

# Task 3. GO Enrichment For DMD Genes in Homo Sapiens

---

Author: Kewei Zhao

Date: 2021-4-10

## Abstract

---

Gene ontology database is a set of concepts. In this task, we want to do GO enrichment for genes which is concerning to DMD in Homo sapiens in order to find out concepts that have close links to these genes

## Principle

---

*Hypergeometric Distribution*

$$X \sim H(n, M, N)$$

*This Distribution describes the number of times that  $k$  special objects are successfully extracted from a finite number of  $N$  objects (including  $M$  special objects)*

$$P(x = k) = \frac{C_M^k \cdot C_{N-M}^{n-k}}{C_N^n}$$

$$P\_value = \sum_k^M P(k)$$

*When  $P\_value$  is very small, this will reflect significant difference between our samples and random samples*

## Measure

---

### 1. Download Gene Symbols

Download Gene Symbols about DMD in Homo sapiens from: <https://www.ncbi.nlm.nih.gov/gene>

Clean the data of search information file by Excel to only save gene symbols

### 2. Download and import Essential Packages

```
install.packages('devtools', repos =  
'https://mirrors.tuna.tsinghua.edu.cn/CRAN')  
install.packages('BiocManager', repos =  
'https://mirrors.tuna.tsinghua.edu.cn/CRAN')  
install.packages("cli", repos = 'https://mirrors.tuna.tsinghua.edu.cn/CRAN')  
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", "europemc", "ggplotify", "gggraph", "ggribes", "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")
B_I <- c("DOSE", "topGO", "clusterProfiler", "pathview")
install.packages(Needed, repos = 'https://mirrors.tuna.tsinghua.edu.cn/CRAN')
BiocManager::install(B_I)
library(DOSE)
library(org.Hs.eg.db)
library(topGO)
library(clusterProfiler)
library(pathview)

```

### 3. GO Enrichment

#### 1) Read Data from CSV File

```

MyGeneSet <- read.table('D:/junior_n/NLP/task_3/data/gene_result.csv')
MyGeneSet <- as.character(MyGeneSet$V1)

```

#### 2) Transform from SYMBOL to ENSEMBL, ENTREZID and GO

```

MyGeneIDSet <- bitr(MyGeneSet,
  fromType="SYMBOL",
  toType=c("ENSEMBL", "ENTREZID", "GO"),
  orgDb="org.Hs.eg.db")

```

#### 3) Use DOSE Package to Get Background Genes for GO Enrichment

```

data(geneList, package="DOSE")

```

#### 4) GO Enrichment in All GO Domains

```

ego_ALL <- enrichGO(gene = MyGeneIDSet$ENTREZID,
  universe = names(geneList),
  orgDb = org.Hs.eg.db,
  ont = "ALL",
  pAdjustMethod = "BH",
  pvalueCutoff = 1,
  qvalueCutoff = 1,
  readable = TRUE)

```

## 5) GO Enrichment in Molecular Function, Cellular Component and Biological Process

```
ego_MF <- enrichGO(gene = MyGeneIDSet$ENTREZID,
                    universe = names(geneList),
                    OrgDb = org.Hs.eg.db,
                    ont = "MF",
                    pAdjustMethod = "BH",
                    pvalueCutoff = 1,
                    qvalueCutoff = 1,
                    readable = TRUE)

ego_CC <- enrichGO(gene = MyGeneIDSet$ENTREZID,
                    universe = names(geneList),
                    OrgDb = org.Hs.eg.db,
                    ont = "CC",
                    pAdjustMethod = "BH",
                    pvalueCutoff = 1,
                    qvalueCutoff = 1,
                    readable = TRUE)

ego_BP <- enrichGO(gene = MyGeneIDSet$ENTREZID,
                    universe = names(geneList),
                    OrgDb = org.Hs.eg.db,
                    ont = "BP",
                    pAdjustMethod = "BH",
                    pvalueCutoff = 1,
                    qvalueCutoff = 1,
                    readable = TRUE)
```

## 4. Visualization

### 1) Draw Dotplot for each GO Enrichment result

```
png("D:/junior_n/NLP/task_3/result/MF_DOT.png",units="in", width=10,
    height=10,res=500)
dotplot(ego_MF,title="EnrichmentGO_MF_dot")
dev.off()
png("D:/junior_n/NLP/task_3/result/CC_DOT.png",units="in", width=10,
    height=10,res=500)
dotplot(ego_CC,title="EnrichmentGO_CC_dot")
dev.off()
png("D:/junior_n/NLP/task_3/result/BP_DOT.png",units="in", width=10,
    height=10,res=500)
dotplot(ego_BP,title="EnrichmentGO_BP_dot")
dev.off()
```

