第五次实验作业

10 MAY 2021

> 实验背景
>
> > 从转录组测序比对数据出发，有多种方法来分析其中的信息[1]。其中一种是差异表达基因分析。在这里，我们根据不同的分组条件设置来对组建的基因表达作假设检验，以发现其中显著差异表达的基因。
>
> 实验目的
>
> > 对比大不列颠群岛和伊巴丹岛不同性别人群的转录组测序结果，发现不同组别下差异表达的基因，并对其进行富集分析以自动注释其中的含义。
>
> 运行环境（见下）

```
platform          x86_64-pc-linux-gnu
arch              x86_64
os                linux-gnu
system            x86_64, linux-gnu
R version         4.0.5 (2021-03-31)
nickname          Lost Library Book
radian version    0.5.10
python version    3.7.3
xonsh version     0.9.27
```

不同人群与性别个体中基因表达差异相关性分析

实验方法

对于上游比对获得的转录本原始计数，使用 `DESeq2` [2] 软件包，分析基因表达对不同变量的敏感性；并且依赖 `clusterProfiler` [3] 对得到的差异表达基因作富集注释分析，与分组条件作交叉验证。

转录本原始数据准备

对十二个样品的原始转录本计数数据重命名后，整合为一个计数矩阵，并滤去在多数在少于一半样本中有表达的基因，将新计数矩阵保存后供下游分析使用。

```python
import os
import pandas as pd

fileList = [fileName for fileName in os.listdir(".") if fileName[:3] == "ERR"]
# fileList
# # ['ERR188044-count.txt',
# #  'ERR188104-count.txt',
# #  'ERR188234-count.txt',
# #  'ERR188245-count.txt',
# #  'ERR188257-count.txt',
# #  'ERR188273-count.txt',
# #  'ERR188337-count.txt',
# #  'ERR188383-count.txt',
# #  'ERR188401-count.txt',
# #  'ERR188428-count.txt',
# #  'ERR188454-count.txt',
# #  'ERR204916-count.txt']
sampleNameList = [name[:9] for name in fileList]
# sampleNameList
# # ['ERR188044',
# #  'ERR188104',
# #  'ERR188234',
# #  'ERR188245',
# #  'ERR188257',
# #  'ERR188273',
# #  'ERR188337',
# #  'ERR188383',
# #  'ERR188401',
# #  'ERR188428',
# #  'ERR188454',
# #  'ERR204916']

total = []
for fileName in fileList:
    with open(fileName) as fin:
        df = pd.read_table(fin, index_col=0, names=[fileName[:9]])
        df = df.sort_index()
        total.append(df)

df = pd.concat(total, axis=1)
df.to_csv("./genes-count.csv")
# df.shape
# # (1618, 12)

filtered = df[df.sum(axis=1) >= 6]
filtered.to_csv("./filtered-count.csv")
# filtered.shape
# # (407, 12)
```

数据准备

在 `R` 中读入转录本数据与样本元数据，并通过 `BioMart` [4] 中各类基因名数据库，将有对应名称的转录本基因名转化为 HGNC 中的名字。

```
cts <- read.csv("./filtered-count.csv")
row.names(cts) <- cts[, 1]
cts <- as.matrix(cts[, -1])
library(httr)
set_config(config(ssl_verifypeer = 0L))
mart <- biomaRt::useMart(
    biomart = "ensembl",
    dataset = "hsapiens_gene_ensembl"
)
namesMapping <- biomaRt::getBM(
    attributes = c("refseq_mrna", "hgnc_symbol"),
    filters = "refseq_mrna",
    values = rownames(cts),
    mart = mart
)
rownames(namesMapping) <- namesMapping[, 1]
rownames(cts) <- ifelse(
    rownames(cts) %in% namesMapping$refseq_mrna,
    namesMapping[rownames(cts), "hgnc_symbol"],
    rownames(cts)
)
meta <- read.csv(
    "./geuvadis_phenodata.csv",
    row.names = "ids"
)
meta$sex <- factor(meta$sex)
meta$population <- factor(meta$population)
if (!all(rownames(meta) == colnames(cts))) {
    cts <- cts[, rownames(meta)]
}
```

