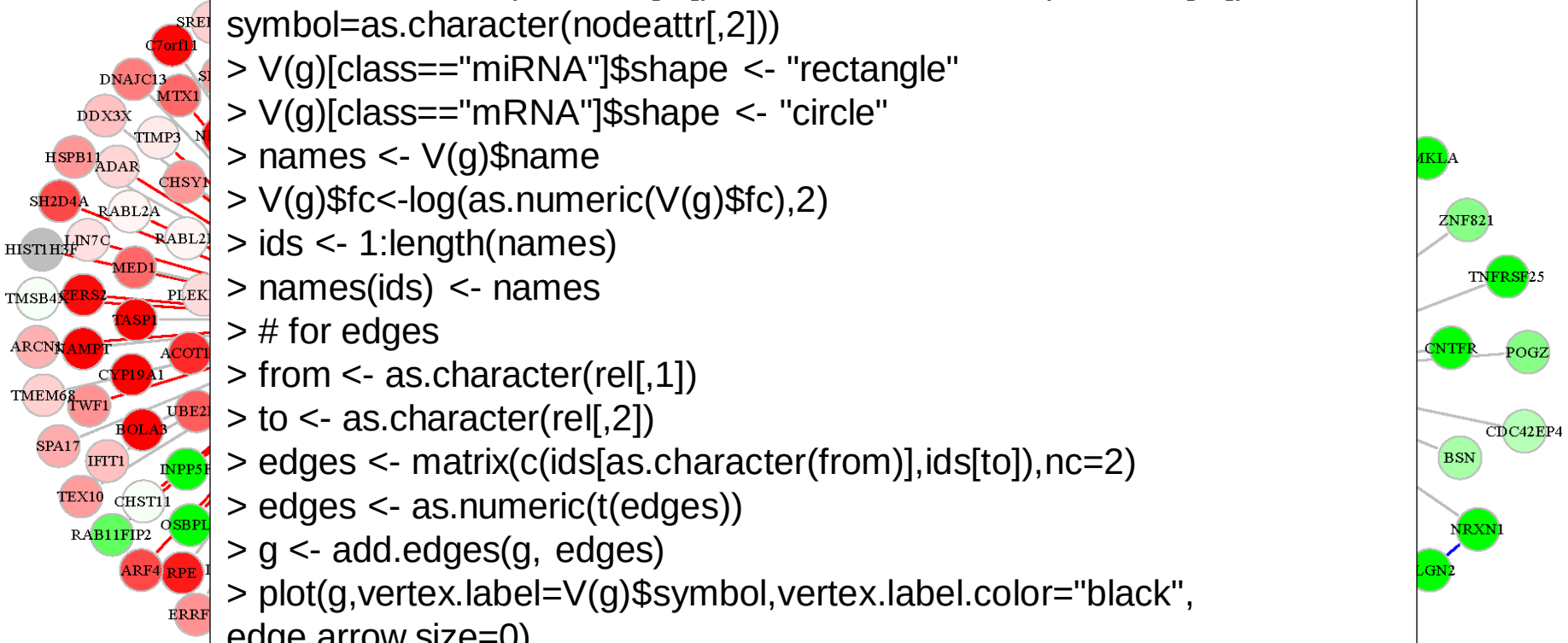
样本无监督聚类




```
> library(gplots)
> data <- read.table(file = "", header = T,quote ="")
> X<-data[,1:6]
> X<-as.matrix(X)
> for(i in 1:dim(X)[1])
   {
   len=max(X[i,])-min(X[i,])
   X[i,]=X[i,]-min(X[i,])
   X[i,]=X[i,]/len
   }
> heatmap.2(X,dendrogram="both",col=greenred,
trace="none", ylab = NULL,margins=c(6,8))
```

