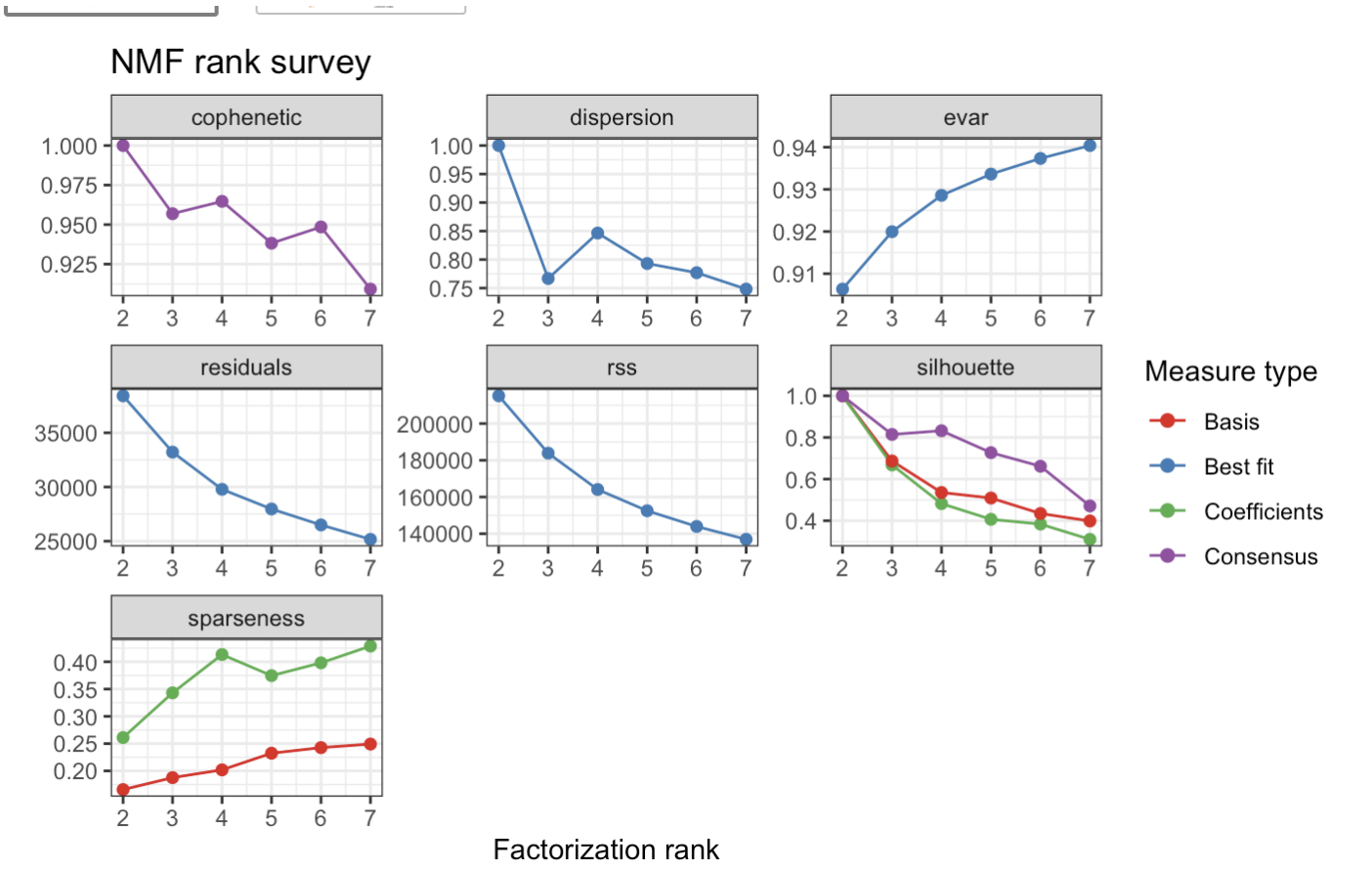
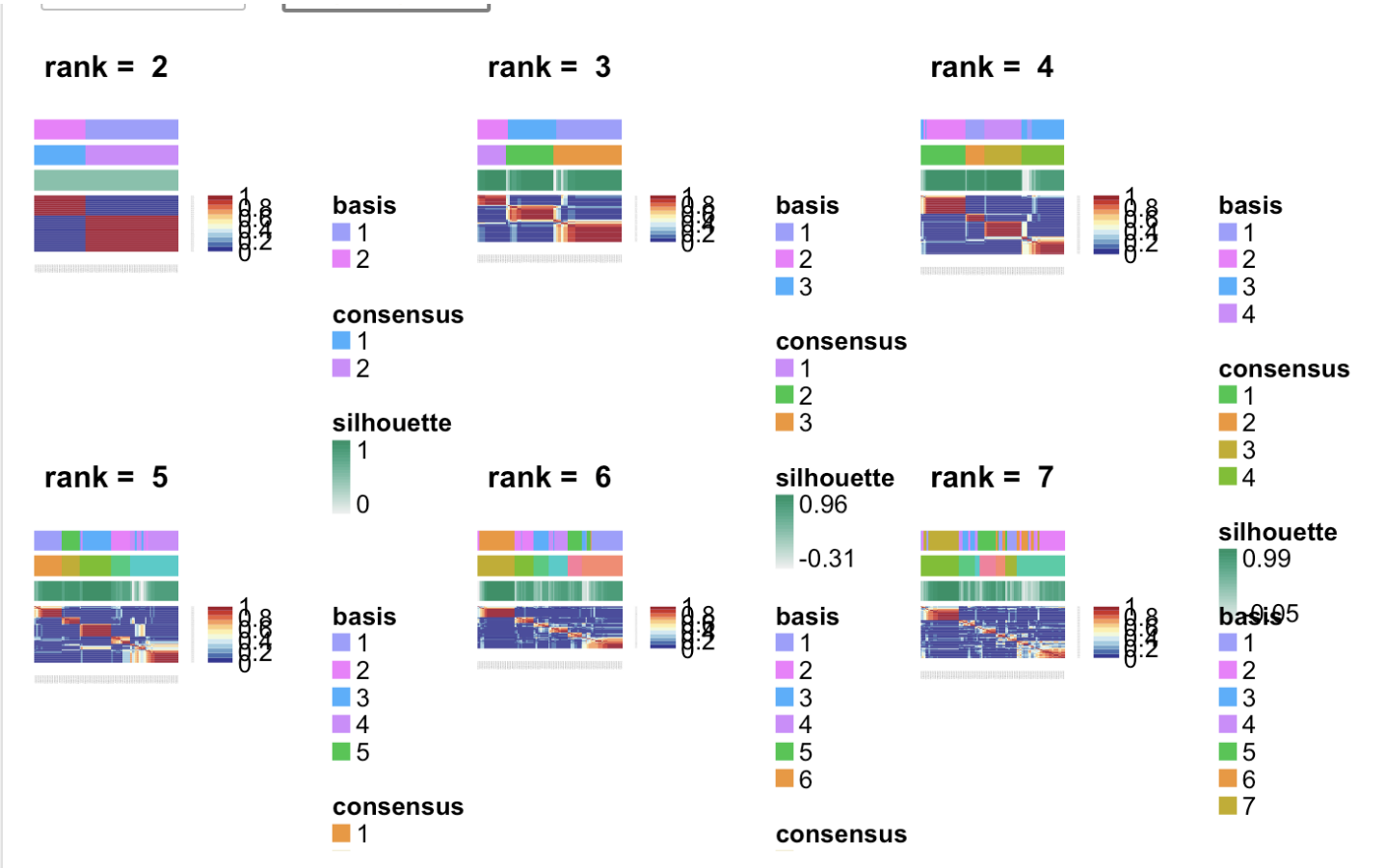
ML_LAQ2_submit

Q1





k=4 gives the best performance based on cophenetic coefficient, and now we perform NMF using k=4

Justification for selecting k = 4 as the optimal number of clusters:

Based on the NMF rank survey plots, we evaluated several metrics including the cophenetic coefficient, dispersion, sparseness, and silhouette scores. While k = 2 gives the highest cophenetic and silhouette values, it represents a very coarse clustering structure that may overlook biologically meaningful subgroups. Among k = 3 to 7, k = 4 consistently shows local peaks or plateau behavior in key metrics (especially cophenetic, dispersion, and sparseness), suggesting a stable and interpretable clustering solution. Therefore, we selected k = 4 as the optimal number of clusters to balance stability, separation, and biological interpretability.

Q2

代码块

```
library(limma)

cluster_assignment <- apply(h, 2, which.max)
cluster_assignment <- as.factor(cluster_assignment)
```

```r

top_genes_list <- list()


for (k in 1:4) {
  group <- ifelse(cluster_assignment == k, "cluster", "others")
  design <- model.matrix(~ factor(group))

  fit <- lmFit(d_exp, design)
  fit <- eBayes(fit)

  top <- topTable(fit, coef=2, number=20, sort.by="P")
  top_genes_list[[k]] <- rownames(top)
}
```

代码块

```r
top_genes_list[[1]]

top_genes_list[[2]]
top_genes_list[[3]]
top_genes_list[[4]]
```

output

```
 [1] "SFN"     "GJB6"    "FGFBP1"  "KRT6A"   "DSG3"    "GJB2"    "PPP2R2C"
"KRT6B"   "CXCL14"  "IL36G"   "FABP5"   "ANXA6"
[13] "CERS3"   "S100A2"  "PGLYRP3" "SLC2A1"  "ELN"     "CLIP3"   "SPRR1B"
"HAS3"

 [1] "REPIN1"   "ALDH1A1"  "ITGA3"    "MIR936"   "PLEK2"    "ARHGEF26"
"COL17A1"  "ATP6V0E2" "KRT14"    "LAMC2"    "LRIG1"
[12] "LAMA3"    "HLF"      "EPS8"     "EPHB6"    "RGS20"    "MYLIP"
"FAM171A1" "CAV1"     "PBX1"

 [1] "FUT3"      "PRSS27"    "CLCA4"     "SPRR2C"    "SCNN1B"    "KRT78"
"PSCA"      "SCEL"      "DUOXA2"    "PPL"
[11] "TMPRSS11F" "C2orf54"   "SPINK7"    "A2ML1"     "TTC9"      "CCDC64B"
"CEACAM5"   "NCCRP1"    "FMO2"      "CD24"

 [1] "OSGIN1"       "CYP4F3"       "ALDH3A1"      "LOC344887"    "CYP4F11"
 "DMRT2"        "AKR1C3"       "ABCC1"
 [9] "TSPAN7"       "TRIM16L"      "MDGA1"        "MRAP2"        "CBR1"
 "PTGR1"        "LOC100133286" "UGT1A1"
[17] "AK126334"     "TXNRD1"       "ADAM23"       "SLC7A11"
```

