作业三. 宫颈癌相关人乳头瘤病毒整合基因位点的GO富集和绘图

生信1802-余思克-2018317220208

实验目的

基因本体是一种系统地对物种基因及其产物属性进行注释的方法和过程。基因本体的注释主要围绕细胞组件、分子功能、生物过程三个方面。宫颈癌是最常见的妇科恶性肿瘤，已有的研究表明：宫颈癌的发生与人乳头瘤病毒（HPV）基因片段的插入整合有着密切的联系本次实验的目的是通过R包对人类基因组中发生了HPV整合事件的一组基因进行GO富集分析，探究这一组基因与宫颈癌有着怎样的关联。

实验步骤

1. 相关R包的加载

本次实验使用R语言完成，不同于网页版的GO富集工具，R语言中提供了相当丰富的GO工具包辅助我们进行GO富集分析，本步骤将逐一安装并加载GO富集分析所需的R包。

```
install.packages("devtools")
install.packages("BiocManager")
library("devtools")
BiocManager::install(version = "3.12")
Needed=c("bit", "formatR", "hms", "triebeard", "tweenr", "polyclip",
"RcppEigen", "RcppArmadillo", "zlibbioc", "bit64", "blob", "plogr", "lambda.r",
"futile.options", "progress", "urltools", "gridGraphics", "ggforce", "ggrepel",
"viridis", "tidygraph", "graphlayouts", "bitops", "XVector", "IRanges",
"RSQLite", "futile.logger", "snow", "data.table", "gridExtra", "fastmatch",
"cowplot", "europepmc", "ggplotify", "ggraph", "ggridges", "igraph", "dplyr",
"tidyselect", "RCurl", "Biostrings", "AnnotationDbi", "BiocParallel", "DO.db",
"fgsea", "GOSemSim", "qvalue", "S4Vectors", "BiocGenerics", "graph", "Biobase",
"GO.db", "SparseM", "matrixStats", "DBI", "enrichplot", "rvcheck", "tidyr",
"org.Hs.eg.db", "KEGGgraph", "XML", "Rgraphviz", "png", "KEGGREST")
install.packages(Needed)
BiocManager::install(c("DOSE","topGO","clusterProfiler","pathview")) # 安装GO富集分析所需的包

library(DOSE)
library(org.Hs.eg.db)
library(topGO)
library(clusterProfiler)
library(pathview) # 加载包
```

2. 读入待分析基因

将感兴趣的一组宫颈癌相关HPV整合相关基因名读入到R语言中，并获取基因id。

```
MyGeneSet = c("POU5F1B", "FHIT", "KLF12", "KLF5", "LRP1B", "HMGA2", "DLG2",
"SEMA3D") # 读入基因名信息
MyGeneIDSet = bitr(MyGeneSet, fromType="SYMBOL", toType="ENTREZID",
OrgDb="org.Hs.eg.db") # 针对读入的基因名信息匹配其对应的基因id
```

3. GO富集分析

针对富集分析的背景基因集，对读入的基因进行富集分析。

```
data(geneList, package="DOSE") # 富集分析的背景基因集
ego_ALL = enrichGO(gene = MyGeneIDSet$ENTREZID, universe = names(geneList),
OrgDb = org.Hs.eg.db, ont = "ALL", pAdjustMethod = "BH", pvalueCutoff = 1,
qvalueCutoff = 1, readable = TRUE) # 富集分析（全部）
ego_MF = enrichGO(gene = MyGeneIDSet$ENTREZID, universe = names(geneList),OrgDb
= org.Hs.eg.db,ont = "MF", pAdjustMethod = "BH",pvalueCutoff = 1,qvalueCutoff =
1,readable = TRUE) # 分子功能富集
ego_CC = enrichGO(gene = MyGeneIDSet$ENTREZID, universe = names(geneList),OrgDb
= org.Hs.eg.db,ont = "CC", pAdjustMethod = "BH",pvalueCutoff = 1,qvalueCutoff =
1,readable = TRUE) # 细胞组件富集
ego_BP = enrichGO(gene = MyGeneIDSet$ENTREZID, universe = names(geneList),OrgDb
= org.Hs.eg.db,ont = "BP", pAdjustMethod = "BH",pvalueCutoff = 1,qvalueCutoff =
1,readable = TRUE) # 生物过程富集
```

