人类X染色体基因表达的生物信息学分析实验

**摘要**

人类X染色体是一个相对基因丰富的染色体，本实验对人类X染色体基因进行生物信息学的分析。对人类X染色体基因样本进行差异表达分析，发现在不同地区不同性别之间人类X染色体的基因表达差异并不大，但其转录本阶段却十分不同，很有可能和性别有关，但最后基因表达仍没有太大差异，应该是在翻译过程中不同性别的机制也有所不同；对人类X染色体基因样本进行聚类分析和PCA分析，发现同一地区男性X染色体基因表达差异不大，但是女性的样本差异却比较大，这有可能是男性X染色体只来源于母亲，而女性的X染色体即来源于父亲也来源于母亲的缘故，并且其中一条X染色体失活可能造成了更多的变异。

**介绍**

X染色体是XY型性别决定生物染色体组中的一种特殊的性染色体，人类X染色体碱基覆盖达151Mb，占整个基因组的5%。对一般人类来说，女性的一对性染色体是两条大小，形态相似的X染色体，男性则X、Y染色体各有一条。本次实验数据来源于2016年发表在Nature Protocols上的《Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown》，从英格兰岛和约鲁巴住民的12个X号染色体样本中选取8个样本来进行本次生物信息学分析。

差异表达分析通常作为根据基因表达矩阵进行生物信息学分析的第一步，有助于我们观察基因在不同样本中的表达差异，从而研究差异表达基因的生物学意义。本次实验是比较两个变量（性别、地区）影响下的人类X染色体基因表达差异，通过一定的统计学方法，过滤低丰度基因，从中识别出与变量条件相关的特异性差异表达基因，作图，然后进一步分析这些特异性基因。

聚类分析用于判断差异基因在不同实验条件下的表达模式，将表达模式相同或相近的基因聚集成类，进而识别基因功能，同类基因可能具有相似的功能。对不同条件下多个基因的表达进行聚类分析可以帮助快速的选择共表达基因。正确的聚类分析，不但有助于推断基因的功能，还可以有效的发现基因之间存在的关系，距离度量是聚类方法的核心。这里的距离度量，是指用来衡量两个基因的表达模式之间的相似程度。本实验对人类X染色体差异表达基因进行聚类分析，用欧式距离(Euclidean distance )评估样本间关系(即总体相似度)，用热图展示样品表达模式及样品间关系。

主成分分析 ( Princ ipal Component Analysis ， PCA ) 是一种掌握事物主要矛盾的统计分析方法，它可以从多元事物中解析出主要影响因素，揭示事物的本质，简化复杂的问题。PCA分析帮助我们确定影响人类X性染色体基因差异表达的关键变量，进一步解释基因表达模式，并能利用散点图实现多维数据可视化。

**实验数据来源**

本次实验数据来源：上次韩老师做差异表达分析的X号染色体相关文件

基因注释文件：chrX.gtf

基因序列文件：chrX.fa

已经建好的索引文件夹：indexes（chrX\_tran.1.ht2 chrX\_tran.3.ht2 chrX\_tran.5.ht2 chrX\_tran.7.ht2 chrX\_tran.2.ht2 chrX\_tran.4.ht2 chrX\_tran.6.ht2 chrX\_tran.8.ht2）

样本文件夹：samples（ERR188\*.fastq.gz ...）

本次从12个样本中选取8个尝试课程所学的差异表达分析，聚类分析及PCA主成分分析。

8个samples：

geuvadis\_phenodata1.csv

|  |  |  |
| --- | --- | --- |
| ids | sex | population |
| ERR188044 | male | YRI |
| ERR188104 | male | YRI |
| ERR188234 | female | YRI |
| ERR188337 | female | GBR |
| ERR188383 | male | GBR |
| ERR188401 | male | GBR |
| ERR188428 | female | GBR |
| ERR204916 | female | YRI |