Q3

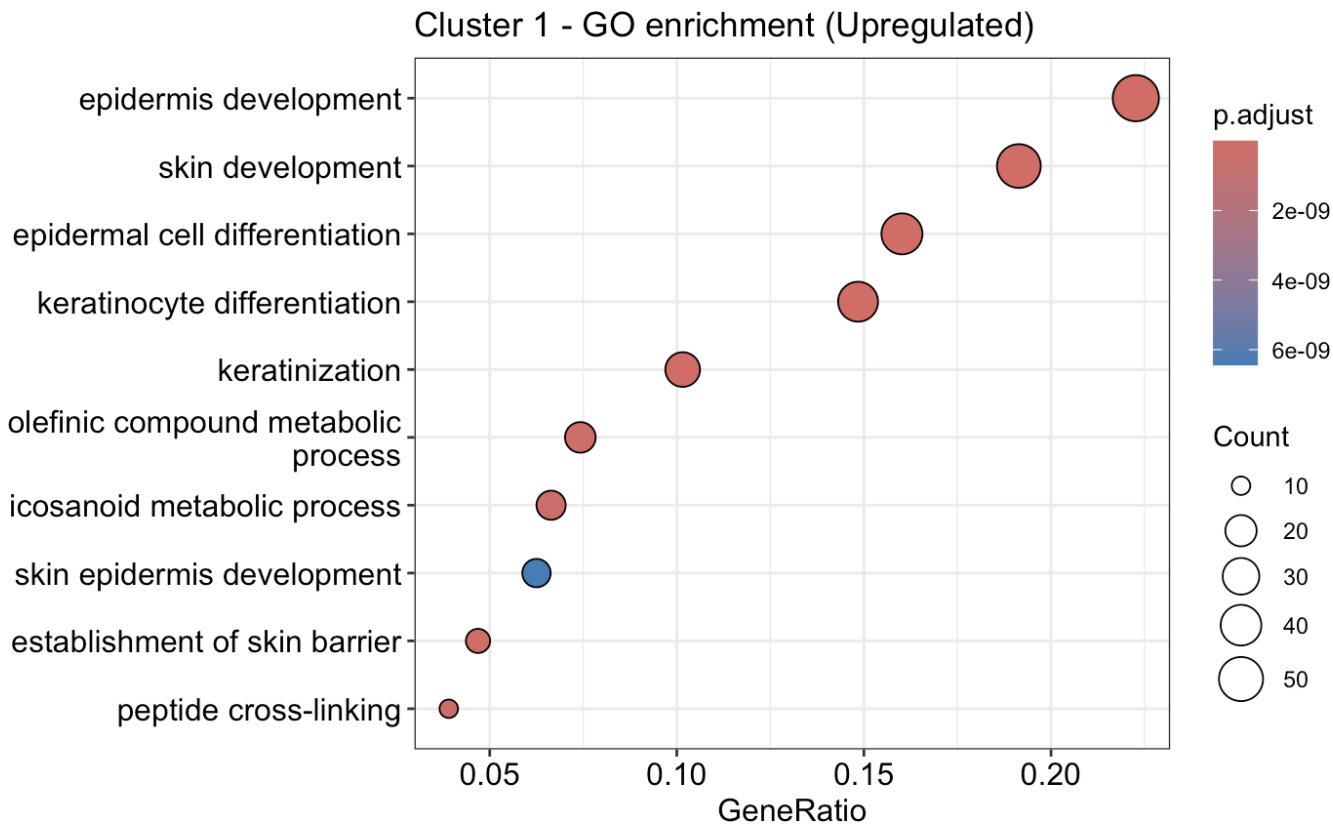
I selected upregulated genes based on these two criteria:

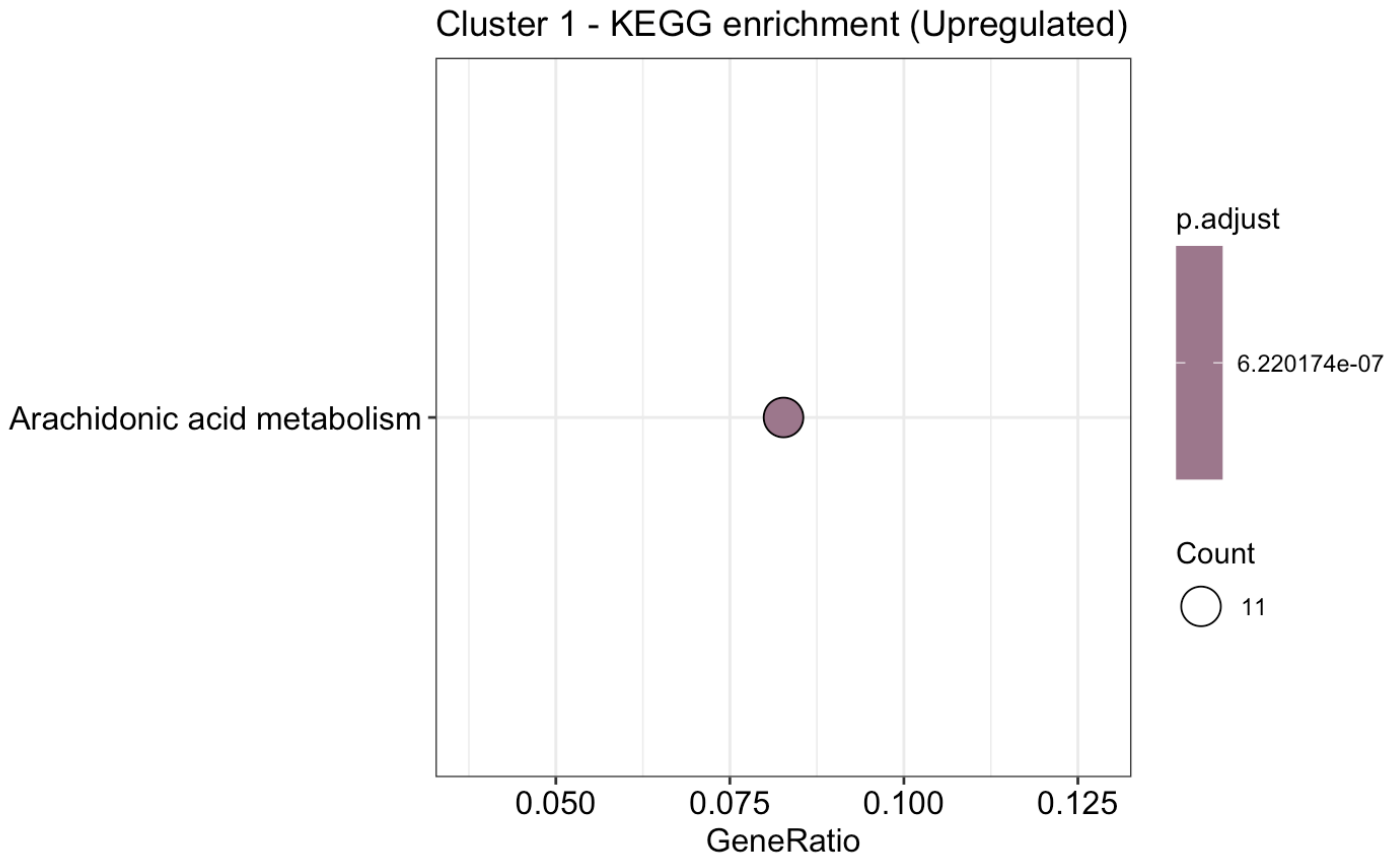
`$logFC > 0 & $adj.P.Val < 0.05`.

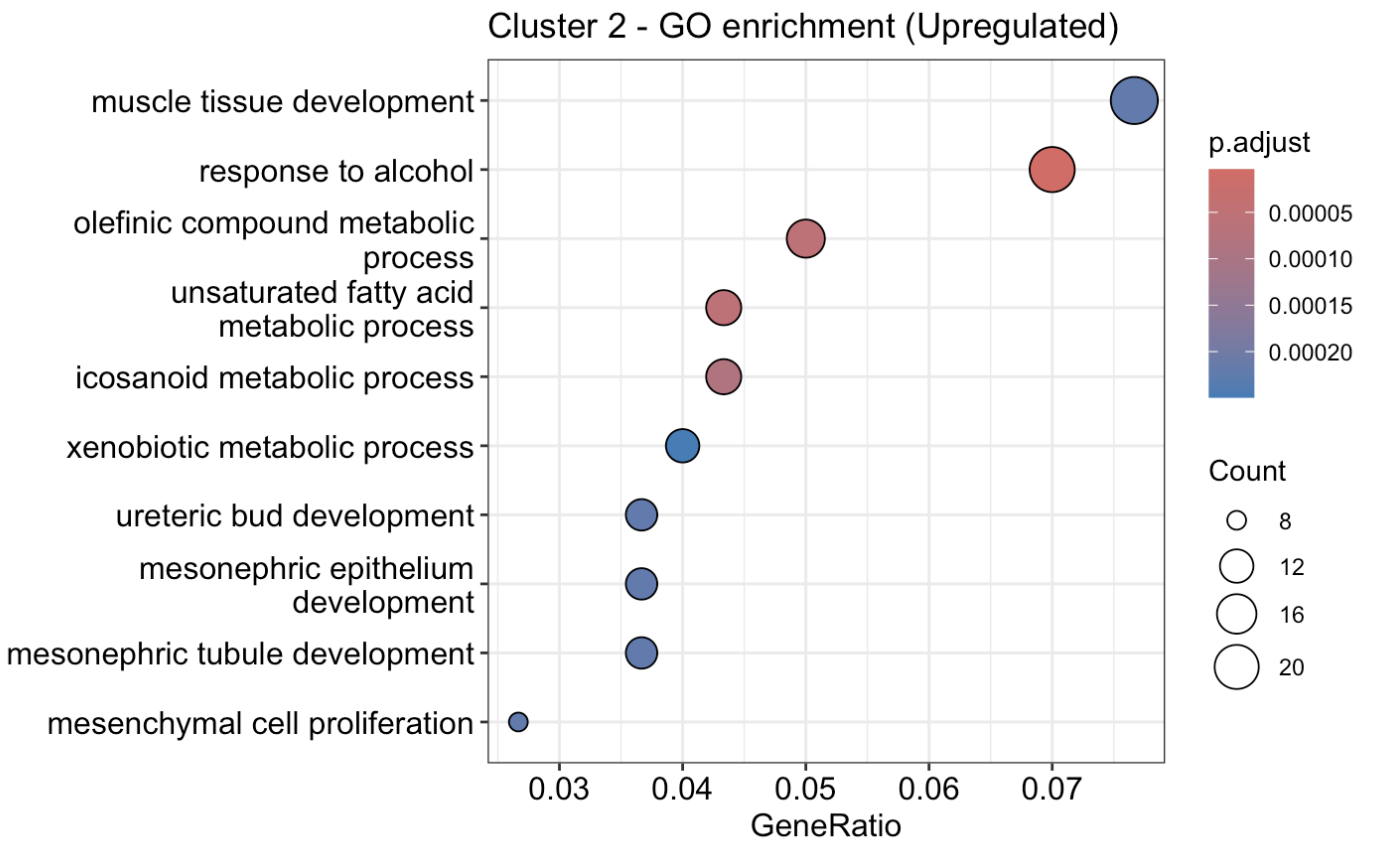
I selected downregulated genes based on these two criteria.