4. 分析结果可视化

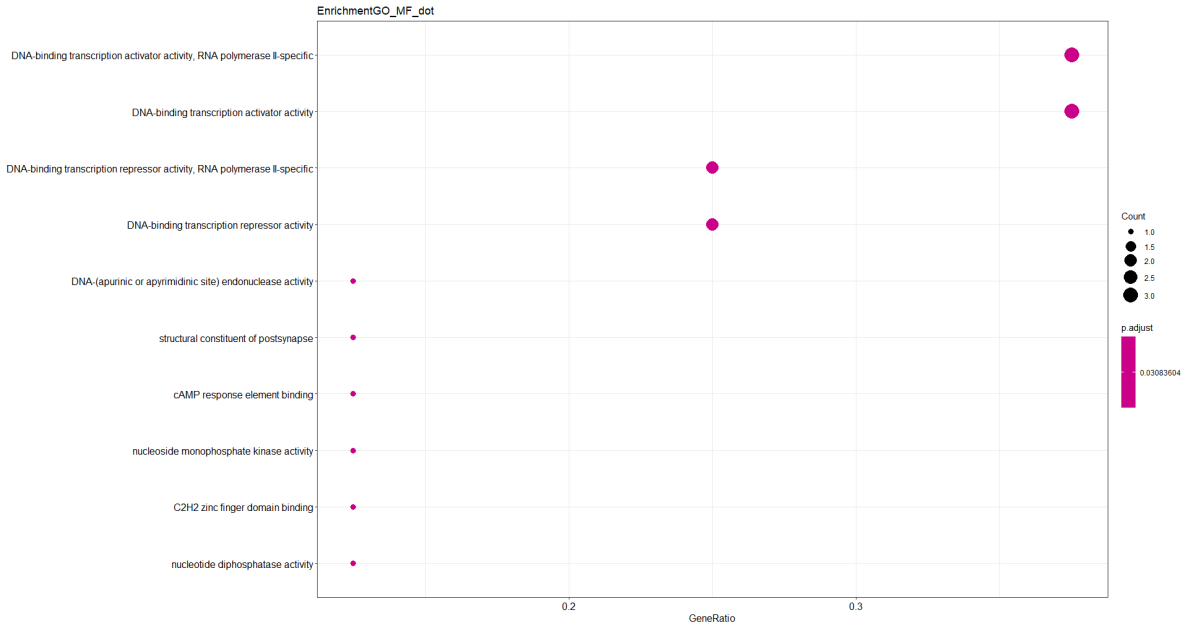
本步骤将针对分子功能、细胞组件、生物过程三个方面，分别绘制点图、柱状图与概念层次树，从而观察分析基因集的GO富集结果。

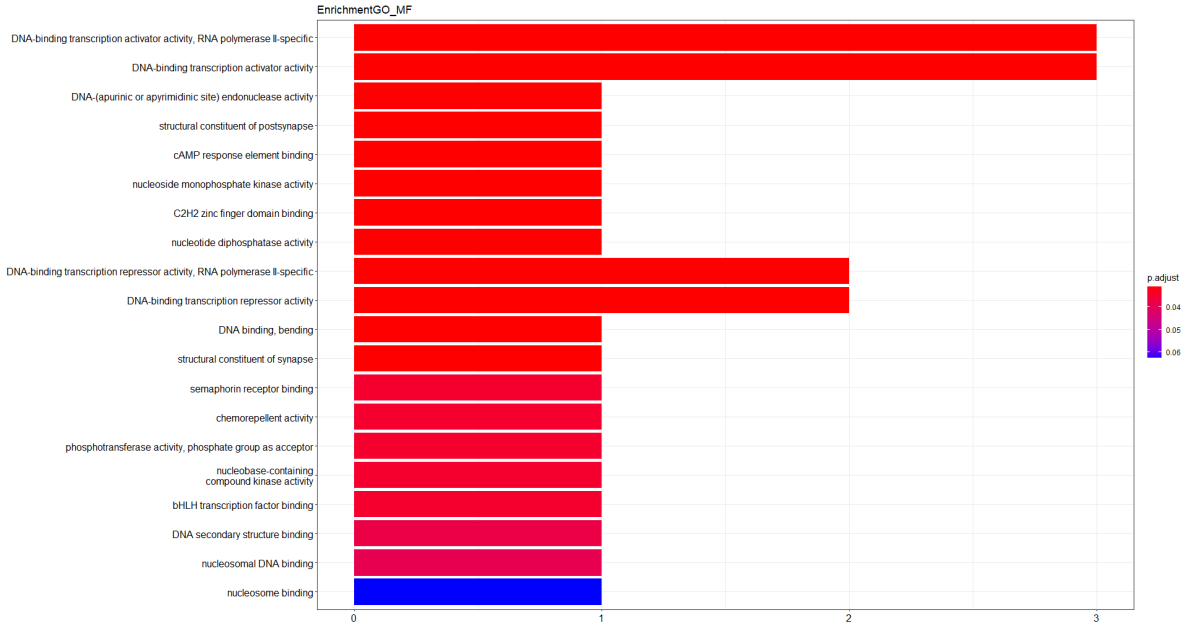
```
dotplot(ego_MF,title="EnrichmentGO_MF_dot")
dotplot(ego_CC,title="EnrichmentGO_CC_dot")
dotplot(ego_BP,title="EnrichmentGO_BP_dot") # 点图的绘制

barplot(ego_MF, showCategory=20,title="EnrichmentGO_MF")
barplot(ego_CC, showCategory=20,title="EnrichmentGO_CC")
barplot(ego_BP, showCategory=20,title="EnrichmentGO_BP") # 柱状图的绘制

plotGOgraph(ego_MF,firstSigNodes = 10, useInfo = "all", sigForAll = TRUE,
useFullNames = TRUE)
plotGOgraph(ego_CC,firstSigNodes = 10, useInfo = "all", sigForAll = TRUE,
useFullNames = TRUE)
plotGOgraph(ego_BP,firstSigNodes = 10, useInfo = "all", sigForAll = TRUE,
useFullNames = TRUE) # 概念层次树的绘制
```