### 2) Draw barplot for each GO Enrichment result

```

png("D:/junior_n/NLP/task_3/result/MF_BAR.png",units="in", width=10,
height=10,res=500)
barplot(ego_MF, showCategory=20,title="EnrichmentGO_MF_bar")
dev.off()
png("D:/junior_n/NLP/task_3/result/CC_BAR.png",units="in", width=10,
height=10,res=500)
barplot(ego_CC, showCategory=20,title="EnrichmentGO_CC_bar")
dev.off()
png("D:/junior_n/NLP/task_3/result/BP_BAR.png",units="in", width=10,
height=10,res=500)
barplot(ego_BP, showCategory=20,title="EnrichmentGO_BP_bar")
dev.off()

```

### 3) Draw GO Graph for each GO Enrichment result

```

png("D:/junior_n/NLP/task_3/result/MF_GG.png",units="in", width=5,
height=5,res=500)
plotGOgraph(ego_MF,firstSigNodes = 10, useInfo = "all", sigForAll =
TRUE,useFullNames = TRUE)
dev.off()
png("D:/junior_n/NLP/task_3/result/CC_GG.png",units="in", width=5,
height=5,res=500)
plotGOgraph(ego_CC,firstSigNodes = 10, useInfo = "all", sigForAll =
TRUE,useFullNames = TRUE)
dev.off()
png("D:/junior_n/NLP/task_3/result/BP_GG.png",units="in", width=5,
height=5,res=700)
plotGOgraph(ego_BP,firstSigNodes = 10, useInfo = "all", sigForAll =
TRUE,useFullNames = TRUE)
dev.off()

```

## Result

### 1. Related GOs' Name and Definition

#### 1) Molecular Function

##### GO:0003779

actin binding

Interacting selectively and non-covalently with monomeric or multimeric forms of actin, including actin filaments.

##### GO:0030165

PDZ domain binding

Interacting selectively and non-covalently with a PDZ domain of a protein, a domain found in diverse signaling proteins.

**GO:0002020**

protease binding

Interacting selectively and non-covalently with any protease or peptidase.

**GO:0005178**

integrin binding

Interacting selectively and non-covalently with an integrin

**GO:0005125**

cytokine activity

The activity of a soluble extracellular gene product that interacts with a receptor to effect a change in the activity of the receptor to control the survival, growth, differentiation and effector function of tissues and cells.

**GO:0008307**

structural constituent of muscle

The action of a molecule that contributes to the structural integrity of a muscle fiber.

**GO:0017166**

vinculin binding

Interacting selectively and non-covalently with vinculin, a protein found in muscle, fibroblasts, and epithelial cells that binds actin and appears to mediate attachment of actin filaments to integral proteins of the plasma membrane.

**GO:0004197**

cysteine-type endopeptidase activity

Catalysis of the hydrolysis of internal, alpha-peptide bonds in a polypeptide chain by a mechanism in which the sulfhydryl group of a cysteine residue at the active center acts as a nucleophile.

**GO:0097718**

disordered domain specific binding

Interacting selectively and non-covalently with a disordered domain of a protein.

**GO:0042805**

actinin binding

Interacting selectively and non-covalently with actinin, any member of a family of proteins that crosslink F-actin.

**2) Cellular Component**

**GO:0098797**

plasma membrane protein complex

Any protein complex that is part of the plasma membrane.

**GO:0042383**

sarcolemma

The outer membrane of a muscle cell, consisting of the plasma membrane, a covering basement membrane (about 100 nm thick and sometimes common to more than one fiber), and the associated loose network of collagen fibers.

**GO:0062023**

collagen-containing extracellular matrix

An extracellular matrix consisting mainly of proteins (especially collagen) and glycosaminoglycans (mostly as proteoglycans) that provides not only essential physical scaffolding for the cellular constituents but can also initiate crucial biochemical and biomechanical cues required for tissue morphogenesis, differentiation and homeostasis. The components are secreted by cells in the vicinity and form a sheet underlying or overlying cells such as endothelial and epithelial cells.

**GO:0031012**

extracellular matrix

A structure lying external to one or more cells, which provides structural support, biochemical or biomechanical cues for cells or tissues.

**GO:0030055**

cell-substrate junction

A cell junction that forms a connection between a cell and the extracellular matrix.