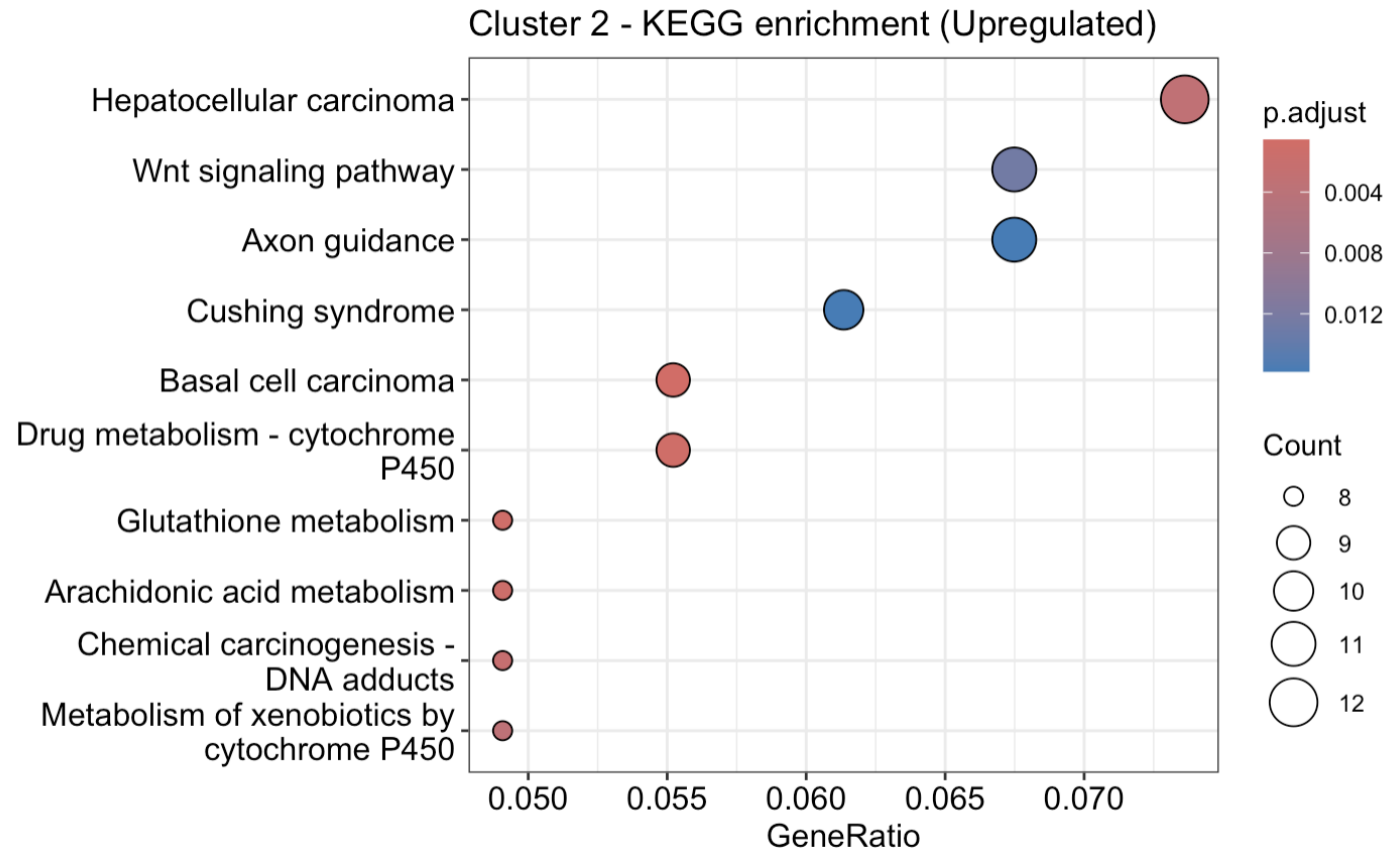
`$logFC < 0 & $adj.P.Val < 0.05`

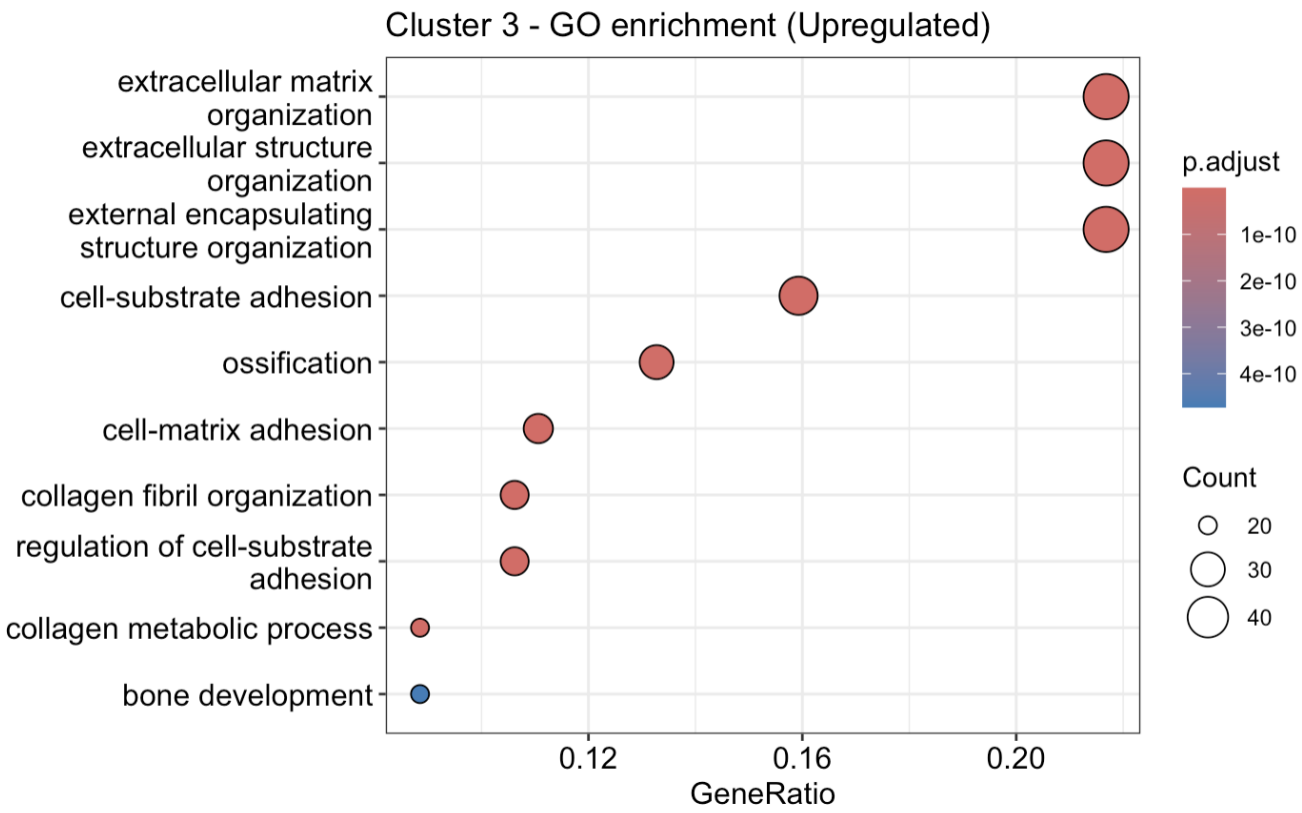
up

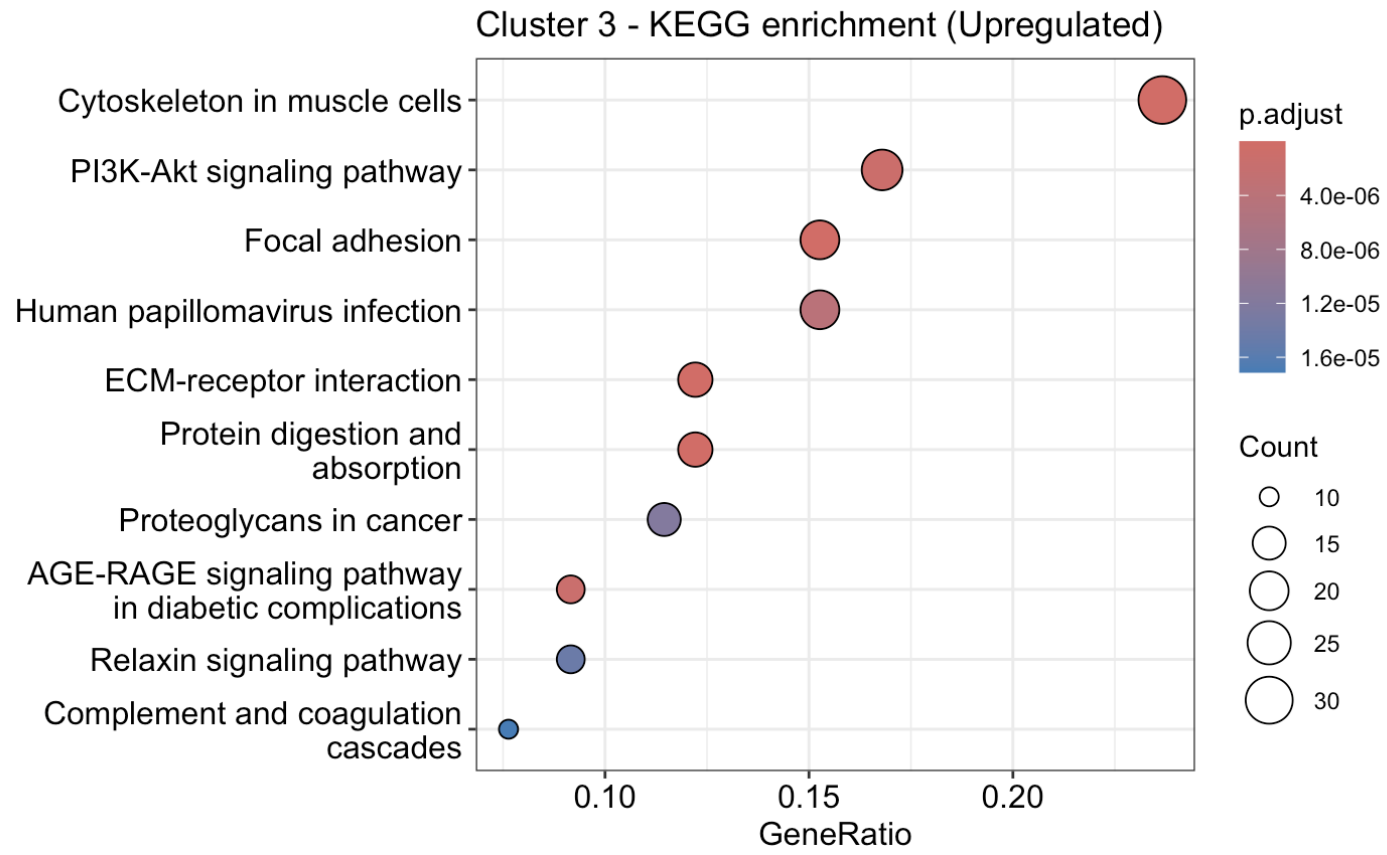


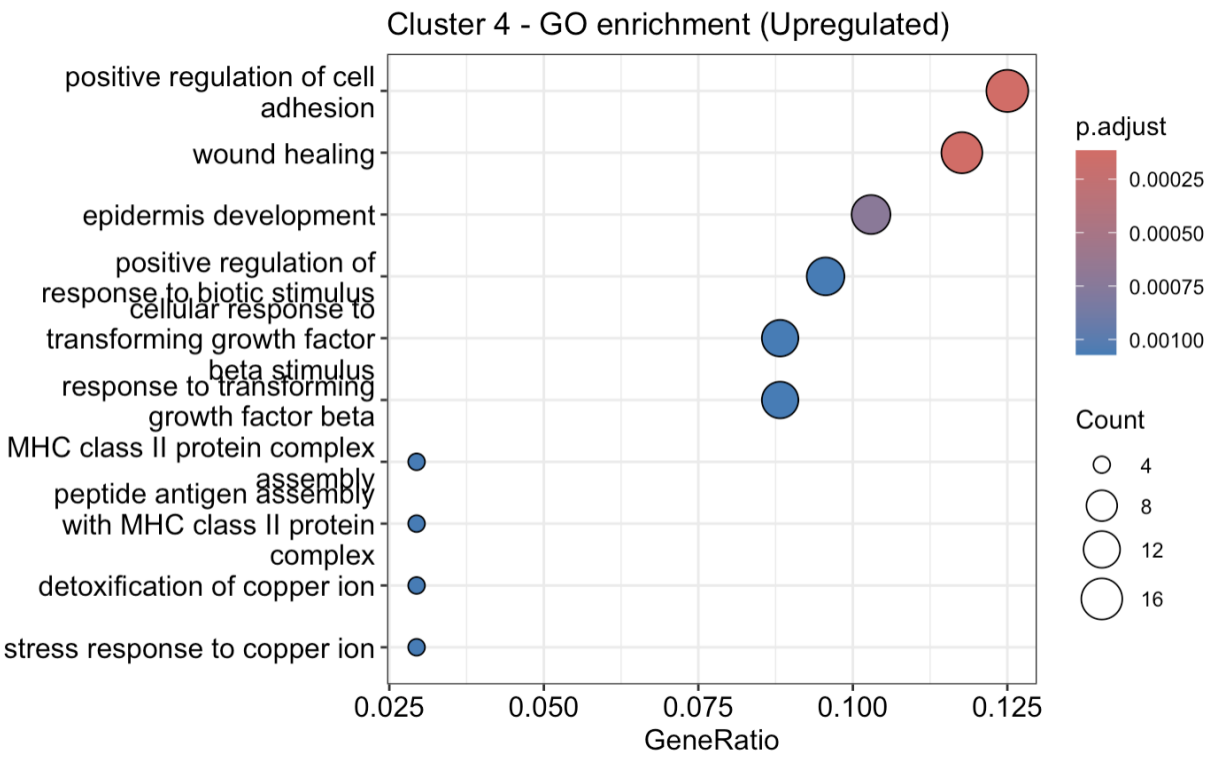

Cluster 1 - KEGG enrichment (Upregulated)
Arachidonic acid metabolism
0.050
0.075
0.100
0.125
GeneRatio
p.adjust
6.220174e-07
Count
11

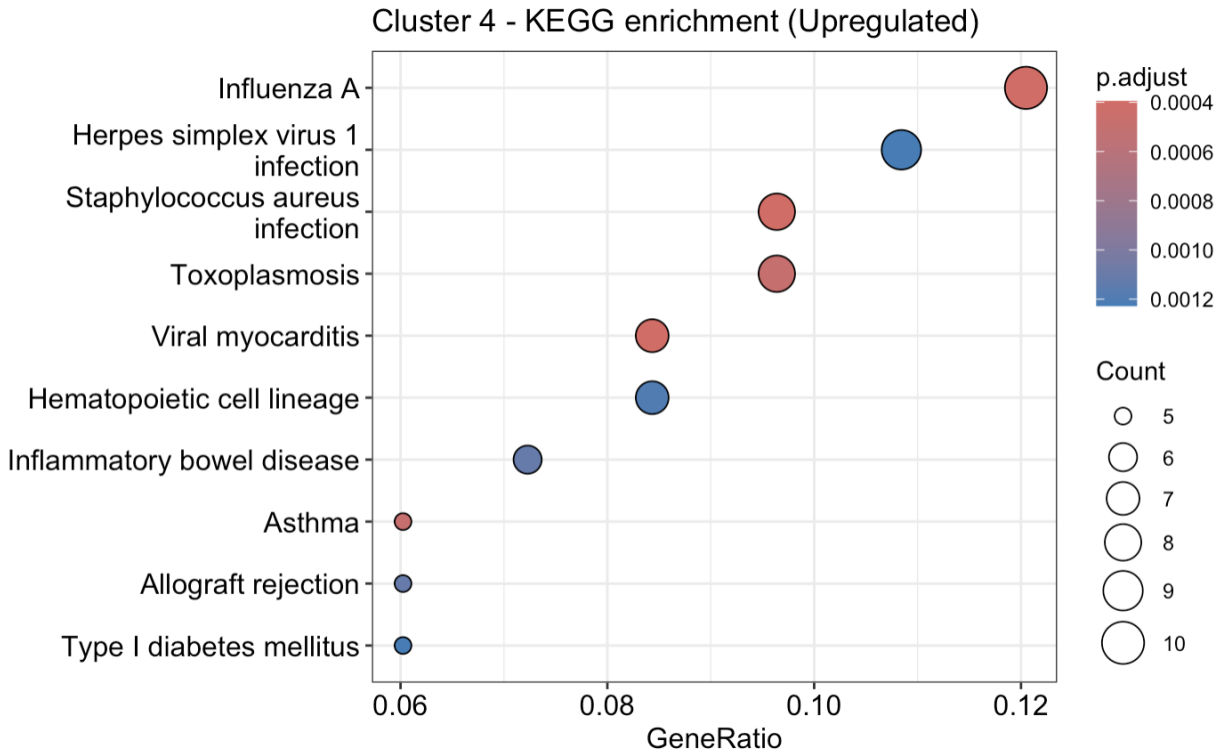

Cluster 2 - GO enrichment (Upregulated)
muscle tissue development
response to alcohol
olefinic compound metabolic process
unsaturated fatty acid metabolic process
icosanoid metabolic process
xenobiotic metabolic process
ureteric bud development
mesonephric epithelium development
mesonephric tubule development
mesenchymal cell proliferation
0.03
0.04
0.05
0.06
0.07
GeneRatio
p.adjust
0.00005
0.00010
0.00015
0.00020
Count
8
12
16
20


Cluster 2 - KEGG enrichment (Upregulated)
Hepatocellular carcinoma
Wnt signaling pathway
Axon guidance
Cushing syndrome
Basal cell carcinoma
Drug metabolism - cytochrome P450
Glutathione metabolism
Arachidonic acid metabolism
Chemical carcinogenesis - DNA adducts
Metabolism of xenobiotics by cytochrome P450
0.050
0.055
0.060
0.065
0.070
GeneRatio
p.adjust
0.004
0.008
0.012
Count
8
9
10
11
12


Cluster 3 - GO enrichment (Upregulated)
extracellular matrix organization