差异基因表达分析

在 `DESeq2` 包的帮助下，我们将对不同样品按照人群与性别分组，并且可以使用该软件包中内置的 Wald 检验[5]来分析转录组差异与单个指标之间的关系；我们也可以通过似然比检验[6]来比较对两个因素同时敏感表达的基因。

```
library(DESeq2)
deseqDataSet <- DESeqDataSetFromMatrix(
    countData = cts,
    colData = meta,
    design = ~ sex + population
)
deseqDataSet
```

```
class: DESeqDataSet
dim: 407 12
metadata(1): version
assays(1): counts
rownames(407): ABCD1 ATP7A ... NHSL2.1 XKRX
rowData names(0):
colnames(12): ERR188044 ERR188104 ... ERR188454 ERR204916
colData names(2): sex population
```

```
WaldTest <- DESeq(deseqDataSet)
WaldTestPopulation <- results(
    WaldTest,
    contrast = c("population", "GBR", "YRI")
)
WaldTestSex <- results(
    WaldTest,
    contrast = c("sex", "female", "male")
)
LRT <- DESeq(deseqDataSet, test = "LRT", reduced = ~1)
LRTResult <- results(LRT)
```

三种假设检验的结果如表 1至表 3所示。

表 1: 按照人群分组对样本 WALD 检验结果

表 2: 按照性别分组对样本 WALD 检验结果

表 3: 依照人群（主）与性别（次）对样本似然比检验结果

显著差异表达基因过滤

在获得了假设检验结果后，我们筛选出显著拒绝零假设的基因（给定 $\alpha = 0.05$），作为我们得到的显著差异表达基因。

```
pValueCutOff <- 0.05
PopulationRelated <-
    WaldTestPopulation[order(WaldTestPopulation$padj), ]
SexRelated <-
    WaldTestSex[order(WaldTestSex$padj), ]
CrossRelated <-
    LRTResult[order(LRTResult$padj), ]
PopulationSignificant <-
    PopulationRelated[!is.na(PopulationRelated$padj), ]
SexSignificant <-
    SexRelated[!is.na(SexRelated$padj), ]
CrossSignificant <-
    CrossRelated[!is.na(CrossRelated$padj), ]
PopulationSignificant <-
    PopulationSignificant[PopulationSignificant$padj <= pValueCutOff, ]
SexSignificant <-
    SexSignificant[SexSignificant$padj <= pValueCutOff, ]
CrossSignificant <-
    CrossSignificant[CrossSignificant$padj <= pValueCutOff, ]
```

我们可也借助 `EnhancedVolcano`[7] 包对于原始结果直观作图展示，图 1至图 3展示了表达增加一倍以上或减少一半以上的显著差异基因（红色圆点）。

```
library(EnhancedVolcano)
EnhancedVolcano(
    PopulationRelated,
    lab = rownames(PopulationRelated),
    x = "log2FoldChange",
    y = "padj",
    ylim = c(0, 4),
    title = "Population: GBR versus YRI",
    pCutoff = pValueCutOff,
    FCcutoff = 0.5,
    pointSize = 2,
    labSize = 4
)
```

