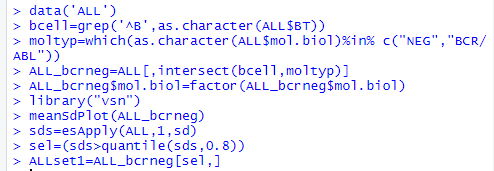
**基因的表达差异分析**

基地一班 刘利生 201730152009

**用t检验进行差异基因的筛选：**





library(“ALL”)

#导入ALL数据包

data("ALL")

bcell=grep(“^B”,as.character(ALL$BT))

#选取ALL$BT中开头为B的字符，即选出B型白血病

moltyp=which(as.character(ALL$mol.biol)%in% c("NEG","BCR/ABL"))

#输出ALL$mol.biol字符串中为"NEG","BCR/ABL"两类的字符的位置

ALL\_bcrneg=ALL[,intersect(bcell,moltyp)]

# intersect函数输出bcell与moltyp中的并集，即选取ALL中B型白血病的"NEG","BCR/ABL"两类病人的数据

ALL\_bcrneg$mol.biol=factor(ALL\_bcrneg$mol.biol)

#提取ALL\_bcrneg中mol.biol类型的数据因子

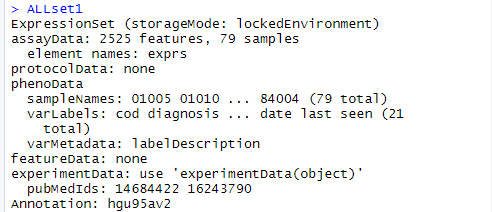
library("vsn")

meanSdPlot(ALL\_bcrneg)

sds=esApply(ALL,1,sd)

sel=(sds>quantile(sds,0.8))

用标准差的方法进行非特异性筛选，分离差是0.8

ALLset1=ALL\_bcrneg[sel,] 

分析：选取ALL\_bcrneg即B细胞白血病中的两个亚组：NEG与BCR/AB数据进行处理，最后筛选出标准差分离度大于0.8，得到的数据命名为ALLset1，输入ALLset1可以看到，它目前还是expressionset对象，有2525个基因数据，79个样本，而不再是处理前的33个样本。

t检验

需要用genefilter包中的rowttests函数





library(genefilter)

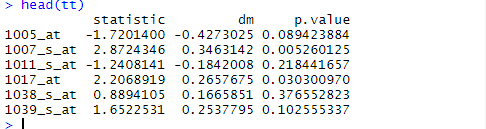
#载入genefilter程辑包

tt=rowttests(ALLset1,’mol.biol’)

#rowtests函数进行t检验，mol.biol代表因子水平，能自动识别一列中的因子个数，本次处理为NEG与BCR/AB两种，所以会进行两种样品的t检验。

names(tt)

#查看tt有哪些内容，可见其包含了statistic（t检验统计量）、dm（倍数值）、p.value（P值）三列

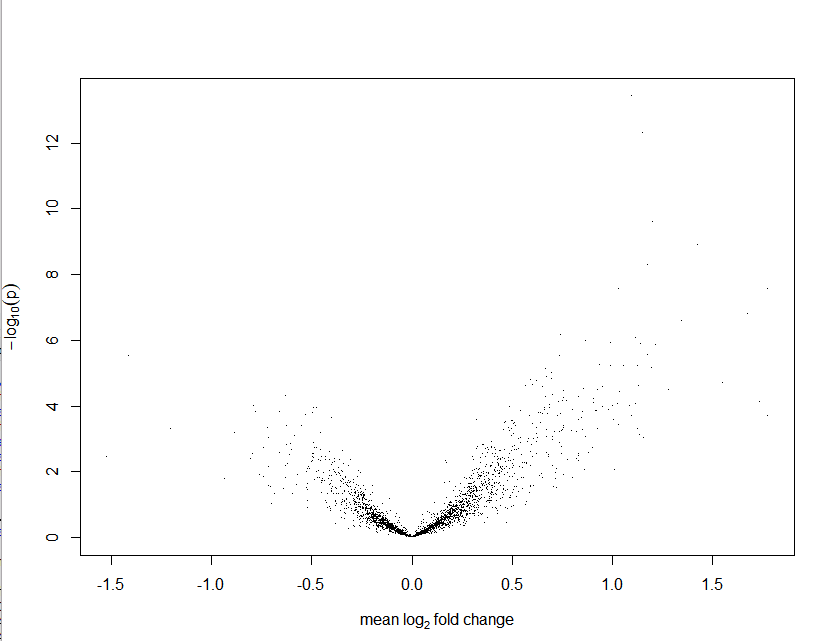


head(tt)

