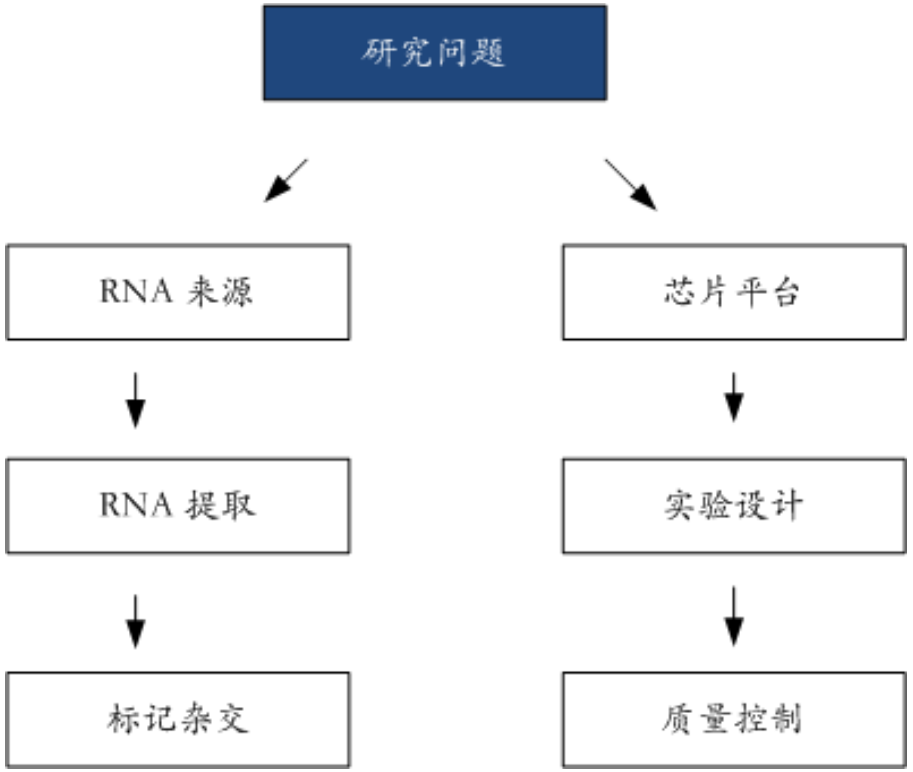
R/BioConductor在斑马鱼心脏再生领域的应用

内容摘要：通过对既往文献斑马鱼心脏再生芯片的数据分析，我们希望找到与再生过程非常密切的基因，用于后续的生物功能学研究。我们对数据采用常规的质控分析步骤，采用Limma进行差异表达基因的分析，对数据的聚类采用Cluster3.0。我们得到的结果与发表文献相似，说明我们的分析方法是准确的，为我们今后的工作奠定了基础。

关键词:斑马鱼；生物芯片；心脏再生

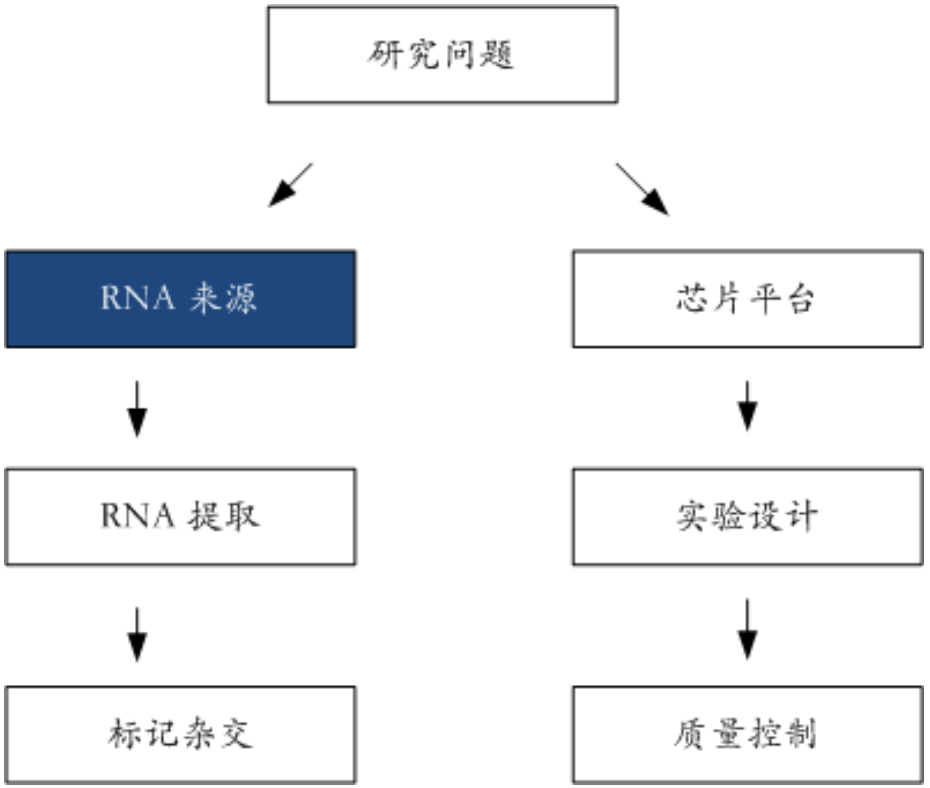
甄一松博士，中国医学科学院　阜外心血管病医院　　教育部基因与临床重点实验室，北京 100037
zhenyisong@cardiosignal.org