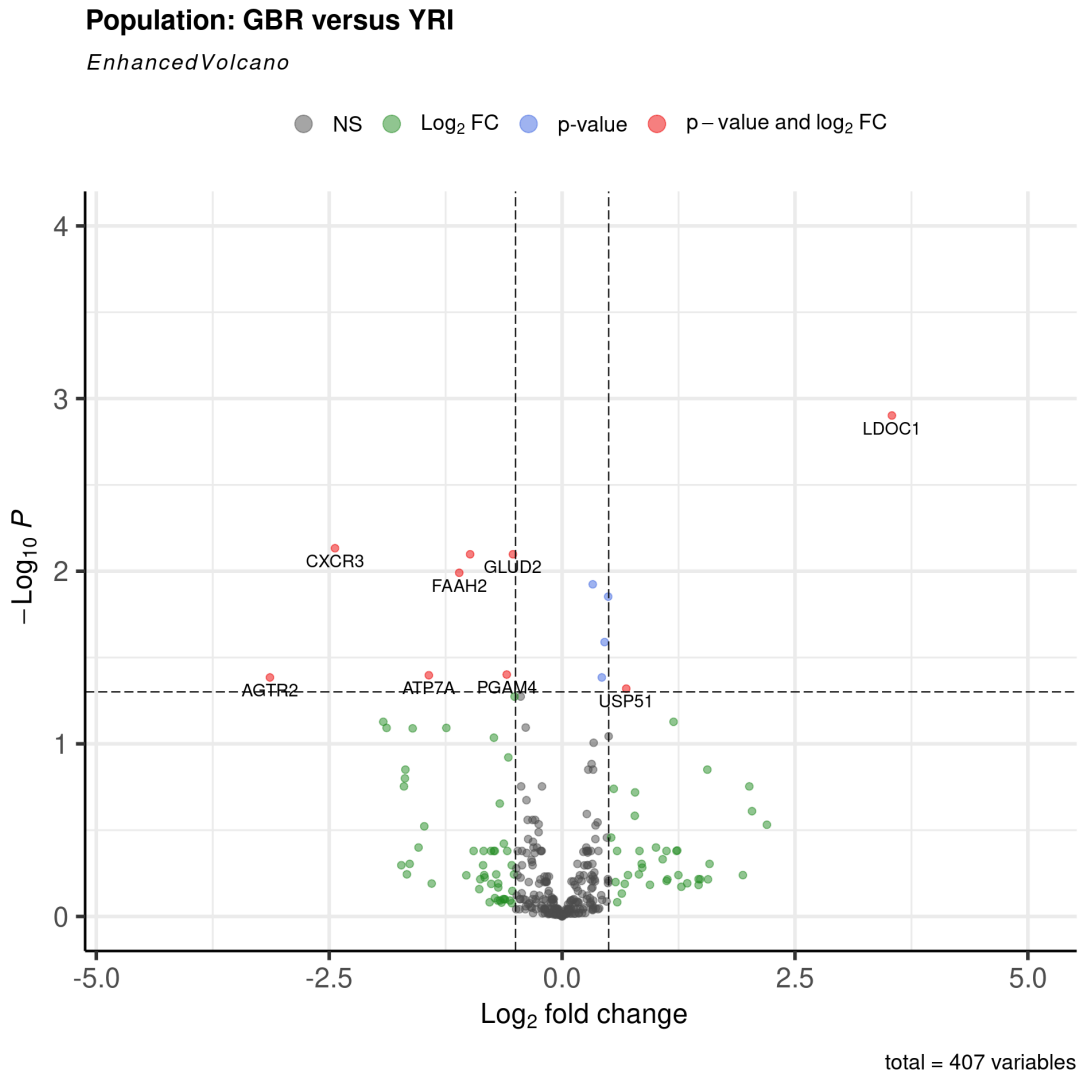

图 1: 大不列颠群岛与伊巴丹岛人群基因表达差异

```
EnhancedVolcano(
    SexRelated,
    lab = rownames(SexRelated),
    x = "log2FoldChange",
    y = "padj",
    ylim = c(0, 4),
    title = "Sex: Female versus Male",
    pCutoff = pValueCutOff,
    FCcutoff = 0.5,
    pointSize = 2,
    labSize = 4
)
```

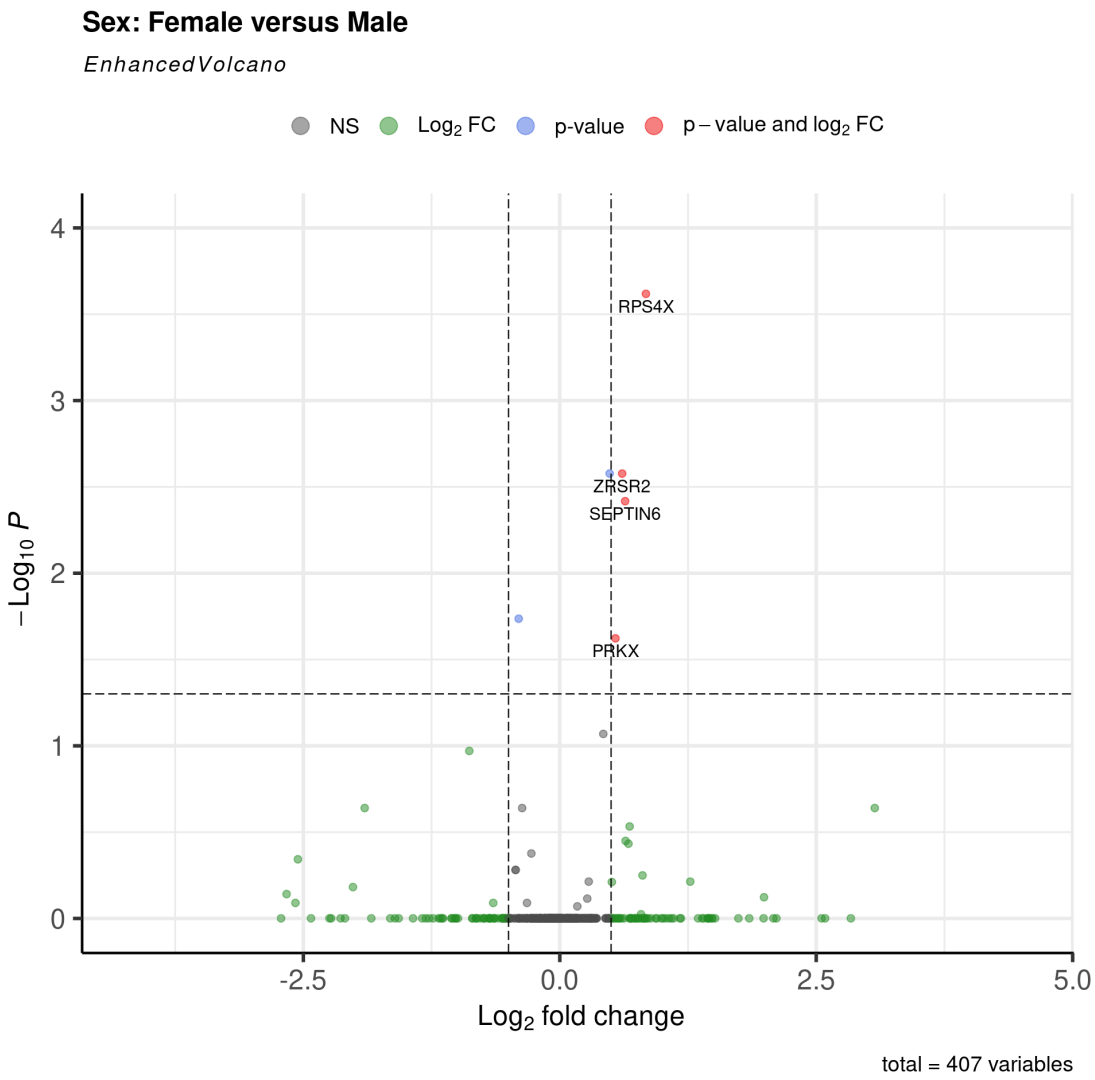

图 2: 女性与男性基因表达差异

```
EnhancedVolcano(
    CrossRelated,
    lab = rownames(CrossRelated),
    x = "log2FoldChange",
    y = "padj",
    ylim = c(0, 4),
    title = "Likelihood ratio analysis: Sex versus Population",
    pCutoff = pValueCutOff,
    FCcutoff = 0.5,
    pointSize = 2,
    labSize = 4
)
```

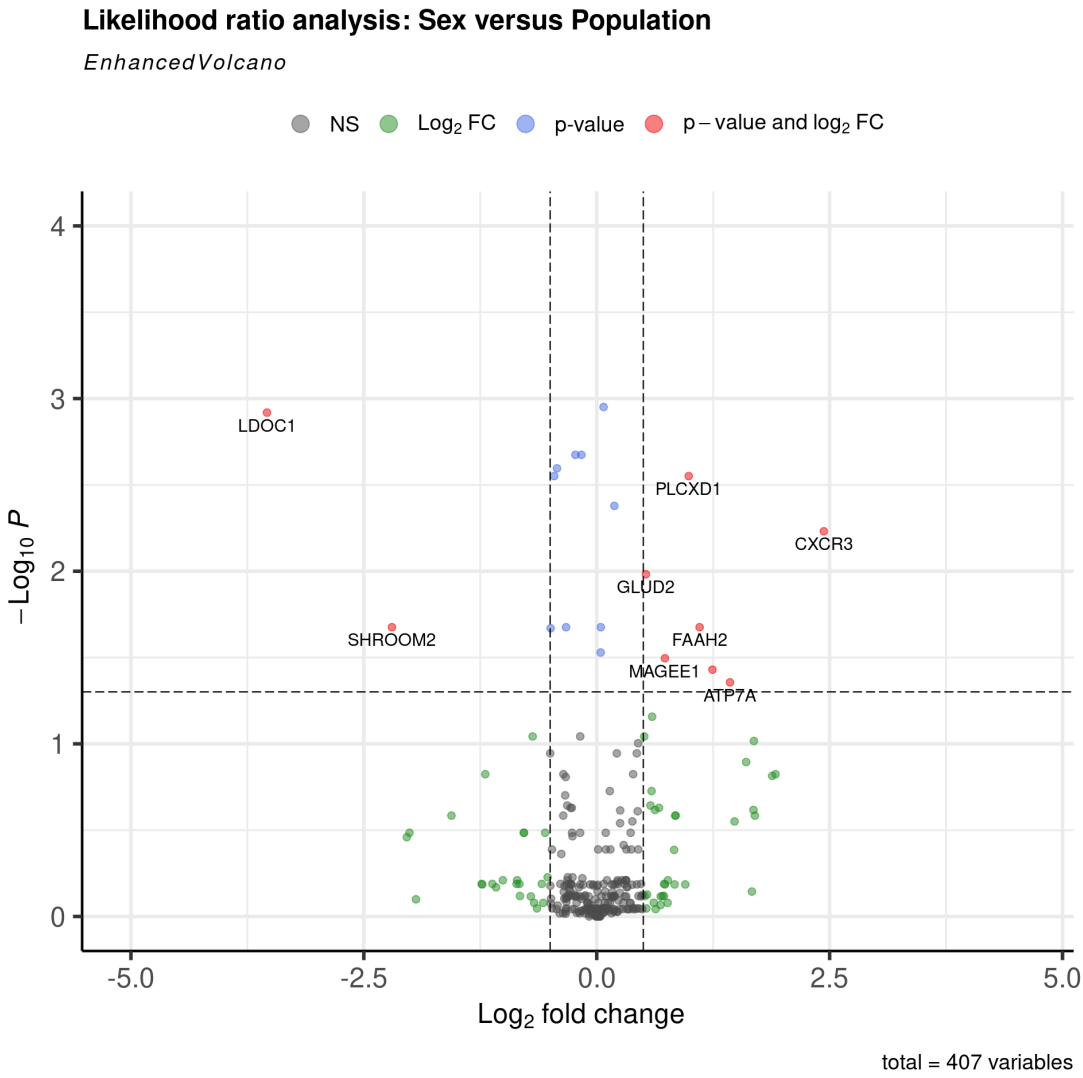

图 3: 人群与性别双因素作用下基因表达差异

差异表达基因组别分析

在三次独立的假设检验中，我们获得了不相同的系列差异表达基因，我们可以将这些结果看作三个集合，通过分析其之间交叉关系，分析不同因素之间对于基因表达的影响是否有一定相关性。

```
library(VennDiagram)
vennDiagramPlot <- venn.diagram(
    x = list(
        rownames(PopulationSignificant),
        rownames(SexSignificant),
        rownames(CrossSignificant)
    ),
    category.names = c("Population", "Sex", "Cross"),
    filename = "PopSexCrossVennDiagram.png",
    imagetype = "png",
    compression = "lzw",
    height = 1200,
    width = 1200,
    resolution = 500,
    lwd = 2,
    lty = "blank",
    fill = RColorBrewer::brewer.pal(3, "Pastel2"),
    cex = .6,
    cat.cex = .6,
    cat.pos = c(-20, 20, 135),
    cat.dist = c(0.05, 0.05, 0.01)
)
```

如图 4所示，我们可以看到双因素检验得到的结果与两个因素单独检验的结果基本重合，但也有三个基因对双因素的同时变化敏感，而三个基因仅对人群因素敏感。

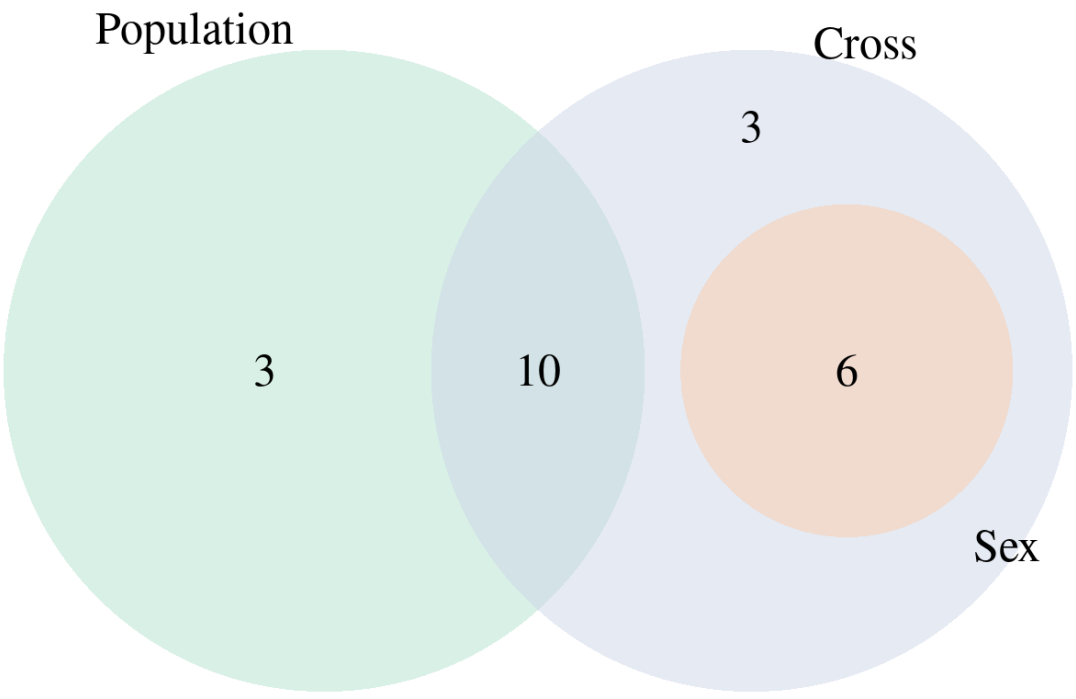

图 4: 不同假设检验中基因集合关系

为了避免这样的结果是由于在假设检验模型设计中人为因素引入的误差，我们可以交换两个因素的顺序

作交叉验证，将再次计算得到的结果同时并入结果，见图 5。我们可以看见集合的重叠与图 4无差异，因此说明此结果是完备可靠的。

```
validationDataSet <- DESeqDataSetFromMatrix(
    countData = cts,
    colData = meta,
    design = ~ population + sex
)
LRTVal <- DESeq(deseqDataSet, test = "LRT", reduced = ~1)
LRTValResult <- results(LRTVal)
CrossRelatedVal <-
    LRTValResult[order(LRTValResult$padj), ]
CrossSignificantVal <-
    CrossRelatedVal[!is.na(CrossRelatedVal$padj), ]
CrossSignificantVal <-
    CrossSignificantVal[CrossSignificantVal$padj <= pValueCutOff, ]
vennDiagramPlot <- venn.diagram(
    x = list(
        rownames(PopulationSignificant),
        rownames(SexSignificant),
        rownames(CrossSignificant),
        rownames(CrossSignificantVal)
    ),
    category.names = c("Population", "Sex", "Cross", "Cross-val"),
    filename = "PopSexCrossValVennDiagram.png",
    imagetype = "png",
    compression = "lzw",
    height = 1200,
    width = 1200,
    resolution = 500,
    lwd = 2,
    lty = "blank",
    fill = RColorBrewer::brewer.pal(4, "Pastel2"),
    cex = .6,
    cat.cex = .6,
    cat.pos = c(-20, 20, 135, -135),
    cat.dist = c(0.05, 0.05, 0.05, 0.05)
)
```

