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

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）。

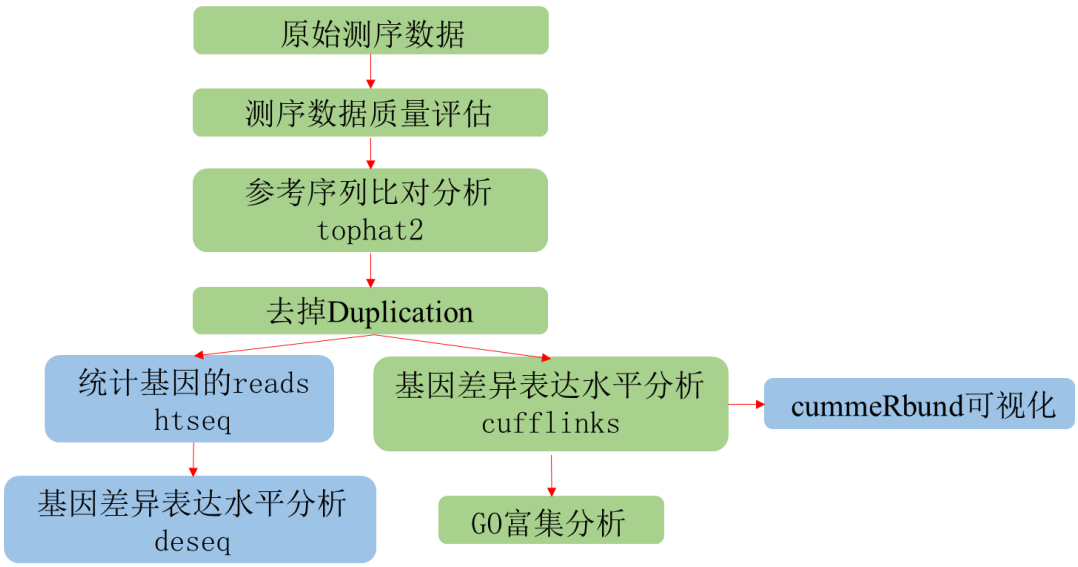

图 1 数据处理基本流程

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。

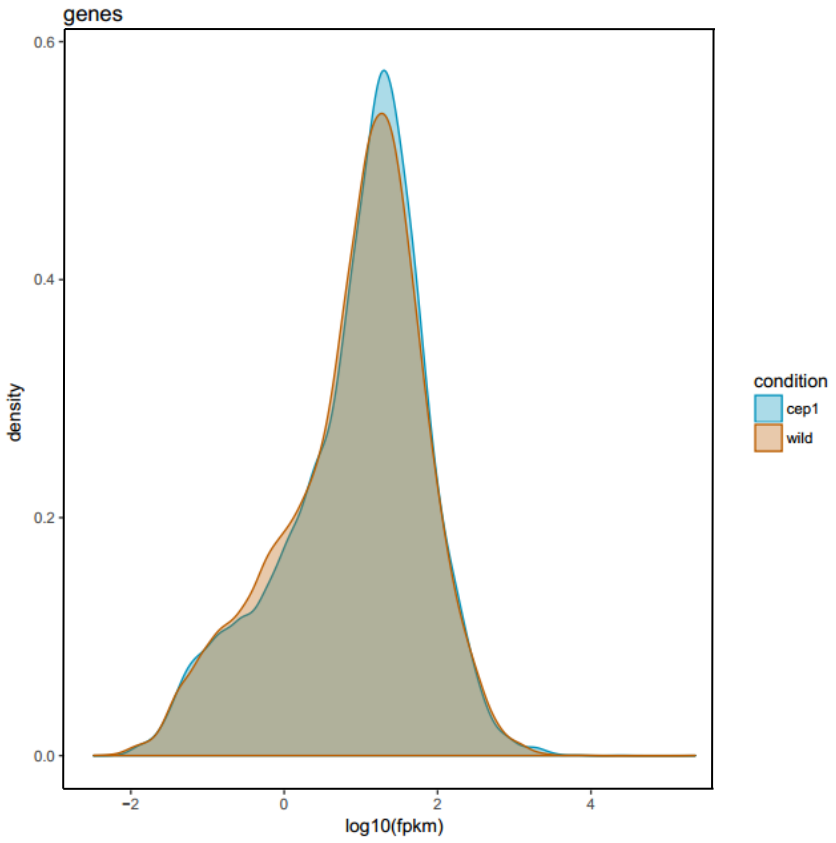

图 2 cep1 和野生型的表达水平分布

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

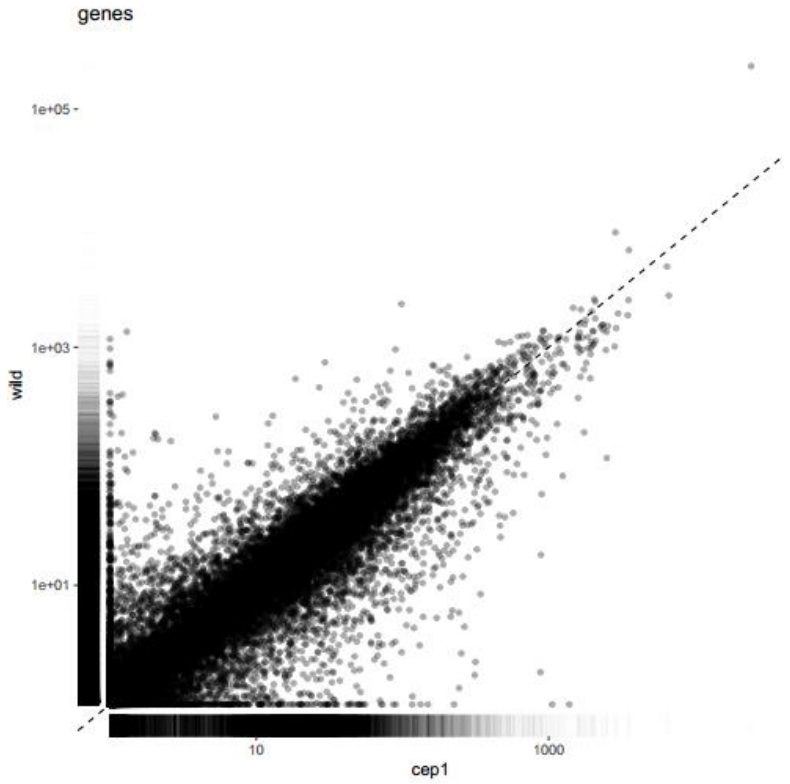

图 3 用散点图比较两种条件下每个基因的表达

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