**GO:0016010**

dystrophin-associated glycoprotein complex

A multiprotein complex that forms a strong mechanical link between the cytoskeleton and extracellular matrix; typical of, but not confined to, muscle cells. The complex is composed of transmembrane, cytoplasmic, and extracellular proteins, including dystrophin, sarcoglycans, dystroglycan, dystrobrevins, syntrophins, sarcospan, caveolin-3, and NO synthase.

**GO:0090665**

glycoprotein complex

A protein complex containing at least one glycosylated protein, may be held together by both covalent and noncovalent bonds.

**GO:0030016**

myofibril

The contractile element of skeletal and cardiac muscle; a long, highly organized bundle of actin, myosin, and other proteins that contracts by a sliding filament mechanism.

**GO:0043292**

contractile fiber

Fibers, composed of actin, myosin, and associated proteins, found in cells of smooth or striated muscle.

**GO:0005796**

Golgi lumen

The volume enclosed by the membranes of any cisterna or subcompartment of the Golgi apparatus, including the cis- and trans-Golgi networks.

**3) Biological Process****GO:0003012**

muscle system process

A organ system process carried out at the level of a muscle. Muscle tissue is composed of contractile cells or fibers.

**GO:0006936**

muscle contraction

A process in which force is generated within muscle tissue, resulting in a change in muscle geometry. Force generation involves a chemo-mechanical energy conversion step that is carried out by the actin/myosin complex activity, which generates force through ATP hydrolysis.

**GO:0007517**

muscle organ development

The process whose specific outcome is the progression of the muscle over time, from its formation to the mature structure. The muscle is an organ consisting of a tissue made up of various elongated cells that are specialized to contract and thus to produce movement and mechanical work.

**GO:0030198**

extracellular matrix organization

A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of an extracellular matrix.

**GO:0043062**

extracellular structure organization

A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of structures in the space external to the outermost structure of a cell. For cells without external protective or external encapsulating structures this refers to space outside of the plasma membrane, and also covers the host cell environment outside an intracellular parasit.

**GO:0060249**

anatomical structure homeostasis

A homeostatic process involved in the maintenance of an internal steady state within a defined anatomical structure of an organism, including control of cellular proliferation and death and control of metabolic function. An anatomical structure is any biological entity that occupies space and is distinguished from its surroundings. Anatomical structures can be macroscopic such as a carpal, or microscopic such as an acrosome.

**GO:0050900**

leukocyte migration

The movement of a leukocyte within or between different tissues and organs of the body.

**GO:0048771**

The reorganization or renovation of existing tissues. This process can either change the characteristics of a tissue such as in blood vessel remodeling, or result in the dynamic equilibrium of a tissue such as in bone remodeling.

**GO:1904645**

response to amyloid-beta

Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a amyloid-beta stimulus.