实验的前期框架
-问题的背景知识

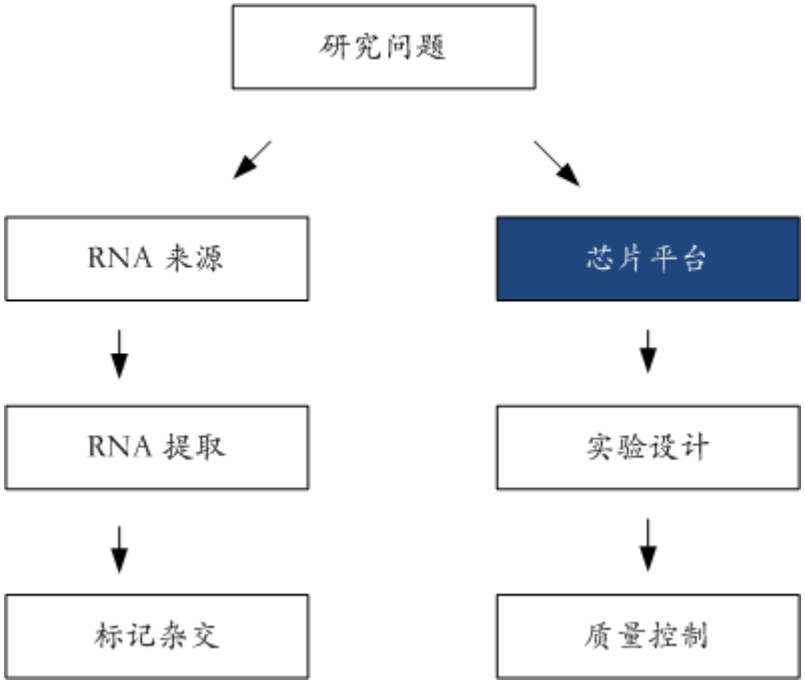


1. 寻找在斑马鱼心脏再生过程中起到关键作用的基因

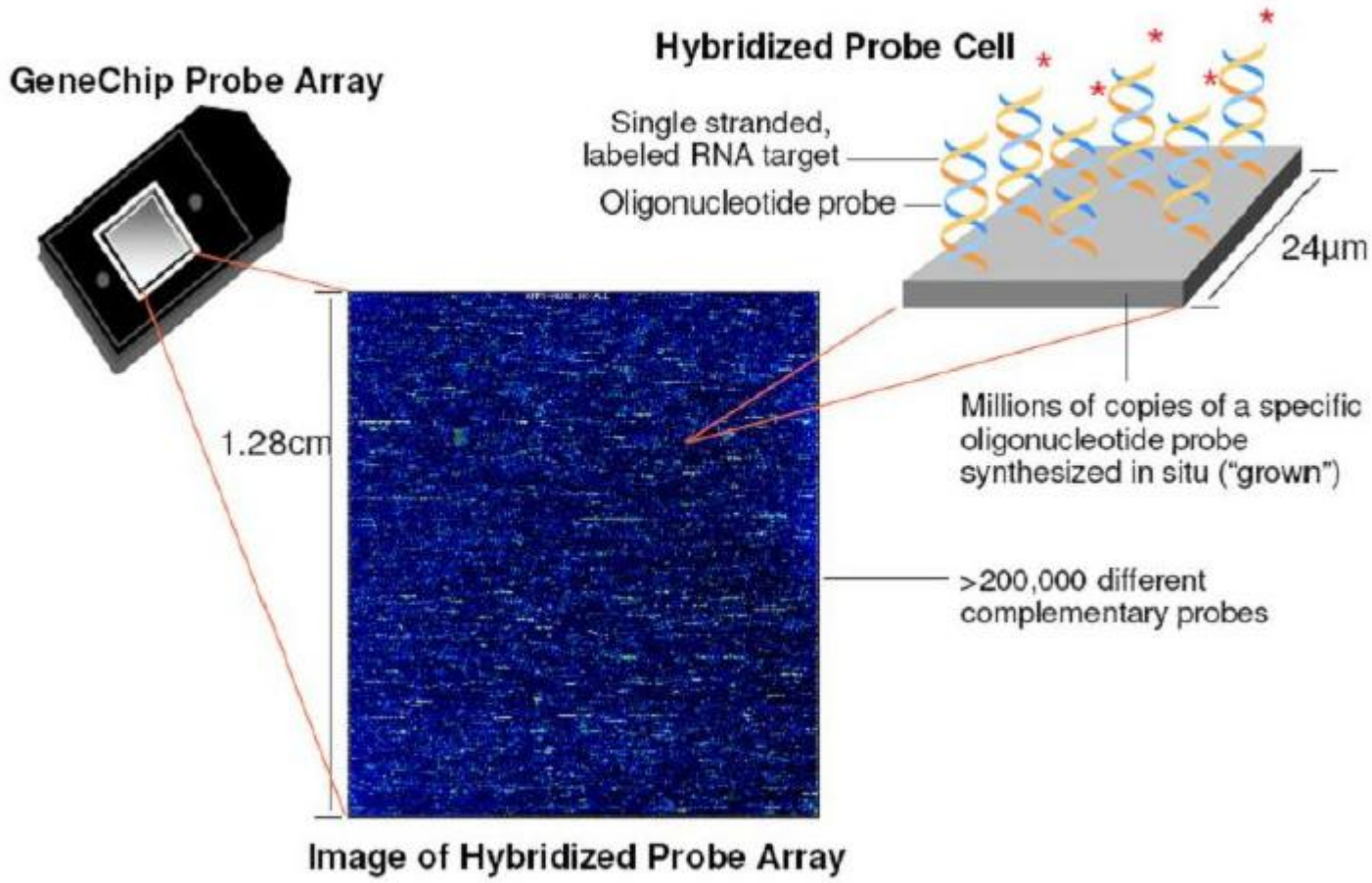
2.相互印证已发表的结果同我们分析的异同。



斑马鱼(zebra fish)，原产于印度、孟加拉国。斑马鱼基因与人类基因的相似度达到87%，这意味着在其身上做药物实验所得到的结果在多数情况下也适用于人体，因此它受到生物学家的重视。


