作业 6
姓名：许芸阁
学号：2019012302
组别：第 4 组

Part I 3.1 R Basics

代码如下：

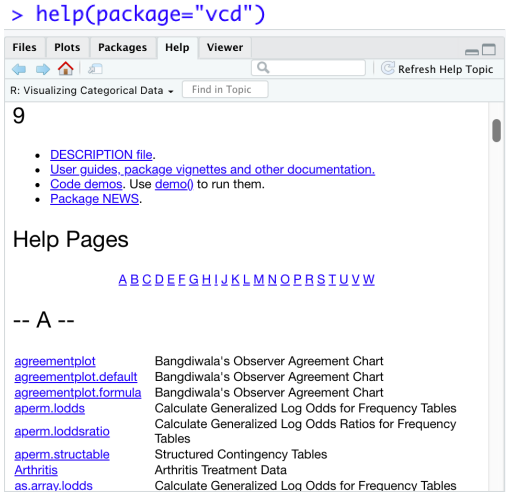
安装vcd包（一个用于可视化类别数据的包）
列出此包中可用的函数和数据集。
载入这个包并阅读数据集Arthritis的描述。
显示数据集Arthritis的内容（直接输入一个对象的名称将列出它的内容）
运行数据集Arthritis自带的示例（尝试用example命令）

```
install.packages("vcd")

help(package="vcd") # 打开的是vcd包的说明（网页）
# 或
library(help="vcd") # 打开的是vcd包的说明（文档）

library(vcd) # 会自动加载vcd依赖的grid包

Arthritis
example(Arthritis)

q()
```

每一步输出的结果如下：

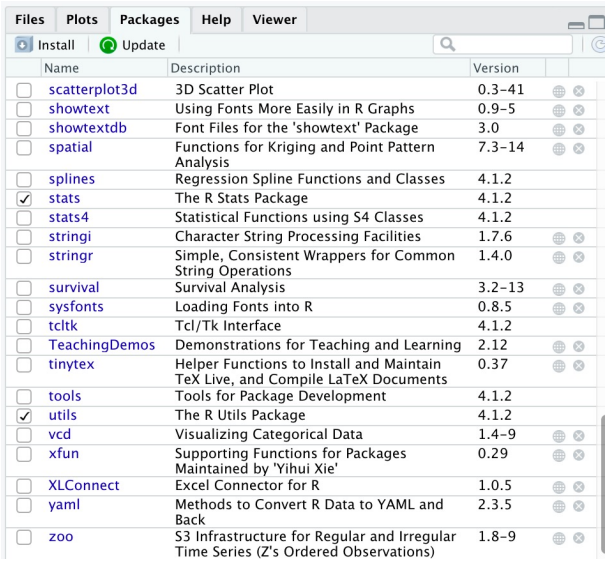
```
> install.packages("vcd")
试开URL'https://cran.rstudio.com/bin/macosx/contrib/4.1/vcd_1.4-9.tgz'
Content type 'application/x-gzip' length 1291646 bytes (1.2 MB)
==================================================
downloaded 1.2 MB


The downloaded binary packages are in
	/var/folders/py/_lkxgrwx2jn664p6fkpj23580000gn/T//RtmpaVDyeu/downloaded_packages
```

```
> help(package="vcd")
```




```
> library(help="vcd")
```