#查看前几个基因探针相关类型的数值，我们可以选择p=0.05来筛选所需要的差异基因

**制作非常简单的火山图**





plot(tt$dm,-log10(tt$p.value),pch=".",xlab=expression(mean~log[2]~fold~change),ylab=expression(-log[10](p)))

#横坐标为倍数以2为底的对数即dm，纵坐标为p值以10为底的负对数，

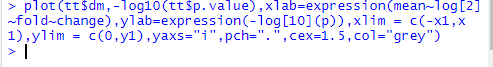
通过图表我们可以通过倍数来选择差异基因一个阈值，也可通过P值来选择差异基因，综合倍数法与t检验，克服各自单独使用存在的缺点（通过代码标出阈值、颜色表现出差异基因，直观）

改进：绘制带有阈值与颜色差异的火山图

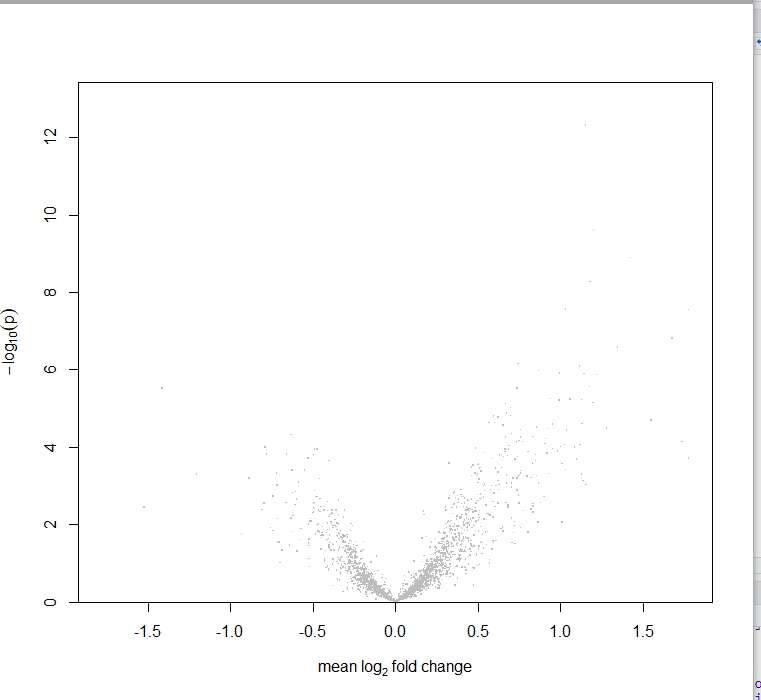


x1=max(abs(tt$dm))

y1=max(-log10(tt$p.value))



plot(tt$dm,-log10(tt$p.value),xlab=expression(mean~log[2]~fold~change),ylab=expression(-log[10](p)),xlim = c(-x1,x1),ylim = c(0,y1),yaxs="i",pch=".",cex=1.5,col="grey")



#绘制图表







tt1=subset(tt,tt$p.value<0.05&tt$dm>0.5)

tt2=subset(tt,tt$p.value<0.05&tt$dm<(-0.5))

#设置筛选的参数，即P值小于0.05，dm大于0.5和dm小于-0.5

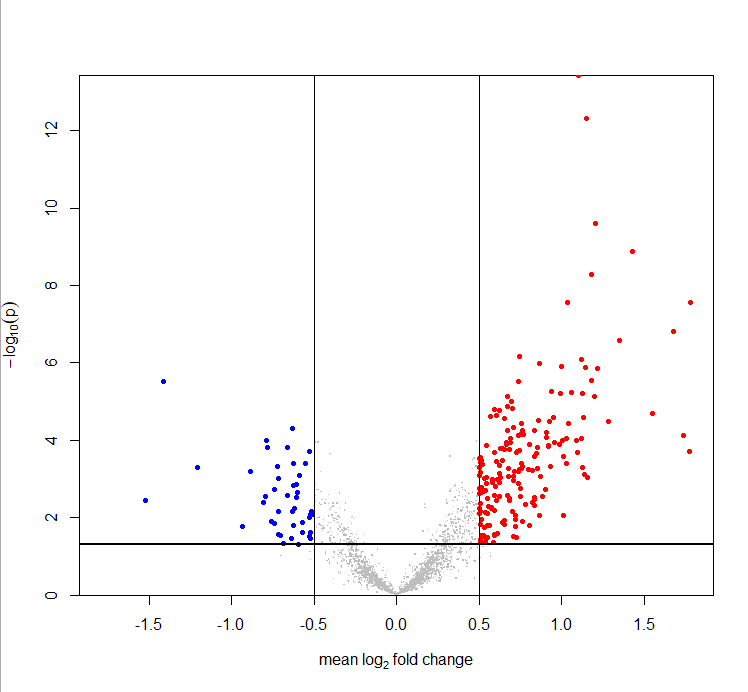
points(tt1$dm,-log10(tt1$p.value),pch=20,col="red",cex=1)