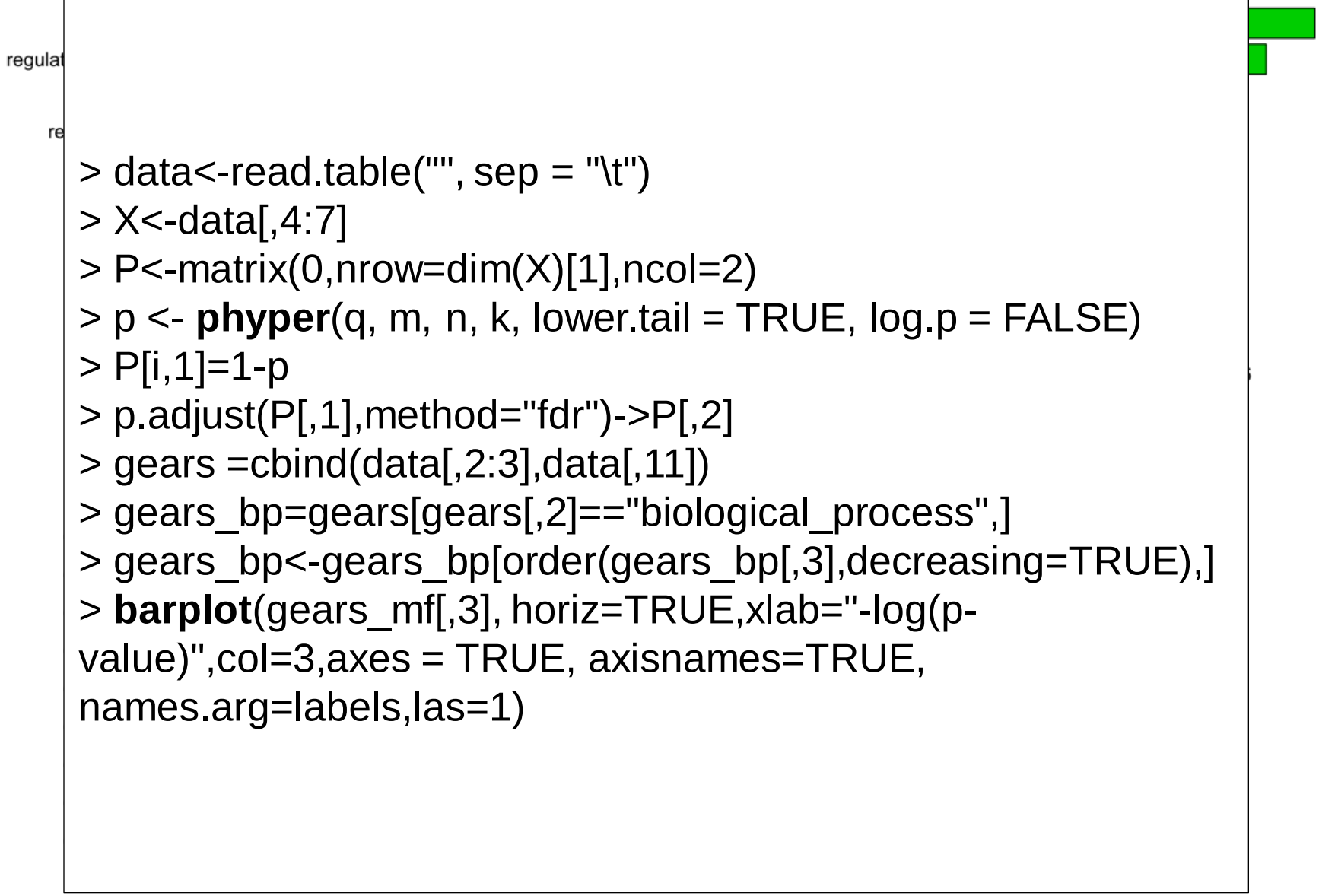


网络构建和模块分析




```
> library(igraph)
> g <- graph.empty()
> g <- add.vertices(g, nrow(nodeattr), name=as.character(nodeattr[,1]),
> fc=as.character(nodeattr[,3]), class=as.character(nodeattr[,4]),
symbol=as.character(nodeattr[,2]))
> V(g)[class=="miRNA"]$shape <- "rectangle"
> V(g)[class=="mRNA"]$shape <- "circle"
> names <- V(g)$name
> V(g)$fc<-log(as.numeric(V(g)$fc),2)
> ids <- 1:length(names)
> names(ids) <- names
> # for edges
> from <- as.character(rel[,1])
> to <- as.character(rel[,2])
> edges <- matrix(c(ids[as.character(from)],ids[to]),nc=2)
> edges <- as.numeric(t(edges))
> g <- add.edges(g, edges)
> plot(g,vertex.label=V(g)$symbol,vertex.label.color="black",
edge.arrow.size=0)
```

