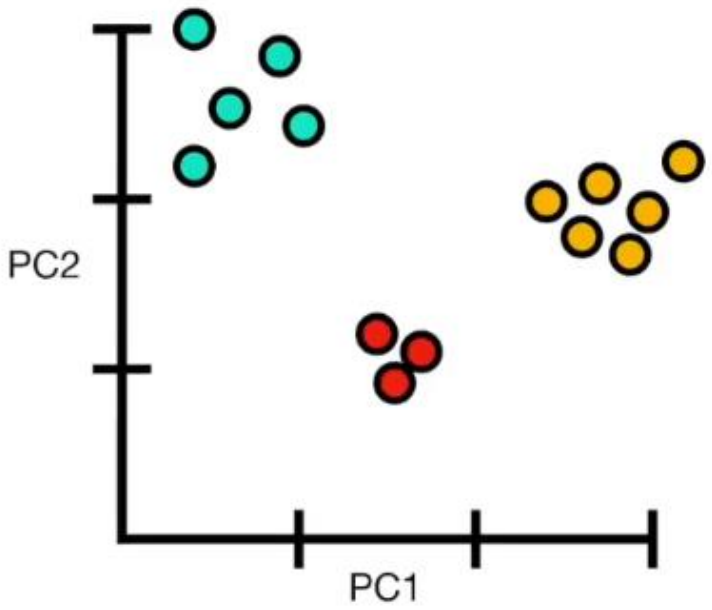
微生信
Wei Sheng Xin

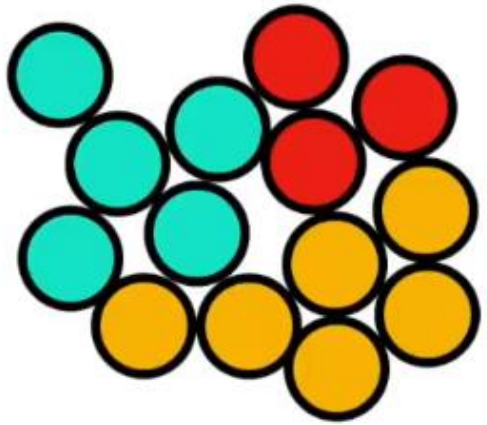
PCA、相关性、热图和火山图

陈明杰202411

PCA降维

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...



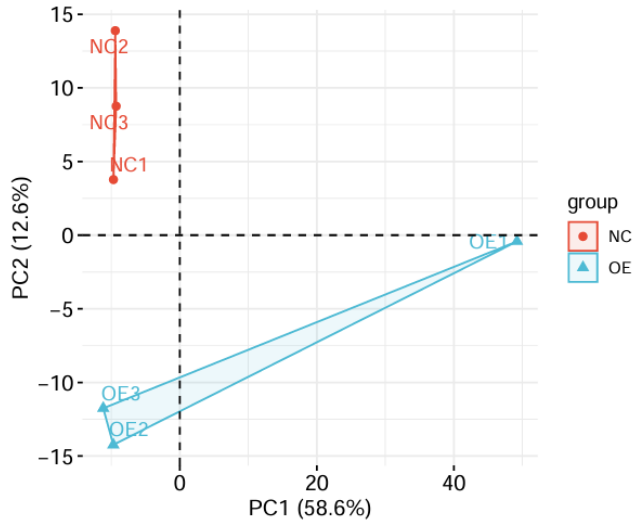


PCA图

	A	B	C	D	E	F	G
1	gene_id	NC1	NC2	NC3	OE1	OE2	OE3
2	group	NC	NC	NC	OE	OE	OE
3	ENSG00000139364	0.039618	0.039606	0.029925	0.036882	0	0
4	ENSG00000163554	0.289785	0.236232	0.342498	0.114713	0.271017	0.340651
5	ENSG00000184347	0	0	0	0.233723	0.051854	0
6	ENSG00000080854	0	0	0.079787	0.033393	0	0
7	ENSG00000115474	0	0.16819	0.246376	0.29874	1.197007	0
8	ENSG00000078295	0.036385	0	0.106897	0	0.161737	0.279819
9	ENSG00000171631	0	0.058958	0	0.054928	0	0.070327
10	ENSG00000184261	0	0	0	0.237534	0.152806	0.054359