points(tt2$dm,-log10(tt2$p.value),pch=20,col="blue",cex=1)

#将符合tt1（即P值小于0.05 dm大于0.5）的点标记为红色，符合tt2（即P值小于0.05

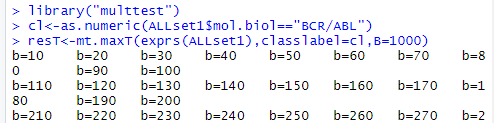
abline(h=-log10(0.05),lty=1,lwd=2,col="black")

abline(v=c(-0.5,0.5,lty=2,lwd=2,col="black"))



#绘制阈值线，y轴为- log10(0.05)即p=0.05，x轴为0.5、-0.5。

**多重比较检验**





library("multtest")

#载入multtest程辑包

cl<-as.numeric(ALLset1$mol.biol=="BCR/ABL")

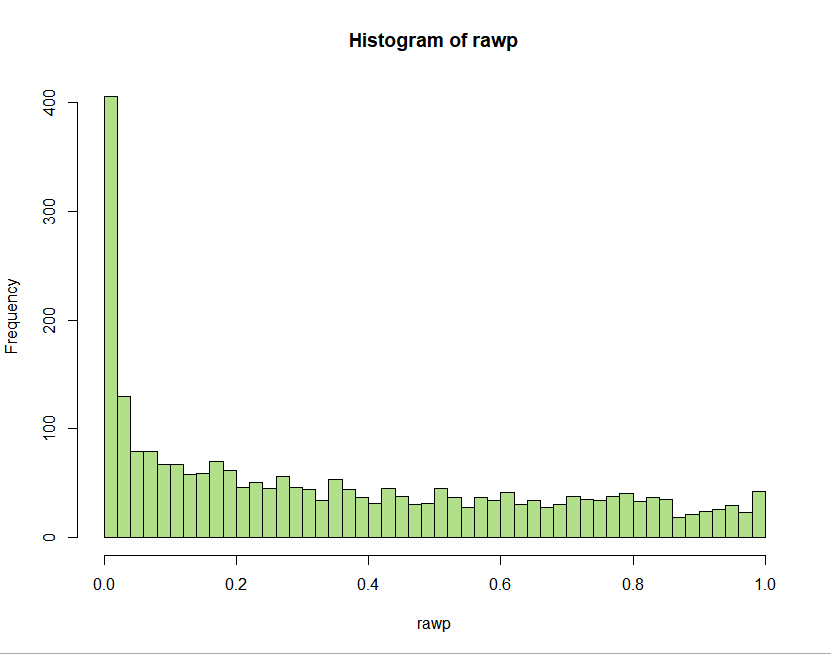
#将ALLset1中BCR/ABL类型的数据转化为数值型（numeric）

resT<-mt.maxT(exprs(ALLset1),classlabel=cl,B=1000)

ord<-order(resT$index)##the original gene order

rawp<-resT$rawp[ord]##permutation p-values

hist(rawp,breaks=50,col="#B2DF8A")



#p值分布图

图中可见大部分数据p值接近于0，数据普遍不大，代表前期数据筛选的结果具有一定效果。

当我们用t检验（或其他检验）时，设定显著性水平P=0.05（每100次检验就可能有5次是错误的），而芯片上的数据是数以万计的，筛选结果就会出现大量的假阳性。所以，需要进行多重比较校正。

**FWER校正**



sum(resT$adjp<0.05)

#查看FWER校正结果，输出结果34

**FDR校正**

res=mt.rawp2adjp(rawp,proc="BH")

