基因芯片数据的质量控制

基地一班 刘利生 201730152009

**质量控制**

基因芯片数据的质量控制对于后续的分析十分重要，质量控制主要集中在cel文件级别的处理，一共分为三个层次。

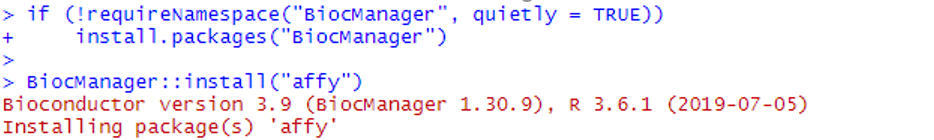
1. 直接查看灰度图；
2. 基于各种平均值指标的方法；
3. 高级的数据拟合的方法

对应的处理工具分别为：

1. image函数
2. simpleaffy包
3. affyPLM包

**导入cel格式的文件**

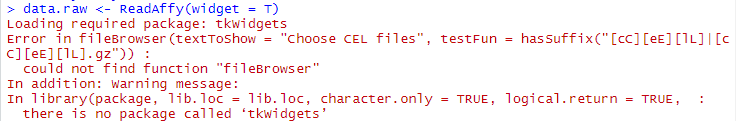
使用Affy包提供的ReadAffy函数



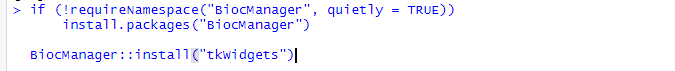
#下载affy包



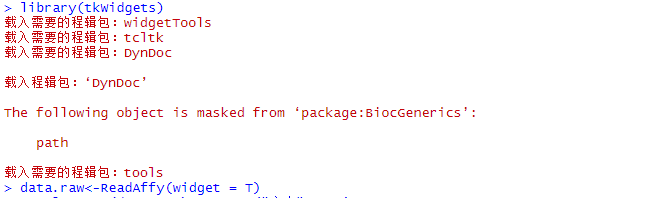
#载入affy包



#错误，提示本电脑未找到tkwidgets程辑包



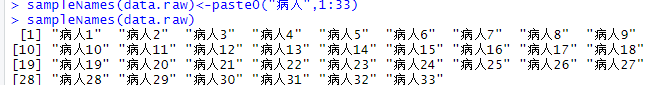
#下载tkwidgets程辑包



#通过ReadAffy函数选择导入所需cel文件，并赋值给data.raw

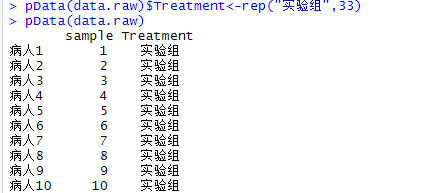