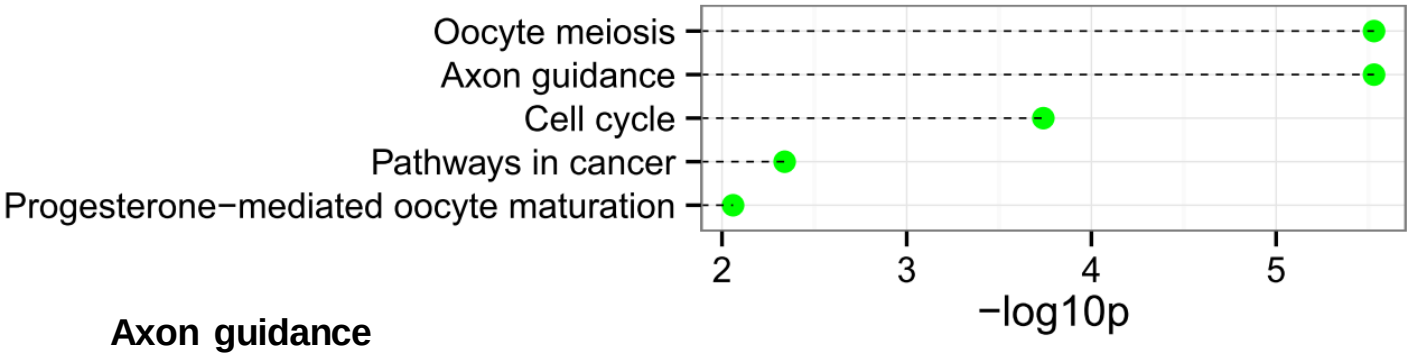
GO功能富集分析

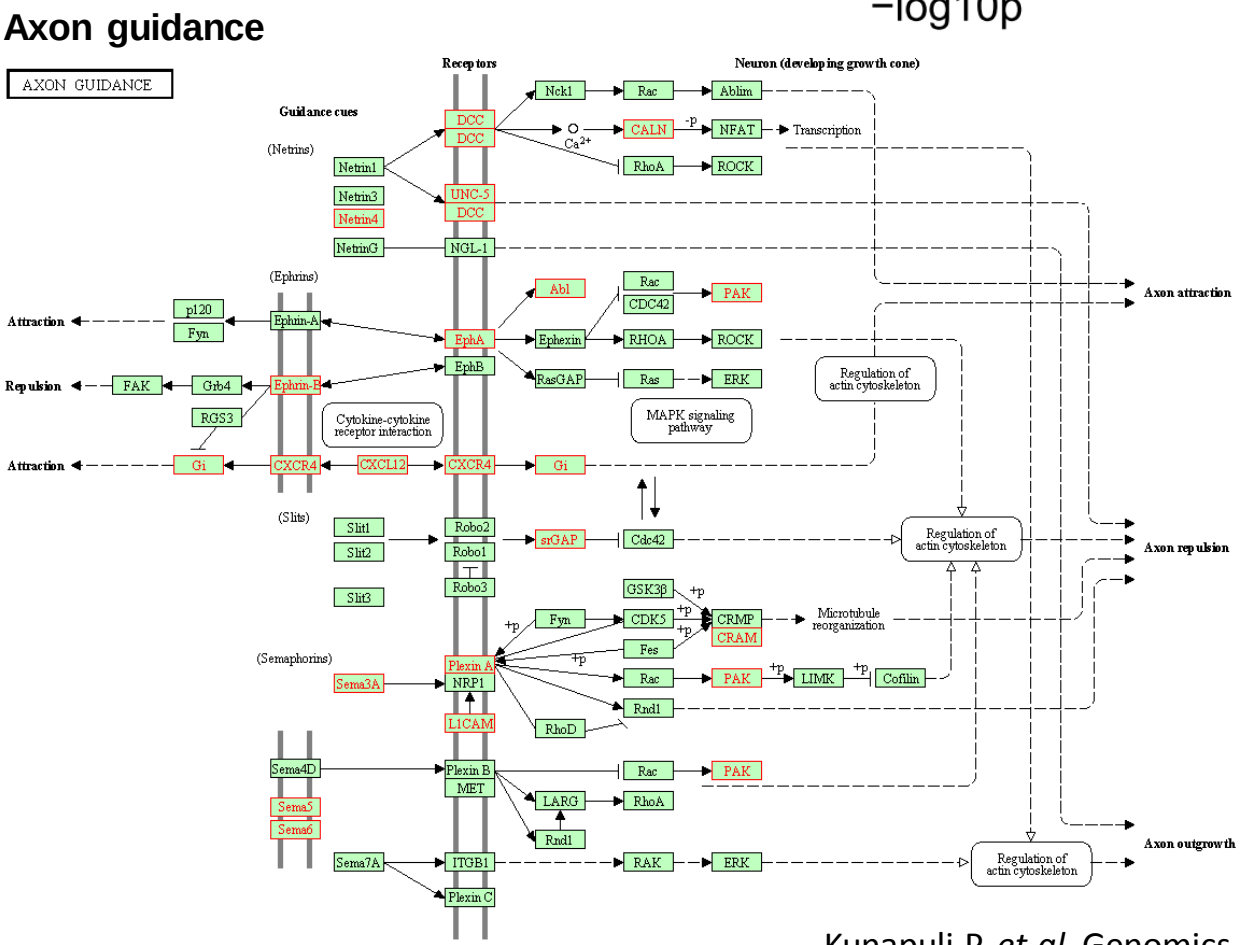


```
> data<-read.table("", sep = "\t")
> X<-data[,4:7]
> P<-matrix(0,nrow=dim(X)[1],ncol=2)
> p <- phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)
> P[i,1]=1-p
> p.adjust(P[,1],method="fdr")->P[,2]
> gears =cbind(data[,2:3],data[,11])
> gears_bp=gears[gears[,2]=="biological_process",]
> gears_bp<-gears_bp[order(gears_bp[,3],decreasing=TRUE),]
> barplot(gears_mf[,3], horiz=TRUE,xlab="-log(p-
value)",col=3,axes = TRUE, axisnames=TRUE,
names.arg=labels,las=1)
```

Song HJ, *et.al*.Nat Neurosci. 2002 May;5(5):438-45.

Pathway功能富集分析





Kunapuli P. *et.al*. Genomics. 2010 ;95(2):93-100.

挑战和展望

- 疾病的复杂性要求
  - 更可靠的临床样本积累；
  - 更真实的科学假设；
  - 更海量的信息和数据；
- 应对生物大数据要求
  - 更高效的算法和程序；
  - 更先进的软件体系(如云或并行构架R)；
  - 更强大的硬件支撑；
  - 学科间的交流和交叉学科人才培养。