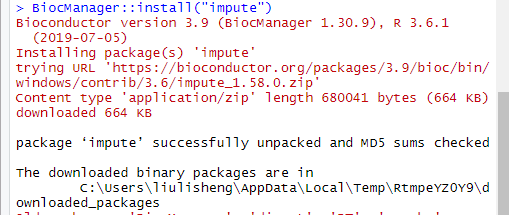
sum(res$adjp[,"BH" ]<0.05)

#查看FDR校正结果输出结果209

上述两种矫正方法代表经过各自的多重比较矫正算法分析后，得出筛选后具有特异性差异的基因个数分别为34与209个，

**用samr包进行差异基因的筛选：**





install.packages(c("matrixStats", "GSA", "shiny", "openxlsx"))

BiocManager::install("impute")

#下载安装包

**数据处理**





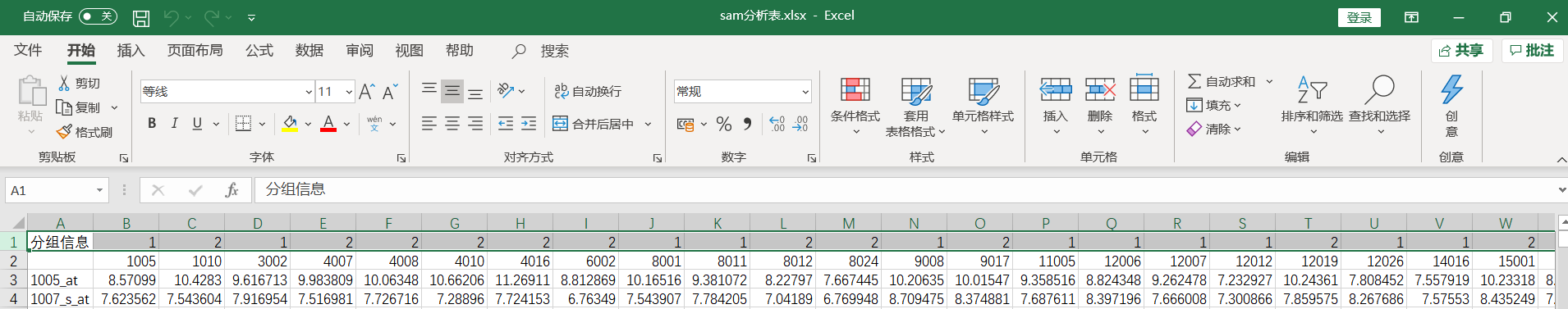
分组信息<-c(as.numeric(as.factor(ALLset1$mol)))

#分组信息，把它作为一组向量（c表示提取为向量），而exprs提取的才是矩阵。

sam<-rbind(exprs(ALLset1),分组信息)

#通过rbind按行组合，将分组信息加到ALLset1表达矩阵中，并命名为sam





write.csv(sam,file="sam分析表.csv")

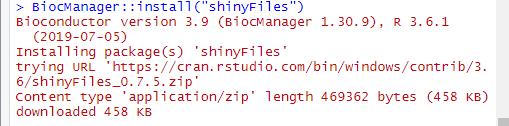
#先将表输出为csv文件（文件名为sam分析表.csv）

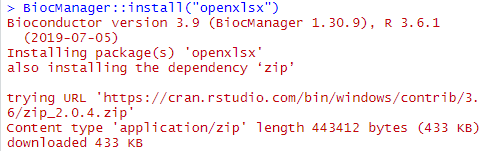
getwd()

#查看文件所在目录，用excle打开

之前加入的分组信息在这个表的最后一行，需要手动把它移动到第一行，然后再另保存为xlsx格式。

**调取sam的交互式页面**

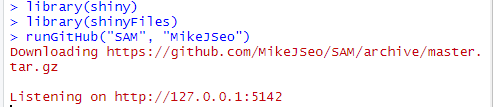




BiocManager::install("shinyFiles")

BiocManager::install("openxlsx")

#下载相关程辑包

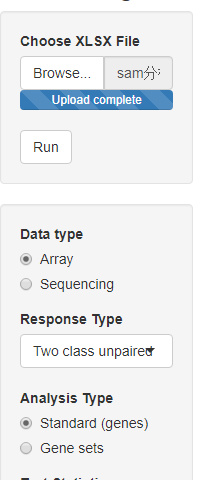
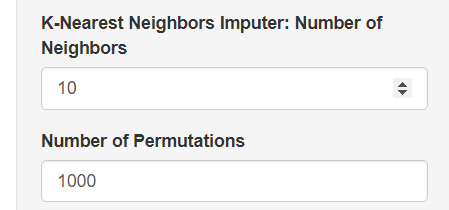


library(shiny)

library(shinyFiles)

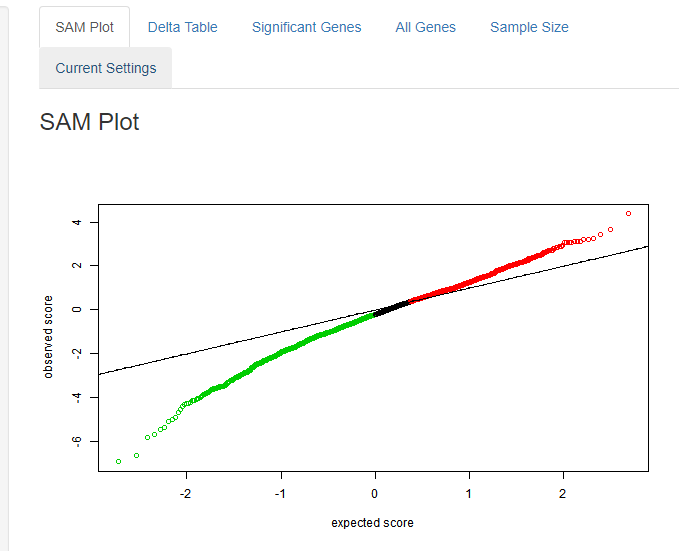
runGitHub("SAM", "MikeJSeo")

#启动SAM交互式页面

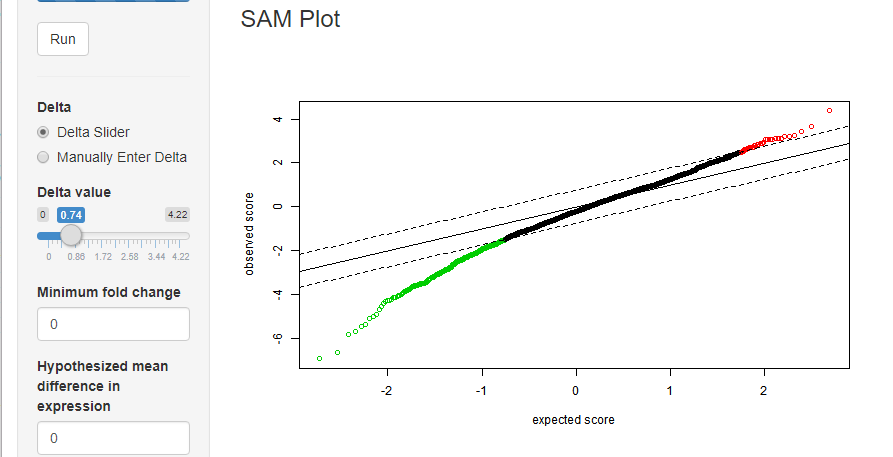


#上传处理好的xlsx文件，设置相关参数类型，Response Type选择两个非配对的独立样本，排列组合(permutation)次数设置为1000。

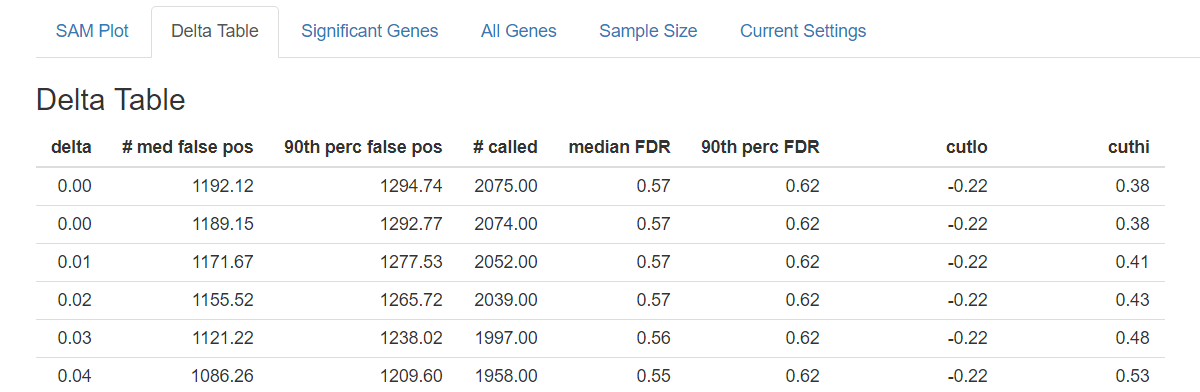
点击RUN



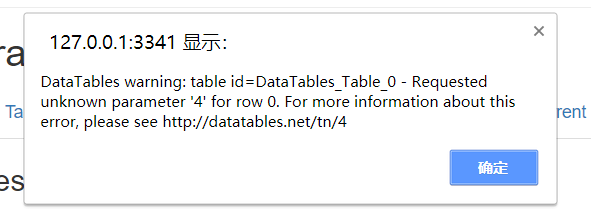
#横坐标代表期望值，纵坐标代表观察值，红色位于斜线上方，表示观察值数据大于期望值代表上调基因，绿色表示观察值数据小于期望值，代表下调基因。



#通过调整delta值（阈值）来调整差异基因范围，delta数值越大，则无差异表达的基因分布越广，差异基因越少。



#Delta表，表示不同delta值对应的一系列参数

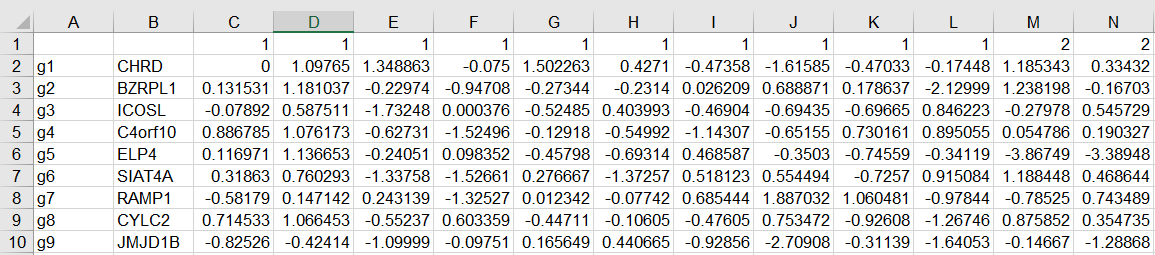


提示出错，DataTables warning: table id=DataTables\_Table\_0 - Requested unknown parameter '4' for row 0. For more information about this error, please see <http://datatables.net/tn/4>

错误分析：table id 为DataTables\_Table\_0-请求的未知参数'4'为第0行。

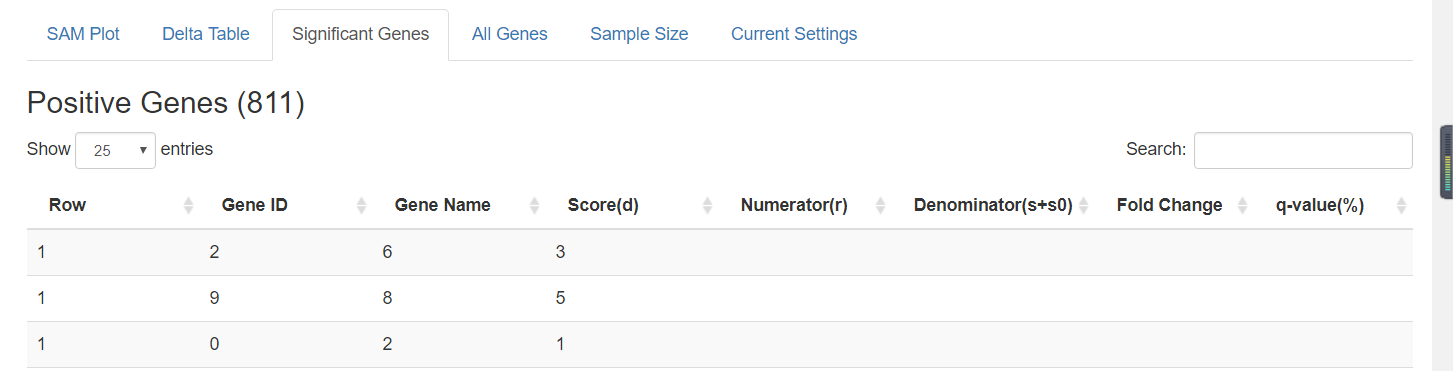
1.初步分析可能为文件格式出错，但将文件格式改成xls发现依旧不能解决该问题，且目前的SAM分析仅支持xlsx格式的文件，故排除这一原因。