**实验方法步骤**

1. 差异表达分析：

1.hisat2 alignment:

for i in {188044,188104,188234,188337,188383,188401,188428,204916}; do

hisat2 -p 8 -x chrX\_data/indexes/chrX\_tran -1 chrX\_data/samples/ERR${i}\_chrX\_1.fastq.gz -2 chrX\_data/samples/ERR${i}\_chrX\_2.fastq.gz -S ERR${i}\_chrX.sam

done

2.samtools sort:

for i in {188044,188104,188234,188337,188383,188401,188428,204916}; do

samtools sort -@ 8 -o ERR${i}\_chrX.bam ERR${i}\_chrX.sam

done

3.stringtie 计算表达量，生成gtf文件

for i in {188044,188104,188234,188337,188383,188401,188428,204916}; do

stringtie ERR${i}\_chrX.bam -p 8 -G chrX\_data/genes/chrX.gtf -l ERR${i} -o ERR${i}\_chrX.gtf

done

4.stringtie merge，合并

stringtie --merge -p 8 -G chrX\_data/genes/chrX.gtf -o stringtie\_merged.gtf chrX\_data/mergelist.txt

mergelist.txt内容：

ERR188044\_chrX.gtf

ERR188104\_chrX.gtf

ERR188234\_chrX.gtf

ERR188337\_chrX.gtf

ERR188383\_chrX.gtf

ERR188401\_chrX.gtf

ERR188428\_chrX.gtf

ERR204916\_chrX.gtf

5.用stringtie生成ballgown文件夹

for i in {188044,188104,188234,188337,188383,188401,188428,204916}; do

stringtie ERR${i}\_chrX.bam -p 8 -G stringtie\_merged.gtf -o ballgown/ERR${i}/ERR${i}\_chrX.gtf -e -B

done

6.R语言处理：R

library(ballgown)

library(RSkittleBrewer)

library(genefilter)

library(dplyr)

library(devtools)

# 读取实验设计的csv文件

pheno\_data = read.csv("geuvadis\_phenodata1.csv")

# 读取ballgown文件

bg\_chrX = ballgown(dataDir = "ballgown", samplePattern = "ERR", pData=pheno\_data)

# 过滤低丰度基因

bg\_chrX\_filt = subset(bg\_chrX,"rowVars(texpr(bg\_chrX)) >1",genomesubset=TRUE)

# 组间有差异的转录本及基因

results\_transcripts = stattest(bg\_chrX\_filt,feature="transcript",covariate="sex",adjustvars =c("population"), getFC=TRUE, meas="FPKM")

results\_genes = stattest(bg\_chrX\_filt,feature="gene",covariate="sex",adjustvars =c("population"), getFC=TRUE, meas="FPKM")

# 添加基因名和id

results\_transcripts =data.frame(geneNames=ballgown::geneNames(bg\_chrX\_filt),geneIDs=ballgown::geneIDs(bg\_chrX\_filt), results\_transcripts)

# 按P值排序

results\_transcripts = arrange(results\_transcripts,pval)

results\_genes = arrange(results\_genes,pval)

# 把结果写入csv文件中

write.csv(results\_transcripts, "chrX\_transcript\_results.csv",row.names=FALSE)

write.csv(results\_genes, "chrX\_gene\_results.csv",row.names=FALSE)

# 筛选q值小于0.05的结果

subset(results\_transcripts,results\_transcripts$qval<0.05)

subset(results\_genes,results\_genes$qval<0.05)

#做图

tropical= c('darkorange', 'dodgerblue','hotpink', 'limegreen', 'yellow')

palette(tropical)

fpkm = texpr(bg\_chrX,meas="FPKM")

fpkm = log2(fpkm+1)

boxplot(fpkm,col=as.numeric(pheno\_data$sex),las=2,ylab='log2(FPKM+1)')

ballgown::transcriptNames(bg\_chrX)[12]

ballgown::geneNames(bg\_chrX)[12]

plot(fpkm[12,] ~ pheno\_data$sex, border=c(1,2),main=paste(ballgown::geneNames(bg\_chrX)[12],' : ',ballgown::transcriptNames(bg\_chrX)[12]),pch=19, xlab="Sex",ylab='log2(FPKM+1)')

points(fpkm[12,] ~ jitter(as.numeric(pheno\_data$sex)),col=as.numeric(pheno\_data$sex))

plotTranscripts(ballgown::geneIDs(bg\_chrX)[1729], bg\_chrX, main=c('Gene XIST in sample ERR188234'), sample=c('ERR188234'))

plotMeans('MSTRG.56', bg\_chrX\_filt,groupvar="sex",legend=FALSE)