05_pca.r

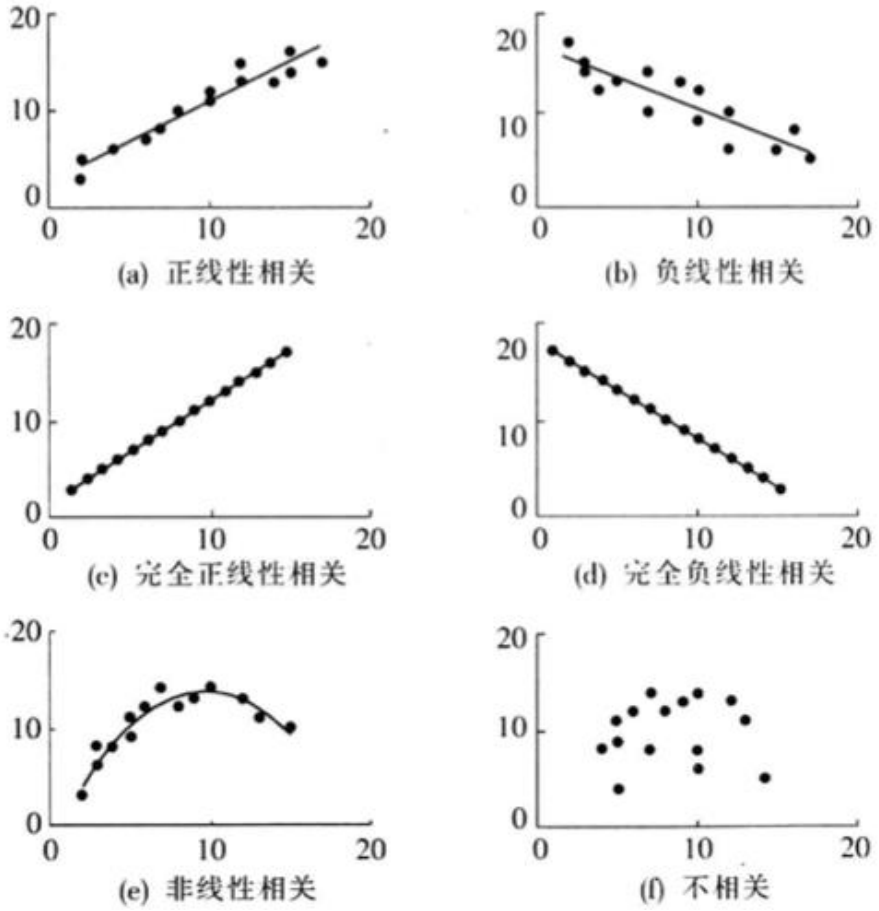
```r
library(FactoMineR)
library(factoextra)
library(ggpubr)

# 行是基因，列是样品，先转置
# 读取data.txt文件中的数据
data <- read.table(file="05_pca.txt", header = FALSE, sep = "\t", row.names=1, check.names = FALSE)

# 转置数据
transposed_data <- t(data)

# 将转置后的数据写入newdata.txt文件
write.table(transposed_data, file="data2.txt", sep = "\t", row.names = FALSE, col.names = T, quote = FALSE)

# 读取转置后的数据，其中第2列为分组
data2 = read.table('data2.txt', header=T, row.names=1, sep='\t', check.names=FALSE, quote="")
newdata = data2[,-1]
newdata = newdata[, which(apply(newdata, 2, var) != 0)]
res.pca <- prcomp(newdata, center=T, scale. = TRUE)
pdf(file = "05_pca.pdf", width = 5, height = 5, onefile = FALSE)
ind.p=fviz_pca_ind(res.pca, repel = TRUE, mean.point = FALSE, geom.ind =c("point", "text"),
     col.ind = as.character(data2$group), palette = c("#E64B35","#4DBBD5"),
     addEllipses = TRUE, ellipse.type = "convex", legend.title = "group", pointsize=2.000000, show.legend = FALSE)+
     font("x", size=12.000000, color="#000000")+ font("y", size=12.000000, color="#000000")+
     font("xy.text", size = 12.000000, color = "#000000")+ theme(aspect.ratio=1)
ggpar(ind.p, title = "", subtitle = "", caption = "", xlab =gsub("Dim","PC", ind.p$labels$x), ylab =gsub("Dim","PC", ind.p$labels$y))
dev.off()
```



相关性

相关系数（correlation coefficient）是一个介于-1和1之间的数，用于衡量两个变量之间的线性关系。如果两个变量之间的变化趋势是一致的，即一个变量增加时另一个也增加，那么它们之间的相关系数为正；如果一个变量增加而另一个减少，则相关系数为负。接近1的相关系数表示强正相关。接近-1的相关系数表示强负相关。接近0的相关系数表示两个变量之间没有或几乎没有线性关系

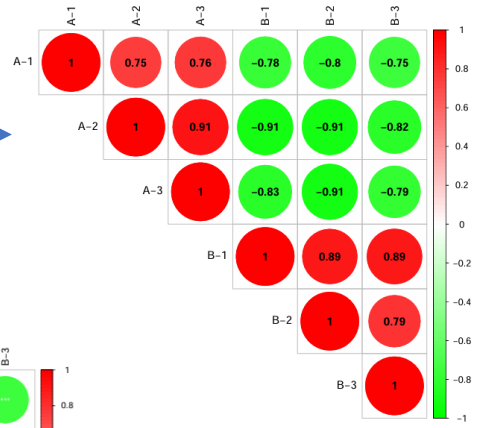
- 皮尔逊相关系数（Pearson correlation coefficient）：适用于测量两个连续变量之间的线性相关性
- 斯皮尔曼等级相关系数（Spearman’s rank correlation coefficient）：适用于测量两个变量的等级（或排序）之间的相关性，不要求数据是线性的

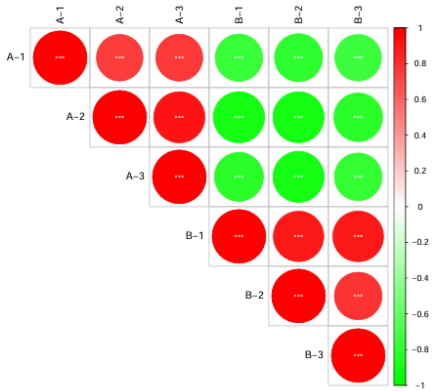


样品相关性图

```r
# 载入包
library("corrplot")

# 读取数据
data = as.matrix(read.table("11_corr.txt", header=TRUE, row.names=1,sep="\t", check.names=FALSE,quote = ""))

# 计算相关性（按列计算）和P值
M = cor(data)

res1 <- cor.mtest(M)

# colorbar颜色
cols = colorRampPalette(c("#00ff00", "#ffffff", "#fe0000"))

# 绘图
pdf(file = "corr1.pdf")
# 仅绘制相关性
corrplot(corr=M,
    method = 'circle',
    type="upper",
    order="original",
    tl.col="black",
    is.corr=T,
    col=cols(100),
    pch.cex=0.8,
    pch.col="white",
    col.lim=c(-1, 1),

    addCoef.col="#000000")

dev.off()

pdf(file="corr2.pdf")
# p值以*表示
corrplot(corr=M, p.mat=res1$p,
    method = 'circle',
    type="upper",
    order="original",
    tl.col="black",
    is.corr=T,
    col=cols(100),
    pch.cex=0.8,
    pch.col="white",
    insig="label_sig",
    sig.level=c(0.001, 0.01, 005))
dev.off()


#    method = c("circle", "square", "ellipse", "number", "shade", "color", "pie"),
#    type = c("full", "lower", "upper"),

# https://www.jianshu.com/p/6436792323b7
```

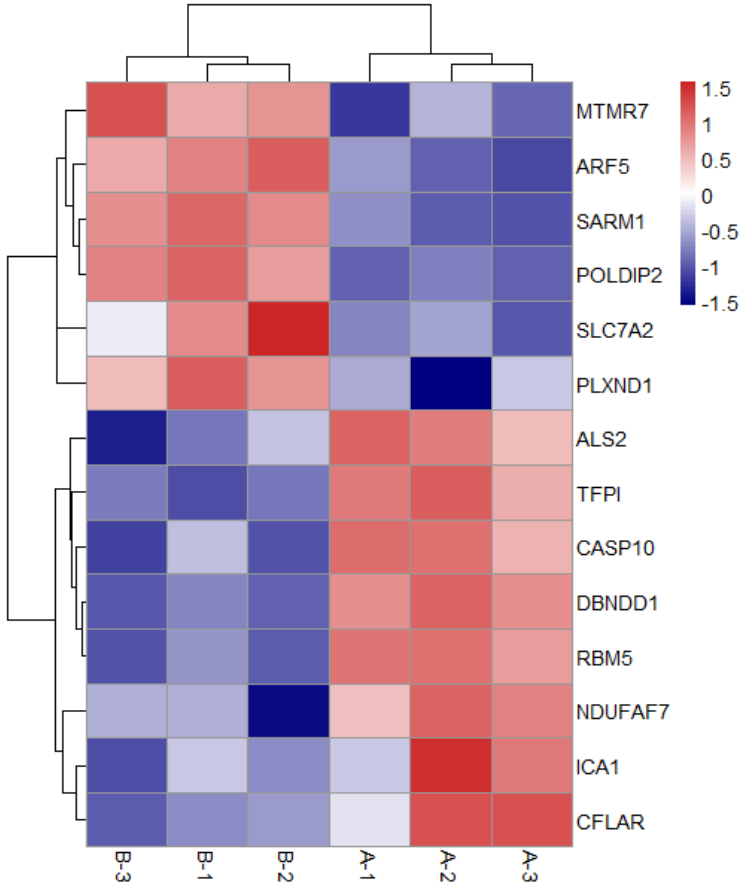
	A	B	C	D	E	F	G
1	sample	A-1	A-2	A-3	B-1	B-2	B-3
2	ICA1	5.7	11.7	9.9	5.7	4.5	3.2
3	DBNDD1	9.43	10.67	9.39	3.4	2.5	2.3
4	ALS2	10.59	9.89	8.5	4.2	5.75	2.5
5	CASP10	10.38	10.2	8.5	5.1	2.8	2.4
6	CFLAR	5.75	10.85	10.9	3.9	4.2	2.8
7	TFPI	9.82	10.45	8.5	2.5	3.4	3.5
8	NDUFAF7	8.9	11.02	10.33	5.75	2.4	5.75
9	RBM5	10.59	10.67	9.94	6.55	5.75	5.55
10	MTMR7	2.5	5.2	3.5	9.21	9.76	11.47
11	SLC7A2	3.5	4.2	2.5	9.21	11.78	5.75





热图简介

1. **行和列**：**行**通常代表基因，**列**代表不同的样本。
2. **颜色编码**：颜色的深浅通常表示基因表达水平的高低，通常使用蓝色到红色的渐变，其中蓝色可能代表低表达，红色代表高表达。
3. **聚类分析**：热图通常包括行聚类（基因聚类）和/或列聚类（样本聚类），以展示基因表达模式的相似性或样本之间的差异。
4. **标准化**：数据通常需要进行标准化处理，以确保不同基因和样本之间的可比性。
5. **数值显示**：有时在热图的单元格内显示具体的数值，以提供更精确的信息。
6. **注释**：热图可以包含行注释和列注释，提供额外的信息，如样本的分组信息、基因的功能分类等。



pheatmap绘制热图

05_heatmap.r

```r
# 安装pheatmap包
# install.packages("pheatmap")

# https://www.rdocumentation.org/packages/pheatmap/versions/1.0.12/topics/pheatmap

# 加载pheatmap包
library(pheatmap)

# 设置工作目录
#setwd("C:\\Users\\uuu\\Desktop\\培训材料\\05_RNAseq数据处理\\06_热图和火山图")

# 读取数据
data <- read.table("01_heatmapdata.txt", header = TRUE, row.names = 1, check.names=F, sep = "\t")

# 检查数据
head(data)

# 列注释-分组信息
col_group <- data.frame(Group=factor(c("groupA","groupA", "groupA", "groupB", "groupB","groupB")))
rownames(col_group) <- colnames(data)

# 行注释
row_group <- data.frame(Regulation = factor(c("down","down","down","down","down","down","down","down",
"up","up","up","up","up","up")))

rownames(row_group) <- rownames(data)

# 定义分组信息的颜色
annotation_row_colors <- list(groupA = "red", groupB = "blue")
annotation_col_colors <- list(up = "red", down = "green")

pdf(file='05_heatmap.pdf', width=6, height=6)

# 绘制热图
pheatmap(data,
         scale = "row", # 按行标准化
         cluster_rows = TRUE, # 行聚类
         cluster_cols = TRUE, # 列聚类
         #cellwidth = 10, # 单元格宽度
         #cellheight = 10, # 单元格高度
         border_color = NA, #"grey60", # 单元格边框颜色
         show_rownames = TRUE, # 显示行名
         show_colnames = TRUE, # 显示列名
         annotation_row = row_group, # 行注释
         annotation_col = col_group, # 列注释
         #annotation_colors = list(row = annotation_row_colors, col = annotation_col_colors), # 注释颜色
         color = colorRampPalette(c("blue", "white", "#ff0000"))(100)) # 颜色渐变
dev.off()
```