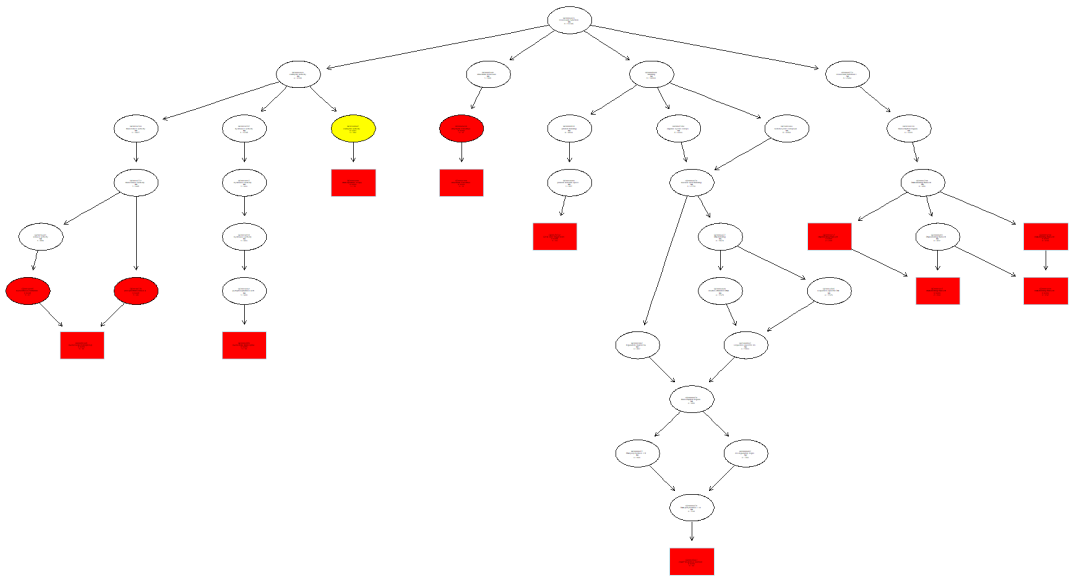
实验结果

1. 分子功能

基因集的分子功能富集在DNA结合转录因子的活性、DNA结合转录阻遏物的活性，II型RNA聚合酶、DNA核酸内切酶、cAMP反应原件等功能上。

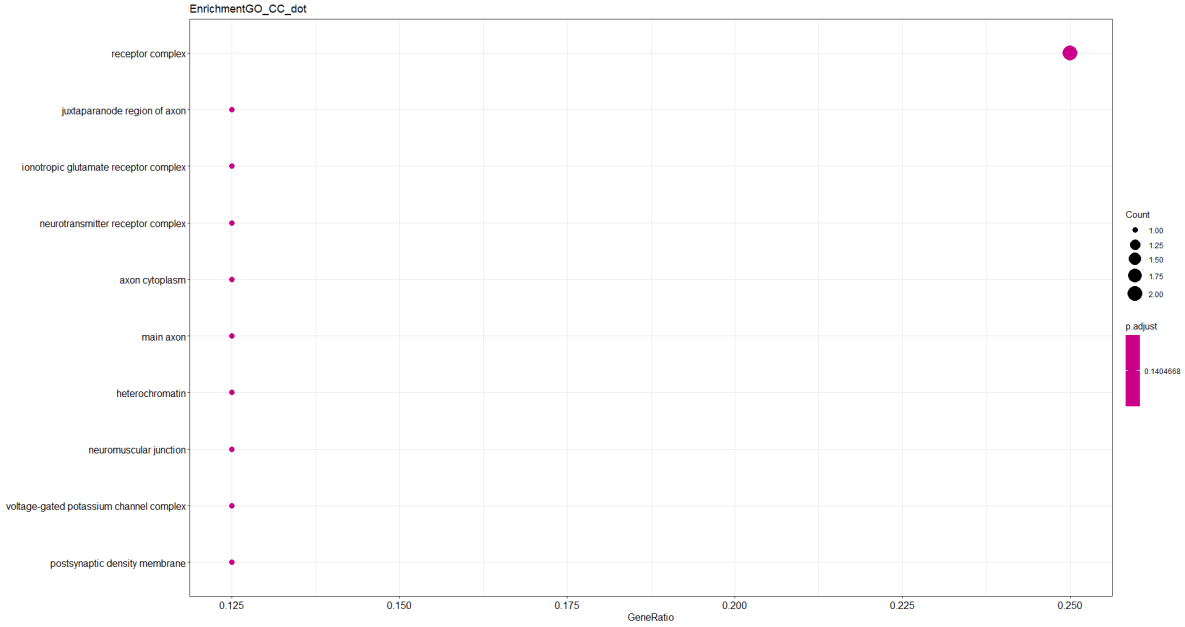

EnrichmentGO_MF_dot
DNA-binding transcription activator activity, RNA polymerase II-specific
DNA-binding transcription activator activity
DNA-binding transcription repressor activity, RNA polymerase II-specific
DNA-binding transcription repressor activity
DNA-(apurinic or apyrimidinic site) endonuclease activity
structural constituent of postsynapse
cAMP response element binding
nucleoside monophosphate kinase activity
C2H2 zinc finger domain binding
nucleotide diphosphatase activity
0.2
0.3
GeneRatio
Count
1.0
1.5
2.0
2.5
3.0
p.adjust
0.03083604

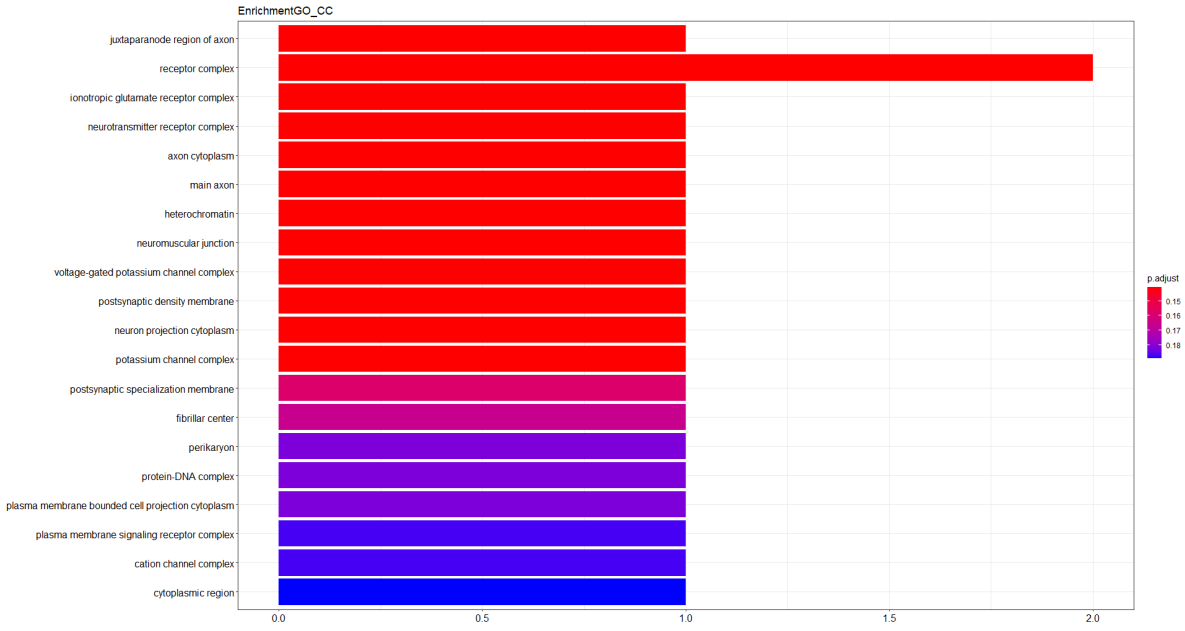

EnrichmentGO_MF
DNA-binding transcription activator activity, RNA polymerase II-specific
DNA-binding transcription activator activity
DNA-(apurinic or apyrimidinic site) endonuclease activity
structural constituent of postsynapse
cAMP response element binding
nucleoside monophosphate kinase activity
C2H2 zinc finger domain binding
nucleotide diphosphatase activity
DNA-binding transcription repressor activity, RNA polymerase II-specific
DNA-binding transcription repressor activity
DNA binding, bending
structural constituent of synapse
semaphorin receptor binding
chemorepellent activity
phosphotransferase activity, phosphate group as acceptor
nucleobase-containing compound kinase activity
bHLH transcription factor binding
DNA secondary structure binding
nucleosomal DNA binding
nucleosome binding
0
1
2
3
p.adjust
0.04
0.05
0.06