2.再次分析可能由于表格数据中未包含相关基因注释名称造成参数错误，但在官网下载参考数据文件（格式如下）

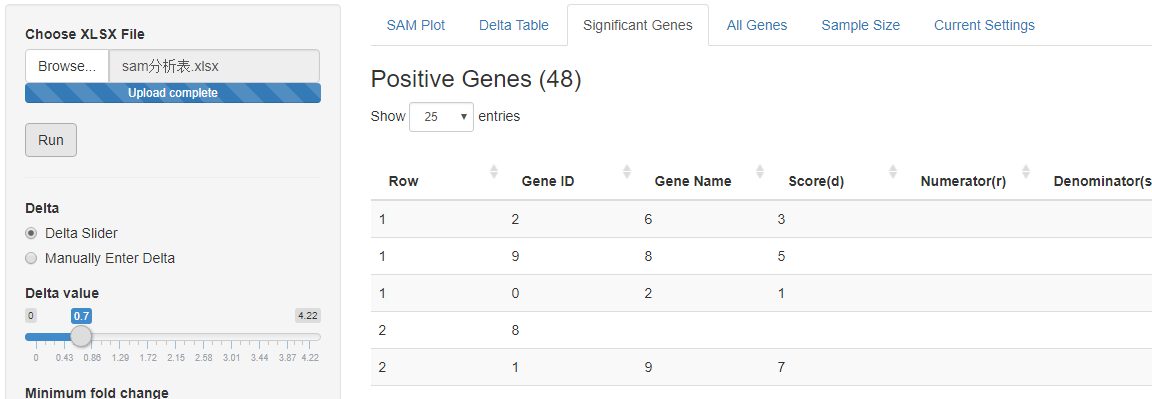


提交检验时依旧报错，则表示此次错误与文件本身格式无直接联系。

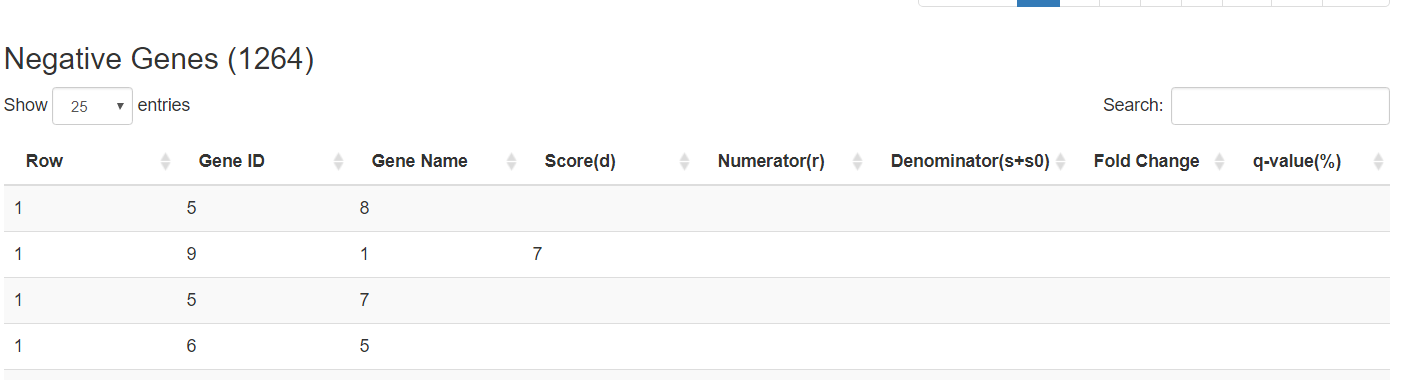
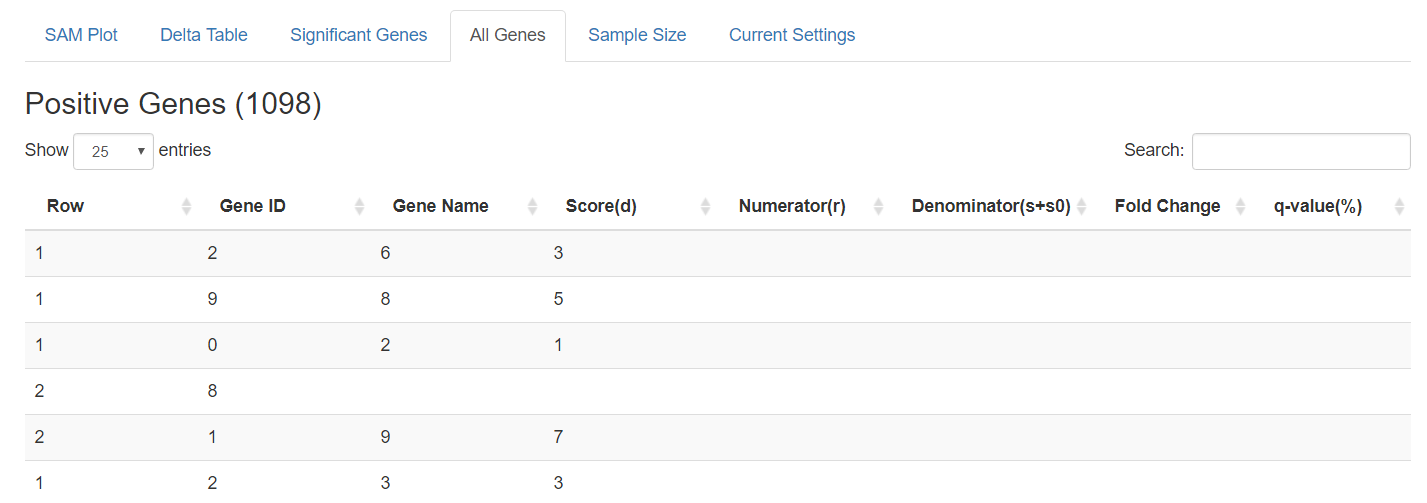
3.进一步了解得知SAM类似于内嵌了data table这一基于jQuery框架的临时保存数据的网格虚拟表插件，分析可能为data table更新后SAM未能及时更新相关算法，导致数据处理过程中出错。