**对导入的文件进行处理**

修改表型相关数据，方便之后调用



#将导入的cel文件依次命名为病人1到33，

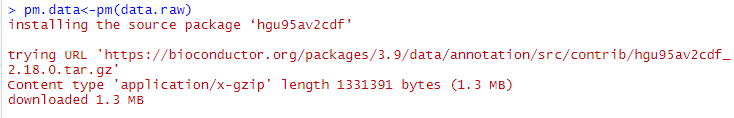
#查看数据名称，可见修改后的文件名称



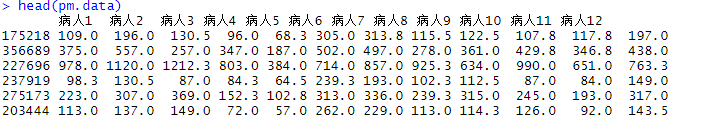
#使用pData函数将33个样本中的Treatment数据修改为实验组。

#调取整个芯片的表型相关数据，查看具体表型，可见所有Treatment数据修均改为实验组

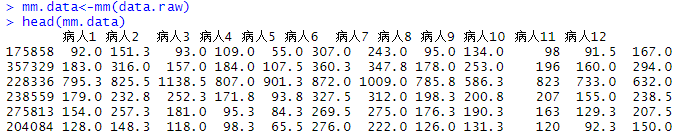
查看PM与MM的信息



#将data.raw中的PM（探针与基因完全匹配）数据赋值给pm.data



#使用head函数显示部分pm.data中的数据

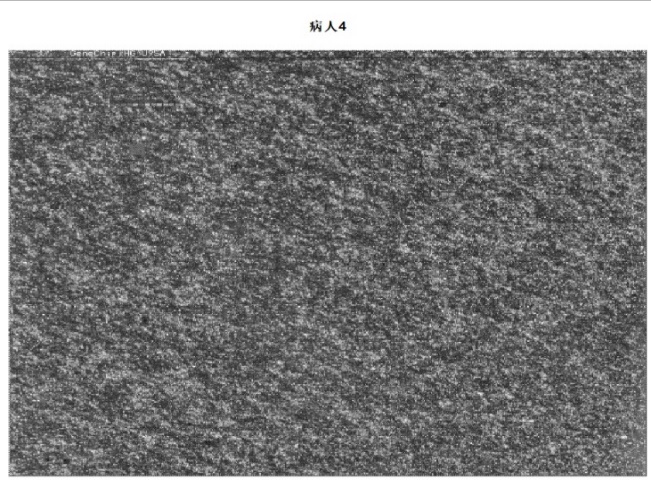
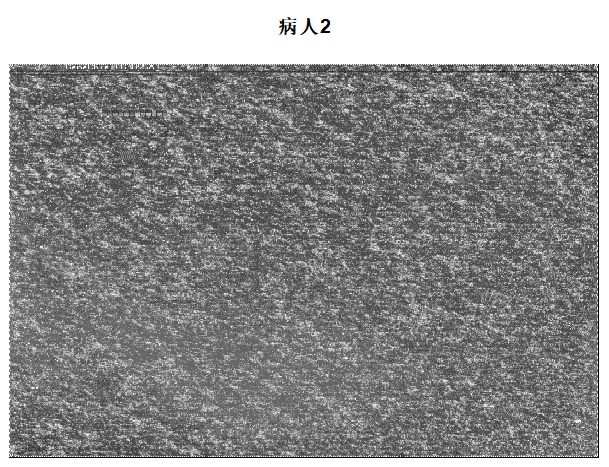
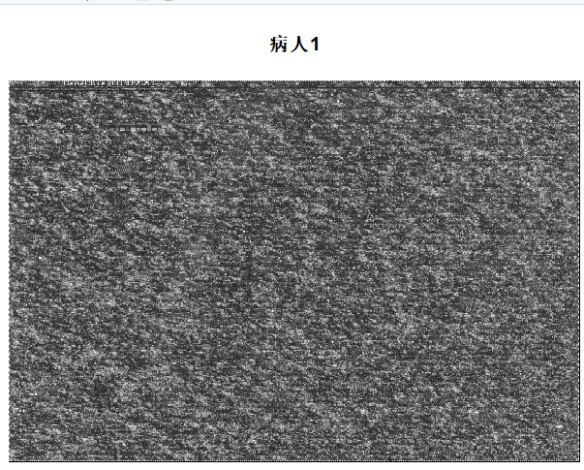
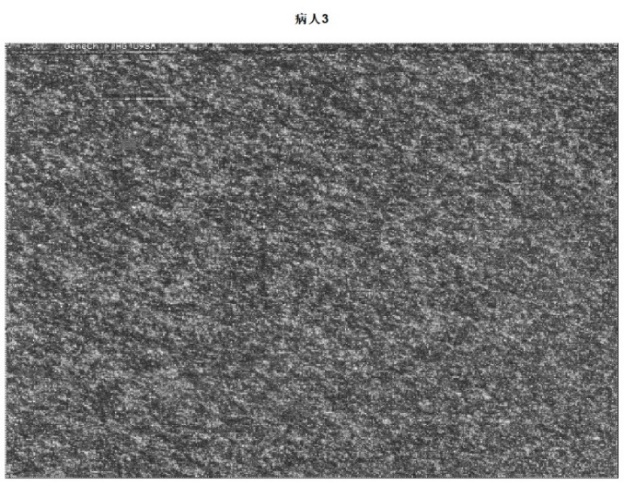
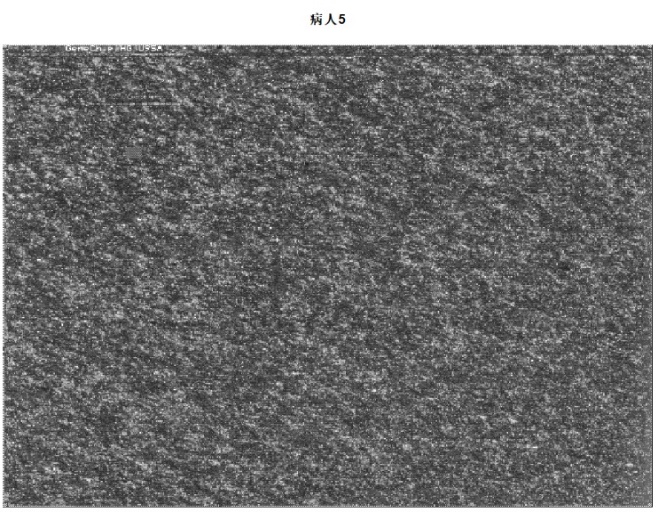