研究问题
RNA 来源
RNA 提取
标记杂交
芯片平台
实验设计
质量控制

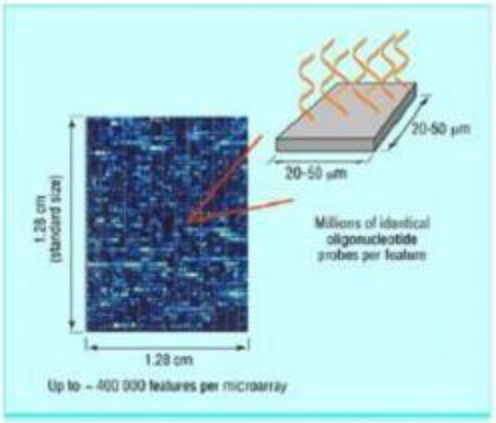
Affymetrix GeneChips



Affymetrix Technology

several *probe pairs*
(perfect match PM
and mismatch MM)
per *probeset*

PM: ATGAGCTGTACCAATGCCAACCTGG
MM: ATGAGCTGTACCTATGCCAACCTGG



64 pixels; Signal intensity is upper quartile of the 36 inner pixels

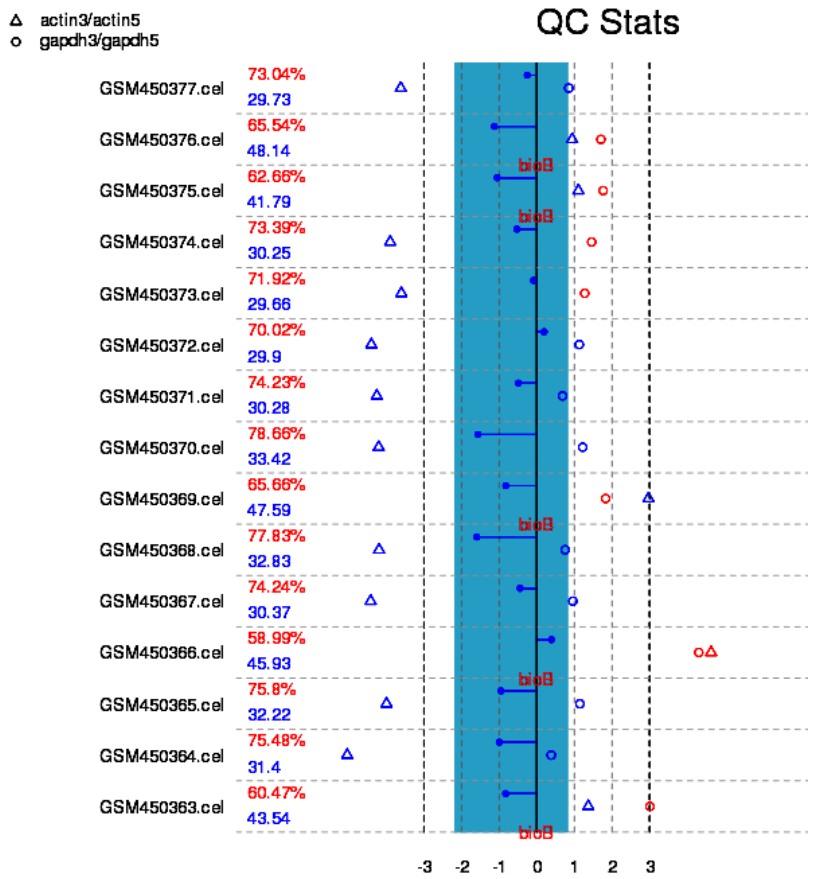
Stored in CEL file

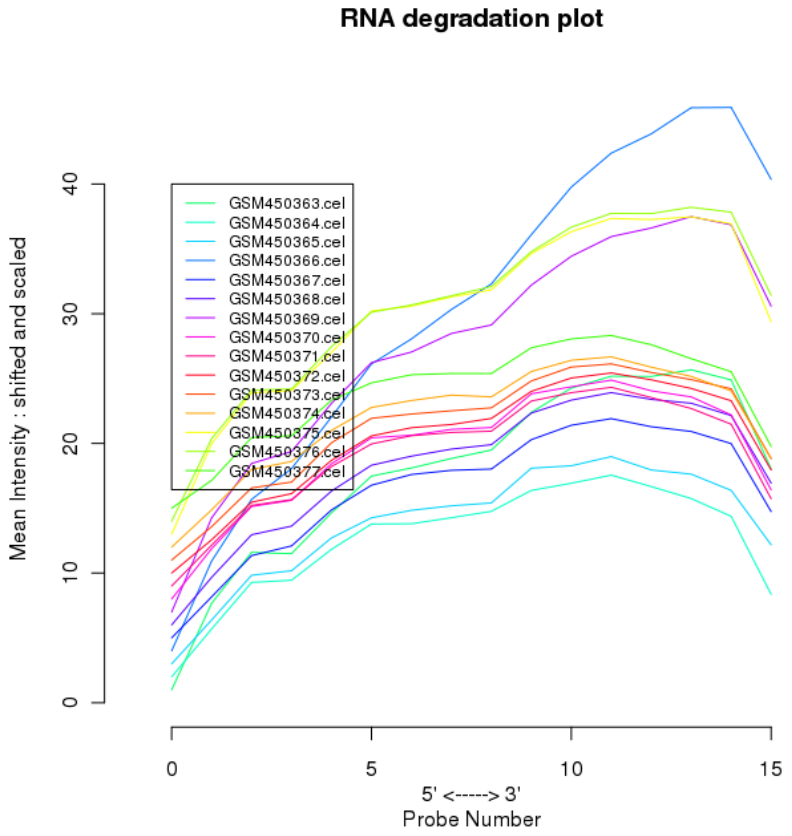
16-20 probe pairs: HG-U95a
11 probe pairs: HG-U133

数据来源

我们首先从欧洲的ArrayExpress数据库（通过关键词zebrafish/heart搜索数据)，然后下载原始（raw data）数据，所需的斑马鱼心脏再生的数据库ID值是：E-GEOD-17993。