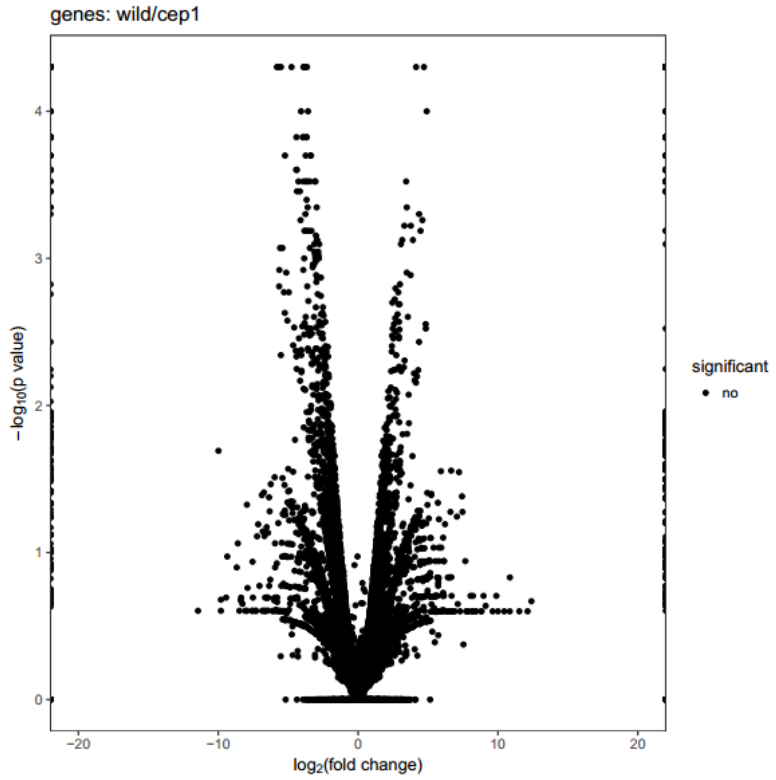

图 4 创建一个火山图检查差异表达基因

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

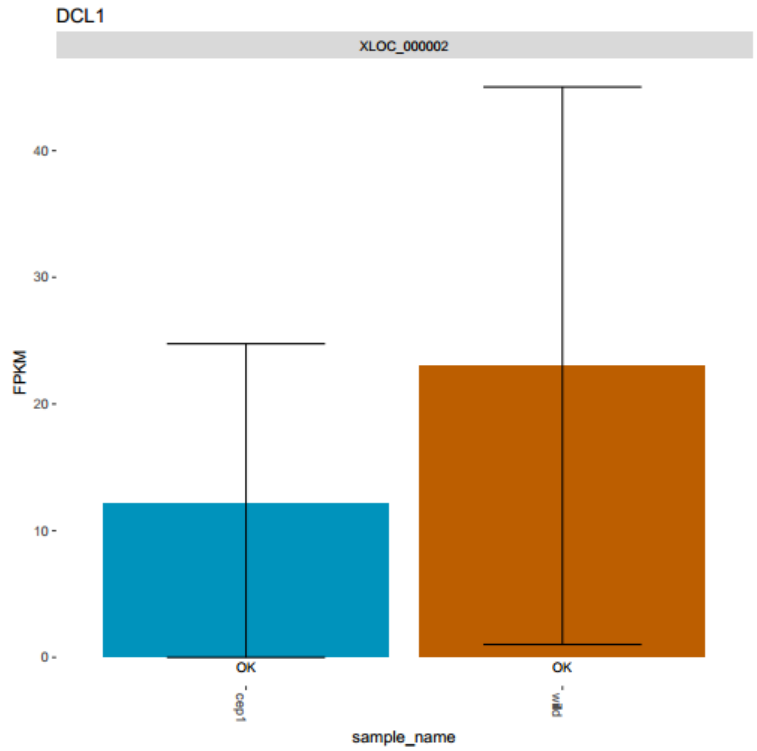

图 5 用柱状图来描述 DCL1 基因的表达水平

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

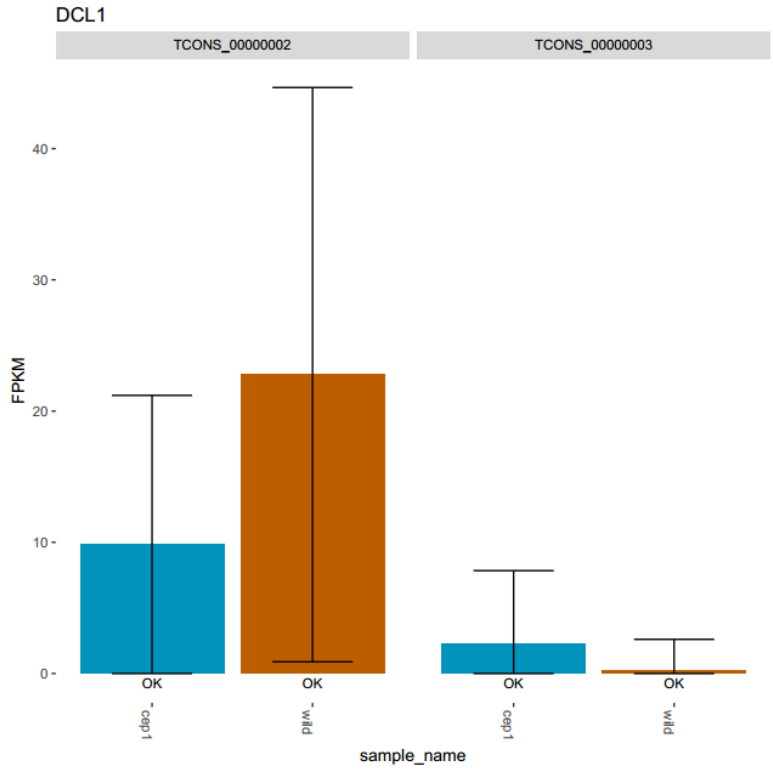

图 6 绘制条状图描述 DCL1 基因的同源异构体的表达水平

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

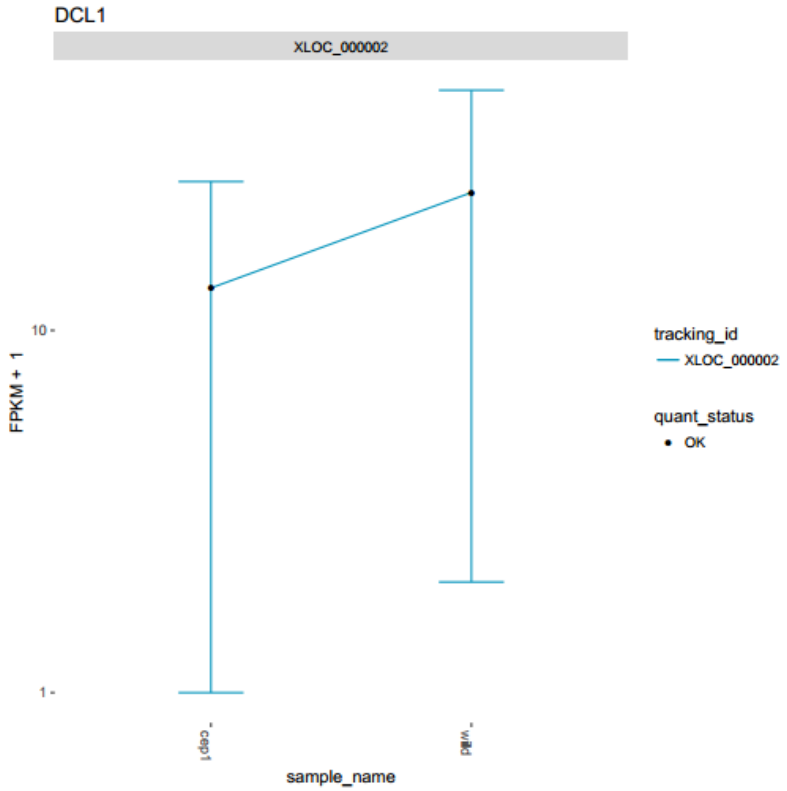

图 7 用折线图来描述 DCL 基因的差异表达趋势