extracellular structure organization
external encapsulating structure organization
cell-substrate adhesion
ossification
cell-matrix adhesion
collagen fibril organization
regulation of cell-substrate adhesion
collagen metabolic process
bone development
0.12
0.16
0.20
GeneRatio
p.adjust
1e-10
2e-10
3e-10
4e-10
Count
20
30
40


Cluster 3 - KEGG enrichment (Upregulated)
Cytoskeleton in muscle cells
PI3K-Akt signaling pathway
Focal adhesion
Human papillomavirus infection
ECM-receptor interaction
Protein digestion and absorption
Proteoglycans in cancer
AGE-RAGE signaling pathway in diabetic complications
Relaxin signaling pathway
Complement and coagulation cascades
0.10
0.15
0.20
GeneRatio
p.adjust
4.0e-06
8.0e-06
1.2e-05
1.6e-05
Count
10
15
20
25
30

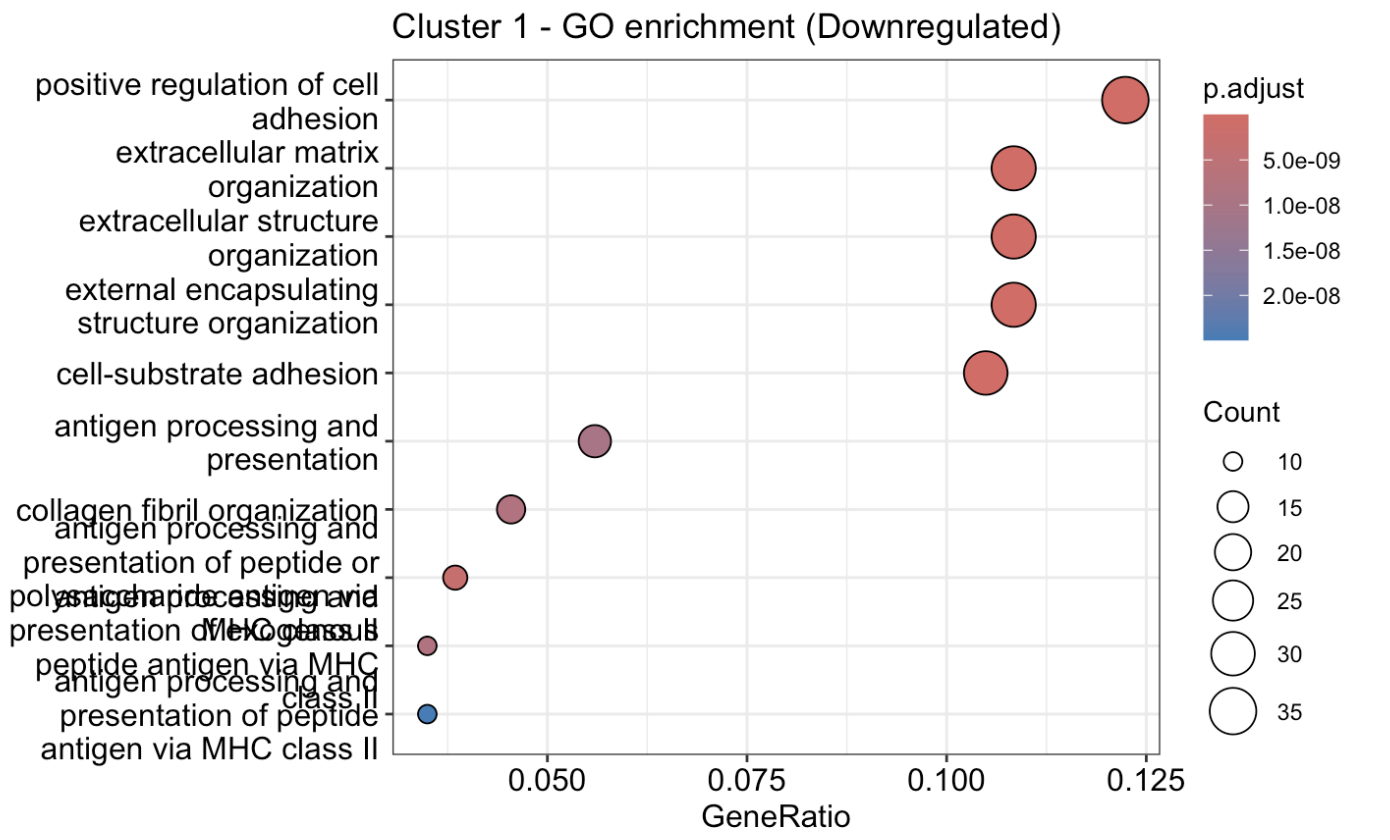

Cluster 4 - GO enrichment (Upregulated)
positive regulation of cell adhesion
wound healing
epidermis development
positive regulation of response to biotic stimulus
cellular response to transforming growth factor beta stimulus
response to transforming growth factor beta
MHC class II protein complex assembly
peptide antigen assembly with MHC class II protein complex
detoxification of copper ion
stress response to copper ion
0.025
0.050
0.075
0.100
0.125
GeneRatio
p.adjust
0.00025
0.00050
0.00075
0.00100
Count
4
8
12
16