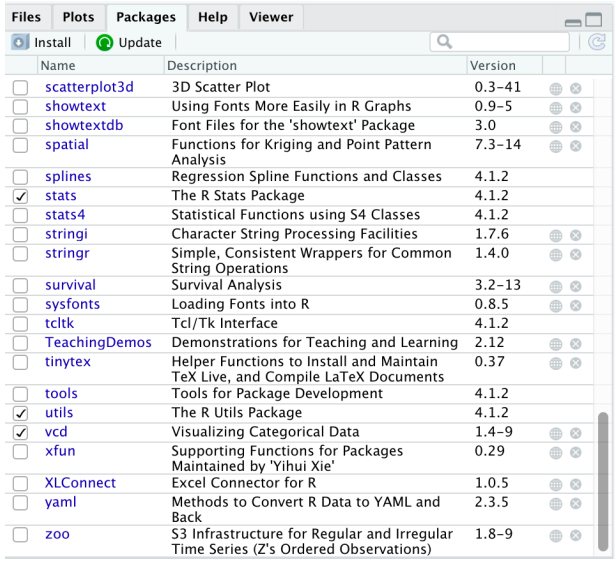
```
> library(vcd)
载入需要的程辑包：grid
```



Files | Plots | Packages | Help | Viewer

Install | Update

	Name	Description	Version
☐	scatterplot3d	3D Scatter Plot	0.3-41
☐	showtext	Using Fonts More Easily in R Graphs	0.9-5
☐	showtextdb	Font Files for the 'showtext' Package	3.0
☐	spatial	Functions for Kriging and Point Pattern Analysis	7.3-14
☐	splines	Regression Spline Functions and Classes	4.1.2
☑	stats	The R Stats Package	4.1.2
☐	stats4	Statistical Functions using S4 Classes	4.1.2
☐	stringi	Character String Processing Facilities	1.7.6
☐	stringr	Simple, Consistent Wrappers for Common String Operations	1.4.0
☐	survival	Survival Analysis	3.2-13
☐	sysfonts	Loading Fonts into R	0.8.5
☐	tcltk	Tcl/Tk Interface	4.1.2
☐	TeachingDemos	Demonstrations for Teaching and Learning	2.12
☐	tinytex	Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents	0.37
☐	tools	Tools for Package Development	4.1.2
☑	utils	The R Utils Package	4.1.2
☐	vcd	Visualizing Categorical Data	1.4-9
☐	xfun	Supporting Functions for Packages Maintained by 'Yihui Xie'	0.29
☐	XLConnect	Excel Connector for R	1.0.5
☐	yaml	Methods to Convert R Data to YAML and Back	2.3.5
☐	zoo	S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)	1.8-9



Files | Plots | Packages | Help | Viewer

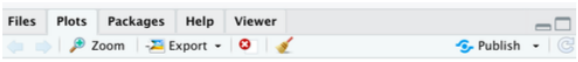
Install | Update

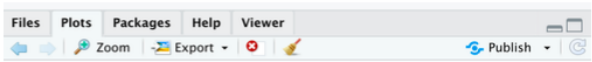
	Name	Description	Version
☐	scatterplot3d	3D Scatter Plot	0.3-41
☐	showtext	Using Fonts More Easily in R Graphs	0.9-5
☐	showtextdb	Font Files for the 'showtext' Package	3.0
☐	spatial	Functions for Kriging and Point Pattern Analysis	7.3-14
☐	splines	Regression Spline Functions and Classes	4.1.2
☑	stats	The R Stats Package	4.1.2
☐	stats4	Statistical Functions using S4 Classes	4.1.2
☐	stringi	Character String Processing Facilities	1.7.6
☐	stringr	Simple, Consistent Wrappers for Common String Operations	1.4.0
☐	survival	Survival Analysis	3.2-13
☐	sysfonts	Loading Fonts into R	0.8.5
☐	tcltk	Tcl/Tk Interface	4.1.2
☐	TeachingDemos	Demonstrations for Teaching and Learning	2.12
☐	tinytex	Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents	0.37
☐	tools	Tools for Package Development	4.1.2
☑	utils	The R Utils Package	4.1.2
☑	vcd	Visualizing Categorical Data	1.4-9
☐	xfun	Supporting Functions for Packages Maintained by 'Yihui Xie'	0.29
☐	XLConnect	Excel Connector for R	1.0.5
☐	yaml	Methods to Convert R Data to YAML and Back	2.3.5
☐	zoo	S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)	1.8-9