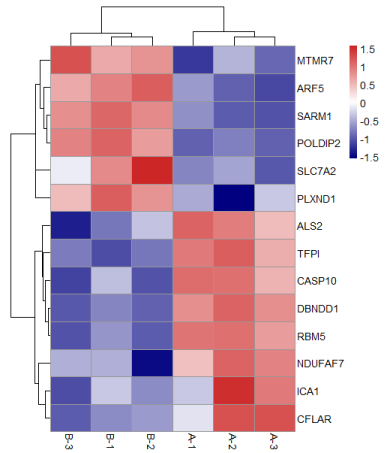
	A	B	C	D	E	F	G
1	sample	A-1	A-2	A-3	B-1	B-2	B-3
2	ICA1	5.7	11.7	9.9	5.7	4.5	3.2
3	DBNDD1	9.43	10.67	9.39	3.4	2.5	2.3
4	ALS2	10.59	9.89	8.5	4.2	5.75	2.5
5	CASP10	10.38	10.2	8.5	5.1	2.8	2.4

列注释：分组

行注释

注释方块的颜色

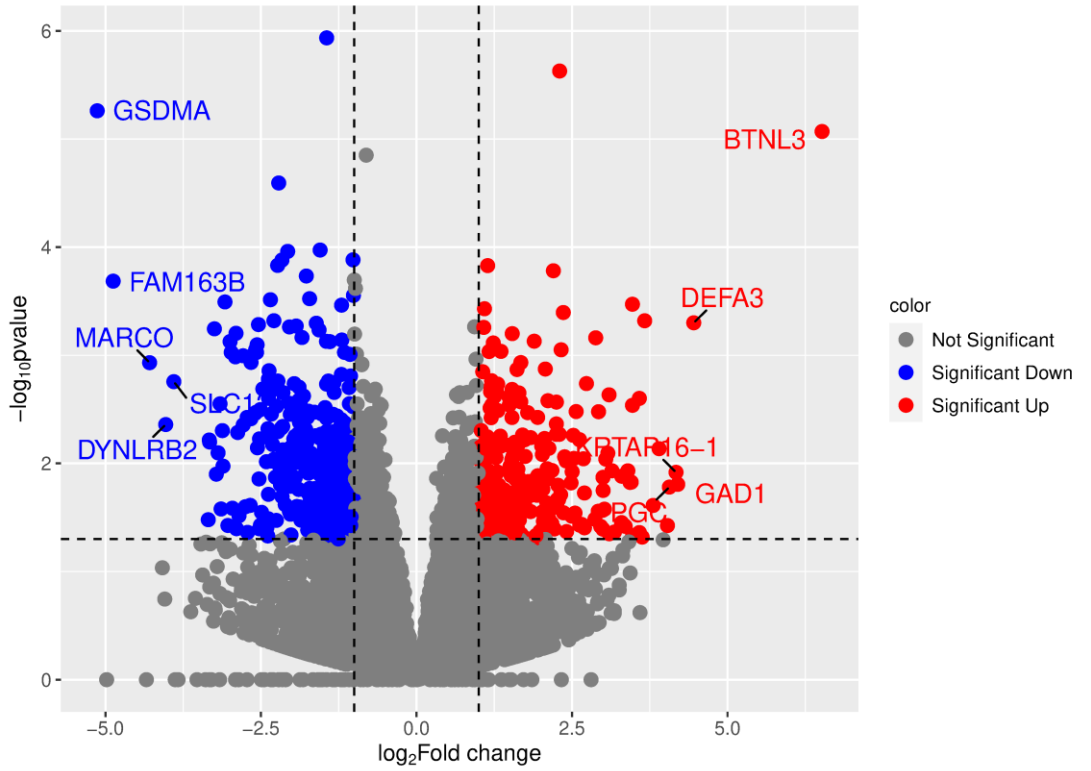
绘图参数



火山图简介

火山图（Volcano Plot）是一种常用于基因表达数据分析的可视化工具，它结合了差异表达的统计显著性和表达变化的幅度，用于快速识别具有显著变化的基因。火山图因其形状类似火山而得名。