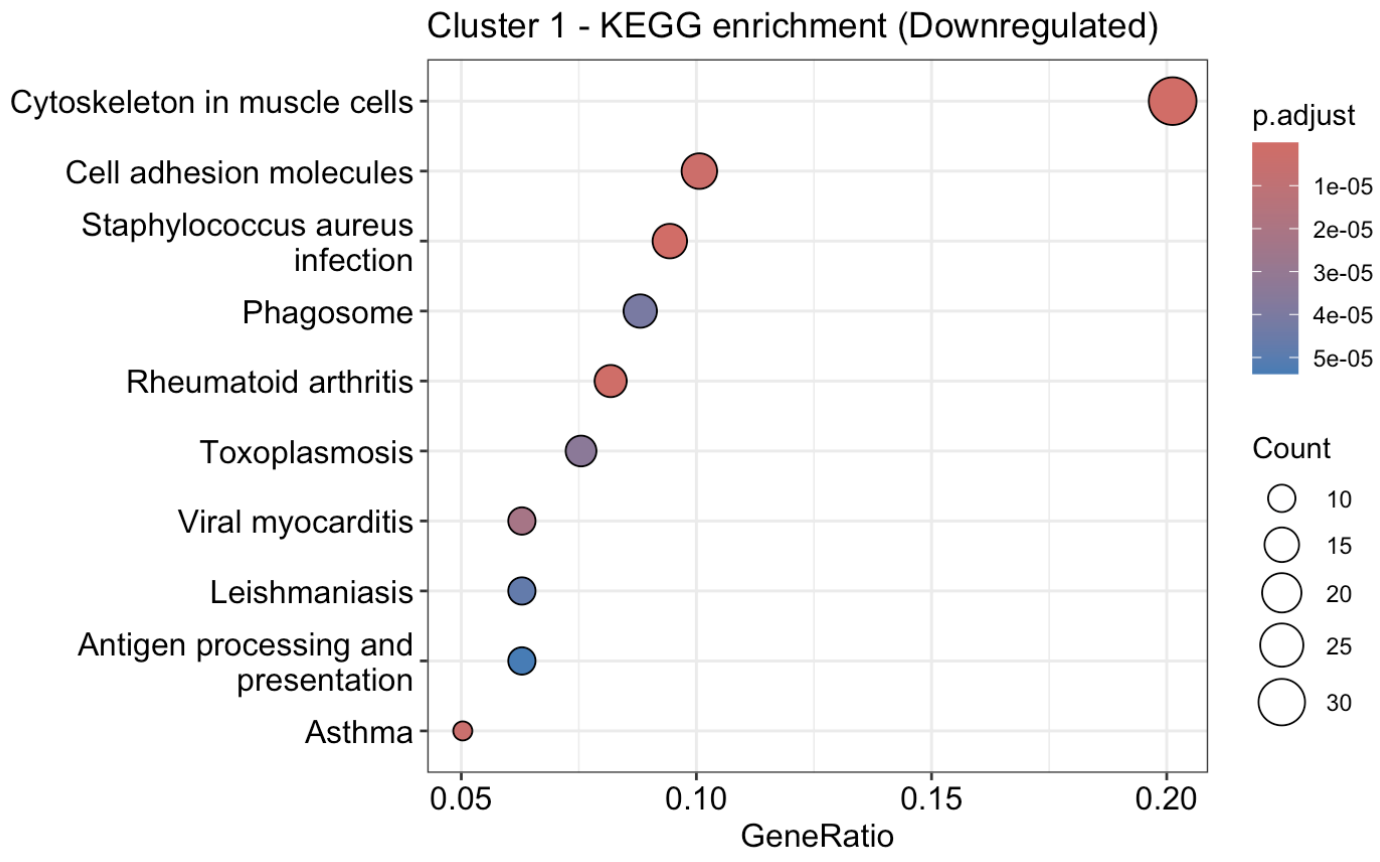


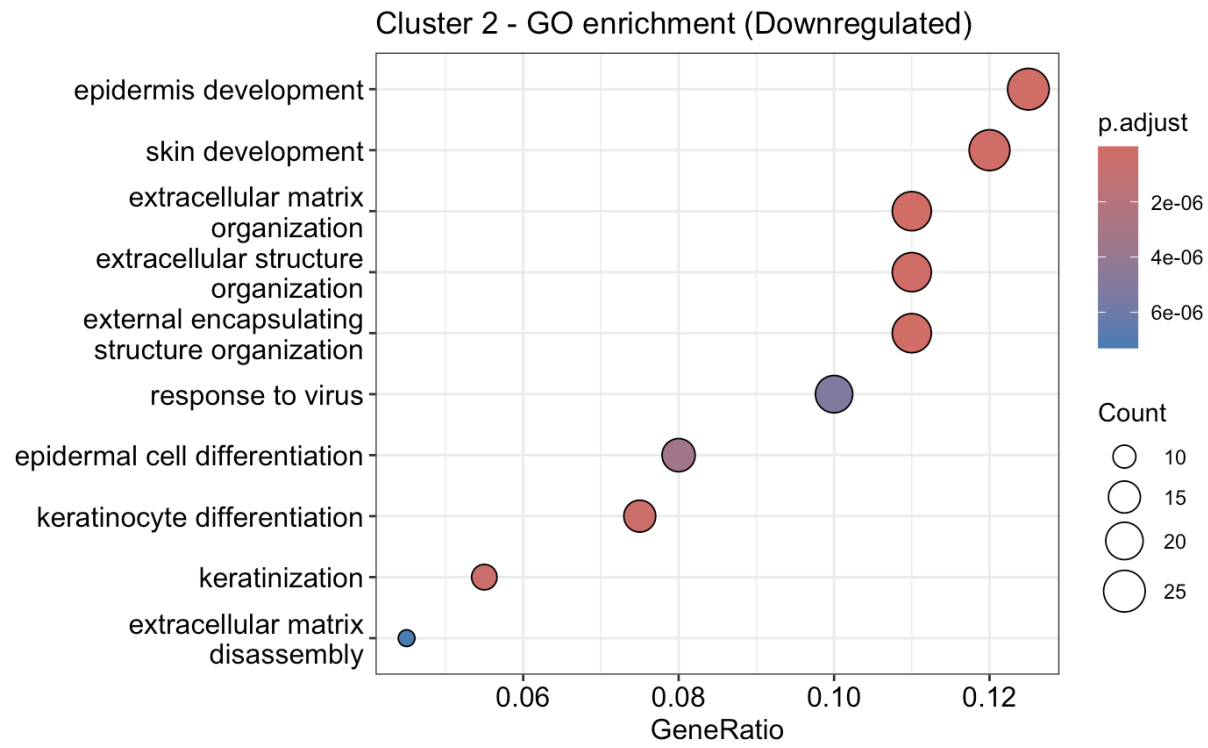
conclusion

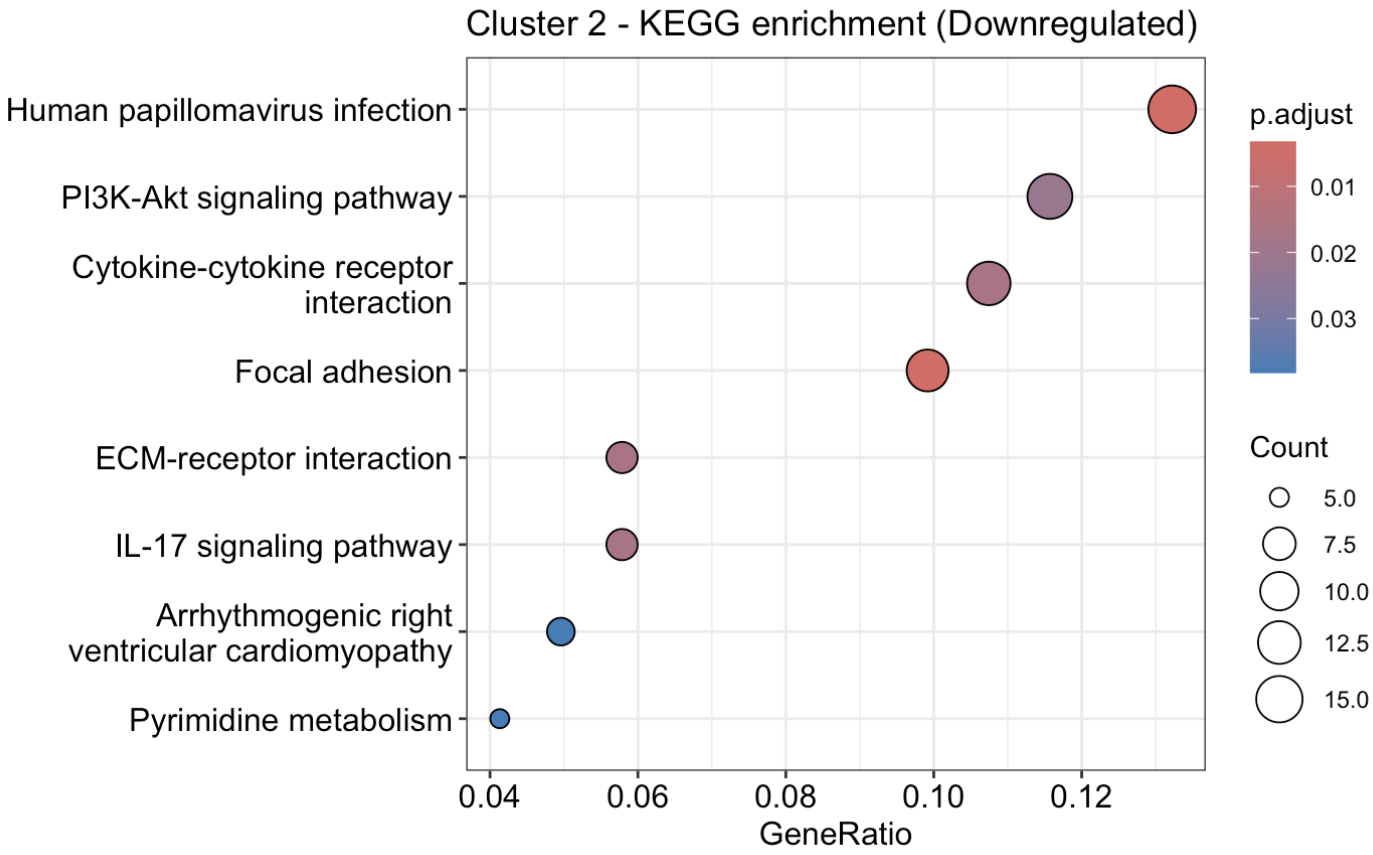
Cluster	Top 3 GO Biological Process (Upregulated)	Top 3 KEGG Pathways (Upregulated)
Cluster 1	epidermis development skin development epidermal cell differentiation	Arachidonic acid metabolism
Cluster 2	muscle tissue development response to alcohol olefinic compound metabolic process	Hepatocellular carcinoma Wnt signaling pathway Axon guidance
Cluster 3	extracellular matrix organization extracellular structure organization external encapsulating structure organization	Cytoskeleton in muscle cells PI3K-Akt signaling pathway Focal adhesion
Cluster 4	positive regulation of cell adhesion wound healing epidermis development	Influenza A Herpes simplex virus 1 infection Staphylococcus aureus infection