**GO:0048641**

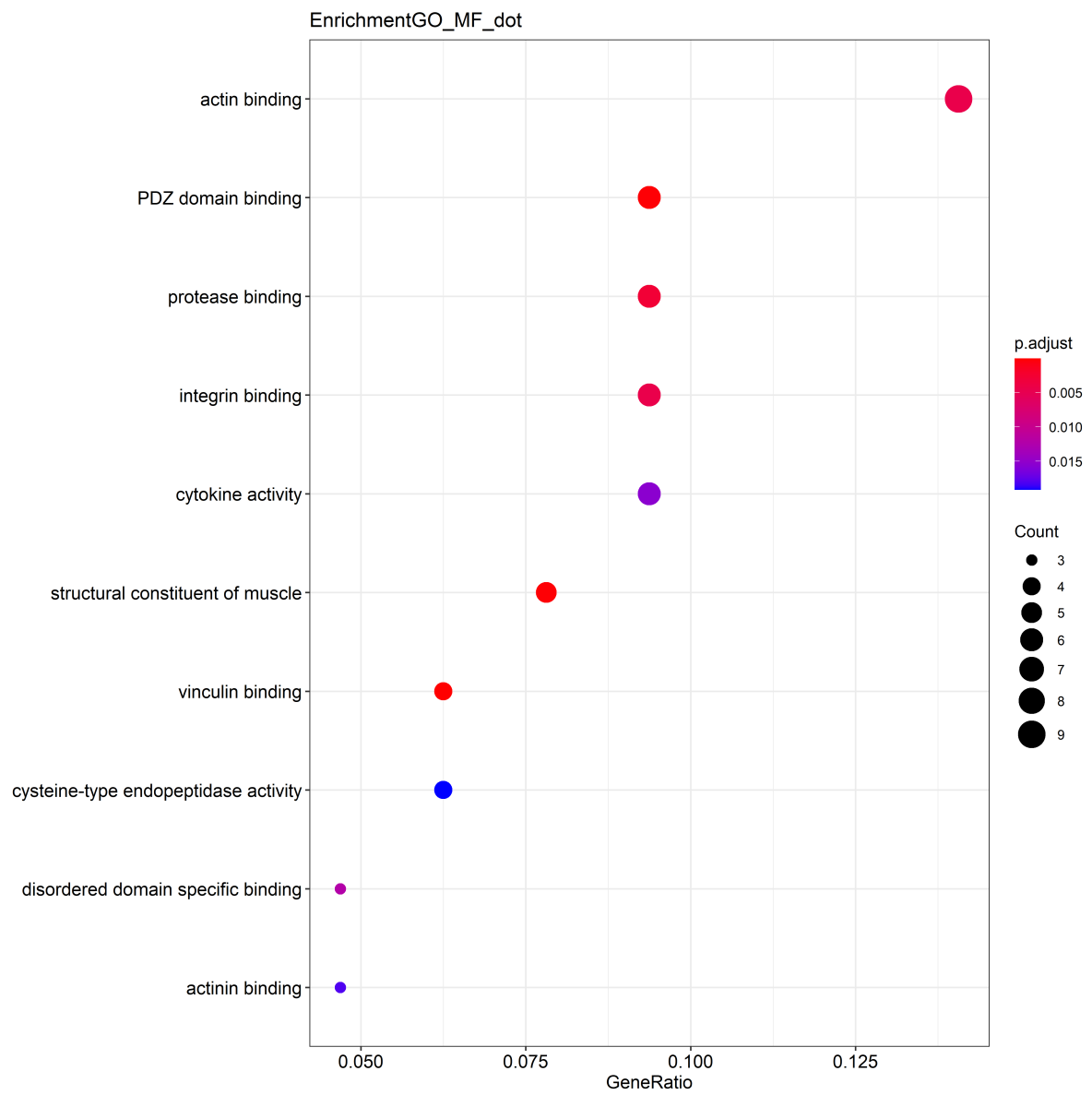
regulation of skeletal muscle tissue development

Any process that modulates the frequency, rate or extent of skeletal muscle tissue development.