#使用mm函数查看mm探针（非特异性杂交）数据，对计算芯片的平均背景噪声有用。

**一、查看灰度图，粗略地估计芯片质量**



#设置调色板的颜色为灰色

#逐个作图，直接查看1到5号病人的灰度图，粗略地查看芯片质量

评价灰度图的两个标准：

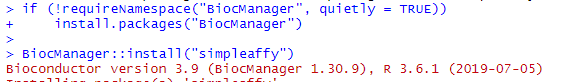
一、是否能分辨四角的花纹和芯片的名称，若不能，则数据可能有问题。

二、图像的明暗让我们对芯片信号强度有一个总体的认识，太暗了表示信号强度低，太亮了表示信号强度太高。

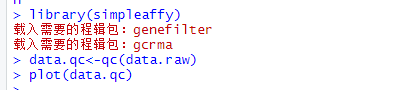
**分析：五个图片中均能分辨出四角的花纹以及在左上边缘的芯片名称，代表数据正常；对比来看，病人二的芯片图像比其他四个更为发白，表示信号强度低。**

**二、基于各种平均值的评估方法（simpleaffy包实现）**

包括尺度因子、检测值与检出率、平均背景噪声、标准内参，这些方法都基于一个假设：一次实验中所有芯片对于每个上述的平均值指标都相差不大，这也是这种方法的局限性。



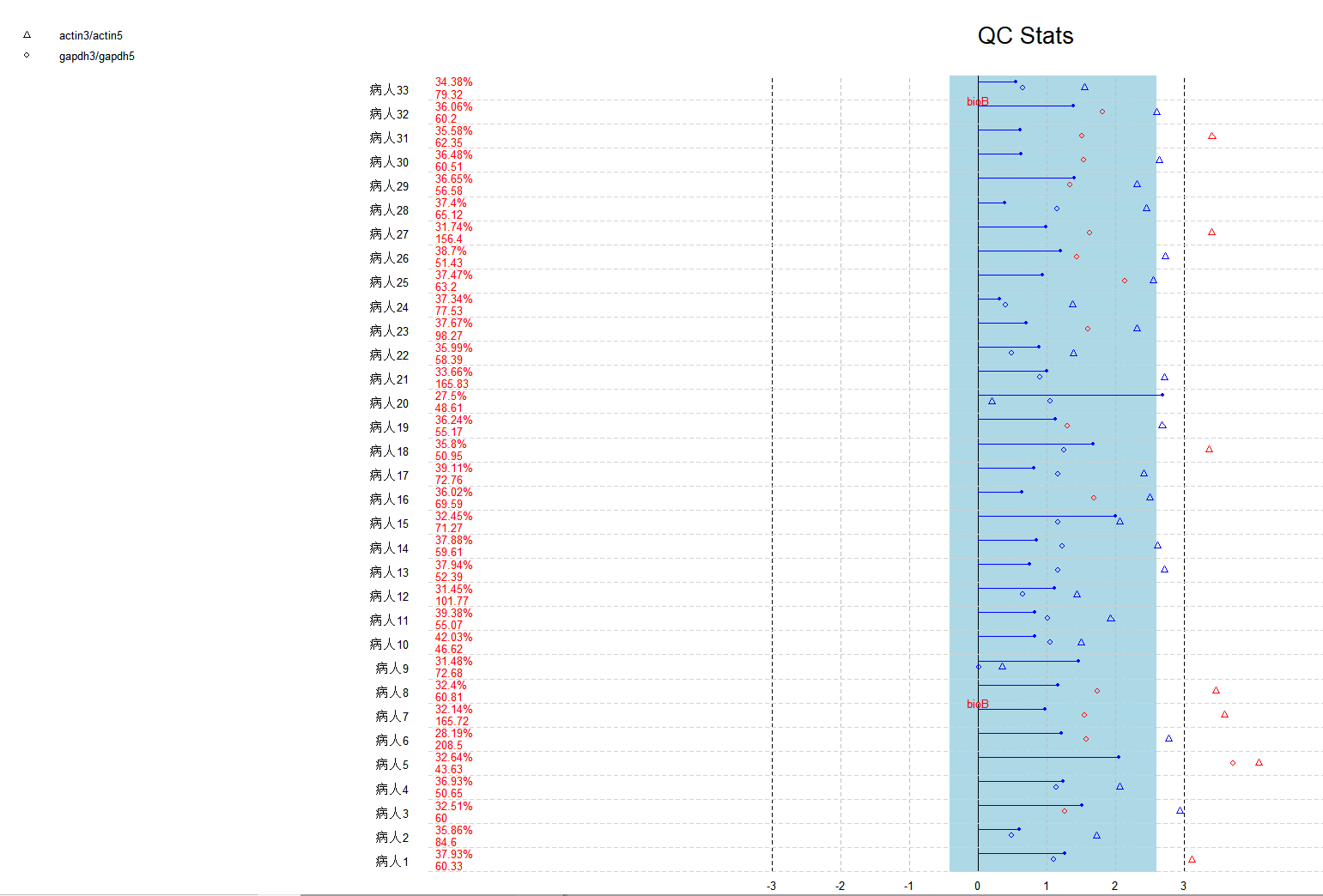
#下载simpleaffy包



#载入simpleaffy包

#通过qc函数进行数据的质量评估，并赋值给data,qc

#利用plot函数作图，得到质量控制总览图



第一列：样品名称

第二列：上面的百分数该芯片的检出率，下面的百分数为平均背景噪音。

第三列：用到了三个指标，分别是：尺度因子( scaling factor )、 标准内参GAPDH和actin的3’/5’的比值(示RNA的降解程度)。所有指标出现蓝色表示正常，红色代表可能存在质量问题。浅蓝色的竖条代表着尺度因子正常的取值范围，它会依照实验具体数据来计算出这个范围，通常它应该是在三倍以内。横轴所标记的数字就是尺度因子的座标。如果所有的一组需要相互比较的芯片间的尺度因子都落在了蓝色范围内，它会以蓝色线条及蓝色端点显示，图中所有病人的尺度因子均在三以内，表明这些芯片可以相互比较，如果标记为红色，那就意味着它们不能相互比较。

**分析：图中第7号与32号病人芯片出现红色的BioB字样，说明该样品嵌入探针未能检测到BioB，该芯片可能存在问题，数据不可靠。**

**actin和GAPDH 3’/5’比值 也分别以△和○表示出来。对于actin的3’/5’应该落在3以内，图中2、3、4、6号等病人符合要求，而对于GAPDH应该落在1左右。如果超过了设定的标准，就会以红色显示。如第3、5、6、7、8号等病人芯片存在RNA降解过多的问题。**

**三、高级的数据拟合方法**

实际情况下，每一块芯片的质量不一定是均匀的，所以还需要多种能反应芯片数据全貌的分析指标来得出最后的结论，并以图的方式表现。包括：1、权重&残差图、2、相对对数表达（RLE）箱线图、3,相对标准差(NUSE)箱线图、4.RNA降解曲线、5.聚类分析、6.主成分分析（PCA）图、7.信号强度分布图、8.MA图等，以上均通过affyPL,M包来实现的。

1.权重&残差图

理论上，芯片的权重与残差图是随机均匀分布的

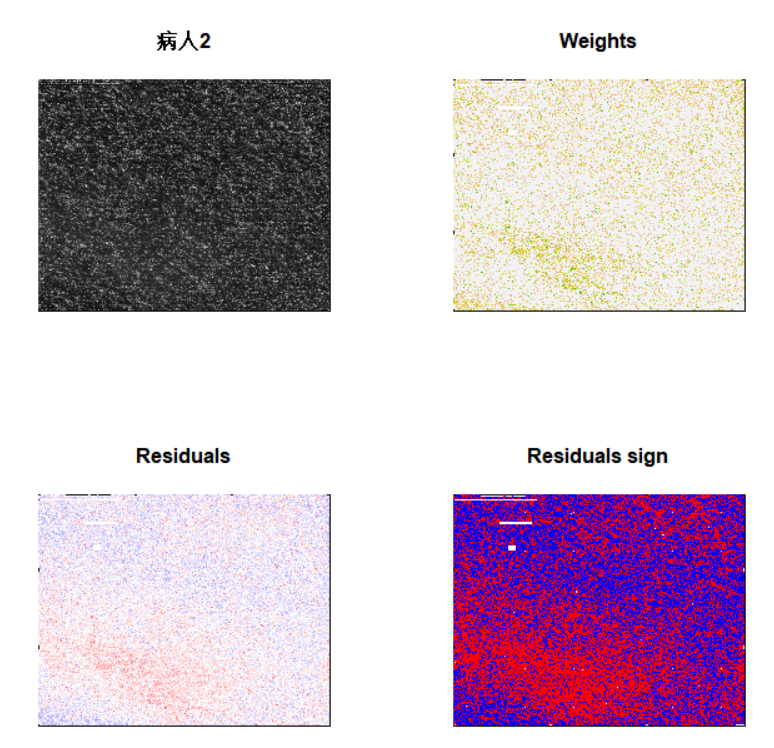




#导入affyPL,M包

#使用fitPLM做权重图、

#调整参数，行列各设置为二

#使用image函数，选取第二位病人的数据作图

右上、左上、右下、左下依次为灰度图，权重图、残差图、残差符号图；

权重图中，绿色代表较低的权重（接近0），白色、灰色代表较高的权重（接近1）

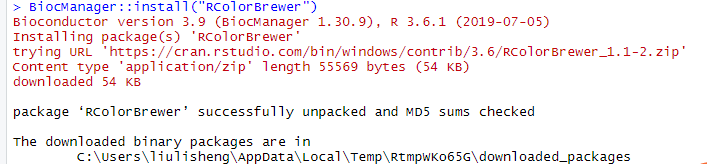
残差图中，红色代表正的高残差，白色代表低残差，蓝色代表负残差

sign.resids图使用红色表达正的残差，蓝色表达负的残差。

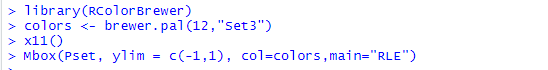
**分析：权重和残差都是随机分布的，所以应该看到的是绿色均匀分布的权重图，或者红蓝均匀分布的残差图，从Residuals sign中我们可得知这个此图红蓝分布不均匀，左下角红色分布密集，说明这张芯片的质量很可能有问题。**

2.相对对数表达（RLE）箱线图

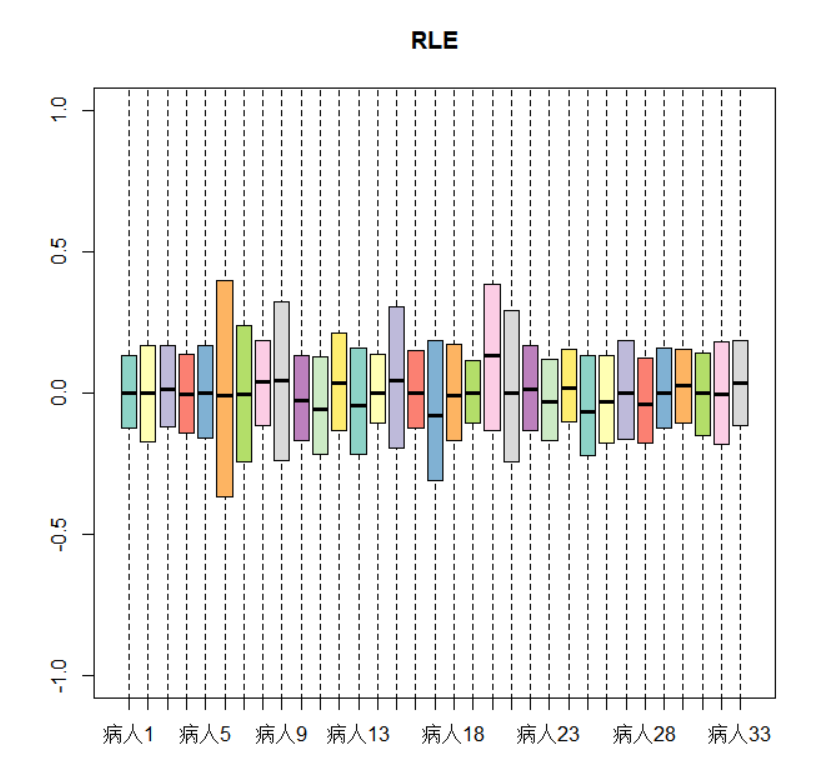
理想情况是每个样品的中心都接近0且各芯片的分布比较一致。



#下载RColorBrewer数据包



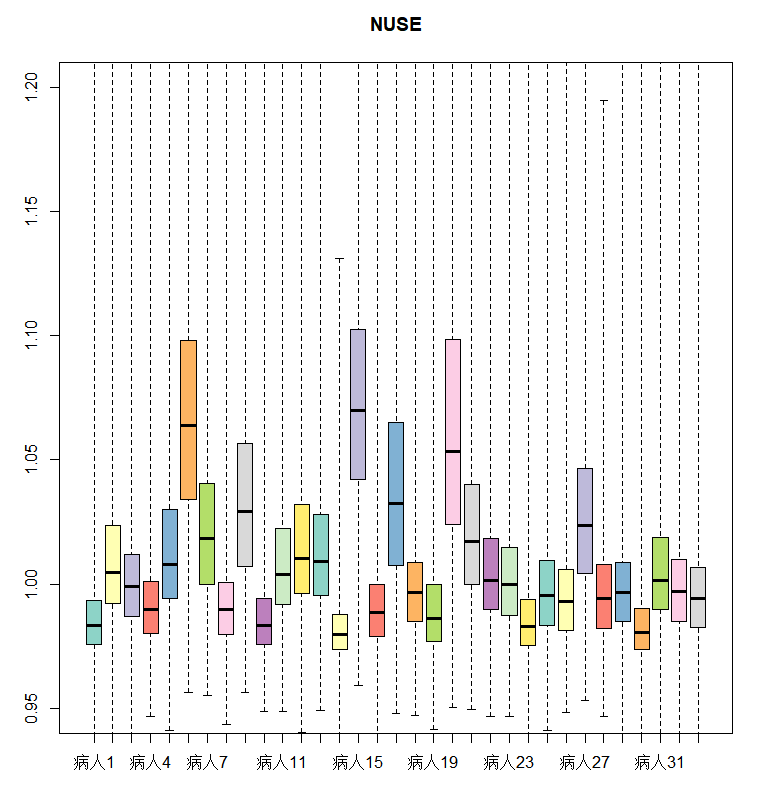
#设置相关颜色，作图



**分析：正常情况下，样本的中心应趋近0且分布较为集中，如病人1、2、3等，图中大部分样本数据正常，病人6、9、15、17、20、21样本数据分布过大、其中第9、15、20号数据中心偏离0，代表这些芯片可能存在问题，数据不太可靠。**

3.相对标准差(NUSE)箱线图

#作图



NUSE是某芯片基因标准差相对于全组标准差的比值

全组芯片都是质量可靠的话，那么，它们的标准差会十分接近，于是它们的NUSE值就会都在1左右如果有实验芯片质量有问题的话，它就会严重的偏离1，进而影响其它芯片的NUSE值偏向相反的方向。

**分析：图中病人6、15、20的相对标准差明显大于1且分布范围较广，代表其可能为存在质量问题的芯片。**

**结合RLE与NUSE箱线图，可以得出结论即第6、15、20号病人的芯片存在质量问题，数据不可靠，应当剔除。**

4.RNA降解曲线

RNA通常都是从5’端开始降解的，所以5’端的探针荧光强度应该低于3’端的荧光强度

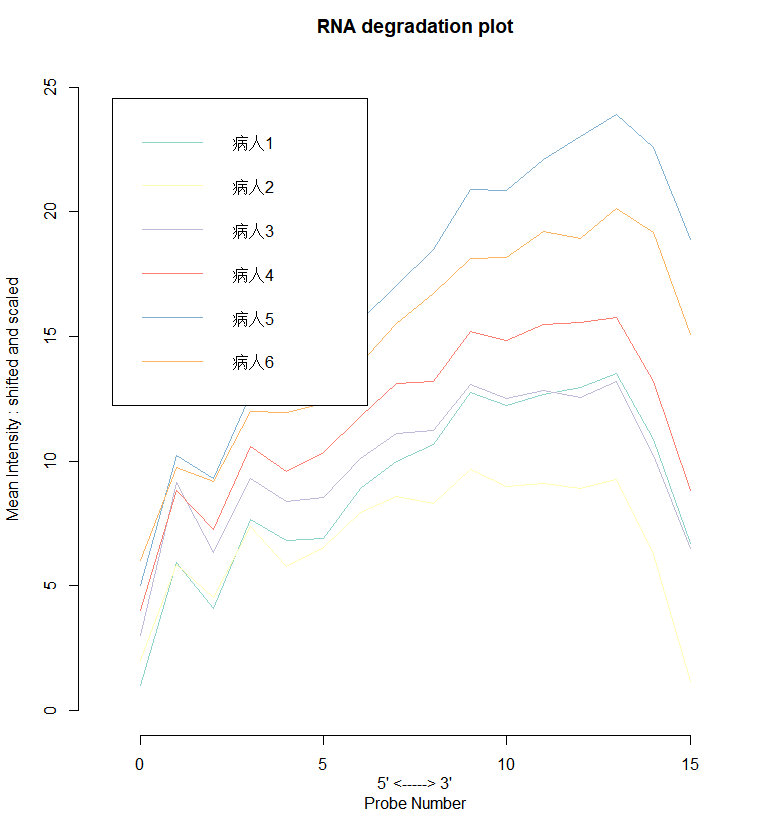




#使用AffyRNAdeg函数处理1至6号病人的相关数据，赋值定义为data.deg

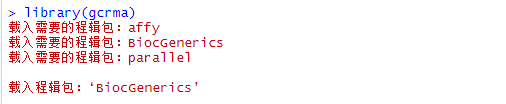
#使用plotAffyRNAdeg函数画出曲线图

#使用legend添加相关图例

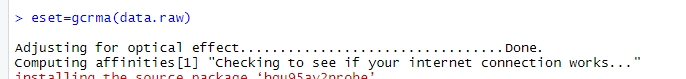


**分析：图像左边代表5’端探针信号强度，右边代表3’端信号强度。斜率越高代表RNA降解地越多，斜率接近0说明降解较少（一般不太可能）或者全部被降解，如图中病人2的曲线斜率较低，说明其可能存在RNA全部被降解的问题。**

5.聚类分析



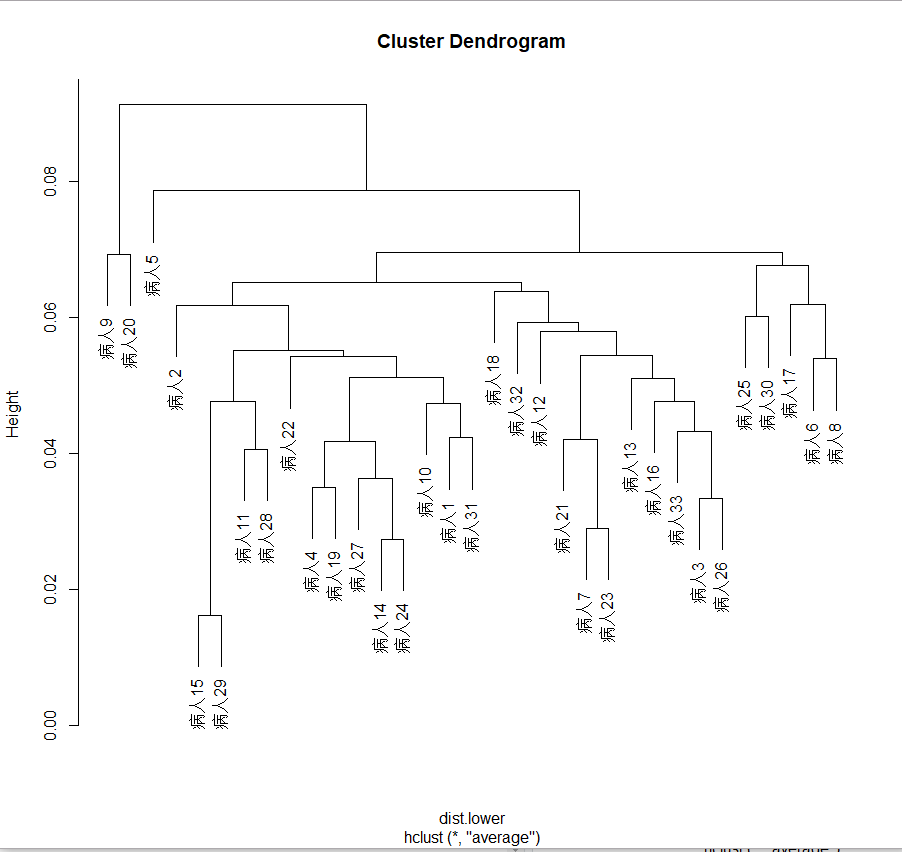
#载入gcrma程辑包







#作图



**分析：图中末端分支临近的病人代表其数据更为相近，如病人9与20、病人6与8等，由于数据中不包含对照组，故在此该聚类分析图意义不大。**