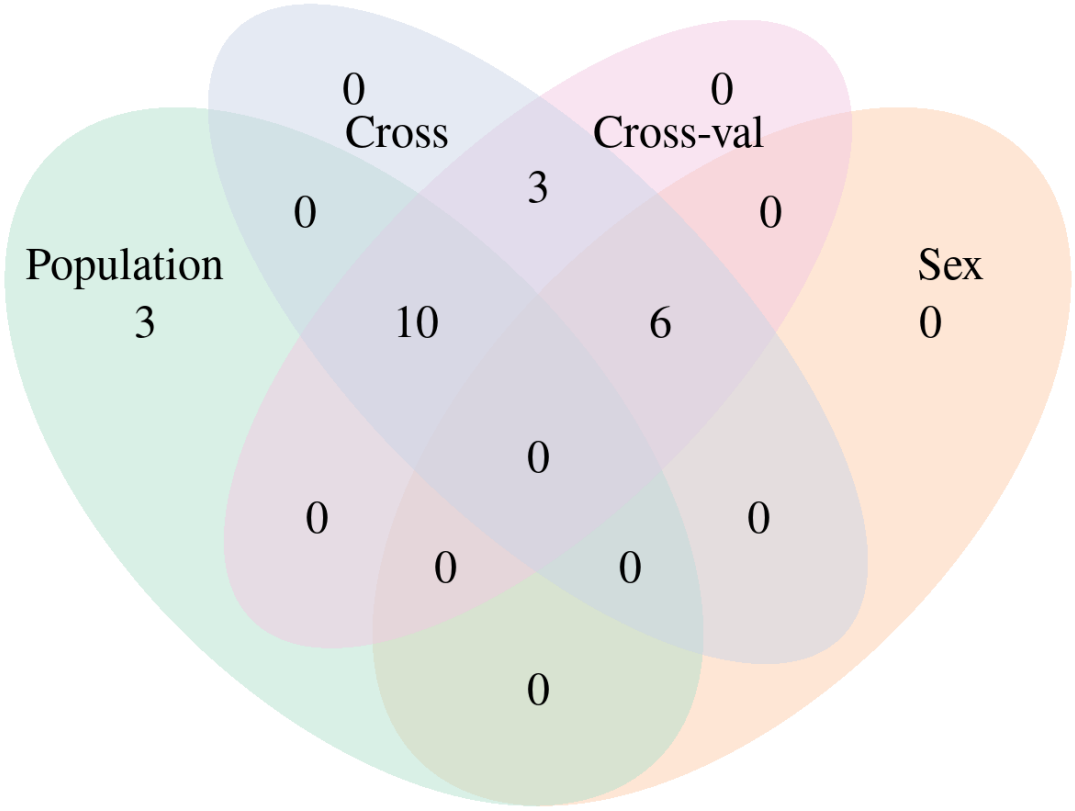

图 5: 不同假设检验中基因集合关系（交叉验证）

差异表达基因热图

针对显著差异表达的基因，我们可以绘制其表达热图，如图 6展示的归一化后的表达热图结果。由于此实验中样品均为健康人群，故基因表达上无显著差异，因此在热土上结果不形成明显对比。

```
library(genefilter)
library(pheatmap)
uniformedResult <- rlogTransformation(LRT)
significantGenes <- unique(c(
    rownames(PopulationSignificant),
    rownames(SexSignificant),
    rownames(CrossSignificant)
))
expressionMatrix <- assay(uniformedResult)[significantGenes, ]
colnames(expressionMatrix) <- paste(meta$population, meta$sex)
pheatmap(expressionMatrix)
```