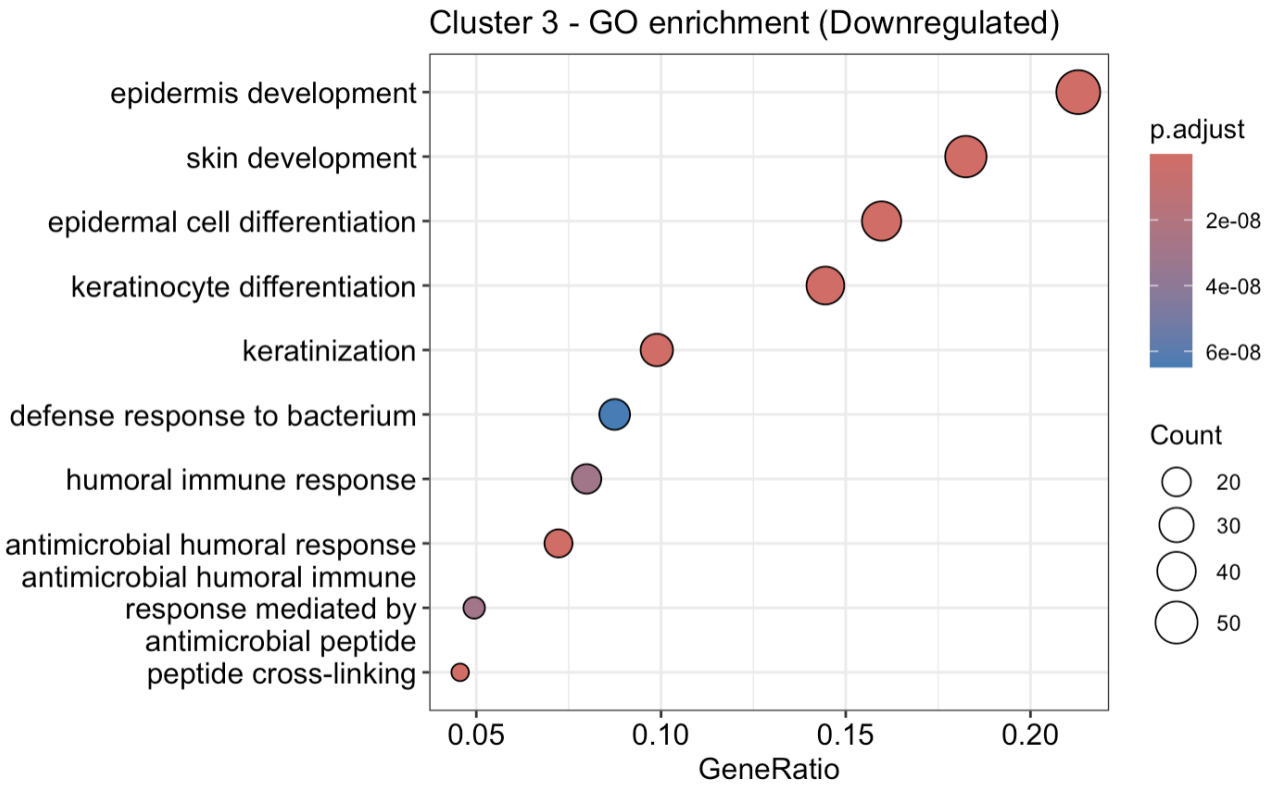
down

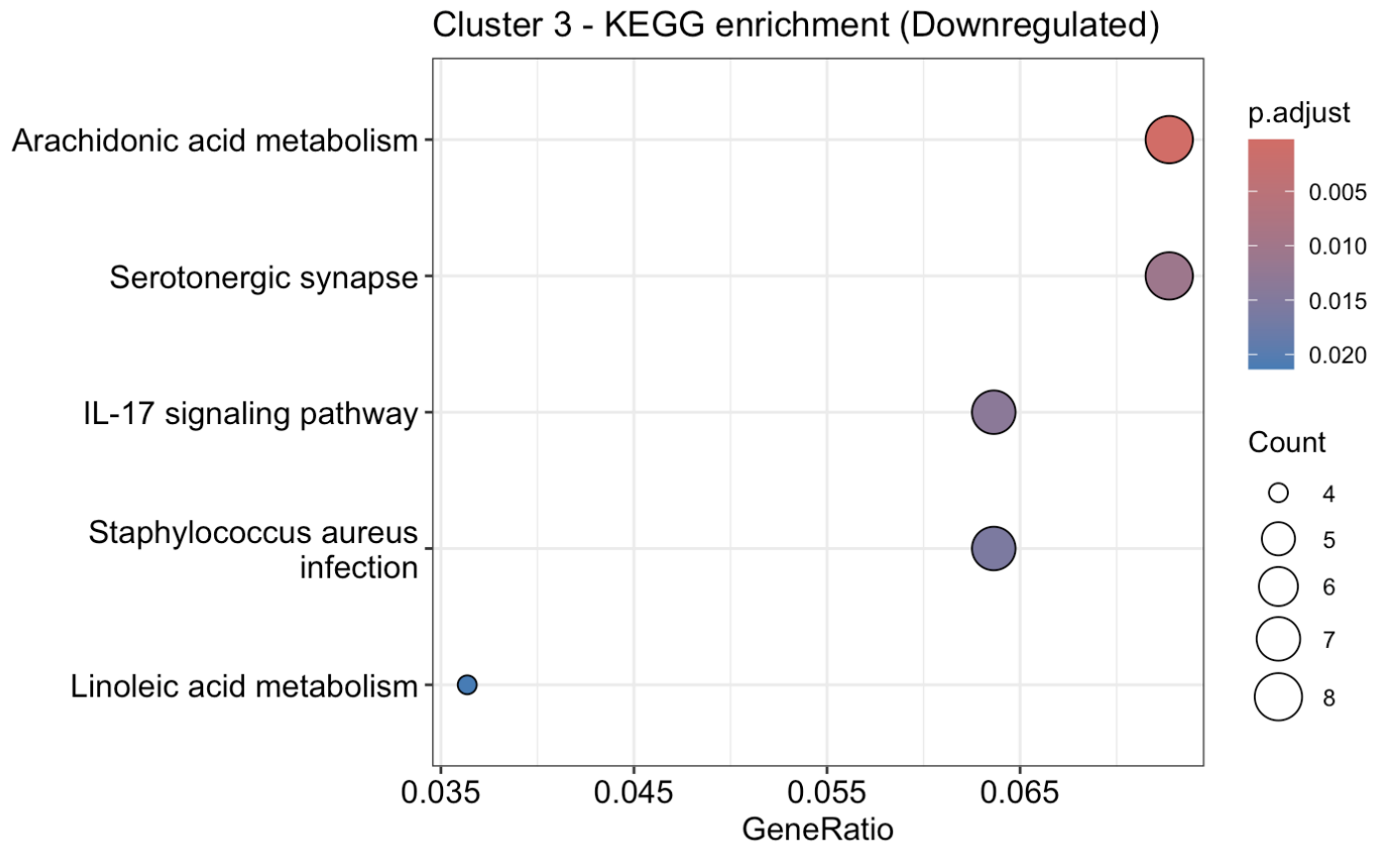

Cluster 1 - GO enrichment (Downregulated)
positive regulation of cell adhesion
extracellular matrix organization
extracellular structure organization
external encapsulating structure organization
cell-substrate adhesion
antigen processing and presentation
collagen fibril organization
antigen processing and presentation of peptide or
peptide antigen via MHC class II
antigen processing and presentation of peptide antigen via MHC class II
0.050
0.075
0.100
0.125
GeneRatio
p.adjust
5.0e-09
1.0e-08
1.5e-08
2.0e-08
Count
10
15
20
25
30
35

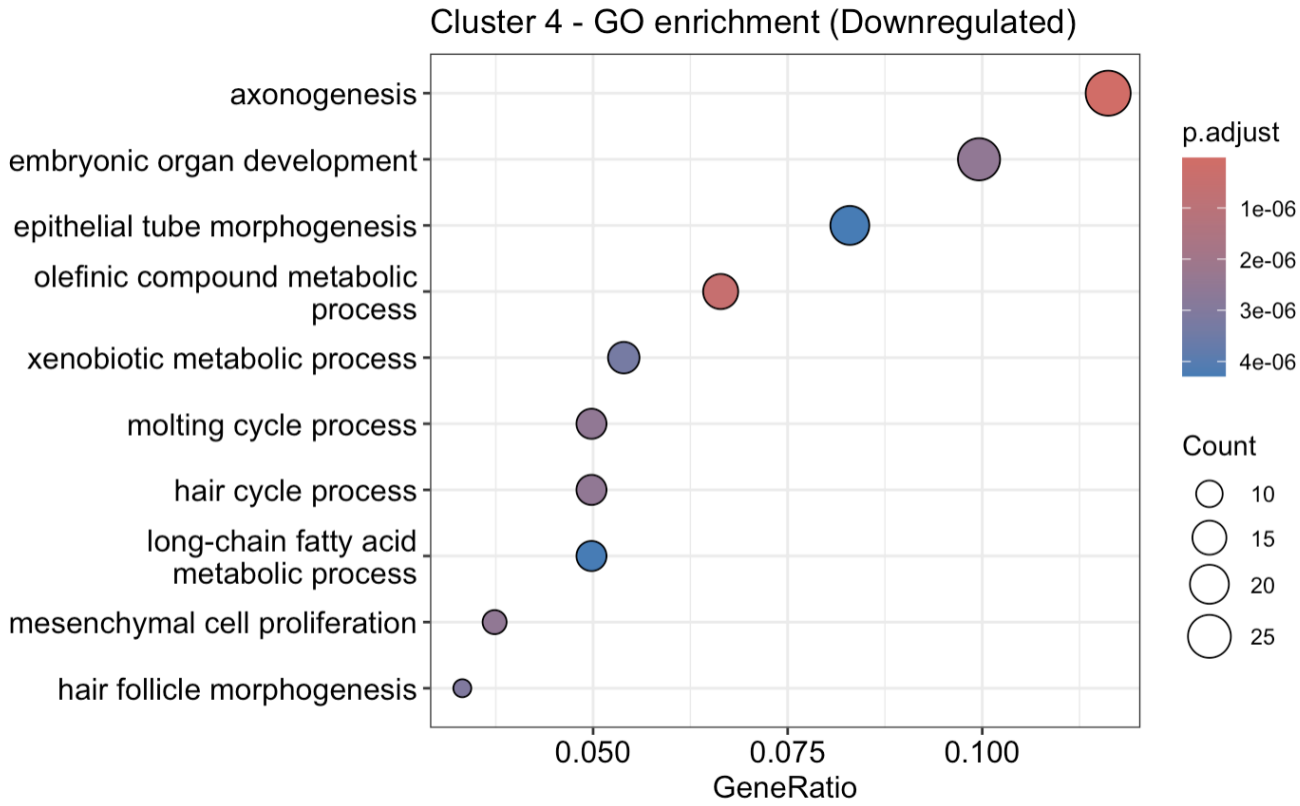

Cluster 1 - KEGG enrichment (Downregulated)
Cytoskeleton in muscle cells
Cell adhesion molecules
Staphylococcus aureus infection
Phagosome
Rheumatoid arthritis
Toxoplasmosis
Viral myocarditis
Leishmaniasis
Antigen processing and presentation
Asthma
0.05
0.10
0.15
0.20
GeneRatio
p.adjust
1e-05
2e-05
3e-05
4e-05
5e-05
Count
10
15
20
25
30

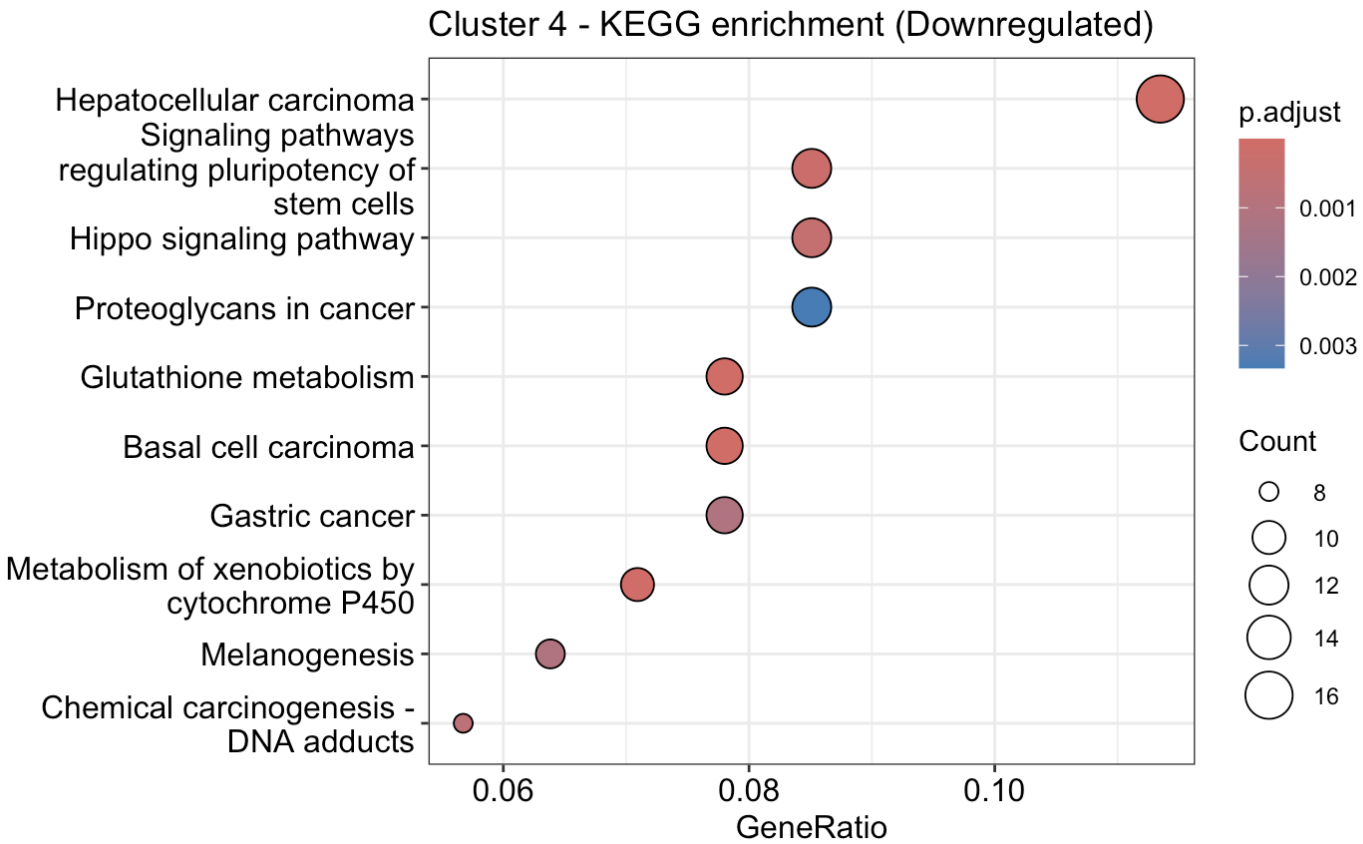

Cluster 2 - GO enrichment (Downregulated)
epidermis development
skin development
extracellular matrix organization
extracellular structure organization
external encapsulating structure organization
response to virus
epidermal cell differentiation
keratinocyte differentiation
keratinization
extracellular matrix disassembly
0.06
0.08
0.10
0.12
GeneRatio
p.adjust
2e-06
4e-06
6e-06
Count
10
15
20
25


Cluster 2 - KEGG enrichment (Downregulated)
Human papillomavirus infection
PI3K-Akt signaling pathway
Cytokine-cytokine receptor interaction
Focal adhesion
ECM-receptor interaction
IL-17 signaling pathway
Arrhythmogenic right ventricular cardiomyopathy
Pyrimidine metabolism
0.04
0.06
0.08
0.10
0.12
GeneRatio
p.adjust
0.01
0.02
0.03
Count
5.0
7.5
10.0
12.5
15.0


Cluster 3 - GO enrichment (Downregulated)
epidermis development
skin development
epidermal cell differentiation
keratinocyte differentiation
keratinization
defense response to bacterium
humoral immune response
antimicrobial humoral response
antimicrobial humoral immune response mediated by antimicrobial peptide
peptide cross-linking
0.05
0.10
0.15
0.20
GeneRatio
p.adjust
2e-08
4e-08
6e-08
Count
20
30
40
50


Cluster 3 - KEGG enrichment (Downregulated)
Arachidonic acid metabolism
Serotonergic synapse
IL-17 signaling pathway
Staphylococcus aureus infection
Linoleic acid metabolism
0.035
0.045
0.055
0.065
GeneRatio
p.adjust
0.005
0.010
0.015
0.020
Count
4
5
6
7
8


Cluster 4 - GO enrichment (Downregulated)
axonogenesis
embryonic organ development
epithelial tube morphogenesis
olefinic compound metabolic process
xenobiotic metabolic process
molting cycle process
hair cycle process
long-chain fatty acid metabolic process
mesenchymal cell proliferation
hair follicle morphogenesis
0.050
0.075
0.100
GeneRatio
p.adjust
1e-06
2e-06
3e-06
4e-06
Count
10
15
20
25


Cluster 4 - KEGG enrichment (Downregulated)
Hepatocellular carcinoma
Signaling pathways regulating pluripotency of stem cells
Hippo signaling pathway
Proteoglycans in cancer
Glutathione metabolism
Basal cell carcinoma
Gastric cancer
Metabolism of xenobiotics by cytochrome P450
Melanogenesis
Chemical carcinogenesis - DNA adducts
0.06
0.08
0.10
GeneRatio
p.adjust
0.001
0.002
0.003
Count
8
10
12
14
16

conclusion

Cluster	Top 3 GO Biological Process (Downregulated)	Top 3 KEGG Pathways (Downregulated)
Cluster 1	positive regulation of cell adhesion extracellular matrix organization extracellular structure organization	Cytoskeleton in muscle cells Cell adhesion molecules Staphylococcus aureus infection
Cluster 2	epidermis development skin development extracellular matrix organization	Human papillomavirus infection PI3K-Akt signaling pathway Cytokine-cytokine receptor interaction
Cluster 3	epidermis development skin development epidermal cell differentiation	Arachidonic acid metabolism Serotonergic synapse IL-17 signaling pathway
Cluster 4	axonogenesis embryonic organ development epithelial tube morphogenesis	Hepatocellular carcinoma Signaling pathways regulating pluripotency of stem cells Hippo signaling pathway