RNA有没有降解？





Code Chunk:

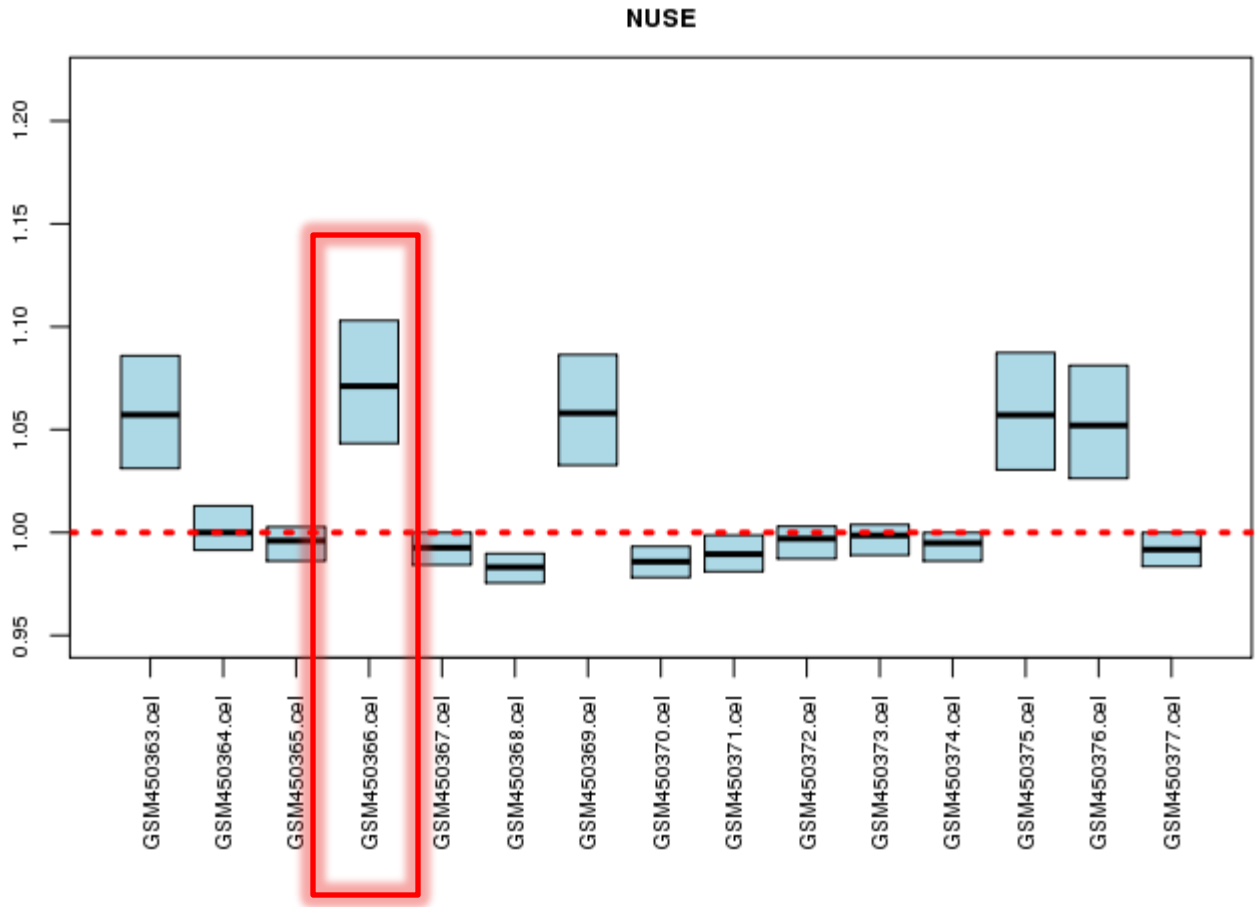
```
library(affyQCReport);
zebrafishQC=qc(raw.data);
plot(zebrafishQC);
```

```
getRNADegradMap<-function(x) {
   rawArray.data<-x;
   degrad_data<-AffyRNAdeg(rawArray.data);
   summaryAffyRNAdeg(degrad_data);
   length=dim(exprs(rawArray.data))[2];
   par(mfrow=c(1,1));
   plotAffyRNAdeg(degrad_data,col=rainbow(length, start=.4, end=.3));
      legend(legend=sampleNames(rawArray.data),x=0,y=40,lty=1,col=rainbow(length,
start=.4, end=.3),cex=0.75);
}
```