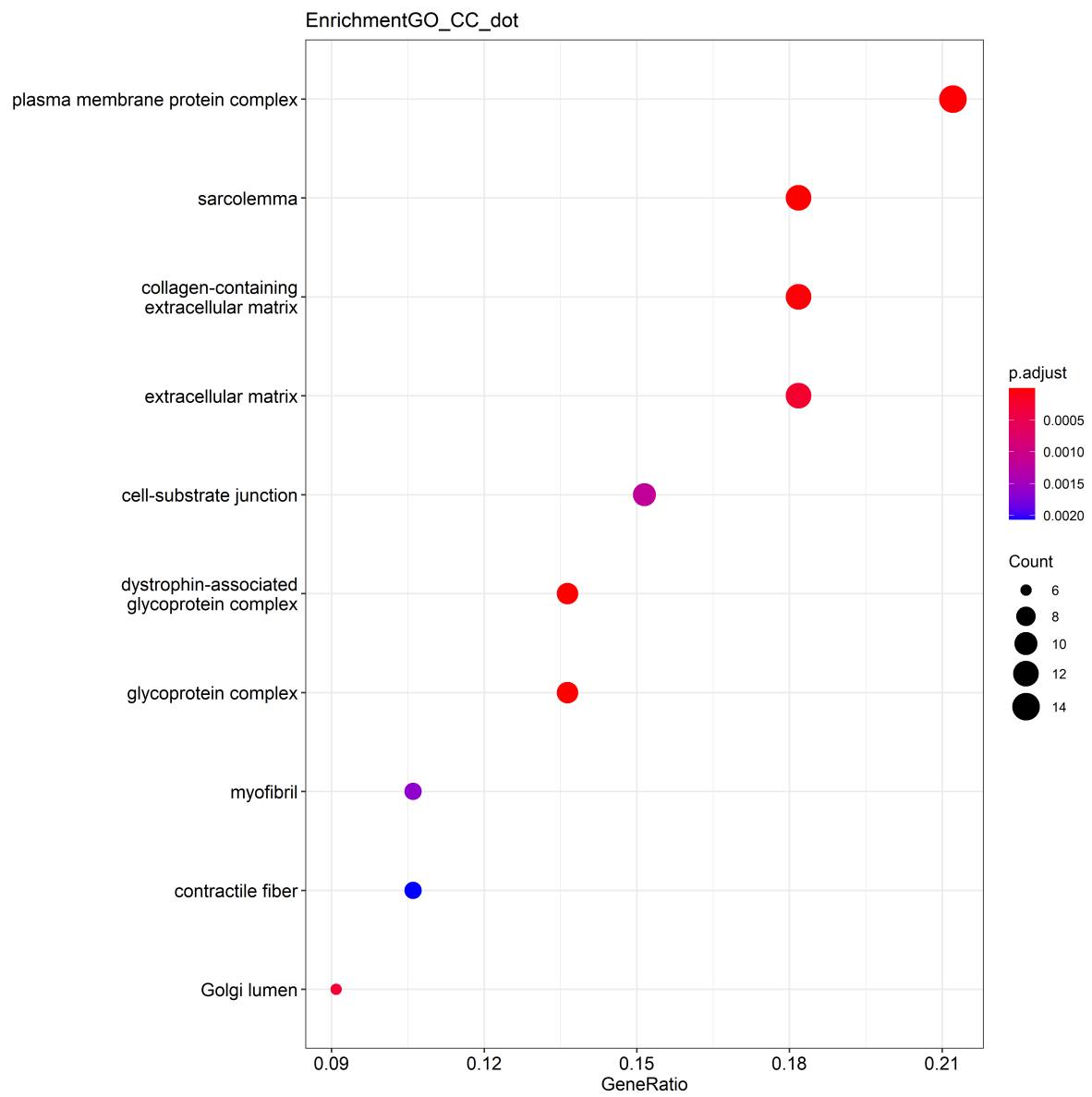
## **2. Visualization Result**

### **1) GO Enrichment\_MF Dotplot**

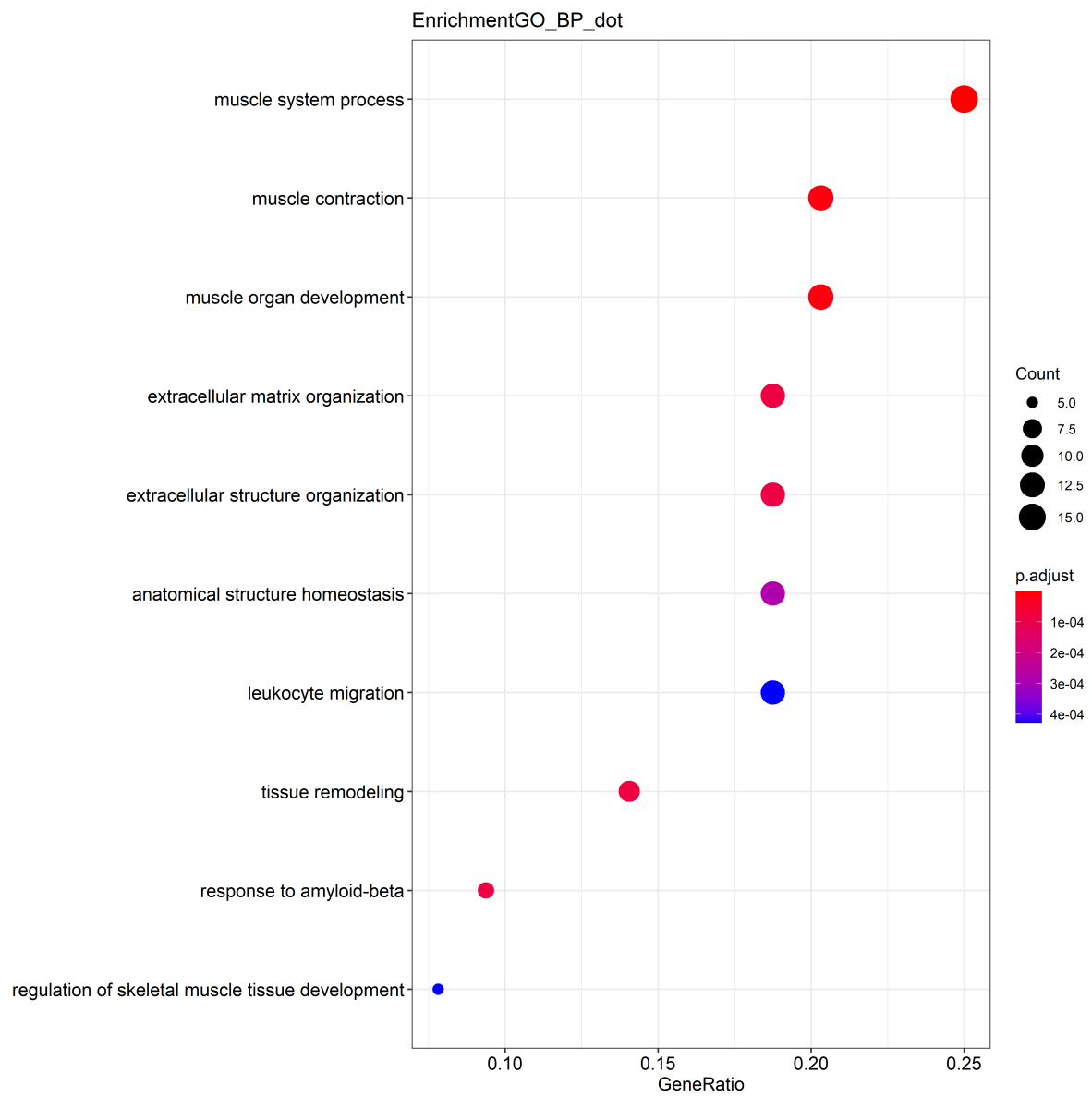




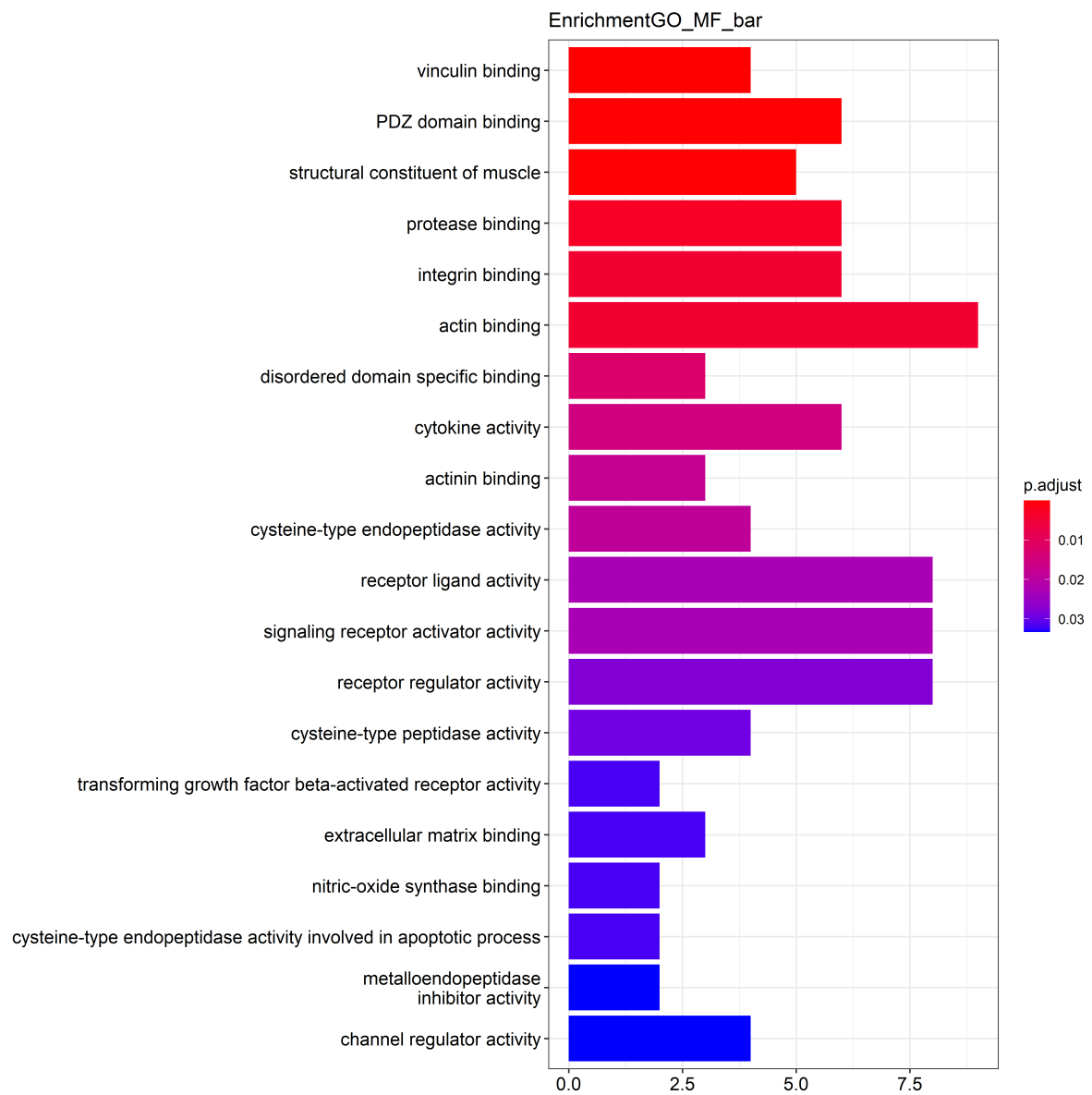
## 2) GO Enrichment\_CC Dotplot



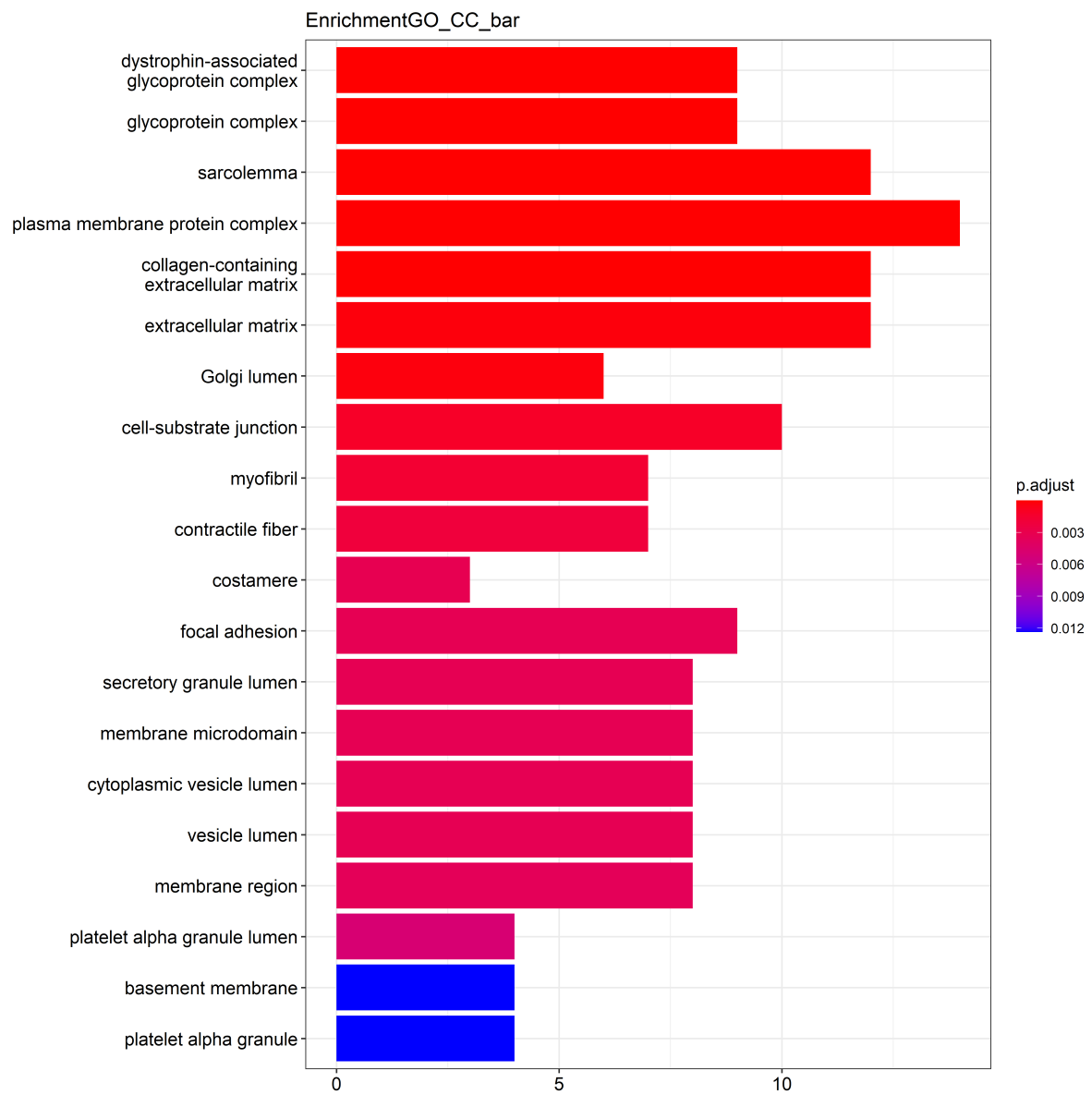
### 3) GO Enrichment\_BP Dotplot



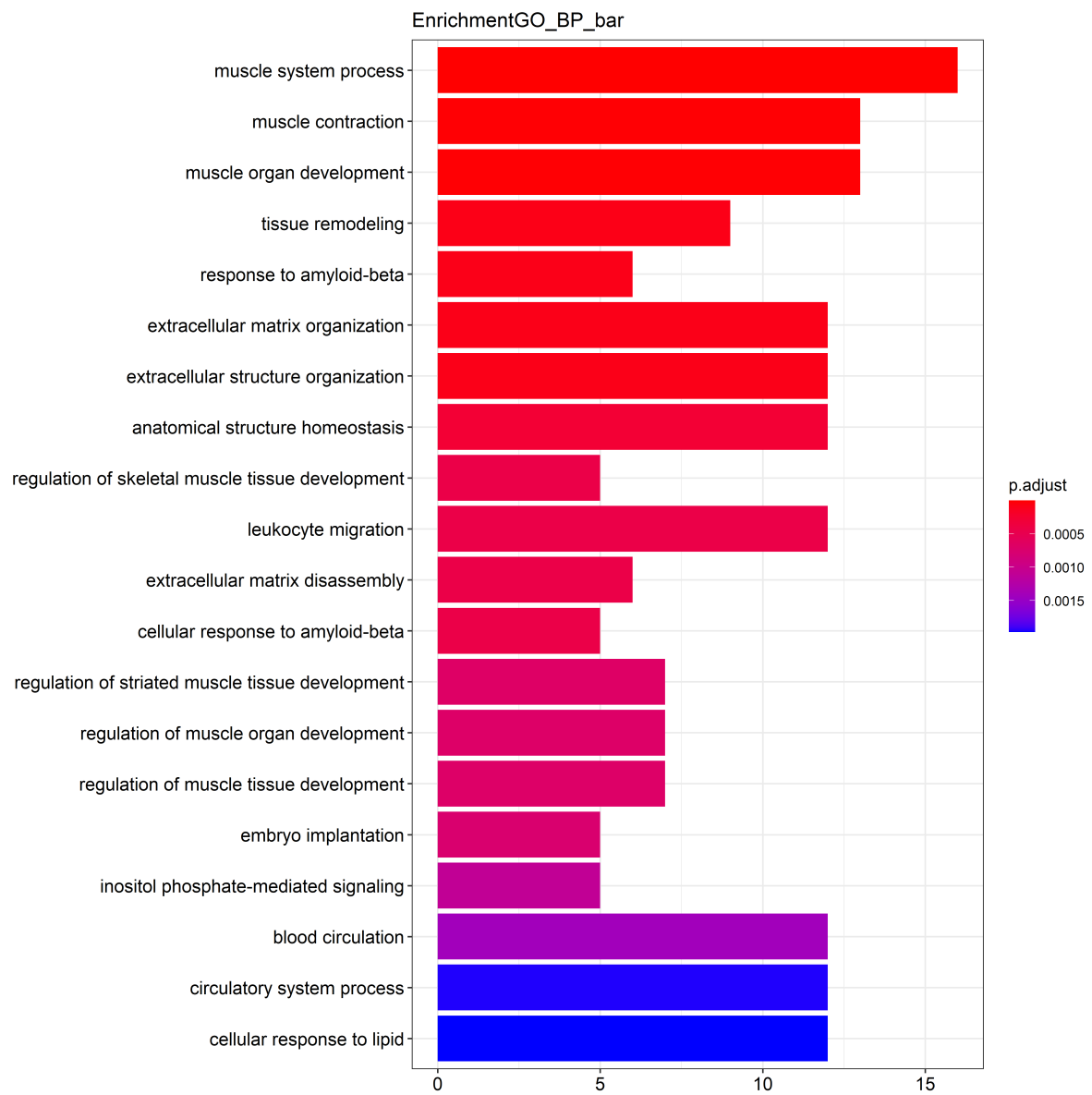
#### 4) GO Enrichment\_MF Barplot



## 5) GO Enrichment\_CC Barplot



## 6) GO Enrichment\_BP Barplot



## 7) GO Enrichment\_MF GO graph









In Cellular Component enrichment, result in this aspect mainly concerns system of intracellular membranes, cells interacting with extracellular components and some glycoprotein complex. The GO graph of cellular component shows that because of abnormal function of Golgi apparatus and Endoplasmic reticulum, glycoprotein complex cannot produce or have the same function as in the healthy body. As is know to all, adhesion and communication is mainly undertake by glycoprotein complex, so the abnormal interaction is closely related to glycoprotein complex.

In Biological Process enrichment, just like the symptoms of this disease, result of this part is mainly about muscle movement and muscle growth. Not only muscle movement and muscle growth itself, but their regulation mechanism also exist in our result. Additionally, GO graph about biological process is the most time-consuming part in my script in the reason of complex relationship and infections between growth, movement and regulation.

Taking all of these into consideration, we infer that:

