**拟南芥RNA-seq数据分析**

**Data analysis of RNA-seq in Arabidopsis thaliana**

|  |  |  |
| --- | --- | --- |
| **姓名** | **:** | **王欢欢** |
| **学号** | **:** | **2017317110025** |
| **专业** | **:** | **生物信息学** |
|  |  |  |

**华中农业大学信息学院**

**中国·武汉**

**目录**

[摘要 i](#_Toc496278752)

[关键词： i](#_Toc496278753)

[1 前言 1](#_Toc496278754)

[1.1目的 1](#_Toc496278755)

[1.2 技术路线 1](#_Toc496278756)

[2 数据来源与数据处理方法 2](#_Toc496278757)

[2.1 数据来源 2](#_Toc496278758)

[2.2 数据处理方法 2](#_Toc496278759)

[2.2.1 数据预处理 2](#_Toc496278760)

[2.2.2 序列比对 2](#_Toc496278761)

[2.2.3 基因差异表达水平分析 3](#_Toc496278762)

[2.2.4 GO富集分析 3](#_Toc496278763)

[3 结果与分析 4](#_Toc496278764)

[3.1 CUFFLINKS差异表达水平结果 4](#_Toc496278765)

[3.2 HTSEQ-DESEQ差异表达水平结果 8](#_Toc496278766)

[3.3 GO结果 10](#_Toc496278767)

[4 讨论 12](#_Toc496278768)

[附录 13](#_Toc496278769)

# 摘要

为了检查拟南芥茎发育过程中cep1突变体植物中差异表达的转录组，收集开花后25天的cep1突变体和野生型植物的茎进行RNA-seq分析；每个样品中不包括重复。

本次实验的基本流程为：通过fastqc、trimmomatic进行数据质量评估；使用tophat2对评估后的数据进行比对分析；利用samtools rmdup去除比对结果的duplication；基因差异表达使用2种方法分析：一是使用cufflinks套件对数据进行差异分析，再使用R包cummeRbund对结果进行可视化，二是先使用python包htseq统计基因的reads，再使用R包deseq进行基因差异表达分析；最后利用GO在线分析网站进行差异基因富集分析。

# 关键词：拟南芥；cep1；RNA-seq；基因差异表达水平分析

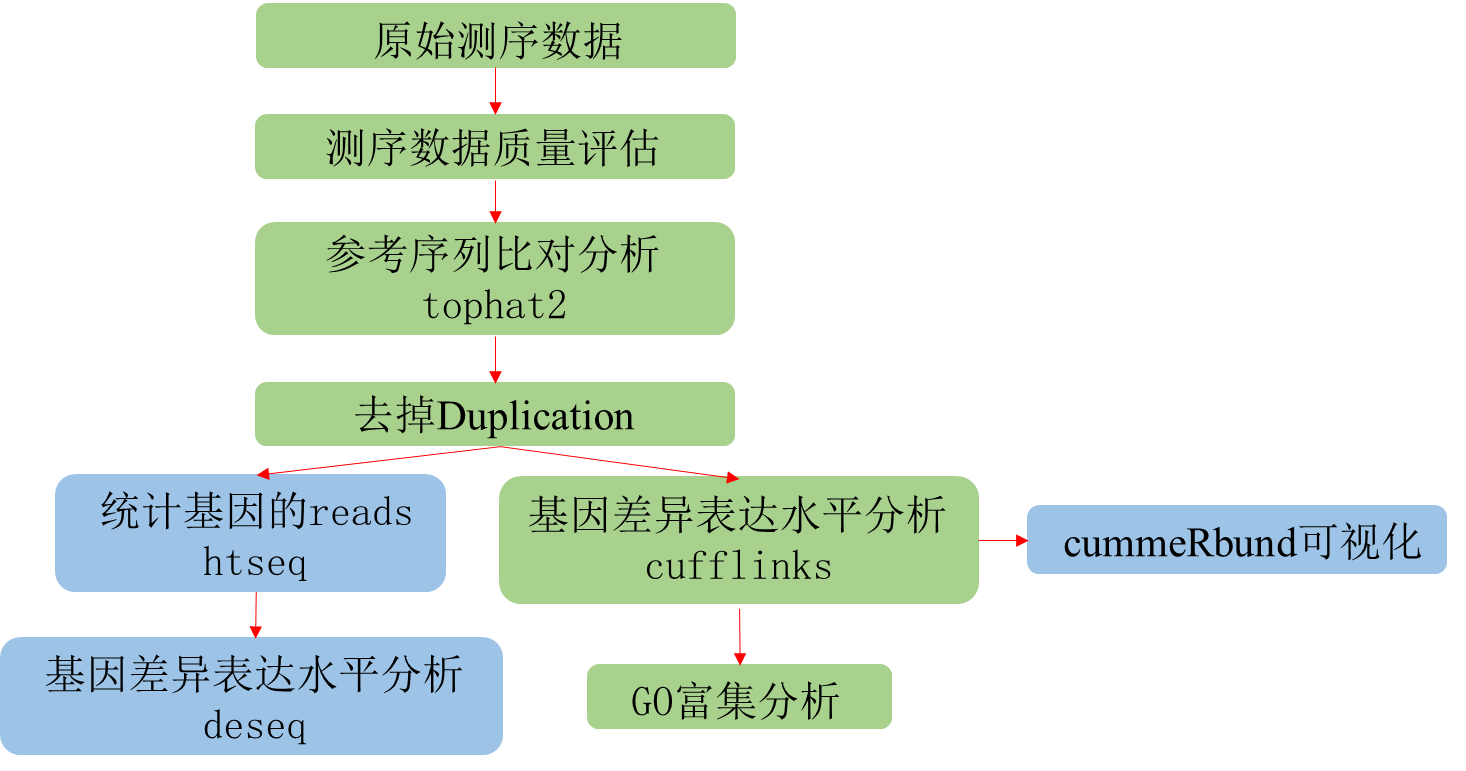
# 1 前言

## 1.1目的

为了检查拟南芥茎发育过程中cep1突变体植物中差异表达的转录组，收集开花后25天的cep1突变体和野生型植物的茎进行RNA-seq分析；每个样品中不包括重复。

## 1.2 技术路线

本次实验的基本流程为：通过fastqc、trimmomatic进行数据质量评估；使用tophat2对评估后的数据进行比对分析；利用samtools rmdup去除比对结果的duplication；基因差异表达使用2种方法分析：一是使用cufflinks套件对数据进行差异分析，再使用R包cummeRbund对结果进行可视化，二是先使用python包htseq统计基因的reads，再使用R包deseq进行基因差异表达分析；最后利用GO在线分析网站进行差异基因富集分析。（图1）。



**图1 数据处理基本流程**

**Fig.1 Data processing basic process**

# 2 数据来源与数据处理方法

## 2.1 数据来源

本文的试验数据为拟南芥（Arabidopsis thaliana）茎的RNA-Seq数据。RNA-Seq数据由Beijing Forestry University于2017年8月16号在NCBI发表；数据可在登陆号为GSE102694的NCBI的GEO数据库中下载，数据链接：ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP115/SRP115495。本文以cep1突变（cep1 mutant），和对应的野生型(wild type)的成熟茎为例，研究成熟茎cep1突变后和野生型之间的表达差异（表1）。表1是本文数据的基本信息。

**表1.** **拟南芥成熟茎cep1突变和野生型RNA-seq数据信息**

**Table 1. Arabidopsis mature stem cep1 mutation and wild type RNA-seq data information**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 登陆号  GEO ID | 物种  Organism | 基因型  Genotype | 组织  Tissue | 状态  Stage | 总测序片段数  Total reads |
| GSM2743110 | Arabidopsis thaliana | cep1 mutant | Stem | 25 days after flowing | 49,634,134 |
| GSM2743111 | Arabidopsis thaliana | wild type | Stem | 25 days after flowing | 50,263,874 |

## 2.2 数据处理方法

### 2.2.1 数据预处理

对于测得的数据，进行数据预处理控制测序片段碱基质量是不可缺少的一步，也是非常重要的一步；否则会影响数据整体的处理效果。例如：测序片段上包含的引物（adapter）会影响其匹配到参考基因组上的准确度，影响数据的后续分析。本文采用fastqc（Andrews et al，2010）和trimmomatic（Bolger A M et al，2014）两个软件来进行序列的质量控制，使得本文所处理的数据碱基质量较高，获得较为准确的结果。

### 2.2.2 序列比对

RNA-sequencing（RNA-seq）是一个重要的转录组学研究技术，数百款分析工具已经开发。已知现在报道的RNA-seq比对软件有Tophat2、STAR、HISAT2、RASER、bwa等。每个比对软件各有优缺点，本文采用的是Tophat2软件以col为参考基因组对数据进行序列比对。另外，RNA-seq call snp的时候，如果某个变异位点的变异碱基都来源于PCR重复，而我们却认为它深度足够判断是真的变异位点，这个结论有很大可能具有假阳性，为了消除这种假阳性，可以使用samtools rmdup和picard软件删除潜在的PCR重复，这里采用了samtools rmdup方法删除潜在的PCR重复。使用samtools flagstat查看比对结果，表2是拟南芥成熟茎cep1突变和野生型删除潜在PCR重复的RNA-seq序列比对结果。

**表2.** **拟南芥成熟茎cep1突变和野生型RNA-seq序列比对结果**

**Table 2. Sequence map of cep1 mutation and wild type RNA-seq sequence in mature stem of Arabidopsis thaliana**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 基因型  Genotype | 总比对数  Total mapped | 完美匹配reads数  Properly paired | 完美匹配百分比  unique Mapping rate | PCR重复  Duplication | 测序深度  Coverage genome |
| cep1 mutant | 34,950,791 | 31,026,030 | 89.67% | 0 | 61.19 |
| wild type | 32,933,960 | 27,528,528 | 84.27% | 0 | 61.97 |

### 2.2.3 基因差异表达水平分析

差异表达分析的目标是突出在不同实验条件下丰度显著变化的基因。很多方法开发出来进行差异表达分析。主要流程有：cufflinks流程、htseq-deseq、StringTie-ballgown等；本文主要做了其中cufflinks流程和htseq-deseq。其中cufflinks流程主要过程有以下几个步骤：cufflinks组装，主要根据比对结果，依托或不依托参考基因组的GTF注释文件，计算各个基因的isoform的FPKM值，并给出transcript.gtf注释结果；cuffmerge将各个cufflinks生成的transcript.gtf文件融合成为一个更加全面的transcript注释结果文件merged.gtf，以利于Cuffdiff分析基因差异表达；cuffdiff用于寻找转录子表达的显著性差异；R包cummeRbund对Cuffdiff得到的差异表达结果进行可视化。Htseq-deseq流程要简单一点，主要有：htseq为python包，其功能为统计基因的reads数目；deseq为R包，其功能为根据基因的reads数目做差异表达分析。

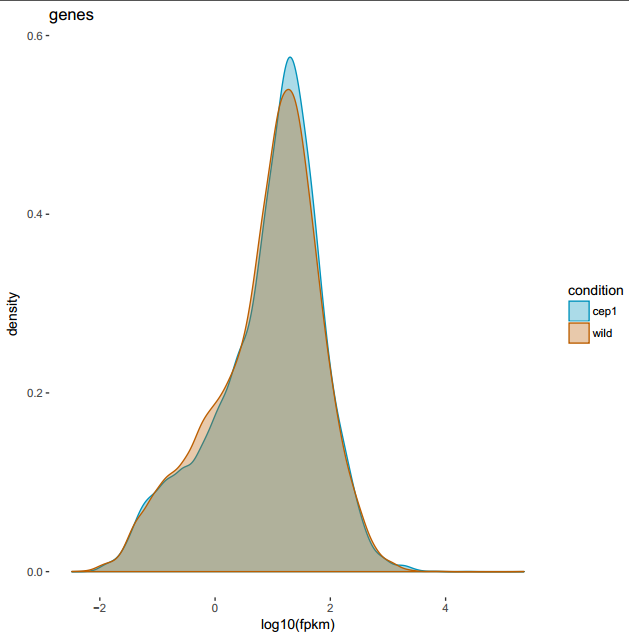
### 2.2.4 GO富集分析

GO是Gene ontology的缩写，Gene Ontology可分为分子功能（Molecular Function），生物过程（biological process）和细胞组成（cellular component）三个部分。蛋白质或者基因可以通过ID对应或者序列注释的方法找到与之对应的GO号，而GO号可对于到Term，即功能类别或者细胞定位。 功能富集分析: 功能富集需要有一个参考数据集，通过该项分析可以找出在统计上显著富集的GO Term。该功能或者定位有可能与研究的目前有关。 GO功能分类是在某一功能层次上统计蛋白或者基因的数目或组成，往往是在GO的第二层次。此外也有研究都挑选一些Term，而后统计直接对应到该Term的基因或蛋白数。做GO富集最简单的方法是用在线网站，动物和人用DAVID，植物可以考虑用AgriGO。本文采用Agrigo2网站，进行RNA-seq差异基因GO富集分析。AgriGO网址：http://systemsbiology.cau.edu.cn/agriGOv2/index.php

# 3 结果与分析

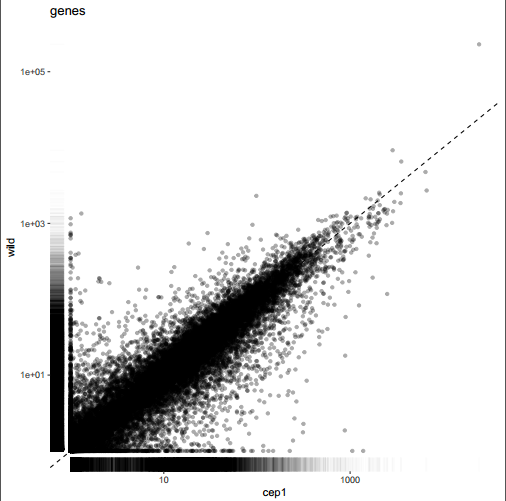
## 3.1 cufflinks差异表达水平结果

Cuffdiff的差异分析结果由R 包cummeRbund进行可视化，图2到图8是可视化的结果。拟南芥茎cep1样本和野生型一共有345个差异表达基因，上调差异表达基因有174个，下调差异表达基因有171个。图2根据FPKM值画出拟南芥cep1和野生型茎的表达水平密度分布。从图中可以看出，二者的重合性很好。图3是用散点图比较两种条件下每个基因的表达情况，从图中可以看出，二种2条件下的每个基因表达情况比较一致。图4是创建一个火山图检查差异表达基因，本来有345个，但图中没显示。图5图6图7描述的是某以个基因的表达水平；图5是用柱状图表示这个基因的表达情况；图6是这个基因的同源异构体的表达情况，图7是用折线图来描述DCL基因的差异表达趋势，本文以DCL1这个基因为例，展示结果。图8是cep1和野生型的差异表达热图。



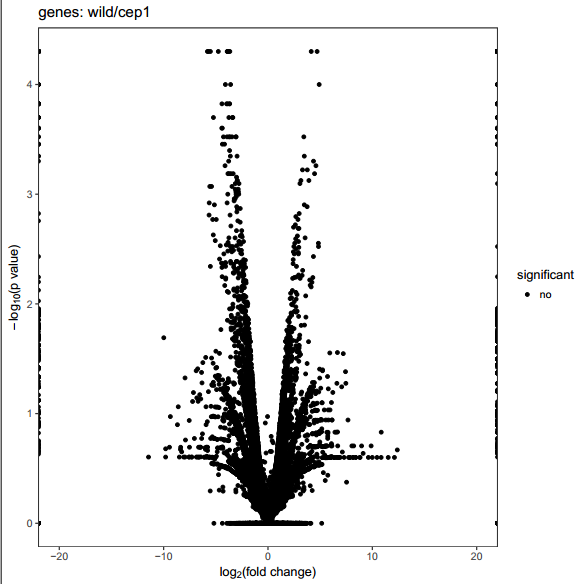
**图2 cep1和野生型的表达水平分布**

**Fig.2 Plot the distribution of expression levels for cep1 and wild**



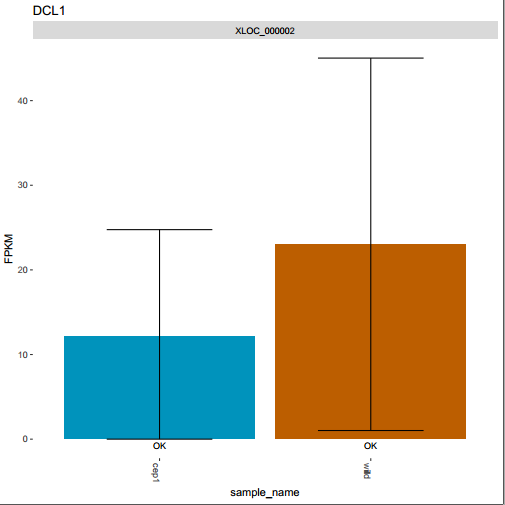
**图3 用散点图比较两种条件下每个基因的表达**

**Fig.3** **Compare the expression of each gene in two conditions with a scatter plot**



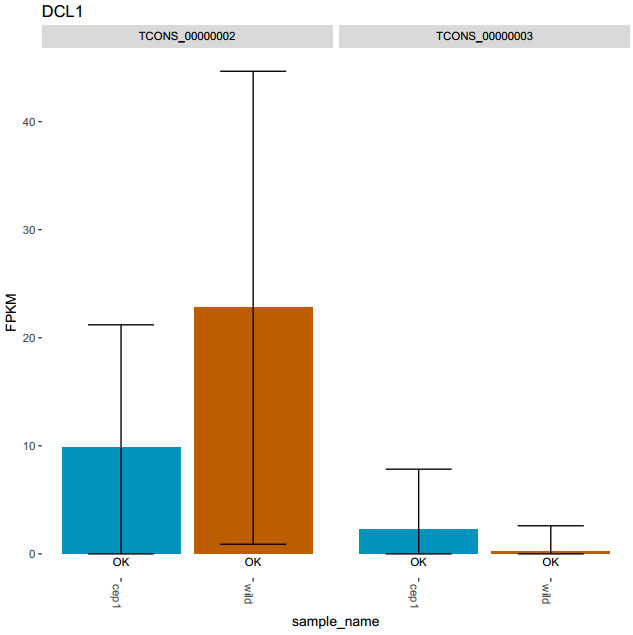
**图4 创建一个火山图检查差异表达基因**

**Fig.4 Creat a volcano plot to inspect differentially expressed genes**



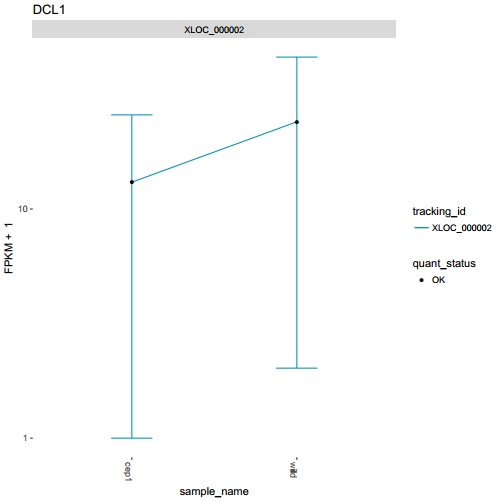
**图5****用柱状图来描述DCL1基因的表达水平**

**Fig.5 Plot expression level of DCL1 gene of interest with bar plots across conditions cep1 and wild, measured in FPKM**

****

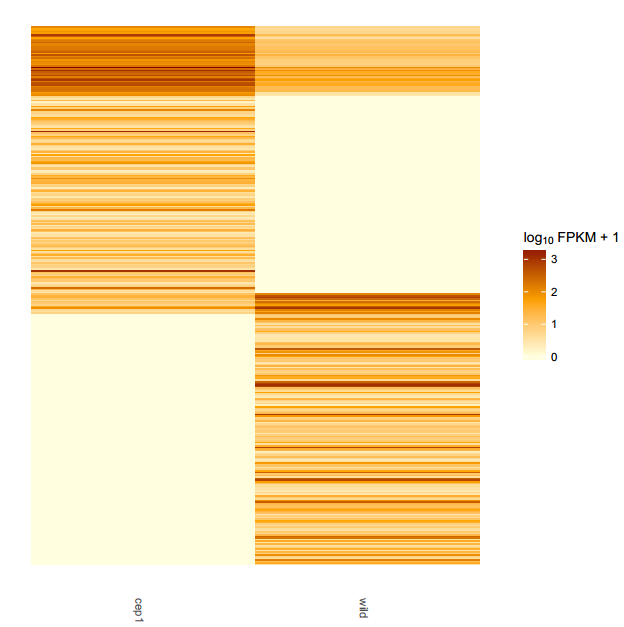
**图6****绘制条状图描述DCL1基因的同源异构体的表达水平**

**Fig.6 Plot individual isoform expression levels of DCL1 gene of interest with bar plots**

****

**图7****用折线图来描述DCL基因的差异表达趋势**

**Fig.7 The differential expression trend of DCL gene was described by line graph**

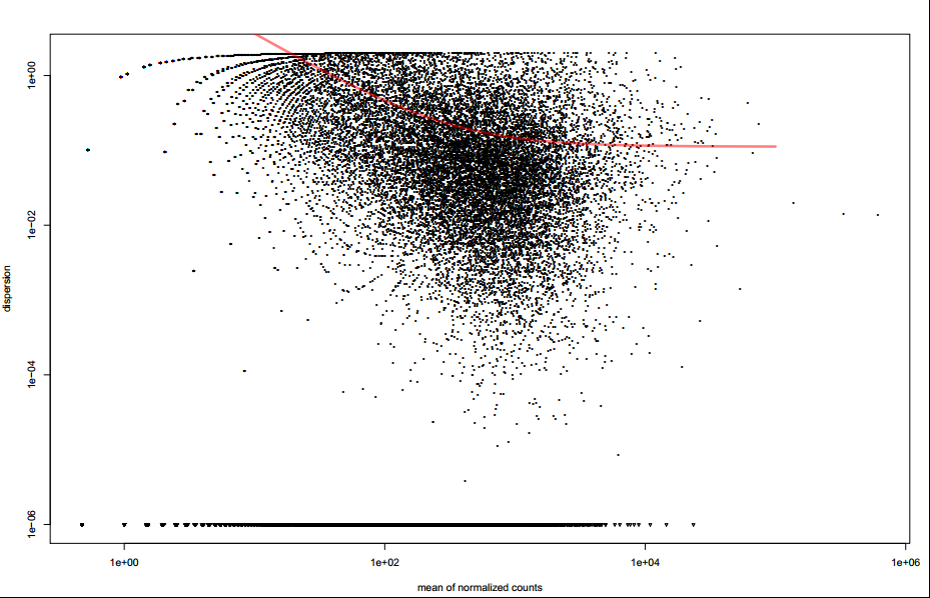
****

**图8** **cep1和野生型的差异表达热图**

**Fig.8 The hotmap of the differential expression level of cep1 and wild type**

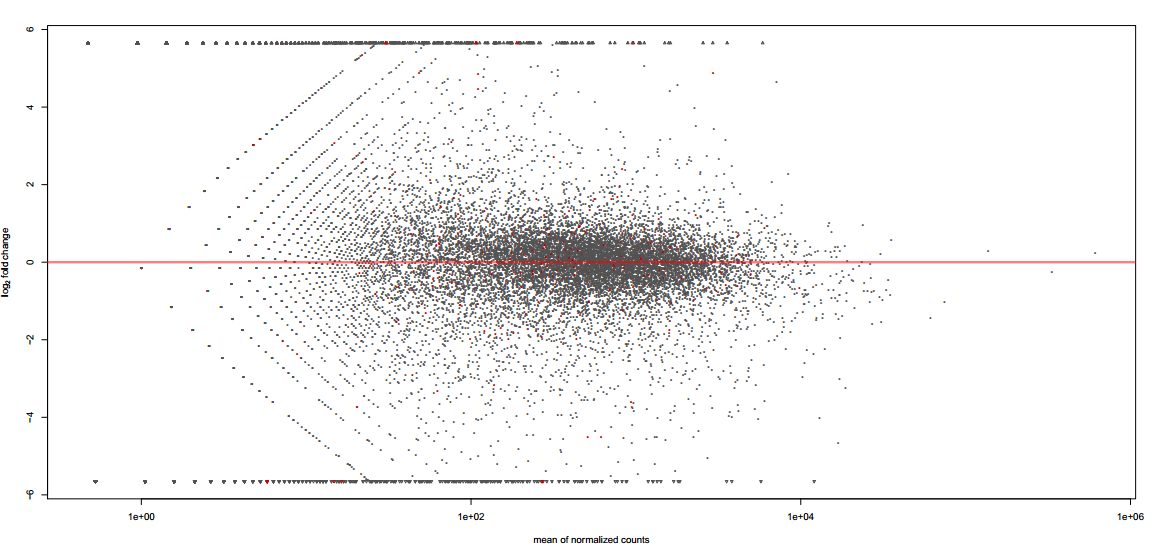
## 3.2 Htseq-deseq差异表达水平结果

这个流程相当于利用了R包deseq对htseq的结果进行了可视化。图9图10图11图12是结果。图9描述的是每条基因的估计值和平均正常统计值的关系。图10是log2折叠变换和平均正常统计量的关系图；图11图12分别是nbinomTest的padj和pval的统计直方图。



**图9****经验的(黑点)和匹配的(红线)离差值与平均正常统计值的关系图**

**Fig.9 The relation between the empirical (black spot) and the matching (Hong Xian) difference with the average normal statistic value**



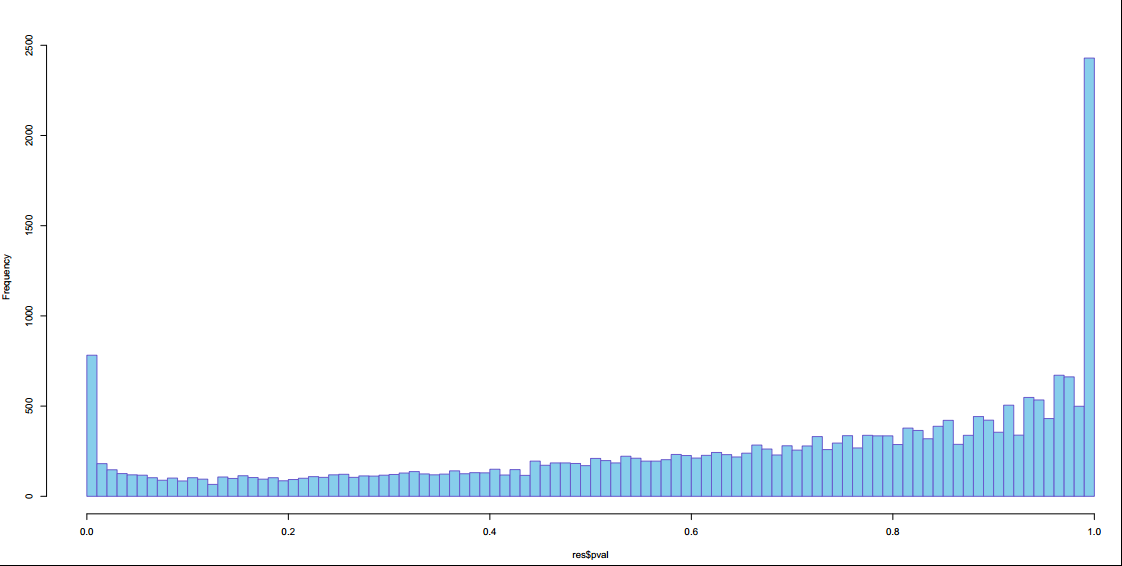
**图10** **log2折叠变换和平均正常统计量的关系图**

**Fig.10 Relation graph between log2 folding transformation and mean normal statistics**

Frequency 
0 
188 
1580 
288 
res$padj 

**图11** **nbinomTest的padj统计直方图**

**Fig.11 Padj statistical histogram of nbinomTest**



**图12 nbinomTest的pval统计直方图**

**Fig.12 Pval statistical histogram of nbinomTest**

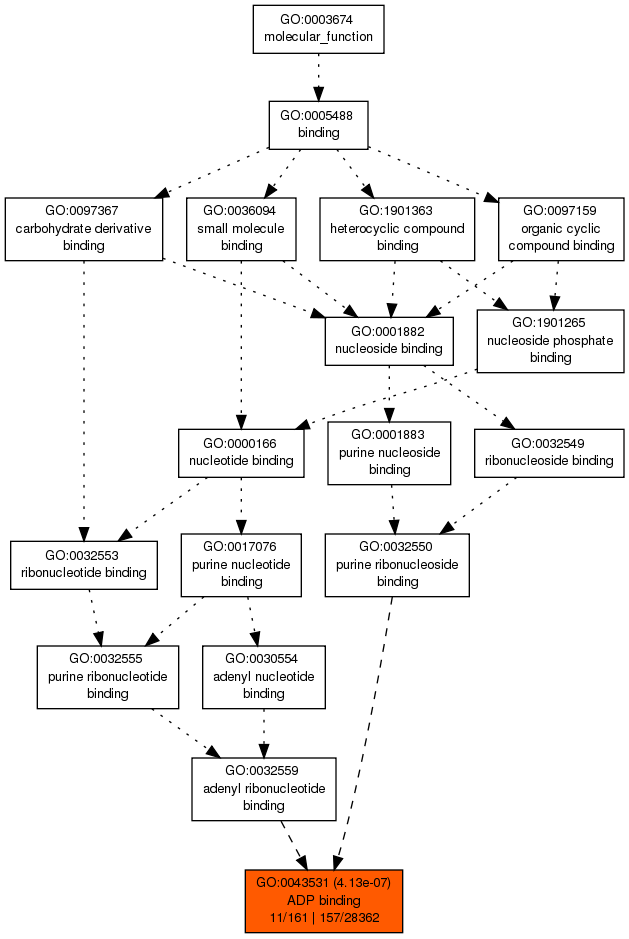
## 3.3 GO结果

将差异表达基因输入到agrigo网站上，得到GO富集情况。图13-15是结果。图13和图14是GO富集分析，分别是在生物学过程和分子功能上的GO富集，而在细胞组成没有富集，颜色越深表示越相关，从2个图中看出，这些差异基因并没有明显的富集情况。图15是在生物学过程、分子功能和细胞组成上的GO注释，绿色表示是背景，蓝色表示的是自己的差异基因，从图中看出绿色比蓝色要高，说明这些差异以及在生物学过程、分子功能和细胞组成这三方面没有明显的富集情况。

GO:0008150 
biolcgical_prmess 
GO:0051179 
malization 
GO:0033036 
macromola:ule Ima ization 
GO:0010876 (0.0496) 
ipid Imalization 
7/161 199,'28362 

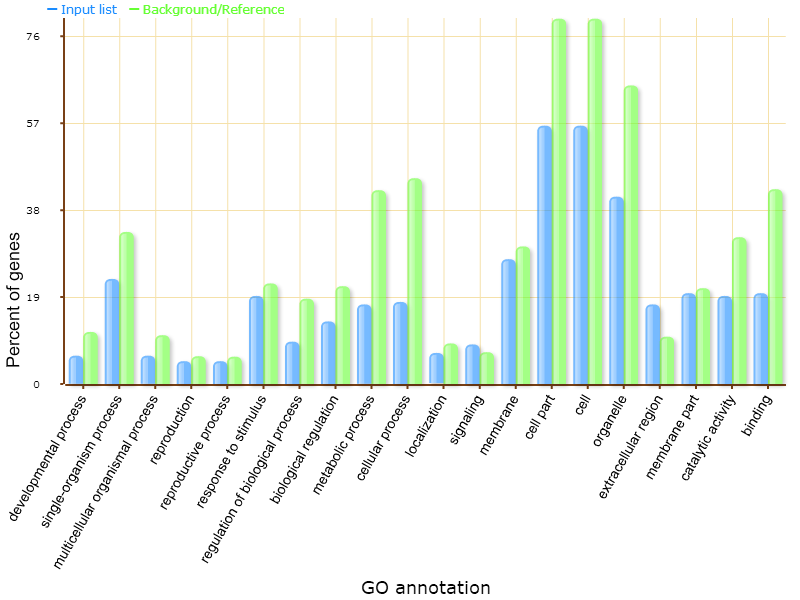
**图13 生物学过程的GO分析**

**Fig.13 GO ontology of biological process**



**图14 分子功能的GO分析**

**Fig.14 GO ontology of mollecular function**



**图15 差异基因的GO注释**

**Fig.15 GO annotation of differeantal genes**

# 4 讨论

为了检查拟南芥茎发育过程中cep1植物中差异表达的转录组，收集开花后25天的cep1突变体和野生型植物的茎进行RNA-seq分析；每个样品中不包括重复。

通过fastqc、trimmomatic进行数据质量评估；使用tophat2对评估后的数据进行比对分析；利用samtools rmdup去除比对结果的duplication；基因差异表达使用2种方法分析：一是使用cufflinks套件对数据进行差异分析，再使用R包cummeRbund对结果进行可视化，二是先使用python包htseq统计基因的reads，再使用R包deseq进行基因差异表达分析；最后利用GO在线分析网站进行差异基因富集分析。

经过上述分析，发现拟南芥cep1突变体和野生型植物之间的基因表达差异并不明显。

# 附录

1、转换数据类型

fastq-dump --split-3 SRR5936240.sra

fastq-dump --split-3 SRR5936241.sra

2、质量控制

#first fastqc

fastqc -o SRR5936240 -t 6 SRR5936240\_1.fastq SRR5936240\_2.fastq

fastqc -o SRR5936241 -t 6 SRR5936240\_1.fastq SRR5936241\_2.fastq

#trimmomatic

java -jar trimmomatic-0.36.jar PE -threads 6 -trimlog log.txt SRR5936240\_1.fastq SRR5936240\_2.fastq SRR5936240\_1\_paired.fastq SRR5936240\_1\_unpaired.fastq SRR5936240\_2\_paired.fastq SRR5936240\_2\_unpaired.fastq ILLUMINACLIP:software/Trimmomatic-0.36/adapters/TruSeq3-PE-2.fa:2:30:10 HEADCROP:15 MINLEN:36

java -jar ~/software/Trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 6 -trimlog SRR5936241/log.txt SRR5936241\_1.fastq SRR5936241/SRR5936241\_2.fastq SRR5936241/SRR5936241\_1\_paired.fastq SRR5936241/SRR5936241\_1\_unpaired.fastq SRR5936241/SRR5936241\_2\_paired.fastq SRR5936241/SRR5936241\_2\_unpaired.fastq HEADCROP:15 MINLEN:36

#second fastqc

fastqc -o SRR5936240 -t 6 SRR5936240\_1\_paired.fastq SRR5936240\_2\_paired.fastq

fastqc -o SRR5936241 -t 6 SRR5936241\_1\_paired.fastq SRR5936241\_2\_paired.fastq

3、比对过程

#索引文件下载：

ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/Arabidopsis\_thaliana/Ensembl/TAIR10/Arabidopsis\_thaliana\_Ensembl\_TAIR10.tar.gz

#tophat2进行比对

tophat2 -p 8 -G genes.gtf -o tophat2\_40 genome SRR5936240\_1\_paired.fastq SRR5936240\_2\_paired.fastq

tophat2 -p 8 -G genes.gtf -o tophat2\_41 genome SRR5936241\_1\_paired.fastq SRR5936241\_2\_paired.fastq

#去掉Duplication

samtools rmdup tophat2\_40/accepted\_hits.bam tophat2\_40/accepted\_rmdup\_hits.bam

result：5520578 / 21878180 = 0.2523 in library

samtools rmdup tophat2\_41/accepted\_hits.bam tophat2\_41/accepted\_rmdup\_hits.bam

result：5099818 / 19639285 = 0.2597 in library

#使用samtools flagstat查看比对结果

samtools flagstat tophat2\_40/accepted\_rmdup\_hits.bam > tophat2\_40/flagstat.txt

result:

[hhwang@login tophat2 more flagstat . txt 
34950791 + in total (QC-passed reads + QC-failed reads) 
352197 + secondary 
@ + @ supplementary 
+ duplicates 
34950791 + mapped : N/A) 
34598594 + paired in sequencing 
17805296 + readl 
16793298 + read2 
31026030 + properly paired (89.67% : N/A) 
32536310 + with itself and mate mapped 
2062284 + singletons (5.96% : N/A) 
62328 + with mate mapped to a different chr 
58570 + with mate mapped to a different chr 

samtools flagstat tophat2\_41/accepted\_rmdup\_hits.bam > tophat2\_41/flagstat.txt

result:

[hhwang@login tophat2 41] $ more flagstat . txt 
32933960 + in total (QC-passed reads + QC-failed reads) 
267820 + @ secondary 
O + @ supplementary 
O + @ duplicates 
32933960 + mapped : N/A) 
32666140 + paired in sequencing 
17054186 + readl 
15611954 + read2 
27528528 + properly paired (84.27% : N/A) 
29019962 + with itself and mate mapped 
3646178 + singletons (11.16% . N/A) 
74638 + with mate mapped to a different chr 
65512 + with mate mapped to a different chr (mapQ>=5) 

4、差异表达过程

#cufflinks 转录组装

cufflinks -p 8 -G genes.gtf -o cufflinks\_40 tophat2\_40/accepted\_rmdup\_hits.bam（-b，-u，是cufflinks算法的纠正，-G/g ）

cufflinks -p 8 -G genes.gtf -o cufflinks\_41 tophat2\_41/accepted\_rmdup\_hits.bam

#建立一个文件命名为：assemblies.txt

vi assemblies.txt

/public/home/hhwang/homework/cufflinks\_40/transcripts.gtf

/public/home/hhwang/homework/cufflinks\_41/transcripts.gtf

#cuffmerge转录合并

cuffmerge -o merged\_asm -g genes.gtf -s genome.fa -p 8 assemblies.txt

#cuffdiff差异表达

cuffdiff -o diff\_out -b genome.fa -p 8 -L cep1,wild -u merged\_asm/merged.gtf tophat2\_40/accepted\_rmdup\_hits.bam tophat2\_41/accepted\_rmdup\_hits.bam

6、R可视化

#R的安装

./configure && make && make check && make install

安装cummeRbund包

> source('http://www.bioconductor.org/biocLite.R')

> biocLite('cummeRbund')

> print(require(cummeRbund))

#运行cummeRbund

> library(cummeRbund)

> cuff\_data <- readCufflinks('diff\_out/')

报错：

Creating database diff/cuffData.db 
Reading 
Writing 
Reading 
Writing 
Reading 
Writing 
Reading 
Run Info File diff/ run . info 
runlnfo Table 
Read Group Info diff/ read groups . info 
replicates Table 
Var Model Info diff/ var model. info 
varModeI Table 
diff/genes . fpkm tracking 
Checking samples table. . 
Populating samples table. 
Error. Column name mismatch . 
In addition: There were 50 or more warnings 
(use warnings() 
to see the first 50) 

解决方法：

####下载低版本的RSQLite

wget -b -c https://launchpad.net/ubuntu/+archive/primary/+files/r-cran-rsqlite\_1.1-2.orig.tar.gz

R CMD INSTALL r-cran-rsqlite\_1.1-2.orig.tar.gz

> cuff\_data <- readCufflinks('diff\_out/')

> cuff\_data

CuffSet instance with:

2 samples

33201 genes

42800 isoforms

34921 TSS

32922 CDS

33201 promoters

34921 splicing

27083 relCDS

1│每个条件表达水平的分布：

> outfile<-paste('csDensity','pdf',sep='.')

> pdf(outfile)

>csDensity(genes(cuff\_data))

> dev.off()

2│比较每个基因在两个条件下表达的散点图：

> outfile<-paste('csScatter','pdf',sep='.')

> pdf(outfile)

> csScatter(genes(cuff\_data), 'cep1', 'wild')

> dev.off()

3│创建一个火山图检查差异表达基因

> outfile<-paste('csVolcano','pdf',sep='.')

> pdf(outfile)

>csVolcano(genes(cuff\_data), 'cep1', 'wild')

> dev.off()

4│查看差异表达基因：分别取出上调和下调的基因

#get the top 100 diff expr genes

> gene.diff <- diffData(genes(cuff\_data))

> gene.diff.top <- gene.diff[order(gene.diff$q\_value),][1:100,]

# gene ids of top 100 diff expr genes

> myGeneIds <- gene.diff.top$gene\_id

5│用柱状图来描述某个基因的表达水平 (图15)

> outfile<-paste('DCL1','pdf',sep='.')

> pdf(outfile)

> mygene<-getGene(cuff\_data, 'DCL1')

> expressionBarplot(mygene)

> dev.off()

6│绘制条状图描述单个选择基因的同源异构体的表达水平

> outfile<-paste('DCL1\_isoforms','pdf',sep='.')

> pdf(outfile)

>expressionBarplot(isoforms (mygene))

> dev.off()

7│用折线图来描述某个基因的差异表达趋势

> outfile<-paste('DCL1\_plot','pdf',sep='.')

> pdf(outfile)

>mygene<-getGene(cuff\_data, 'DCL1')

>expressionPlot(mygene,logMode=T)

> dev.off()

8│热图

> outfile<-paste('hot','pdf',sep='.')

> pdf(outfile)

>gene\_diff\_data<-diffData(genes(cuff\_data))

>sig\_gene\_data<-subset(gene\_diff\_data,(significant=='yes'))

>geneids<-sig\_gene\_data$gene\_id

>sig\_genes<- getGenes(cuff\_data,geneids)

>csHeatmap(sig\_genes,clustering='row',labRow=F)

> dev.off()

9│差异表达基因及差异表达转录本，差异剪切和调控基因等的提取

>gene\_diff\_data<-diffData(genes(cuff\_data))

>sig\_gene\_data<-subset(gene\_diff\_data,(significant=='yes'))

>nrow(sig\_gene\_data)

[1] 345

>isoform\_diff\_data<-diffData(isoforms(cuff\_data),'cep1','wild')

>sig\_isoform\_data<-subset(isoform\_diff\_data,(significant=='yes'))

>nrow(sig\_isoform\_data)

[1] 326

>tss\_diff\_data<-diffData(TSS(cuff\_data),'cep1','wild')

>sig\_tss\_data<-subset(tss\_diff\_data,(significant=='yes'))

>nrow(sig\_tss\_data)

[1] 342

>cds\_diff\_data<-diffData(CDS(cuff\_data),'cep1','wild')

>sig\_cds\_data<-subset(cds\_diff\_data,(significant=='yes'))

>nrow(sig\_cds\_data)

[1] 265

>promoter\_diff\_data<-distValues(promoters(cuff\_data))

>sig\_promoter\_data<-subset(promoter\_diff\_data,(significant=='yes'))

>nrow(sig\_promoter\_data)

[1] 0

>splicing\_diff\_data<-distValues(splicing(cuff\_data))

>sig\_splicing\_data<-subset(splicing\_diff\_data,(significant=='yes'))

>nrow(sig\_splicing\_data)

[1] 0

>relCDS\_diff\_data<-distValues(relCDS(cuff\_data))

>sig\_relCDS\_data<-subset(relCDS\_diff\_data,(significant=='yes'))

>nrow(sig\_relCDS\_data)

[1] 0

7、从cuffdiff的结果中取出差异表达基因shell script

#查看差异表达基因个数

awk '{print $NF}' gene\_exp.diff |sort|uniq -c

$NF是最后一个域的内容，表示把gene\_exp.diff文件的最后一行排序，并打印，uniq -c表示删除重复行并显示重复行的次数，-c：计数

32856 no

1 significant

345 yes（差异表达基因个数）

#取出上调差异表达基因（请确定value1 > value2）

awk '{if($NF=="yes"&&$10>0){print}}' gene\_exp.diff > ./up\_gene

一共有174个gene

#取出前100个上调的差异表达基因

sort -g -k 13 up\_gene > up\_sort

head -n 100 up\_sort > 100up

awk '{print $3}' 100up

#取出下调差异表达基因（请确定value1 < value2

awk '{if($NF=="yes"&&$10<0){print}}' gene\_exp.diff > ./down\_gene

一共有171个基因

sort -g -k 13 down\_gene > down\_sort

head -n 100 down\_sort > 100down

awk '{print $3}' 100down

8、检查比对文件读取比对到每个染色体的数量。

for i in tophat2\_40/accepted\_rmdup\_hits.bam ; do echo $i; samtools index $i; done

for i in tophat2\_40/accepted\_rmdup\_hits.bam ; do echo $i; samtools idxstats $i; done

出现结果如下：

|  |  |  |  |
| --- | --- | --- | --- |
| tophat2\_40/accepted\_rmdup\_hits.bam | | | |
| Chr | Che length | Mapped reads Number | unmapped reads number |
| 1 | 30427671 | 9563643 | 0 |
| 2 | 19698289 | 5188967 | 0 |
| 3 | 23459830 | 6505193 | 0 |
| 4 | 18585056 | 5568125 | 0 |
| 5 | 26975502 | 7949690 | 0 |
| Mt | 366924 | 38105 | 0 |
| Pt | 154478 | 137068 | 0 |

注：idxstats 统计一个表格，4列，分别为”序列名(染色体名)，序列长度(染色体长度)，mapped reads number，unmapped reads number echo显示。

for i in tophat2\_41/accepted\_rmdup\_hits.bam ; do echo $i; samtools index $i; done

for i in tophat2\_41/accepted\_rmdup\_hits.bam ; do echo $i; samtools idxstats $i; done

|  |  |  |  |
| --- | --- | --- | --- |
| tophat2\_41/accepted\_rmdup\_hits.bam | | | |
| Chr | Che length | Mapped reads Number | unmapped reads number |
| 1 | 30427671 | 8623495 | 0 |
| 2 | 19698289 | 4933229 | 0 |
| 3 | 23459830 | 6253512 | 0 |
| 4 | 18585056 | 5490691 | 0 |
| 5 | 26975502 | 7359199 | 0 |
| Mt | 366924 | 48932 | 0 |
| Pt | 154478 | 224902 | 0 |

8、htseq，deseq分析

#安装python包：htseq（统计gene的reads），R包：deseq（根据gene的reads数目做差异表达分析）

htseq：pip install htseq

Deseq2：

> source("http://bioconductor.org/biocLite.R")

> biocLite("DESeq")

#统计gene的reads（htseq）

（如果你是双端测序，必须要对SAM进行排序，推荐samtools sort进行排序 ，-n 按read name 排序）

samtools sort -n tophat2\_40/accepted\_rmdup\_hits.bam -o tophat2\_40/accepted\_rmdup\_hits\_sort.bam

samtools sort -n tophat2\_41/accepted\_rmdup\_hits.bam -o tophat2\_41/accepted\_rmdup\_hits\_sort.bam

htseq-count -f bam -s no tophat2\_40/accepted\_rmdup\_hits\_sort.bam genes.gtf > reads\_40.txt

htseq-count -f bam -s no tophat2\_41/accepted\_rmdup\_hits\_sort.bam genes.gtf > reads\_41.txt

注：-s：是否这个数据是来自链特异性建库（默认 yes)

-m：判断一个reads属于某个基因的模型，用来判断统计reads的时候对一些比较特殊的reads定义是否计入。 <mode> 包括：默认的union和intersection-strict、 intersection-nonempty （默认：union）

#根据gene的reads数目做差异表达分析（deseq2）

#把我们的多个样本计数结果合并起来成数据框，列是不同样本，行是不同基因；加载数据

>library(DESeq)

> cep1 <- read.table("reads\_40.txt",row.name=1)

> wild <- read.table("reads\_41.txt",row.name=1)

> counts <- cbind(cep1,wild)

> dim(counts)

[1] 33607 2

#构造成DESeq的对象，并对分组样本进行基因表达量检验

> colnames(counts)

[1] "V2" "V2"

> colnames(counts)<-c("cep1","wild")

> design<-rep(c("cep1","wild"),each=1)

> design

[1] "cep1" "wild"

#函数estimateSizeFactors估计统计数据的大小因子；如果统计数据的每列除以这列的大小因子，这样统计值就变成同一规模，使它们具有可比性。函数counts可以做这个计算；

> de<-newCountDataSet(counts,design)

>de<-estimateSizeFactors(de)

#函数estimateDispersions做了三步，首先估计每条基因的离差，然后，通过估计匹配一条曲线，最后，每个基因分配一个离差，从每条基因估计值和匹配值选一个

#函数 plotDispEsts可以画出每条基因的估计值和平均正常统计值的关系

#经验的(黑点)和匹配的(红线)离差值与平均正常统计值的关系图

> de<-estimateDispersions(de,method='blind',sharingMode="fit-only")##因为无重复样本，采用blind方法

> outfile<-paste('estimaDispersions','pdf',sep='.')##设置PDF的名字

> pdf(outfile,height=10,width=15)

> plotDispEsts(de)

> dev.off()

#为了看在条件”cep1"和”wild"是否有差异表达，我们简单的使用nbinomTest函数

#我们首先画log2折叠变换和平均正常统计量的关系，红色的点表示在10%FDR的基因

> res<-nbinomTest(de,"cep1","wild")#res就是我们的表达量检验结果

> head(res)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| id | baseMean | baseMeanA | baseMeanB | foldChange | Log2FoldChang3 | pval | padj |
| AT1G01010 | 132.808572 | 123.421522 | 142.1956209 | 1.1521137 | 0.20428304 | 0.8845056 | 1 |
| AT1G01020 | 318.525318 | 290.093322 | 346.9573149 | 1.1960197 | 0.25824110 | 0.7997349 | 1 |
| AT1G01020 | 9.907364 | 8.439078 | 11.3756497 | 1.3479730 | 0.43079157 | 0.9609876 | 1 |
| AT1G01020 | 1976.476364 | 1346.033013 | 2606.9197159 | 1.9367428 | 0.95363236 | 0.2080842 | 1 |
| AT1G01046 | 2.056313 | 3.164654 | 0.9479708 | 0.2995495 | -1.73913343 | 0.9680644 | 1 |
| AT1G01050 | 1360.174172 | 1324.935317 | 1395.4130261 | 1.0531933 | 0.07477028 | 0.9242305 | 1 |

#我们首先画log2折叠变换和平均正常统计量的关系，红色的点表示在10%FDR的基因

> outfile<-paste('nbinomTest','pdf',sep='.')

> pdf(outfile,height=10,width=20)

> plotMA(res)

> dev.off()

#从nbinomTest的P值的统计直方图

> p=outfile<-paste('Histogram.padj','pdf',sep='.')

> pdf(outfile)

> hist(res$padj,breaks=100,col="skyblue",border="slateblue",main="")

> dev.off()

#从nbinomTest的P值的统计直方图

> outfile<-paste('Histogram.pavl','pdf',sep='.')

> pdf(outfile,height=10,width=20)

> hist(res$pval, breaks=100, col="skyblue", border="slateblue", main="")

> dev.off()

#我们可以通过FDR，Chang过滤有效的基因

> resSig=res[res$padj<0.05,]

> resSig=resSig[resSig$foldChang>2,]

#列举最有效的差异表达基因

> head(resSig[order(resSig$pval),])

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| id | baseMean | baseMeanA | baseMeanB | foldChange | Log2FoldChang3 | pval | padj |
| AT4G28520 | 7206.948 | 0 | 14413.896 | Inf | Inf | 1.633514e-26 | 4.079374e-22 |
| AT2G27380 | 5876.471 | 0 | 11752.942 | Inf | Inf | 1.483361e-25 | 1.852199e-21 |
| AT4G27160 | 5256.024 | 0 | 10512.048 | Inf | Inf | 5.076645e-25 | 4.225969e-21 |
| AT1G03880 | 3576.694 | 0 | 7153.388 | Inf | Inf | 4.093131e-23 | 2.044355e-19 |
| AT4G27150 | 3010.281 | 0 | 6020.563 | Inf | Inf | 3.163774e-22 | 1.316815e-18 |
| AT4G27170 | 2801.728 | 0 | 5603.455 | Inf | Inf | 7.527514e-22 | 2.685494e-18 |

>write.table(resSig[order(resSig$pval),],file="most\_significantly\_differentially\_expressed\_genes.csv")

#最显著的下调基因

> head( resSig[ order( resSig$foldChange, -resSig$baseMean ), ] )

>write.table(resSig[order(resSig$foldChange,-resSig$baseMean ), ],file="most\_significantly\_down-regulated\_genes.csv")

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| id | baseMean | baseMeanA | baseMeanB | foldChange | Log2FoldChang3 | pval | padj |
| AT1G65480 | 3192.527 | 874.4995 | 5510.554 | 6.301381 | 2.655668 | 0.0006126461 | 0.03608399 |
| AT1G64780 | 2113.287 | 556.9792 | 3669.595 | 6.588388 | 2.719926 | 0.0006347086 | 0.03686181 |
| AT1G80760 | 2697.191 | 705.7179 | 4688.664 | 6.643821 | 2.732013 | 0.0004962912 | 0.03037716 |
| AT3G02380 | 1379.935 | 358.6608 | 2401.210 | 6.694932 | 2.743069 | 0.0008972920 | 0.04924851 |
| AT2G46830 | 4298.278 | 1079.1472 | 7517.408 | 6.966064 | 2.800344 | 0.0002761070 | 0.01909978 |
| AT4G10380 | 7833.222 | 1938.8783 | 13727.565 | 7.080158 | 2.823782 | 0.0001946221 | 0.01451010 |

#显著上调基因

> head( resSig[ order( -resSig$foldChange, -resSig$baseMean ), ] )

>write.table(resSig[order(-resSig$foldChange,-resSig$baseMean ), ],file="most\_significantly\_up-regulated\_genes.csv")

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| id | baseMean | baseMeanA | baseMeanB | foldChange | Log2FoldChang3 | pval | padj |
| AT4G28520 | 7206.948 | 0 | 14413.896 | Inf | Inf | 1.633514e-26 | 4.079374e-22 |
| AT2G27380 | 5876.471 | 0 | 11752.942 | Inf | Inf | 1.483361e-25 | 1.852199e-21 |
| AT4G27160 | 5256.024 | 0 | 10512.048 | Inf | Inf | 5.076645e-25 | 4.225969e-21 |
| AT1G03880 | 3576.694 | 0 | 7153.388 | Inf | Inf | 4.093131e-23 | 2.044355e-19 |
| AT4G27150 | 3010.281 | 0 | 6020.563 | Inf | Inf | 3.163774e-22 | 1.316815e-18 |
| AT4G27170 | 2801.728 | 0 | 5603.455 | Inf | Inf | 7.527514e-22 | 2.685494e-18 |

> sum(na.omit(res$padj<0.05))

[1] 455

> write.table(res,file="file.csv")