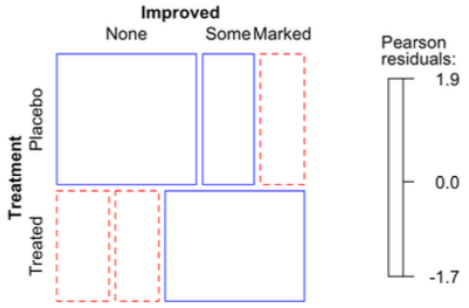
```
> Arthritis
   ID Treatment    Sex Age Improved
1  57   Treated   Male  27     Some
2  46   Treated   Male  29     None
3  77   Treated   Male  30     None
4  17   Treated   Male  32   Marked
5  36   Treated   Male  46   Marked
6  23   Treated   Male  58   Marked
7  75   Treated   Male  59     None
8  39   Treated   Male  59   Marked
9  33   Treated   Male  63     None
10 55   Treated   Male  63     None
11 30   Treated   Male  64     None
12  5   Treated   Male  64     Some
13 63   Treated   Male  69     None
14 83   Treated   Male  70   Marked
15 66   Treated Female  23     None
16 40   Treated Female  32     None
17  6   Treated Female  37     Some
18  7   Treated Female  41     None
19 72   Treated Female  41   Marked
20 37   Treated Female  48     None
21 82   Treated Female  48   Marked
22 53   Treated Female  55   Marked
23 79   Treated Female  55   Marked
24 26   Treated Female  56   Marked
25 28   Treated Female  57   Marked
26 60   Treated Female  57   Marked
27 22   Treated Female  57   Marked
28 27   Treated Female  58     None
29  2   Treated Female  59   Marked
30 59   Treated Female  59   Marked
31 62   Treated Female  60   Marked
32 84   Treated Female  61   Marked
33 64   Treated Female  62     Some
34 34   Treated Female  62   Marked
35 58   Treated Female  66   Marked
36 13   Treated Female  67   Marked
37 61   Treated Female  68     Some
38 65   Treated Female  68   Marked
39 11   Treated Female  69     None
40 56   Treated Female  69     Some
41 43   Treated Female  70     Some
42  9   Placebo   Male  37     None
43 14   Placebo   Male  44     None
44 73   Placebo   Male  50     None
45 74   Placebo   Male  51     None
46 25   Placebo   Male  52     None
47 18   Placebo   Male  53     None
48 21   Placebo   Male  59     None
49 52   Placebo   Male  59     None
50 45   Placebo   Male  62     None
51 41   Placebo   Male  62     None
52  8   Placebo   Male  63   Marked
53 80   Placebo Female  23     None
54 12   Placebo Female  30     None
55 29   Placebo Female  30     None
56 50   Placebo Female  31     Some
57 38   Placebo Female  32     None
58 35   Placebo Female  33   Marked
59 51   Placebo Female  37     None
60 54   Placebo Female  44     None
61 76   Placebo Female  45     None
62 16   Placebo Female  46     None
63 69   Placebo Female  48     None
64 31   Placebo Female  49     None
65 20   Placebo Female  51     None
66 68   Placebo Female  53     None
67 81   Placebo Female  54     None
68  4   Placebo Female  54     None
69 78   Placebo Female  54   Marked
70 70   Placebo Female  55   Marked
71 49   Placebo Female  57     None
72 10   Placebo Female  57     Some
73 47   Placebo Female  58     Some
74 44   Placebo Female  59     Some
75 24   Placebo Female  59   Marked
76 48   Placebo Female  61     None
77 19   Placebo Female  63     Some
78  3   Placebo Female  64     None
79 67   Placebo Female  65   Marked
80 32   Placebo Female  66     None
81 42   Placebo Female  66     None
82 15   Placebo Female  66     Some
83 71   Placebo Female  68     Some
84  1   Placebo Female  74   Marked
```