Normalized Unscaled Standard Errors (NUSE)

$$NUSE(\hat{\theta}_{ig}) = \frac{SE(\hat{\theta}_{ig})}{med_i(SE(\hat{\theta}_{ig}))}$$

```
library(affyPLM);
dataPLM<-fitPLM(raw.data);
generateNUSE(dataPLM );
generateNUSE<-function(x) {
    dataPLM<-x;
    par(ps=8,las=2,mar=c(10,3,2,1)+ 0.1);
    boxplot(dataPLM, main="NUSE", ylim = c(0.95, 1.22),outline = FALSE,
col="lightblue", las=3, whisklty=0, staplelty=0);
    abline(h=1,col="red",lwd=3,lty=3);
```

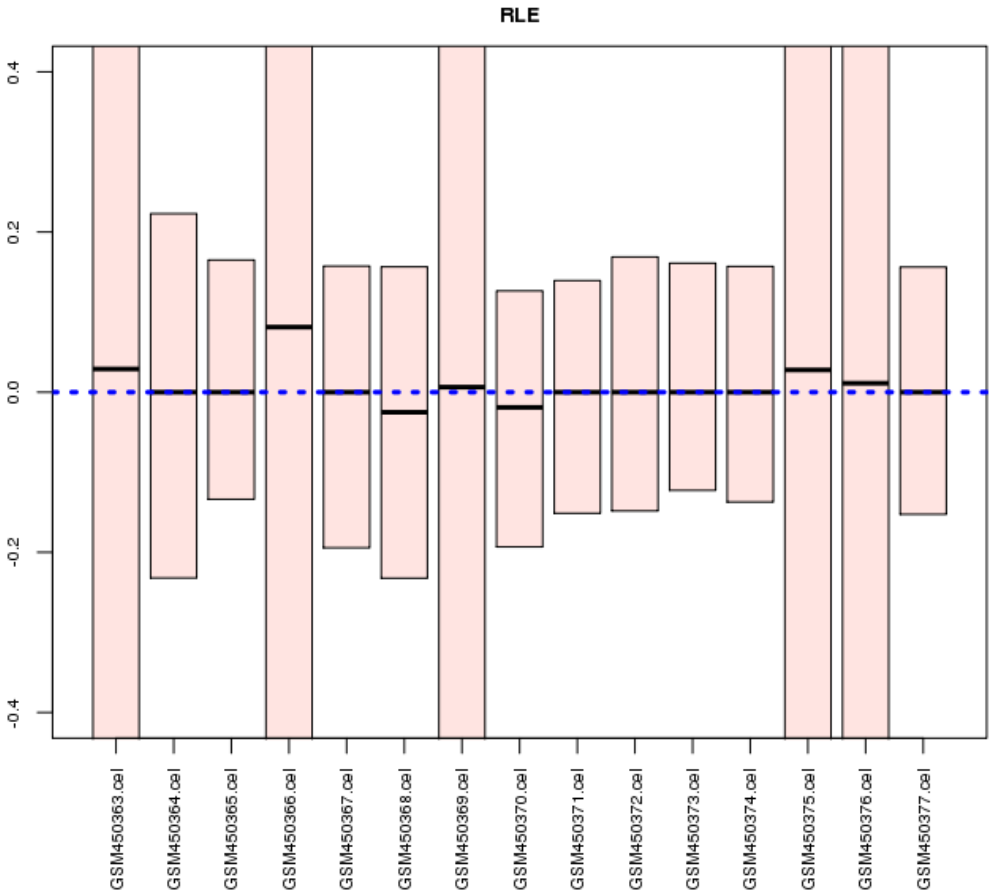

NUSE
1.20
1.15
1.10
1.05
1.00
0.95
GSM450363.cel
GSM450364.cel
GSM450365.cel
GSM450366.cel
GSM450367.cel
GSM450368.cel
GSM450369.cel
GSM450370.cel
GSM450371.cel
GSM450372.cel
GSM450373.cel
GSM450374.cel
GSM450375.cel
GSM450376.cel
GSM450377.cel

}

相对指数表达，Relative Log Expression (RLE)

```
generateRLE<-function(x) {
    dataPLM<-x;
    par(ps=8,las=2,mar=c(10,3,2,1)+ 0.1);
    Mbox(dataPLM, main="RLE", ylim = c(-0.4, 0.4), outline = FALSE,
col="mistyrose", las=3, whisklty=0, staplelty=0);
    abline(h=0,col="blue",lwd=3,lty=3);
}
```

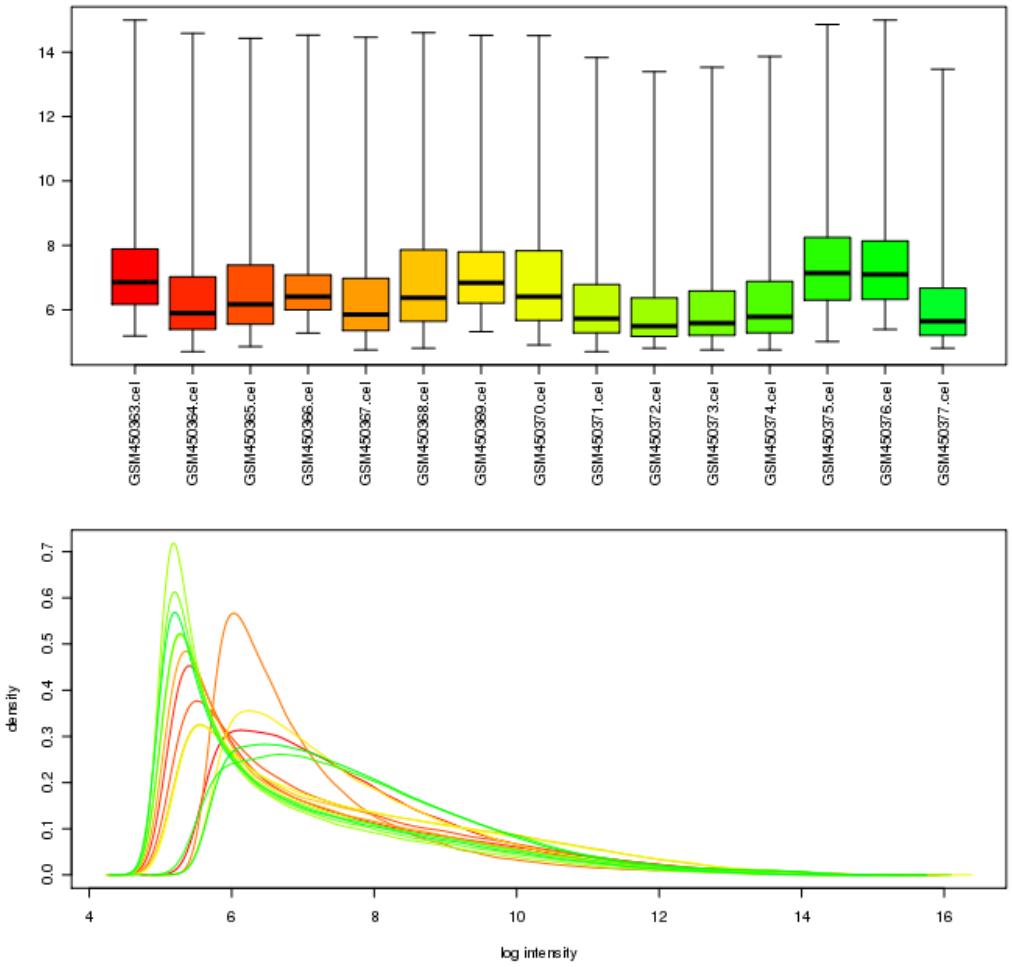
RLE图


RLE
0.4
0.2
0.0
-0.2
-0.4
GSM450363.cel
GSM450364.cel
GSM450365.cel
GSM450366.cel
GSM450367.cel
GSM450368.cel
GSM450369.cel
GSM450370.cel
GSM450371.cel
GSM450372.cel
GSM450373.cel
GSM450374.cel
GSM450375.cel
GSM450376.cel
GSM450377.cel

Affymetrix 表达芯片的预处理

- 背景校正(background correction)
- 芯片间归一化（between-array normalization）
- 基因综合(reporter summarization)