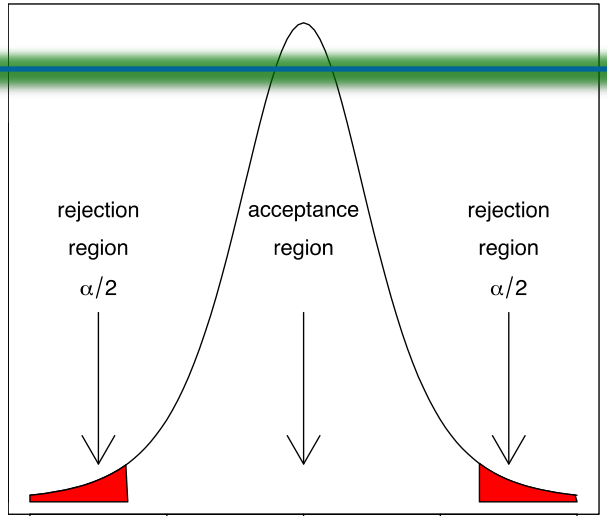
《R语言在生物信息学中的应用》

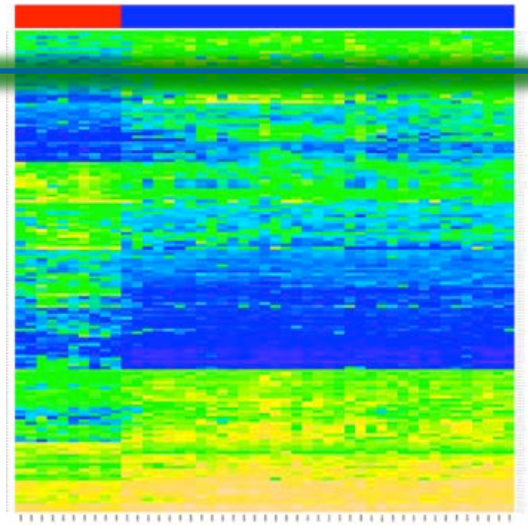
——一本即将面世的R红宝书

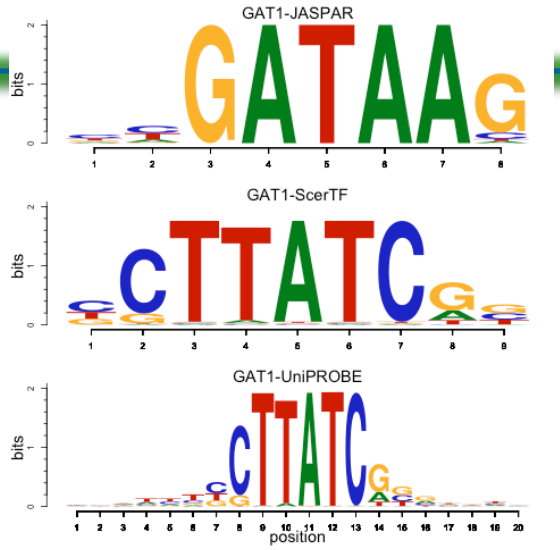
康奈尔大学 高山博士

****

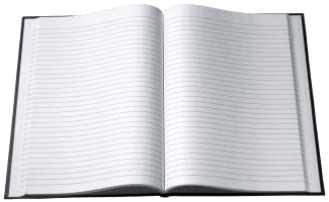
基本统计分析，高通量与多维数据可视化，统计绘图，下一代测序技术数据分析







《R语言在生物信息学中的应用》



1.本书是多名R领域专家通过互联网联手写作。

2. 从实例出发，直接和应用挂钩，不是罗列流水账，不是背课文。

3. 很多例子来自于实际工作，有些工作发表在nature等高水平期刊上。

4. 作者之一参与了R bioconductor的开发，所开发包的内容也包括进本书。

5. 从研究课题出发，讲思路，有具体代码和详细注释，不空洞，学生可以系统掌握如何设计课题，编程实现。

编者阵容

- 高　山（Cornell University，目前主要研究生物信息学算法，专长R与新算法开发。）
- 欧剑虹（Umass Medical School，主要从事R package的开发，曾参与并成功开发了多个Bioconductor包）
- 肖　凯（职业数据分析师，专长于R平台的数据分析）
- 李　勃（重庆大学，曾主编基因工程等教材，现从事生物芯片数据挖掘）
- 施劲松（南京军区总医院）
- 管栋印（Case Western Reserve University）
- 张　洋（University of Illinois)
- 其他。。。

致谢

- 感谢国内首本“R/Bioconductor”的作者谢建明老师的大力支持！
- 欣蒙多家出版商的邀稿，虽未最终确定出版方，但仍表示感谢！

谢 谢！

思博奥科

SysBiomics Bioinformatics (Beijing) Ltd.

思行创新

Consideration, Action and Creation

联系人：杭兴宜 博士
地址：北京市中关村科技园(丰台园)航丰路8号1808室 100070
电话(传真)：+86 10 5805 1799，传真：+86 10 8826 9778
手机：15611223895
电子邮件：xingyi.hang@sysbiomics.com