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

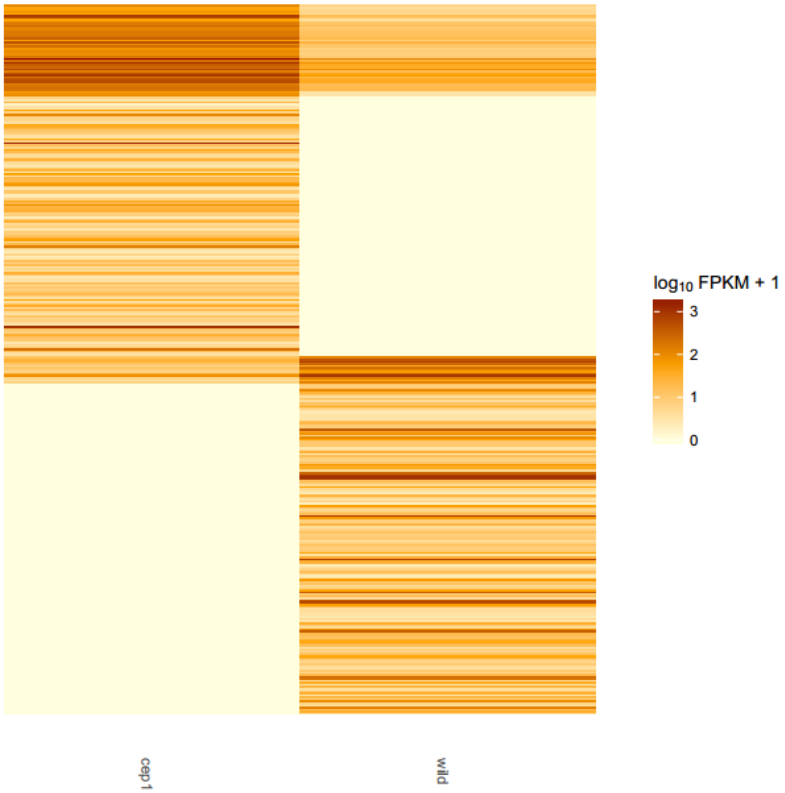

图 8 cep1 和野生型的差异表达热图

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。

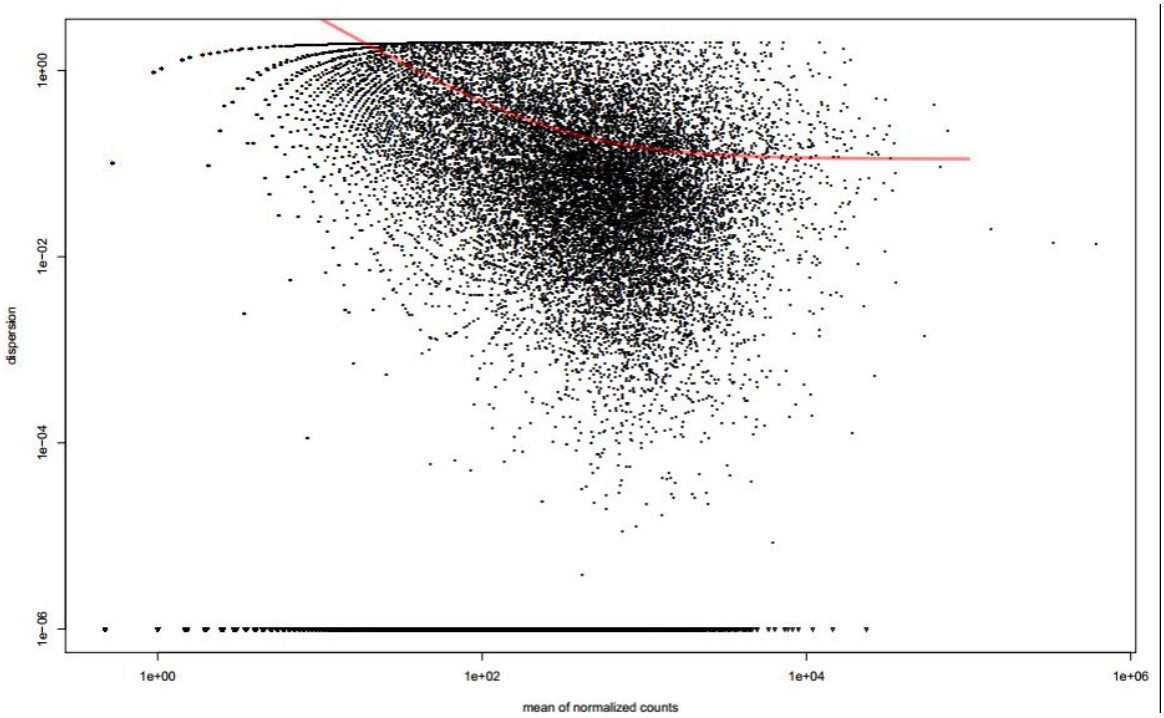

图 9 经验的(黑点)和匹配的(红线)离差值与平均正常统计值的关系图

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

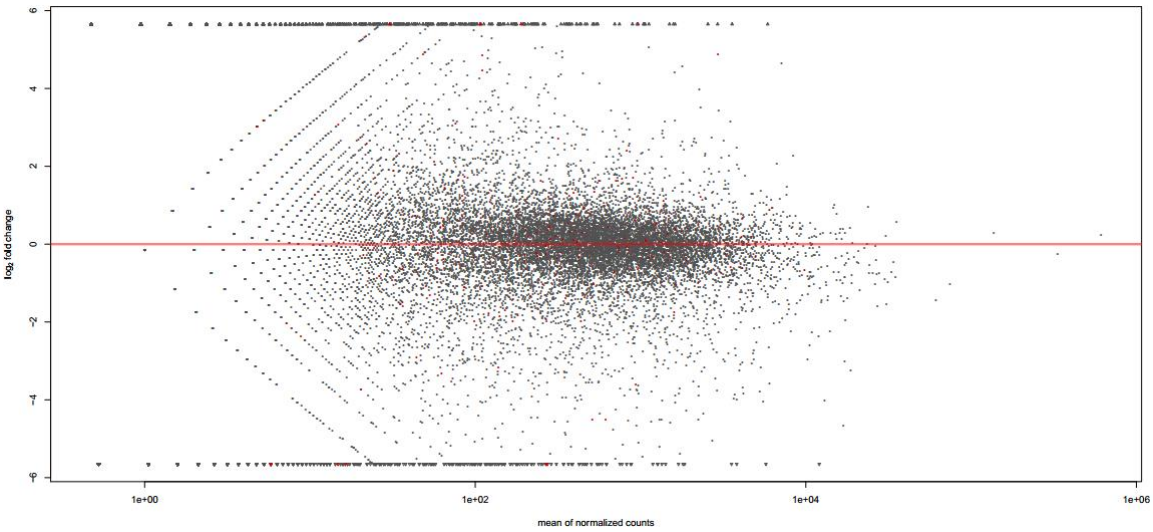

图 10 log2 折叠变换和平均正常统计量的关系图

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

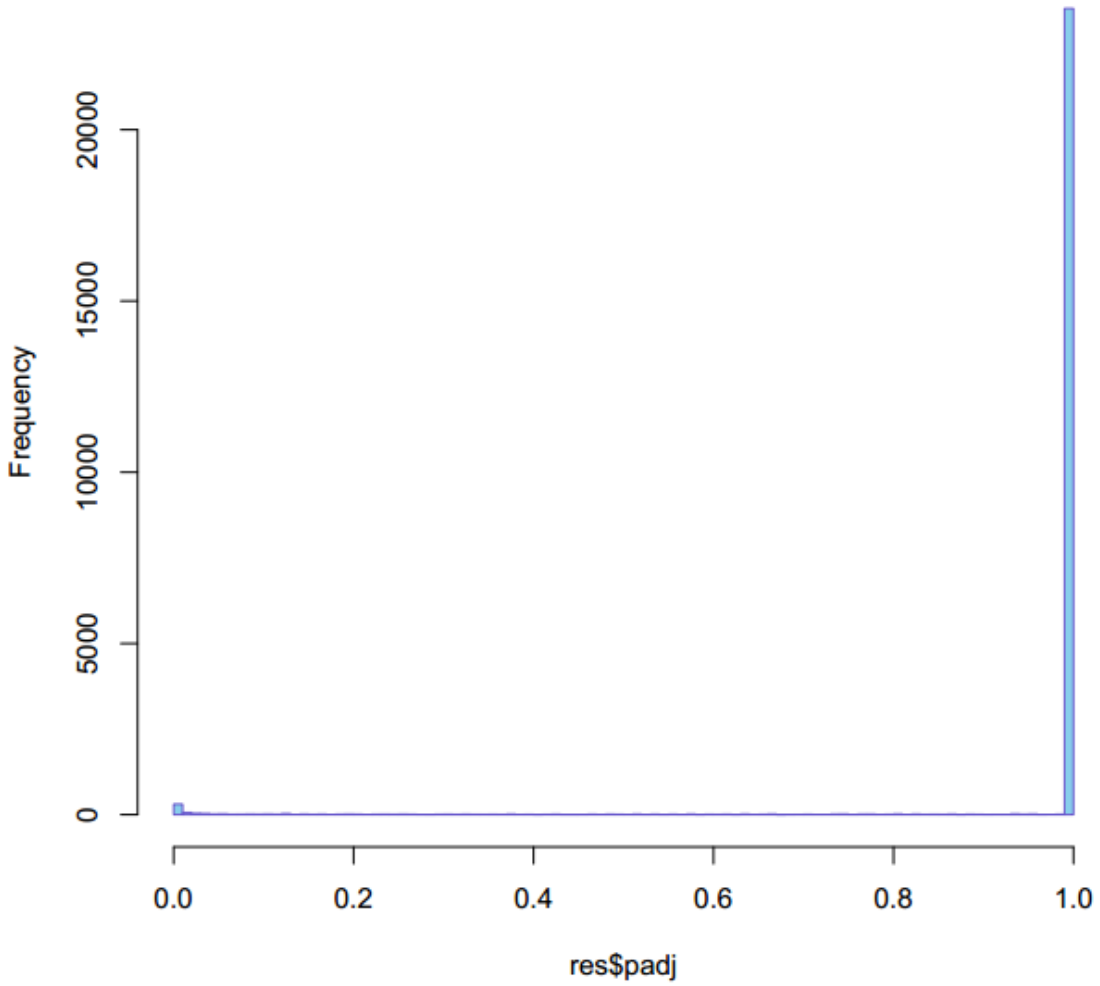

图 11 nbinomTest 的 padj 统计直方图

Fig.11 Padj statistical histogram of nbinomTest

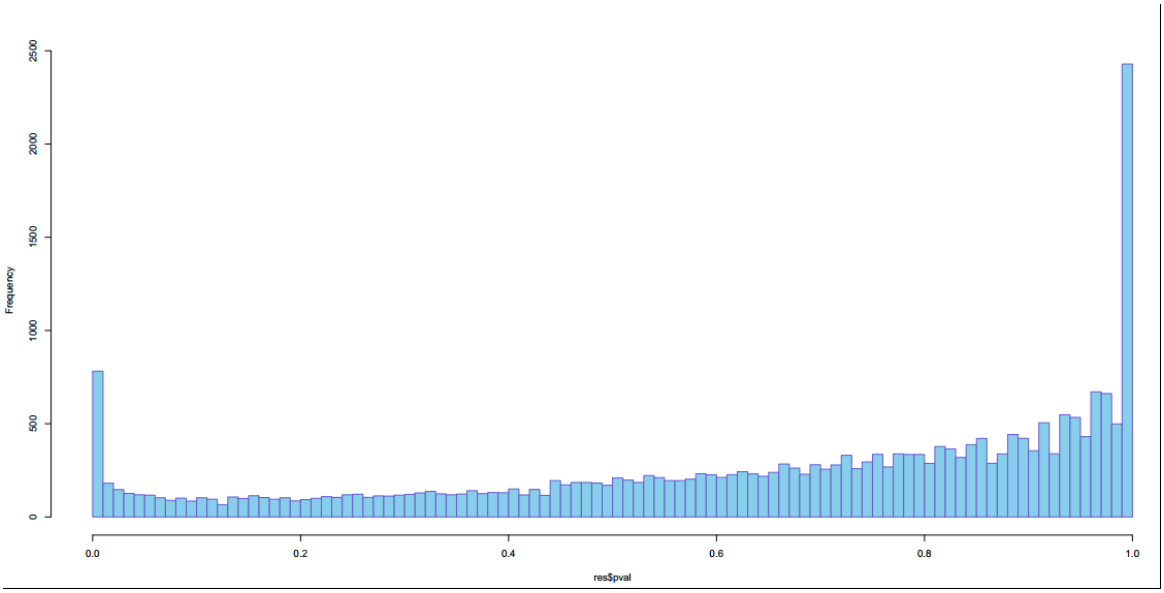

图 12 nbinomTest 的 pval 统计直方图

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。

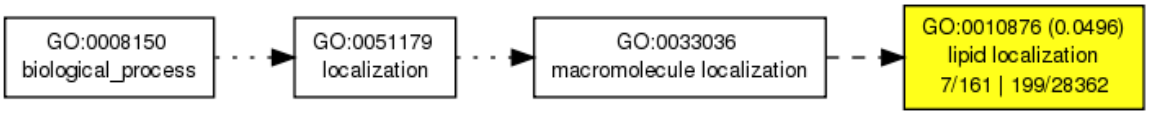

图 13 生物学过程的 GO 分析

Fig.13 GO ontology of biological process

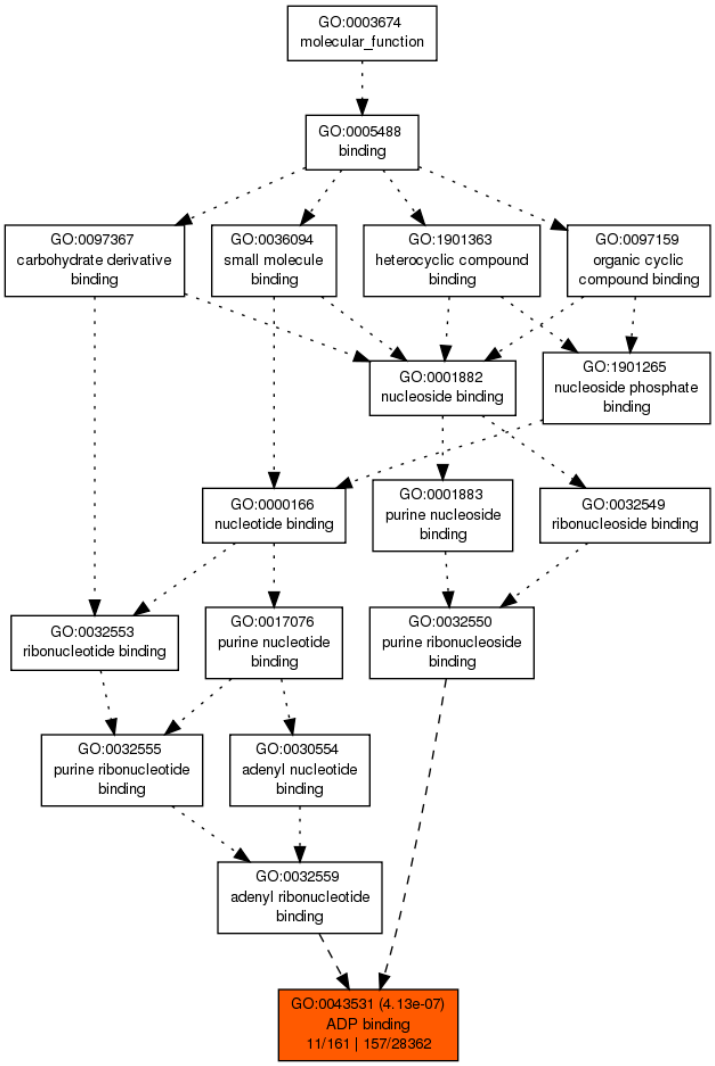

图 14 分子功能的 GO 分析

Fig.14 GO ontology of mollecular function

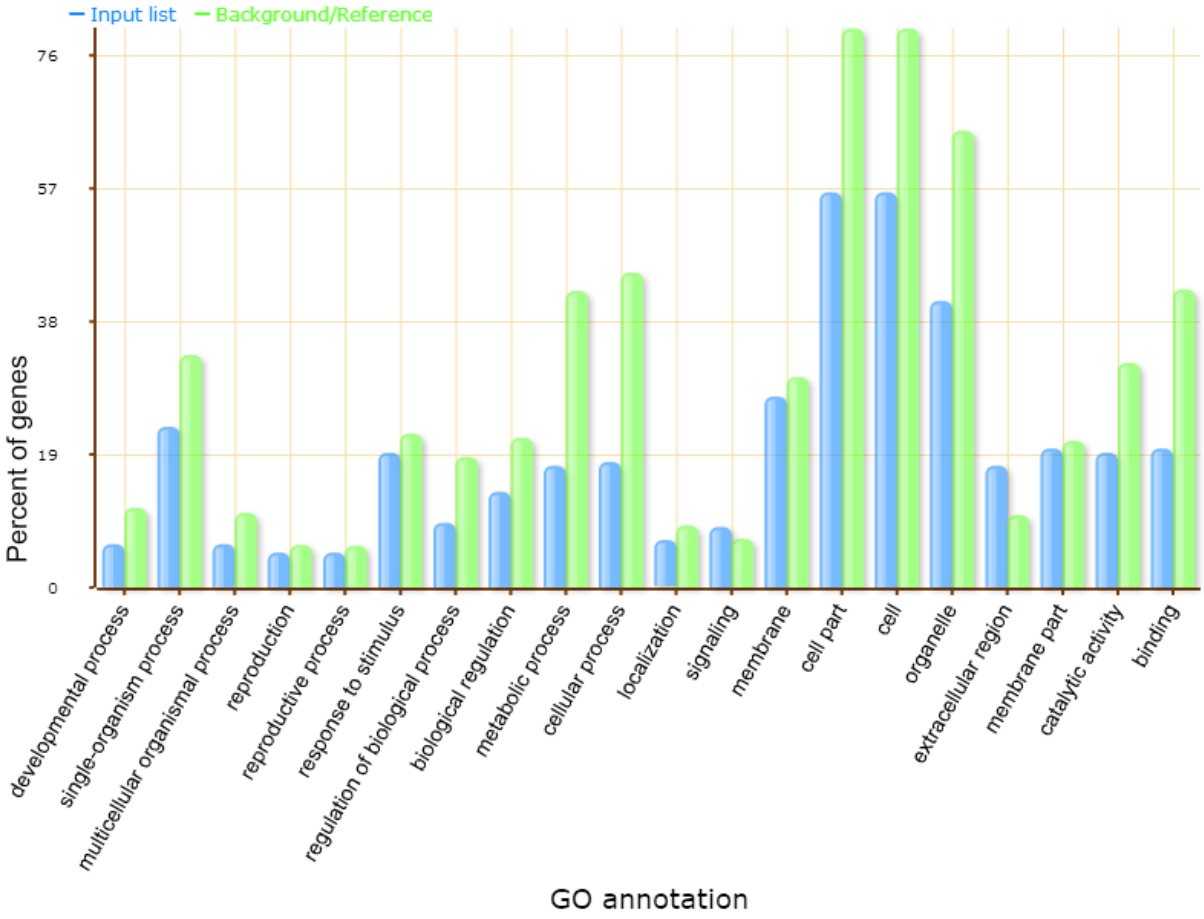

图 15 差异基因的 GO 注释

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_40]$ more flagstat.txt
34950791 + 0 in total (QC-passed reads + QC-failed reads)
352197 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
34950791 + 0 mapped (100.00% : N/A)
34598594 + 0 paired in sequencing
17805296 + 0 read1
16793298 + 0 read2
31026030 + 0 properly paired (89.67% : N/A)
32536310 + 0 with itself and mate mapped
2062284 + 0 singletons (5.96% : N/A)
62328 + 0 with mate mapped to a different chr
58570 + 0 with mate mapped to a different chr (mapQ>=5)
```

samtools flagstat tophat2_41/accepted_rmdup_hits.bam > tophat2_41/flagstat.txt

result:

```
[hhwang@login tophat2_41]$ more flagstat.txt
32933960 + 0 in total (QC-passed reads + QC-failed reads)
267820 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
32933960 + 0 mapped (100.00% : N/A)
32666140 + 0 paired in sequencing
17054186 + 0 read1
15611954 + 0 read2
27528528 + 0 properly paired (84.27% : N/A)
29019962 + 0 with itself and mate mapped
3646178 + 0 singletons (11.16% : N/A)
74638 + 0 with mate mapped to a different chr
65512 + 0 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 Run Info File diff/run.info
Writing runInfo Table
Reading Read Group Info  diff/read_groups.info
Writing replicates Table
Reading Var Model Info  diff/var_model.info
Writing varModel Table
Reading 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：

Chr	tophat2_40/accepted_rmdup_hits.bam		unmapped reads number
	Che length	Mapped 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

Chr	tophat2_41/accepted_rmdup_hits.bam		unmapped reads number
	Che length	Mapped 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	Log2FoldChange3	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")