#查看当前delta值（0）筛选出来的差异基因，其中q-value（类似于t值）越小代表差异越明显

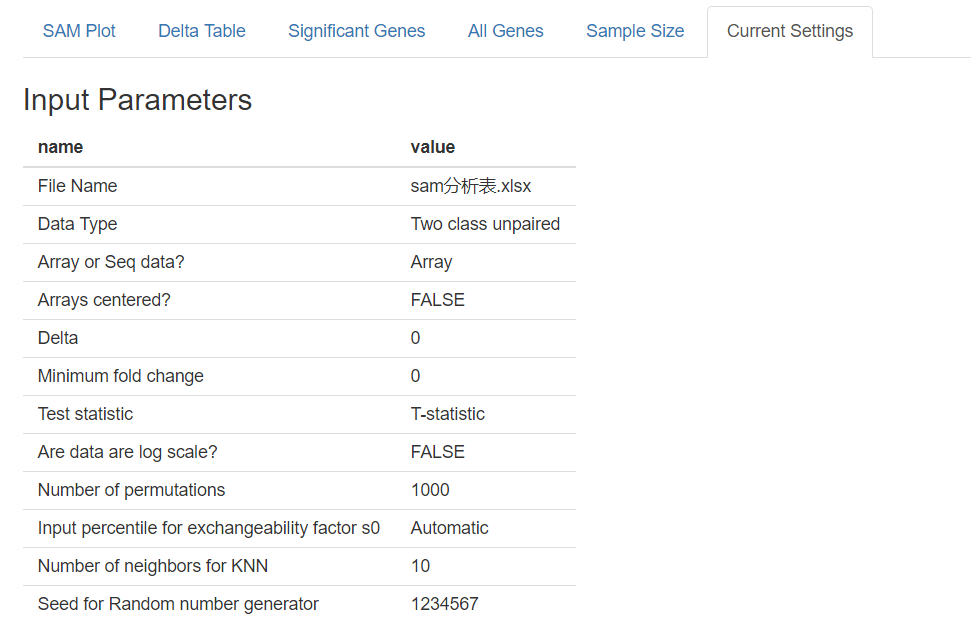


#将delta值改为0.7后，筛选出来的差异基因个数为48。



#查看所有基因信息，positive genes代表阳性、negetive genes为表达差异不显著的基因





#查看当前相关参数，例如文件名为sam分析表.xlsx，数据类型（Array）、Delta值（0）、是否经过log对数处理（FALSE）

#设置保存路径与文件名（result），输出保存结果。