二、聚类分析及PCA主成分分析

所需数据文件：

实验设计csv表格：geuvadis\_phenodata1.csv，右侧加一栏id，与ids相同

geuvadis\_phenodata2.csv。

|  |  |  |  |
| --- | --- | --- | --- |
| ids | sex | population | id |
| ERR188044 | male | YRI | ERR188044 |
| ERR188104 | male | YRI | ERR188104 |
| ERR188234 | female | YRI | ERR188234 |
| ERR188337 | female | GBR | ERR188337 |
| ERR188383 | male | GBR | ERR188383 |
| ERR188401 | male | GBR | ERR188401 |
| ERR188428 | female | GBR | ERR188428 |
| ERR204916 | female | YRI | ERR204916 |

基因注释文件：chrX.gtf

差异表达分析中hisat2+samtools所得8个bam文件

进入R程序：R

library("airway")

vignette("airway")

library("Rsamtools")

library("GenomicFeatures")

library("DESeq2")

library("pheatmap")

library("RColorBrewer")

# 读取csv文件和8个bam文件

csvfile <- file.path("geuvadis\_phenodata2.csv")

(sampleTable <- read.csv(csvfile,row.names=1))

# 新加入的一栏id用于这里：sampleTable$id

filenames <- file.path(paste0(sampleTable$id, "\_chrX.bam"))

bamfiles <- BamFileList(filenames, yieldSize=2000000)

# 读取gtf文件，构建TxDb object对象

gtffile <- file.path("chrX.gtf")

(txdb <- makeTxDbFromGFF(gtffile, format="gtf", circ\_seqs=character()))

(ebg <- exonsBy(txdb, by="gene"))

library("GenomicAlignments")

library("BiocParallel")

# 设置多核运行

register(SerialParam())

se <- summarizeOverlaps(features=ebg, reads=bamfiles,mode="Union",singleEnd=FALSE,ignore.strand=TRUE,fragments=TRUE )

# 将样本实验设计信息传递给colData(se)，方便后续分组及寻找差异基因

(colData(se) <- DataFrame(sampleTable))

# 将male定义为sex因素的第一水平，随后的foldchange 将采用female/male.

se$sex <- relevel(se$sex, "male")

# 控制population，研究sex的影响。

dds <- DESeqDataSet(se, design = ~ population + sex)

countdata <- assay(se)

coldata <- colData(se)

(ddsMat <- DESeqDataSetFromMatrix(countData = countdata,colData = coldata,design = ~ population + sex))

# 筛选

dds <- dds[ rowSums(counts(dds)) > 1, ]

# The rlog transformation

rld <- rlog(dds, blind=FALSE)

head(assay(rld), 3)

par( mfrow = c( 1, 2 ) )

dds <- estimateSizeFactors(dds)

# 比较普通的log2(左)及rlog2(右)的作图结果

plot(log2(counts(dds, normalized=TRUE)[,1:2] + 1),pch=16, cex=0.3)

plot(assay(rld)[,1:2],pch=16, cex=0.3)

# 用欧式距离(Euclidean distance )评估样本间关系(即总体相似度)，采用rlog-transformed data

(sampleDists <- dist( t( assay(rld) ) ))

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | ERR188044 | ERR188104 | ERR188234 | ERR188337 | ERR188383 | ERR188401 | ERR188428 |
| ERR188104 | 5.228711 |  |  |  |  |  |  |
| ERR188234 | 7.288198 | 7.668854 |  |  |  |  |  |
| ERR188337 | 8.875514 | 8.726699 | 7.031264 |  |  |  |  |
| ERR188383 | 6.102277 | 5.86267 | 7.685866 | 8.276389 |  |  |  |
| ERR188401 | 5.873334 | 5.772187 | 8.663745 | 8.17113 | 5.527673 |  |  |
| ERR188428 | 7.719631 | 7.319028 | 5.65862 | 5.631482 | 7.169712 | 7.625595 |  |
| ERR204916 | 8.138408 | 7.941104 | 6.424259 | 6.354825 | 7.143844 | 8.012608 | 5.826696 |

# 用热图展示样品间关系

sampleDistMatrix <- as.matrix( sampleDists )

rownames(sampleDistMatrix) <- paste( rld$dex, rld$cell, sep="-" )

colnames(sampleDistMatrix) <- NULL

colors <- colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)

pheatmap(sampleDistMatrix,clustering\_distance\_rows=sampleDists,clustering\_distance\_cols=sampleDists,col=colors)

# PCA分析：

# 第一种方法：用plotPCA做

plotPCA(rld, intgroup = c("sex", "population"))

(data <- plotPCA(rld, intgroup = c("sex", "population"), returnData=TRUE))

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | group | sex | population | name |
| ERR188044 | -3.254229 | 1.8608593 | male:YRI | male | YRI | ERR188044 |
| ERR188104 | -3.164121 | 0.8863557 | male:YRI | male | YRI | ERR188104 |
| ERR188234 | 2.51583 | 3.6167524 | female:YRI | female | YRI | ERR188234 |
| ERR188337 | 3.921531 | -2.422433 | female:GBR | female | GBR | ERR188337 |
| ERR188383 | -2.342253 | -0.9677665 | male:GBR | male | GBR | ERR188383 |
| ERR188401 | -3.212352 | -2.5520669 | male:GBR | male | GBR | ERR188401 |
| ERR188428 | 2.710774 | 0.1135457 | female:GBR | female | GBR | ERR188428 |
| ERR204916 | 2.82482 | -0.5352467 | female:YRI | female | YRI | ERR204916 |

# 第二种方法：用ggplot做

(data <- plotPCA(rld, intgroup = c( "sex", "population"), returnData=TRUE))

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | group | sex | population | name |
| ERR188044 | -3.254229 | 1.8608593 | male:YRI | male | YRI | ERR188044 |
| ERR188104 | -3.164121 | 0.8863557 | male:YRI | male | YRI | ERR188104 |
| ERR188234 | 2.515830 | 3.6167524 | female:YRI | female | YRI | ERR188234 |
| ERR188337 | 3.921531 | -2.4224330 | female:GBR | female | GBR | ERR188337 |
| ERR188383 | -2.342253 | -0.9677665 | male:GBR | male | GBR | ERR188383 |
| ERR188401 | -3.212352 | -2.5520669 | male:GBR | male | GBR | ERR188401 |
| ERR188428 | 2.710774 | 0.1135457 | female:GBR | female | GBR | ERR188428 |
| ERR204916 | 2.824820 | -0.5352467 | female:YRI | female | YRI | ERR204916 |

percentVar <- round(100 \* attr(data, "percentVar"))

library("ggplot2")

ggplot(data, aes(PC1, PC2, color=sex, shape=population)) + geom\_point(size=3) + xlab(paste0("PC1: ",percentVar[1],"% variance")) + ylab(paste0("PC2: ",percentVar[2],"% variance"))

# Differential expression analysis差异表达分析

dds <- DESeq(dds)

#产生log2 fold change 及p值

(res <- results(dds))