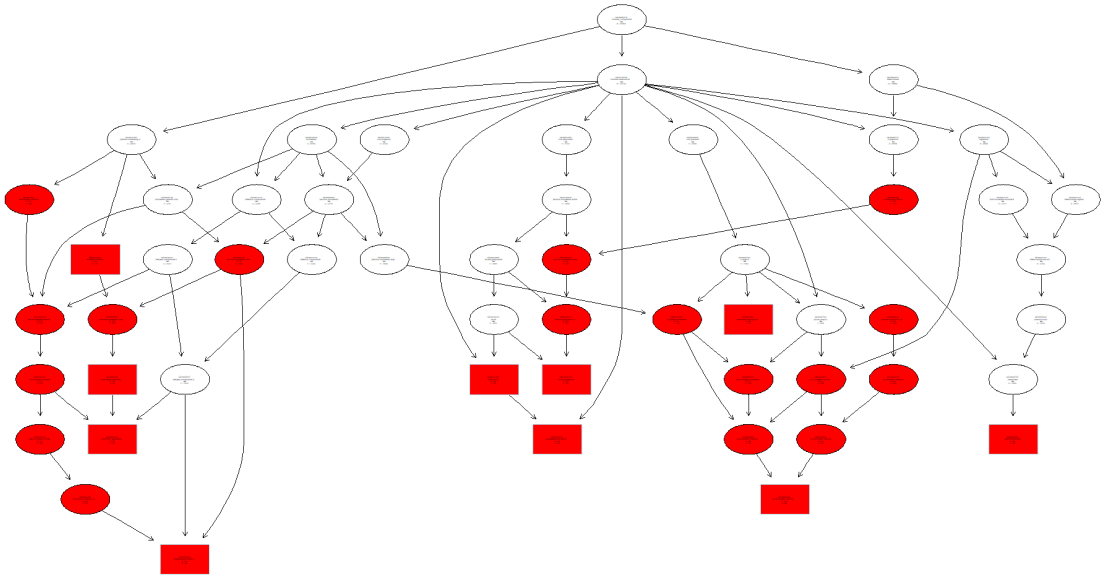


2. 细胞组成

基因集的细胞组成富集在受体复合物、轴突的近旁节区、谷氨酸离子受体复合物、神经递质受体复合物、轴突细胞质等区域。

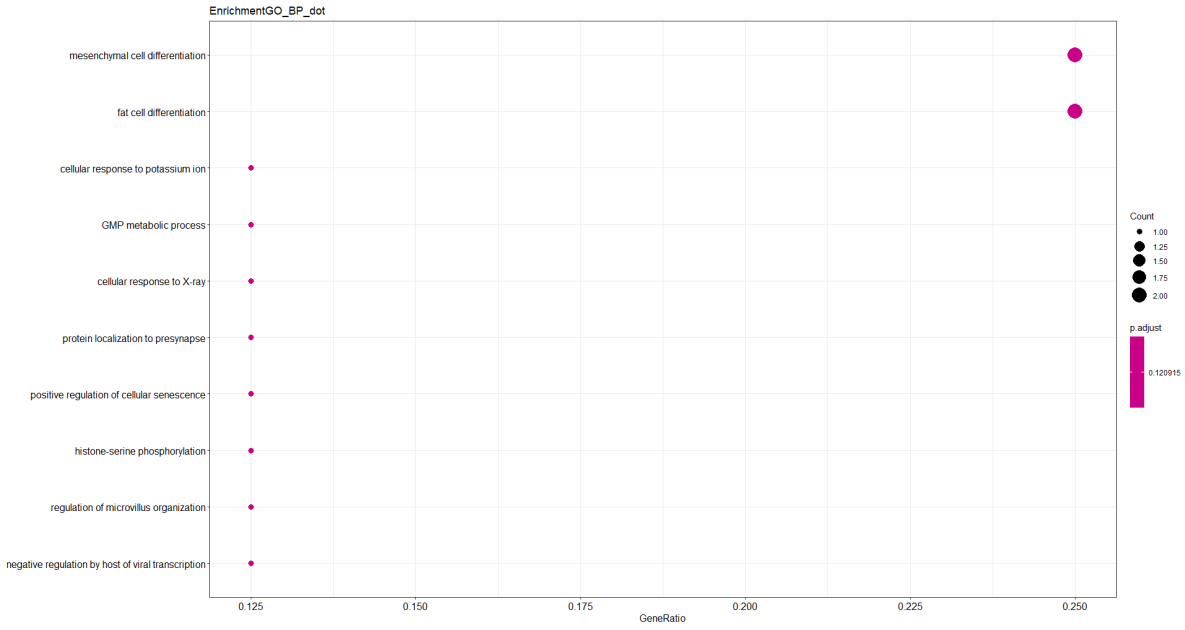


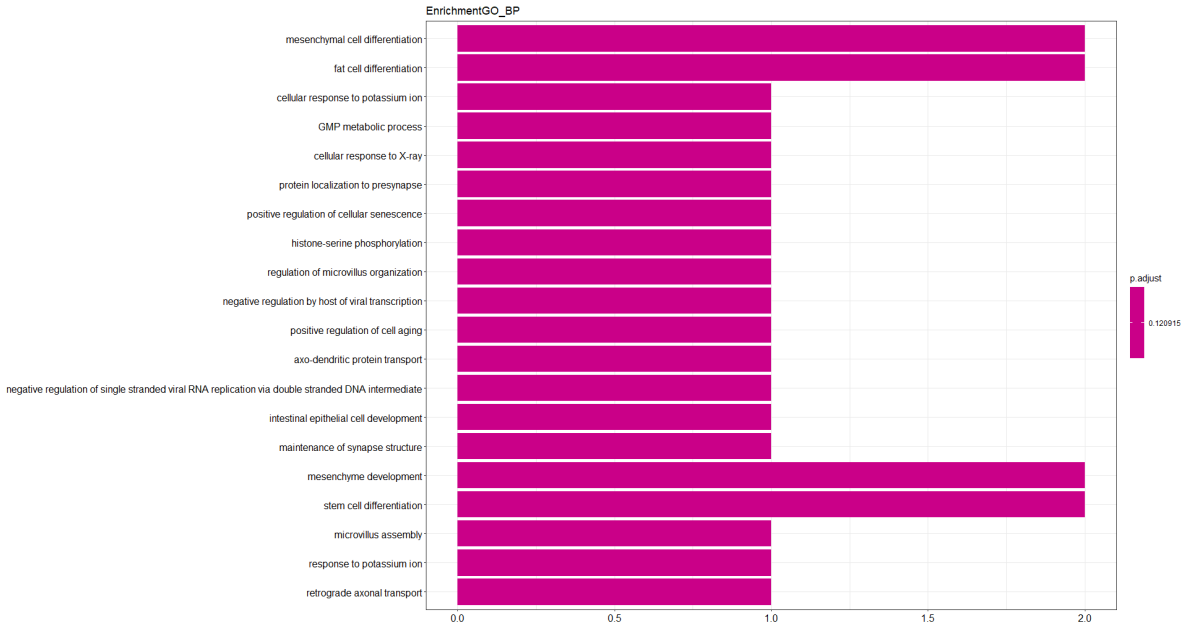


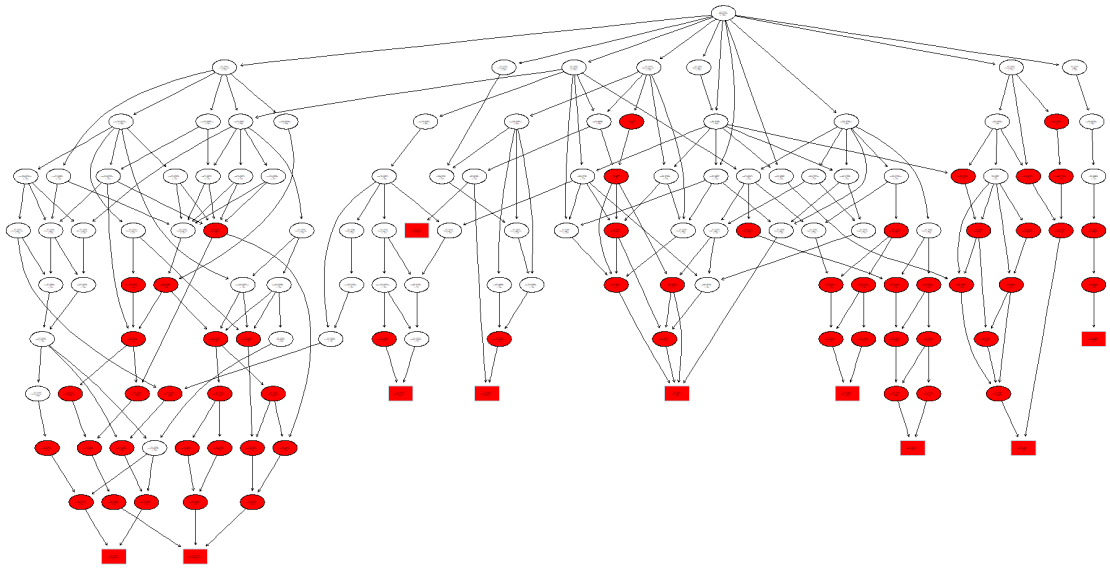


3. 生物过程

基因集的生物过程富集在间充质细胞分化、脂肪细胞分化、干细胞分化、间充质发育、细胞对钾离子的反应、GMP代谢过程等。







讨论