- **x轴**通常表示对数变化倍数（Log2 Fold Change），即基因在两个不同条件下表达水平的对数比值，通常以2为底
- **y轴**表示统计显著性，如负对数p值（-log10(p-value)），表示基因表达差异的显著性
- **颜色**：使用颜色来表示基因表达变化的方向，例如红色表示上调，绿色表示下调，灰色表示不显著
- **阈值线**：图中通常有3条阈值线，一条表示显著性阈值（如p值小于0.05），另一条表示变化倍数的阈值（如变化倍数大于2倍和小于-2倍）。
- **基因点：**每个点代表一个基因，点的位置表示该基因的Log2FC和显著性-log10(pvalue)。



ggplot2绘制火山图

02_volcanodata.r

```r
# 安装包
if (!requireNamespace("ggplot2", quietly = TRUE)) {
    install.packages("ggplot2")
}

# 载入包
library(ggplot2) # 画图
library(ggrepel) # 标注基因
library(dplyr)   # 数据过滤


# 读取数据
volcano_data <- read.table('02_volcanodata.txt', sep='\t', quote="", header=T, check.names=F)

# 将p值转换为负对数10
volcano_data$minusLog10Pvalue <- -log10(volcano_data$pvalue)

# 根据显著性p和倍数变化fc设置颜色
volcano_data$color <- ifelse(volcano_data$log2FoldChange > 1 & volcano_data$pvalue < 0.05, "Significant Up",
                             ifelse(volcano_data$log2FoldChange < -1 & volcano_data$pvalue < 0.05, "Significant Down", "Not Significant"))

# 添加感兴趣基因名
vals <- c("PGC", "GAD1", "BTNL3", "KRTAP16-1", "DEFA3", "MARCO", "SLC12A1", "GSDMA", "DYNLRB2", "FAM163B")

# 使用 filter() 函数根据条件筛选数据
volcano_data_new <- filter(volcano_data, symbol %in% vals)

# 绘制火山图
pdf(file='02_volcano.pdf', width=8, height=6)
volcano_plot <- ggplot(volcano_data, aes(x = log2FoldChange, y = minusLog10Pvalue, color = color)) +
  geom_point(size = 3) +
  scale_color_manual(values = c("Significant Up" = "red", "Significant Down" = "blue", "Not Significant" = "grey50")) +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "black") +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed", color = "black") +
  #theme_minimal() +
  theme(
    text = element_text(size = 12), # 修改全局字体大小
    plot.title = element_text(size = 14, hjust = 0.5), # 修改标题字体大小和水平位置
    axis.title = element_text(size = 12), # 修改轴标题字体大小
    axis.text = element_text(size = 10), # 修改轴文本字体大小
    legend.title = element_text(size = 10), # 修改图例标题字体大小
    legend.text = element_text(size = 10), # 修改图例文本字体大小
  )+
labs(title="",  x = expression("log"[2]*"Fold change"), y = expression("-log"[10]*"pvalue"))+
guides(color = guide_legend(override.aes = list(label = '')))+ # 去掉legend点上的文字
geom_text_repel(data=volcano_data_new, aes(label=vals), size=5.000000,
    segment.color = "#000000", box.padding = unit(0.5, 'lines'))+  # 标注感兴趣基因
theme(legend.position = "right")

# 显示火山图
print(volcano_plot)
dev.off()
```

	A	B	C
1	symbol	log2FoldChange	pvalue
2	TNMD	1.783725956	0.278199
3	DPM1	0.149060686	0.564554
4	SCYL3	0.408288137	0.090887
5	FGR	0.344091407	0.585774
6	CFH	0.055331141	0.915372
7	NFYA	0.403569632	0.047838
8	STPG1	0.54863332	0.169866

标注基因
从上到下，按顺序