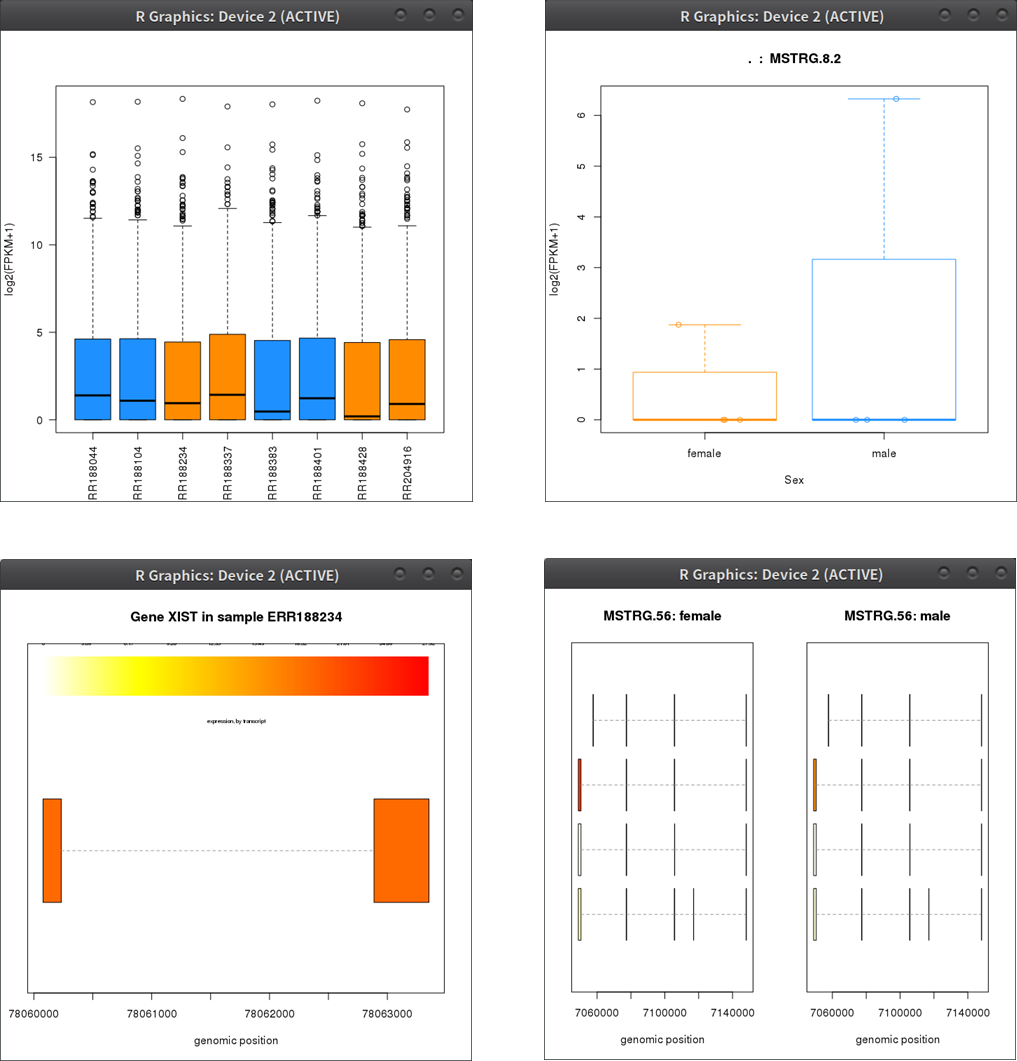
#输出csv文件

write.csv(res, file = 'res.csv')

**结果与讨论**

1. 差异表达分析

作图结果：



将所有数值p值按照从小到大排列，在得到的数据中筛选出q值小于0.05的transcripts和genes，这样剩下的为性别之间表达有明显差异的基因。其中橘色为female，蓝色为male。结合第一张图可以看出，在同一地区，不同性别的基因的表达差异性并不大，且不同区域基因表达差异性相差也很小。

第二张图是就单个转录本观测样品中的分布状况，很明显male性别FPKM值数值更大，波动也更大，因此其差异性显著。

第三张图表示ERR188234基因的转录本状况，基本出现在78060200-78060300以及78062800-78063300这两个片段之间。

第四张图展示了不同性别之间的基因表达情况，结合第三张图一起看，可以发现，不同性别的基因表达基本一致，7050000左右区域有显著的基因表达状况，而在7080000，7105000以及7150000区域有较显著的基因表达状况，7080000以及7120000区域有弱基因表达装备。

总的说来不同地区不同性别之间的基因表达并没有多大的差异性，但是其转录本阶段却存在着很大的不同，图3中转录本在两个区域显著，很有可能和不同性别有关，但最后基因表达仍然相同，应该在翻译过程中，不同性别下，机制也有所不同。当然很大一部分可能是样本太少，数据缺乏普遍性。