在安装`BiocManager`时，出现了报错，原因是不同的R语言版本需要对应到不同的`BiocManager`版本，调整好版本信息之后即可正常安装（R 4.0对应的版本是3.12），此外，若出现链接打开失败的报错，更换国内其他镜像即可解决问题。

通过对基因集在分子功能、细胞组成、生物过程三个方面的富集分析，可以发现，这些基因有着较大的可能具有进行DNA转录调控，信号转导等功能，其富集在细胞的信号传输、调节区域，有较大的可能参与了细胞的生长发育、分化以及GMP代谢过程。这些功能与生物过程对细胞的增殖调控起到了重大的作用。根据基因集的宫颈癌肿瘤以及HPV病毒整合背景，这些基因可能是因为受到了HPV病毒基因片段的整合插入，使得基因的表达紊乱，功能无法正常发挥，进而扰乱了基因的信号转导功能，使得细胞的DNA转录功能紊乱，DNA的复制失去失控，从而使细胞的生长发育失控、代谢过程紊乱，从而使得细胞癌变进而无限增殖，从而促使了宫颈癌的发生。

基因本体是一种系统地对物种基因及其产物属性进行注释的方法和过程。在生物信息领域， 信息的获取有着一个主要的瓶颈：生物学及相关学科的不同领域使用不同的术语，不仅让信息查找变得困难，也使数据的交流和分享更加困难。例如在一些不同的医疗数据库中，可能会存在很多不一致的描述，给数据的挖掘和分享带来很多麻烦。基因本体提供了统一定义的条目来表示基因产物的属性。在本次实验中，通过对一组宫颈癌相关的HPV病毒基因片段整合的人类基因位点进行GO富集分析发现，这一组基因的功

能、细胞组成、生物过程确实与肿瘤的增殖有着紧密的联系，进一步辅助佐证了HPV病毒基因片段的整合与宫颈癌的相关性，并一定程度揭示了基因片段整合对宫颈癌的发生这一事件的作用机理，为后续的深入研究打下了基础。