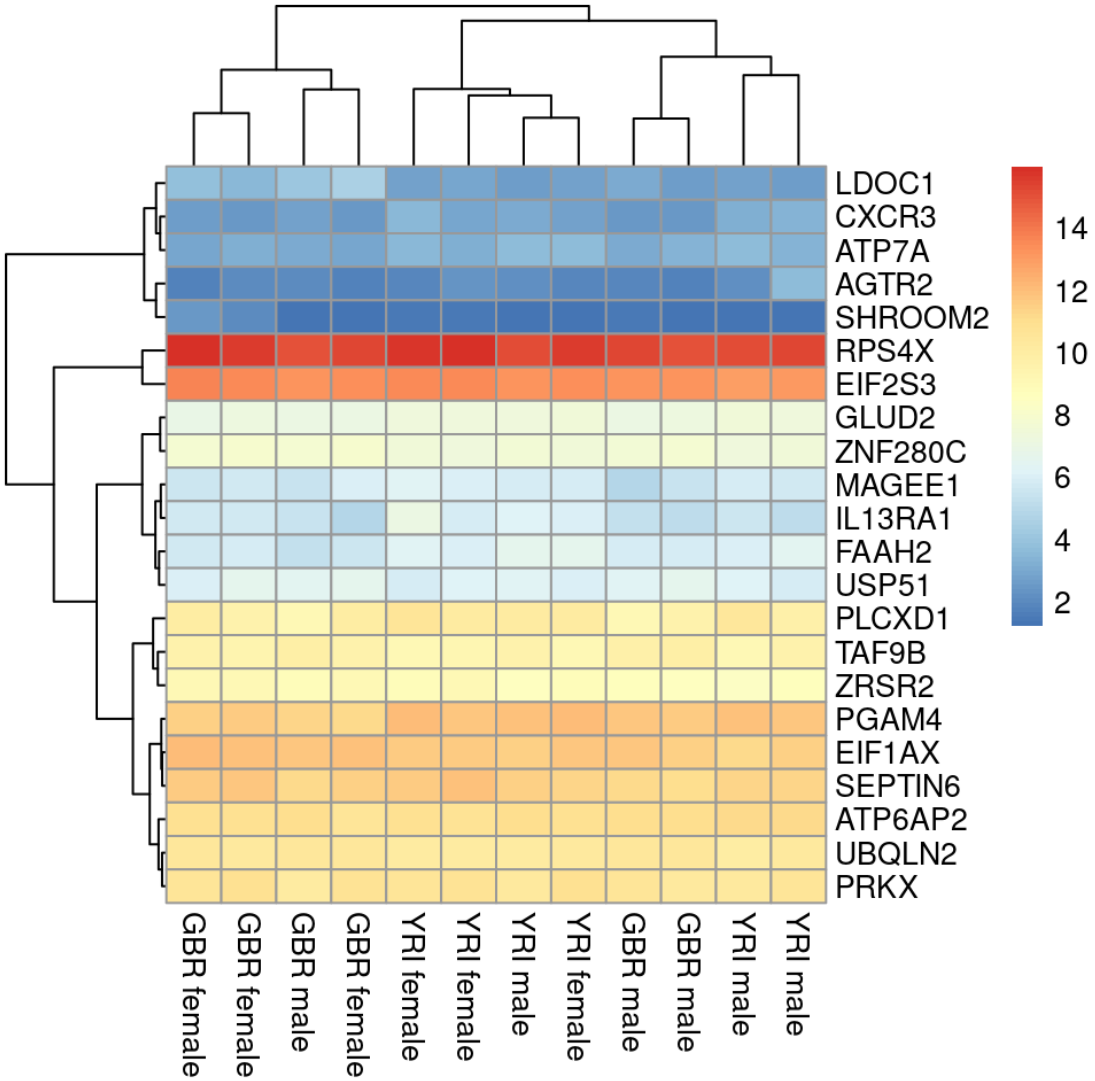

图 6: 差异表达显著基因的表达热图

基因富集分析

基因富集分析（Gene Ontology Analysis）基于对基因的注释信息，将富集的注释类型筛选出来，以帮助我们推断与变量条件相关的差异表达基因的功能。此处由于样本量较小，因此我们将上述过程中所有筛选出的差异表达基因整合，进行基因富集分析，结果见图 7。

```
library(clusterProfiler)
library(org.Hs.eg.db)
eGO <- enrichGO(
    gene = significantGenes,
    OrgDb = org.Hs.eg.db,
    keyType = "SYMBOL",
    ont = "BP",
    pAdjustMethod = "BH",
    pvalueCutoff = pValueCutOff,
    qvalueCutoff = 2 * pValueCutOff
)
GOplot <- plotGOgraph(eGO)
```

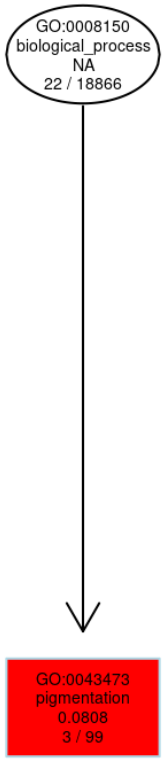

图 7: 差异表达显著基因的富集分析

由此可见，仅有色素沉着相关的基因在次富集出来，故我们可以推断出人群间确有显著差异表达的通路，而在不同性别间并无显著的差异通路，故筛选出性别相关差异表达基因可能是由样本量过小导致的假阳性结果。

结论

通过差异基因表达分析与基因富集分析，我们可以解读出与自变量因素相关的生物学过程等信息，以此解读出数据的生物学意义。

参考资料

[1] HESKETH A R. RNA Sequencing Best Practices: Experimental Protocol and Data Analysis[M/OL]. OLIVER S G, CASTRILLO J I, 编//Yeast Systems Biology: Methods and Protocols. New York, NY: Springer New York, 2019: 113–129. https://doi.org/10.1007/978-1-4939-9736-7_7. DOI:10.1007/978-1-4939-9736-7_7.

[2] LOVE M I, HUBER W, ANDERS S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J/OL]. Genome Biology, 2014, 15(12): 550. https://doi.org/10.1186/s13059-014-0550-8. DOI:10.1186/s13059-014-0550-8.

[3] LIAN Y, WANG Q, MU J, 等. Network pharmacology assessment of Qingkailing injection treatment of cholestatic hepatitis[J]. J Tradit Chin Med, 2021, 41(1): 167–180. DOI:10.19852/j.cnki.jtcm.20201208.001.

[4] DURINCK S, MOREAU Y, KASPRZYK A, 等. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis[J]. Bioinformatics, 2005, 21(16): 3439–40. DOI:10.1093/bioinformatics/bti525.

[5] WIKIPEDAIA. Wald test[Z/OL](2021–05). https://en.wikipedia.org/wiki/Wald_test.

[6] WIKIPEDIA. Likelihood-ratio test[Z/OL](2021–05). https://en.wikipedia.org/wiki/Likelihood-ratio_test.

[7] BLIGHE K, RANA S, LEWIS M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling[M/OL]. https://github.com/kevinblighe/EnhancedVolcano.