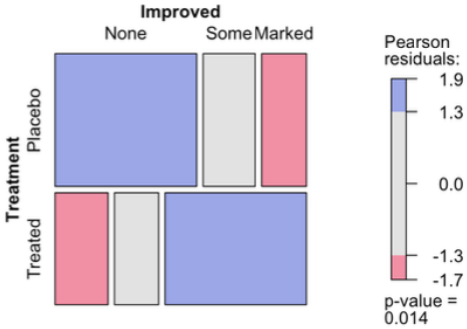
```
> example("Arthritis")

Arthrt> data("Arthritis")

Arthrt> art <- xtabs(~ Treatment + Improved, data = Arthritis, subset = Sex == "Female")

Arthrt> art
         Improved
Treatment None Some Marked
  Placebo   19    7      6
  Treated    6    5     16

Arthrt> mosaic(art, gp = shading_Friendly)
按<Return>键来看下一个图:
```

Data	
Arthritis	84 obs. of 5 variables
Values	
art	'xtabs' int [1:2, 1:3] 19 6 7 5 6 16









Part III 2.1 Expression Matrix

1. E，D，A

- Standard illumina 是 PE、non-strand specific 的建库方法，只要 map 到 geneG 所在区域的 reads 无论方向如何，都归于 geneG，共 13 条；
- Ligation method 是 PE、strand specific 的建库方法，reads 1 是 sense 的（reads 2 是 antisense 的），因此只有与 geneG 同向的 reads 1 和与 geneG 反向的 reads 2 归于 geneG，共 9 条；
- dUTPs method 是 PE、strand specific 的建库方法，reads 2 是 sense 的（reads 1 是 antisense 的），因此只有与 geneG 同向的 reads 2 和与 geneG 反向的 reads 1 归于 geneG，共 4 条；

2. 1) 来自 Paired end、non-strand specific 的测序方法。用 infer_experiment.py 推断采用的测序策略，从结果（见下方代码）中我们可以看到，“1++,1--,2+-,2-+”与“1+-,1-+,2++,2--”的比例几乎相同，因此有很大的把握认定这个数据是由非链特异性建库得到的。

解释：

1、2 表示 reads1、reads2，如果出现，说明这是一个 Paired end 测序

1+-表示如果 reads1 map 到+链上，说明这个 reads 对应的是-链的基因，依此类推可知其余符号组合的含义

1++,1--,2+-,2-+类似于 Ligation method 中的情形

1+-,1-+,2++,2--类似于 dUTPs method 中的情形

如果 1++,1--,2+-,2-+与 1+-,1-+,2++,2--的比例大致相同，表明这很有可能是 non-strand specific 测序得来的结果

如果 1++,1--,2+-,2-+ >> 1+-,1-+,2++,2--，表明这很有可能是 strand specific 测序得来的结果（Ligation method）

如果 1++,1--,2+-,2-+ << 1+-,1-+,2++,2--，表明这很有可能是 strand specific 测序得来的结果（dUTP、PICO）

2)　AT1G09530 基因(PIF3 基因)上的 reads/counts 数目为 891

代码如下：

```
docker exec -it bioinfo_tsinghua_featurecount bash
cd /home/test
```

```
/usr/local/bin/infer_experiment.py \
-r GTF/Arabidopsis_thaliana.TAIR10.34.bed \
-i bam/Shape02.bam
# 结果:
# This is PairEnd Data
# Fraction of reads failed to determine: 0.0277
# Fraction of reads explained by "1++,1--,2+-,2-+": 0.4783
# Fraction of reads explained by "1+-,1-+,2++,2--": 0.4939
```

```
/home/software/subread-1.6.0-Linux-x86_64/bin/featureCounts \
-s 0 \
-p -t exon \
-g gene_id \
-a GTF/Arabidopsis_thaliana.TAIR10.34.gtf \
-o result/Shape02.featurecounts.exon.txt bam/Shape02.bam
echo read_counts success
```

```
echo -e "gene_id\tShape02" > result/Shape02.txt
cat result/Shape02.featurecounts.exon.txt | grep -v -w '#' | \
grep -v -w 'Geneid' | cut -f 1,7 >> result/Shape02.txt
cat result/Shape02.txt | grep -w "AT1G09530"
# 结果: AT1G09530 891
```

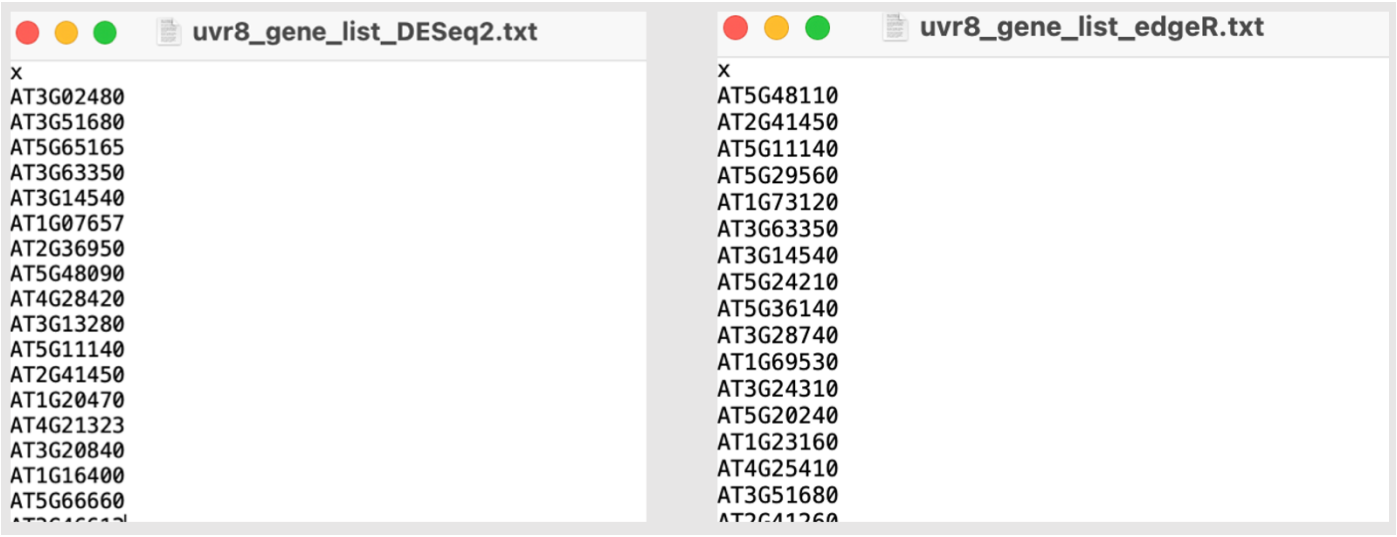
Part III 2.3 Differential Expression with DESeq2 and edgeR

得到的文件：（文件中的基因见后面）

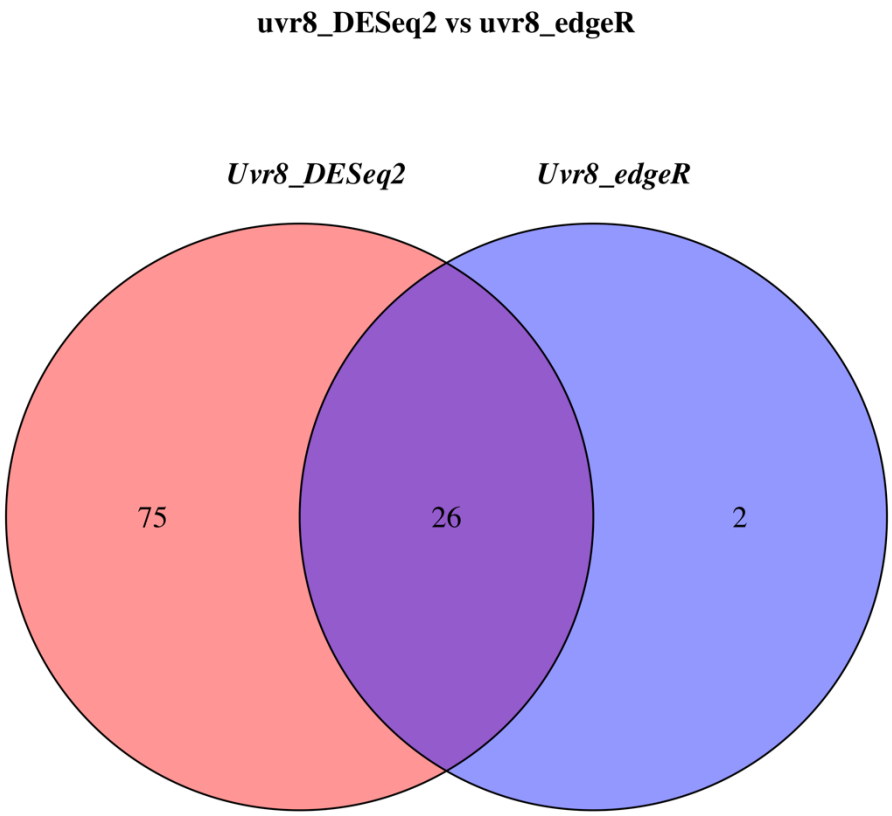
uvr8_gene_list_DESeq2.txt，共 101 个基因

uvr8_gene_list_edgeR.txt，共 28 个基因

局部截图：



Venn 图：



DESeq2 代码：

```
library(DESeq2)

raw_count <- read.table("count_exon.txt", sep='\t', header = T)

#uvr8突变型
uvr8_raw_count <- raw_count[c("gene_id", "UD1_1", "UD1_2", "UD1_3", "UD0_1", "UD0_2", "UD0_3")]
row.names(uvr8_raw_count) <- uvr8_raw_count[, 1]
uvr8_raw_count <- uvr8_raw_count[, -1]

#过滤
countData_uvr8 <- uvr8_raw_count[rowSums(uvr8_raw_count) > 100, ]

condition_merge <- factor(c(rep("Control", 3), rep("Treat", 3)))
colData <- data.frame(row.names = colnames(countData_uvr8), condition_merge)

#从表达矩阵出发初始化DESeqDataSet对象
dds <- DESeqDataSetFromMatrix(countData_uvr8, colData, design = ~condition_merge)

#进行差异分析
dds2 <- DESeq(dds)

#获取结果
res <- results(dds2)
res <- res[order(res$padj), ]

#过滤标准
diff_gene_deseq_uvr8 <- subset(res, padj < 0.05 & abs(log2FoldChange) > 1)

#提取差异基因名称
uvr8_gene_names <- row.names(diff_gene_deseq_uvr8)
write.table(uvr8_gene_names,"uvr8_gene_list_DESeq2.txt", sep='\t', row.names = F, quote = F)
```

edgeR 代码：

```
library(edgeR)

raw_count <- read.table("count_exon.txt",sep='\t',header=T)

#提取uvr8突变型
uvr8_raw_count <- raw_count[c("gene_id", "UD1_1", "UD1_2", "UD1_3", "UD0_1", "UD0_2", "UD0_3")]
row.names(uvr8_raw_count) <- uvr8_raw_count[, 1]
uvr8_raw_count <- uvr8_raw_count[, -1]

#过滤
countData_uvr8 <- uvr8_raw_count[rowSums(uvr8_raw_count) > 100, ]

dgListGroups <- c(rep("Control", 3), rep("Treat", 3))
dgList <- DGEList(counts = countData_uvr8, genes = rownames(countData_uvr8), group = factor(dgListGroups))

#TMM标准化
dgList <- calcNormFactors(dgList, method="TMM")

#获取design矩阵
design.mat <- model.matrix(~dgList$sample$group)
colnames(design.mat) <- levels(dgList$sample$group)

#对负二项分布模型进行参数估计
d2 <- estimateGLMCommonDisp(dgList, design=design.mat)
d2 <- estimateGLMTrendedDisp(d2, design=design.mat)
d2 <- estimateGLMTagwiseDisp(d2, design=design.mat)

#似然比检验
fit <- glmFit(d2, design.mat)
lrt <- glmLRT(fit,coef=2)

edgeR_result <- topTags(lrt, n = nrow(dgList))$table
edgeR_result <- edgeR_result[which(abs(edgeR_result$logFC) > 1 & edgeR_result$FDR < 0.05), ]

#输出差异显著的基因
write.table(edgeR_result$genes, file = 'uvr8_gene_list_edgeR.txt', sep = "\t",
quote = F, row.names = F, col.names = T)
```

两个结果文件中的基因：

uvr8_gene_list_DESeq2.txt

x AT3G02480 AT3G51680 AT5G65165 AT3G63350 AT3G14540 AT1G07657 AT2G36950 AT5G48090 AT4G28420 AT3G13280 AT5G11140 AT2G41450 AT1G20470 AT4G21323 AT3G20840 AT1G16400 AT5G66660 AT3G46613 AT4G25410 AT1G27670 AT5G48110 AT5G51470 AT1G44318 AT3G24310 AT1G23160 AT2G47770 AT2G47780 AT1G77885 AT1G65390 AT1G04187 AT2G05520 AT2G21820 AT5G56400 AT2G27120 AT5G20240 AT1G05557 AT1G69530 AT1G73120 AT1G58050 AT4G32490 AT1G78390 AT2G23800 AT2G43610 AT1G47130 AT5G36140 AT4G17470 AT2G41445 AT4G37800 AT5G47280 AT2G41451 AT5G24210 AT3G19560 AT4G31870 AT2G43050 AT1G70420 AT3G07255 AT1G30190 AT1G07180 AT2G03360 AT4G23590 AT1G49450 AT1G67110 AT1G04370 AT1G09110 AT1G08947 AT3G58480 AT4G27670 AT3G28740 AT5G07480 AT3G07675 AT4G08040 AT4G05200 AT3G15670 AT5G04205 AT3G55240 AT1G03982 AT5G66650 AT2G41260 AT2G40100 AT1G64110 AT1G66100 AT5G16980 AT2G28340 AT3G10950 AT5G47175 AT2G16005 AT3G19610 AT5G45573 AT1G56250 AT2G42540 AT4G32500 AT4G32510 AT5G10510 AT1G32450 AT3G22600 AT1G27135 AT4G31950 AT1G07493 AT4G01780 AT1G71390 AT4G22214

uvr8_gene_list_edgeR.txt

x AT5G48110 AT2G41450 AT5G11140 AT5G29560 AT1G73120 AT3G63350 AT3G14540 AT5G24210 AT5G36140 AT3G28740 AT1G69530 AT3G24310 AT5G20240 AT1G23160 AT4G25410 AT3G51680 AT2G41260 AT3G46613 AT2G36950 AT2G16005 AT2G43050 AT4G25580 AT1G77885 AT4G37800 AT4G17470 AT1G47130 AT4G31870 AT5G07480

制作 Venn 图的代码：

```
#把表格中的基因导入dataframe中存储
df1 <- read.table("uvr8_gene_list_DESeq2.txt", sep='\t', header = T)
df2 <- read.table("uvr8_gene_list_edgeR.txt", sep='\t', header = T)

#把dataframe转为list，用于制作venn图
df1 <- t(df1)
df2 <- t(df2)
df1 <- as.data.frame(df1)
df2 <- as.data.frame(df2)
df1 <- as.list(df1)
df2 <- as.list(df2)

#制作venn图
library(VennDiagram)
venn.diagram(list(Uvr8_DESeq2=df1, Uvr8_edgeR=df2), #数据
scaled=FALSE, #不根据数字大小调整venn图形状
resolution = 300, #清晰度
imagetype = "tiff", #图片类型
alpha=c(0.5, 0.5), #区域透明度
cex=2, #数字字号
fill=c("red","blue"), #区域颜色
cat.fontface=4, #区域标题字体
cat.cex=2, #区域标题字号
cat.just=list(c(-1, -6), c(2, -6)), #区域标题位置
main="uvr8_DESeq2 vs uvr8_edgeR", #venn图标题
main.cex=2, #标题字号
main.fontface=2, #标题字体
filename="uvr8_DESeq2_vs_edgeR.tif" #文件名)
```