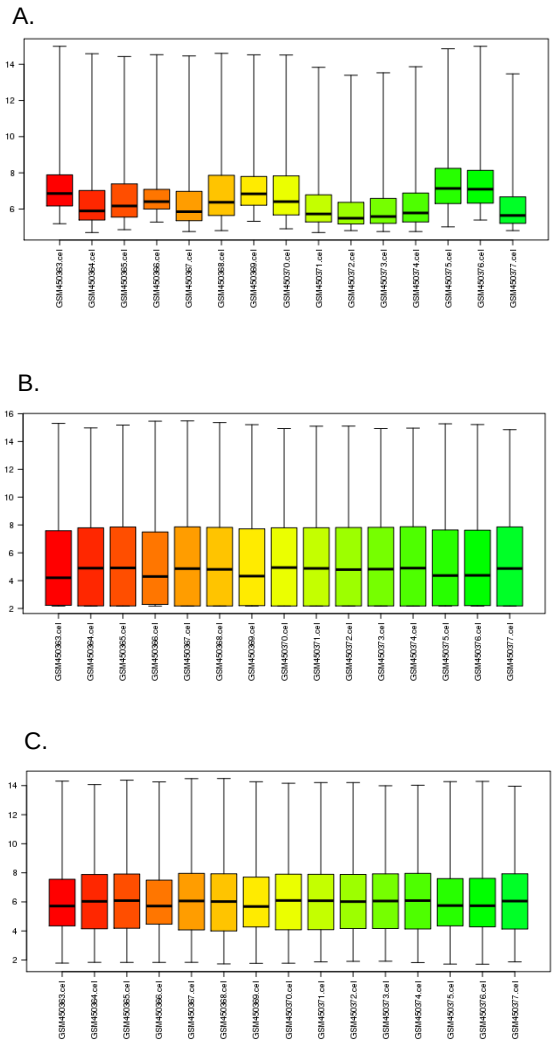
未归一化的原始数据

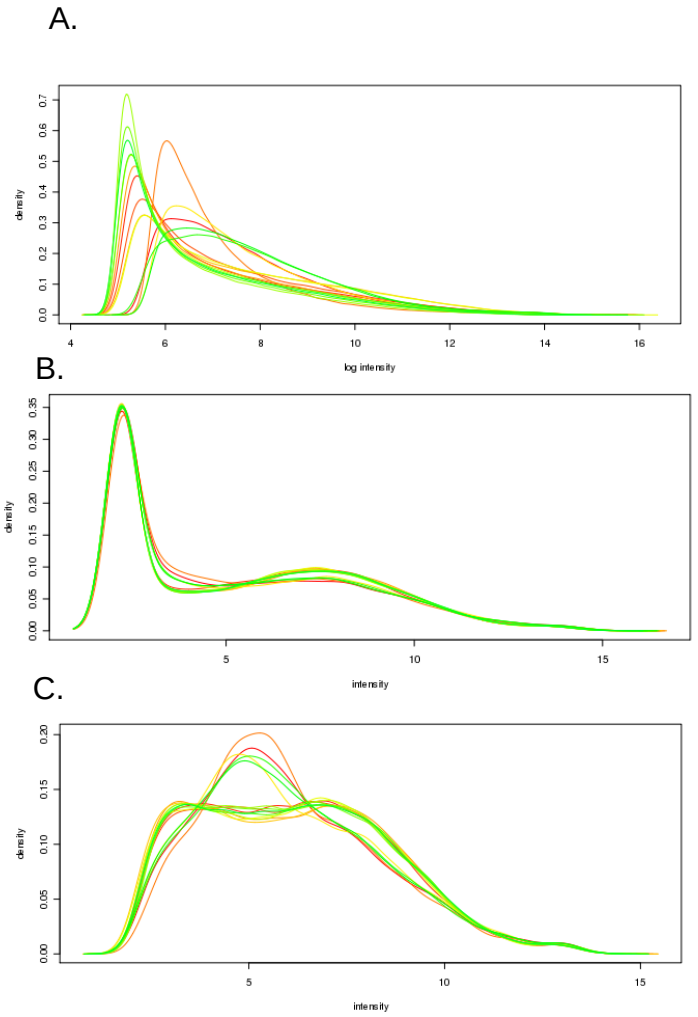


Code Chunk:

```
par(mfrow=c(2,1), cex=0.6);
boxplot(raw.data, col=rainbow(39), lty=1, las=2);
hist(raw.data, col=rainbow(39), lty=1);
```