1. 聚类分析

作图结果：

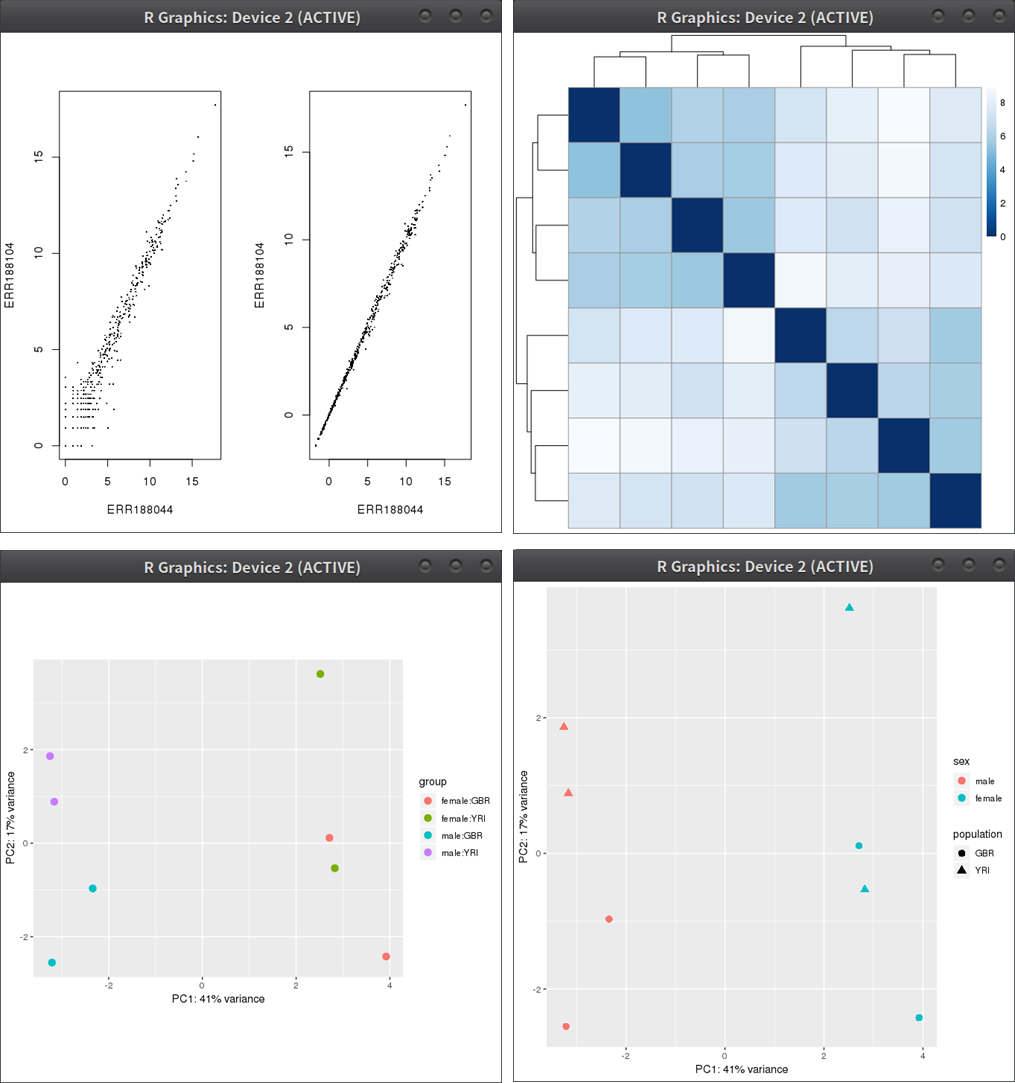


图1：这个图对比了同一地区两个男性之间X染色体之间的差异，左边是一个对照，右边结果显示两个样本的X染色体差异不大。

图2：这是两个地区居民X染色体的热图，每一行代表一个基因，每一列代表一个样本，图上每一格代表一个数值，按色键（右边）给每个数值分配颜色，图外的树状结构代表样本之间相似度聚落关系，第1和第2个样本相似，第7和第8相似，越往右越不相似，这两组具有有显著差异，因为其所属的大类具有显著差异。

图3：这是两个地区居民X染色体的散点图，每个点代表一个样本，如果两个样本距离较远，则说明两个样本遗传背景差异越大，理想情况下，遗传背景相似的个体会在图中聚在一起，图中男性同地区的两个样本差异不大，但是女性的样本差异却很大，这有可能是男性X染色体只来源于母亲，而女性的X染色体即来源于父亲也来源于母亲的缘故，并且其中一条X染色体失活可能造成了更多的变异。

图4：从图4中我们也能发现相同地区男性样本基因表达差异不大，但是女性样本基因差异却很大，结果与图4类似。

**参考文献**

Pertea M , Kim D , Pertea G M , et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown[J]. Nature Protocols, 2016, 11(9):1650-1667.