RMA归一化方法 vs GCRMA归一化方法





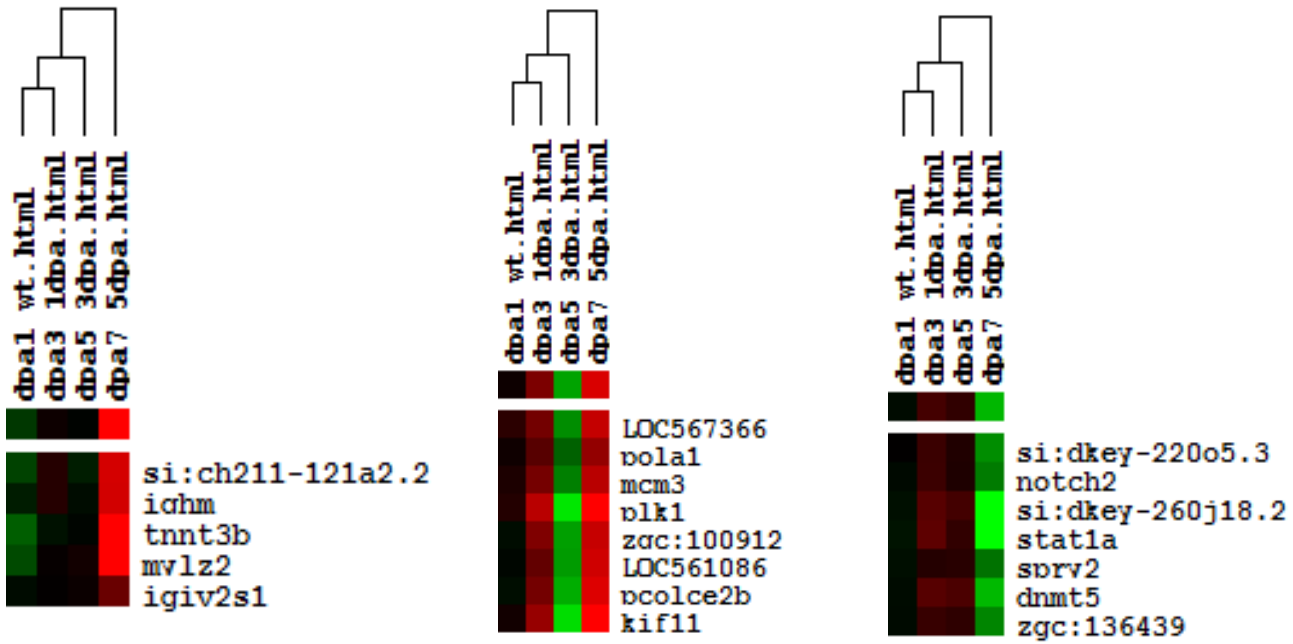
(A)未归一化的芯片结果（B）GCRMA归一化的芯片结果。（C）RMA归一化的结果。

Limma, to find differentially expressed genes

```
design<-model.matrix(~-1+factor(c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5)));

colnames(design)<-c("ControlHeart","Surgical1Day","Surgical3Day",\
"Surgical5Day","Surgical7Day");

contrast.matrix<-makeContrasts(Surgical1Day-ControlHeart,\
Surgical3Day-Surgical1Day,Surgical5Day-Surgical3Day,\
Surgical7Day-Surgical5Day,levels=design);

fit<-lmFit(filter.final.gcrmasubset,design);
fit2<-contrasts.fit(fit, contrast.matrix);
fit2<-eBayes(fit2);
```

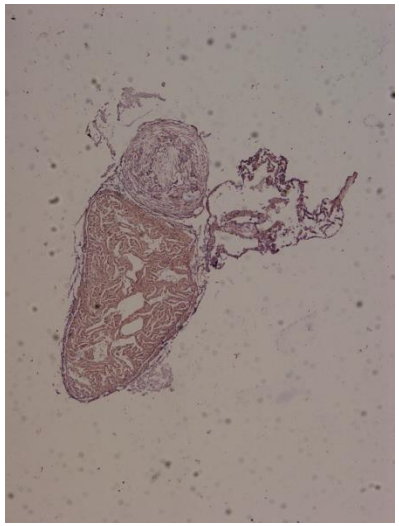
Limma是采用线性模型来分析芯片数据的。它需要两个矩阵：第一个是设计矩阵（design matrix），它指明了芯片和RNA样本之间的关系。例如如下语句：*design<-model.matrix(~-1+factor(c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5)));* 它的意思说明目前有五组芯片需要分析，并且两两间需要比较。每组芯片数据都有三个生物重复。第二个矩阵是比较矩阵（contrast matrix），说明我们希望所取得的两两比较。例如在如下语句中*contrast.matrix<-makeContrasts()*，我们希望得到手术1天后和对照组之间的比较结果，手术3天和手术1天的比较，手术5天和手术3天一级手术7天和手术5天芯片间的比较。

实验结果

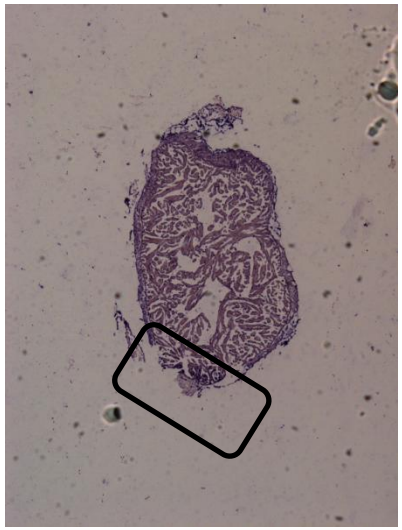


采用聚类分析软件cluster3.0得到的分析结果。我们采用的方法得到的结果与文献发表的结果大体一致（A）与心肌结构相关的一些基因基因在表达下降后又恢复的原有的水平，如tnnt3b。（B）参与细胞周期调节的基因如plk1显著上升。（C）notch信号通络的基因如Notch2表达量有一定的上调。

斑马鱼心脏Notch信号通路上调表达



正常心脏；Notch1在室壁小梁区域表达



手术心脏。2天左右的心脏切口区域Notch1有明显的上调表达

致谢

- 阜外医院教育部基因与临床重点室主任惠汝太教授
- 北京大学分子医学研究所心血管发育研究室熊敬维教授
- 中国R语言会议组委会