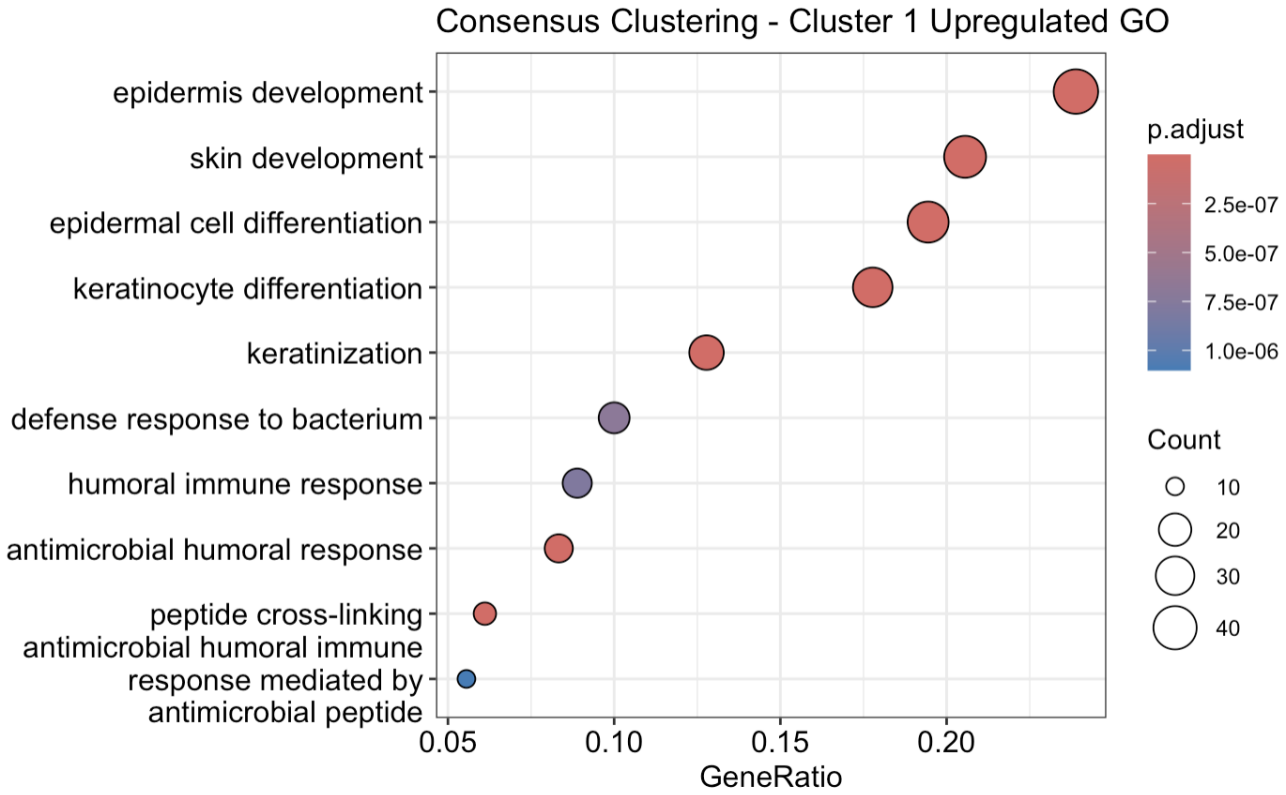
Q4

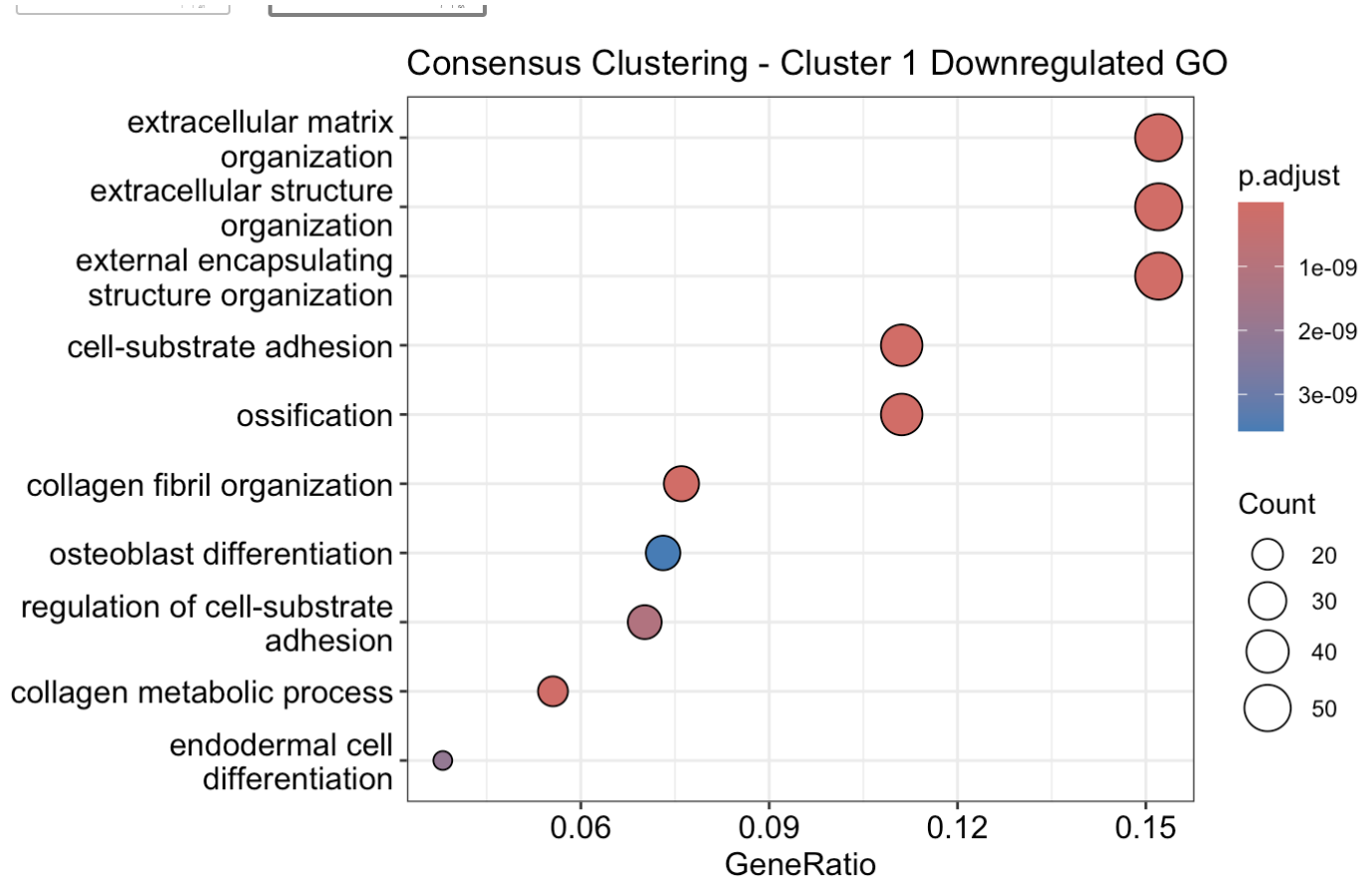
I selected upregulated and downregulated genes based on these two criteria separately

`$logFC > 0 & $adj.P.Val < 0.05`

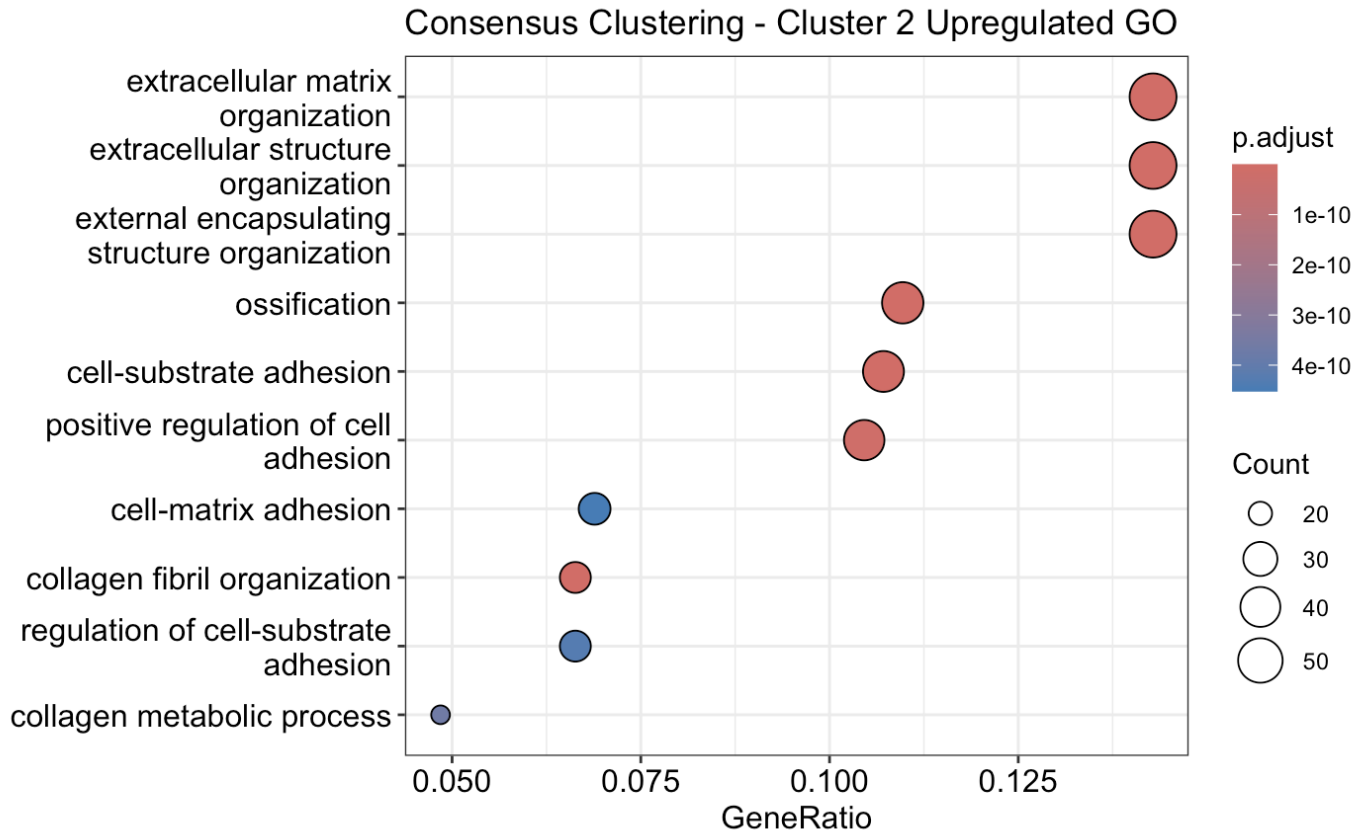
`$logFC < 0 & $adj.P.Val < 0.05`

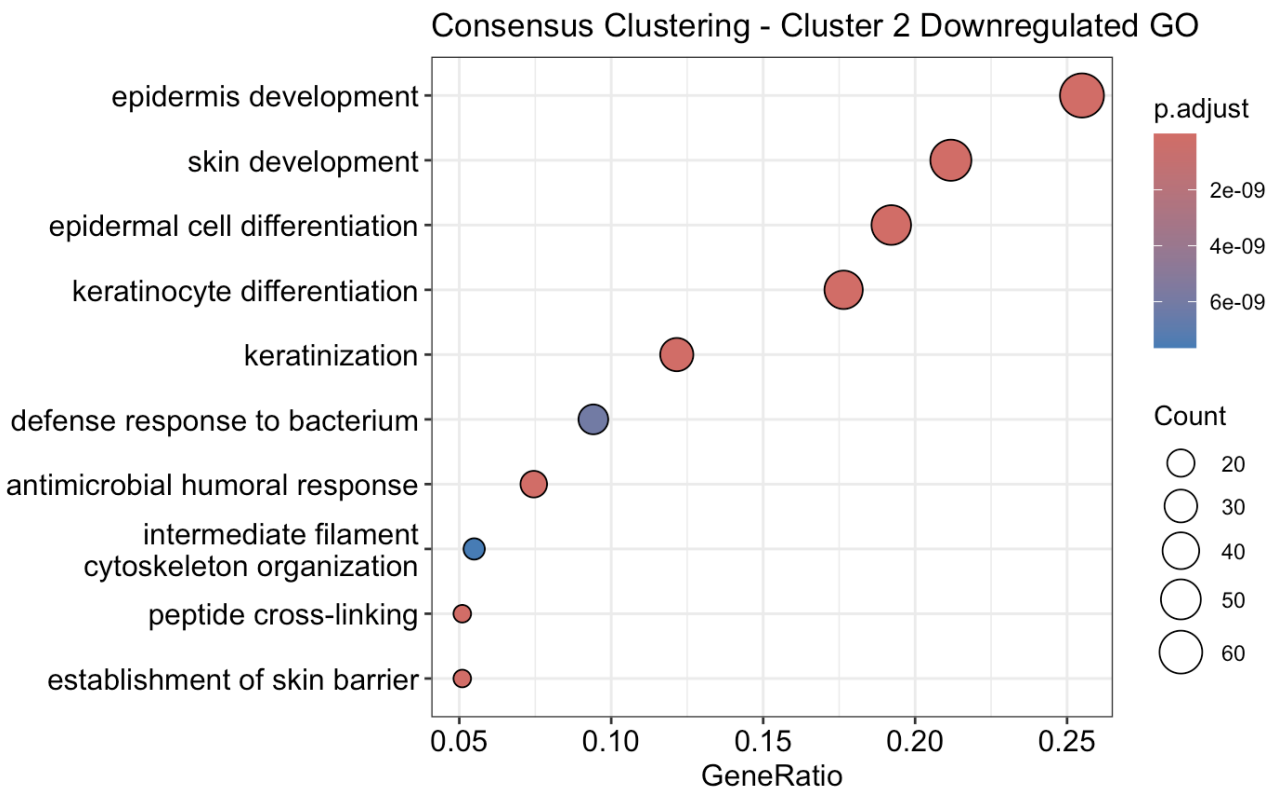
Cluster 1


Consensus Clustering - Cluster 1 Upregulated GO
epidermis development
skin development
epidermal cell differentiation
keratinocyte differentiation
keratinization
defense response to bacterium
humoral immune response
antimicrobial humoral response
peptide cross-linking
antimicrobial humoral immune response mediated by antimicrobial peptide
0.05
0.10
0.15
0.20
GeneRatio
p.adjust
2.5e-07
5.0e-07
7.5e-07
1.0e-06
Count
10
20
30
40

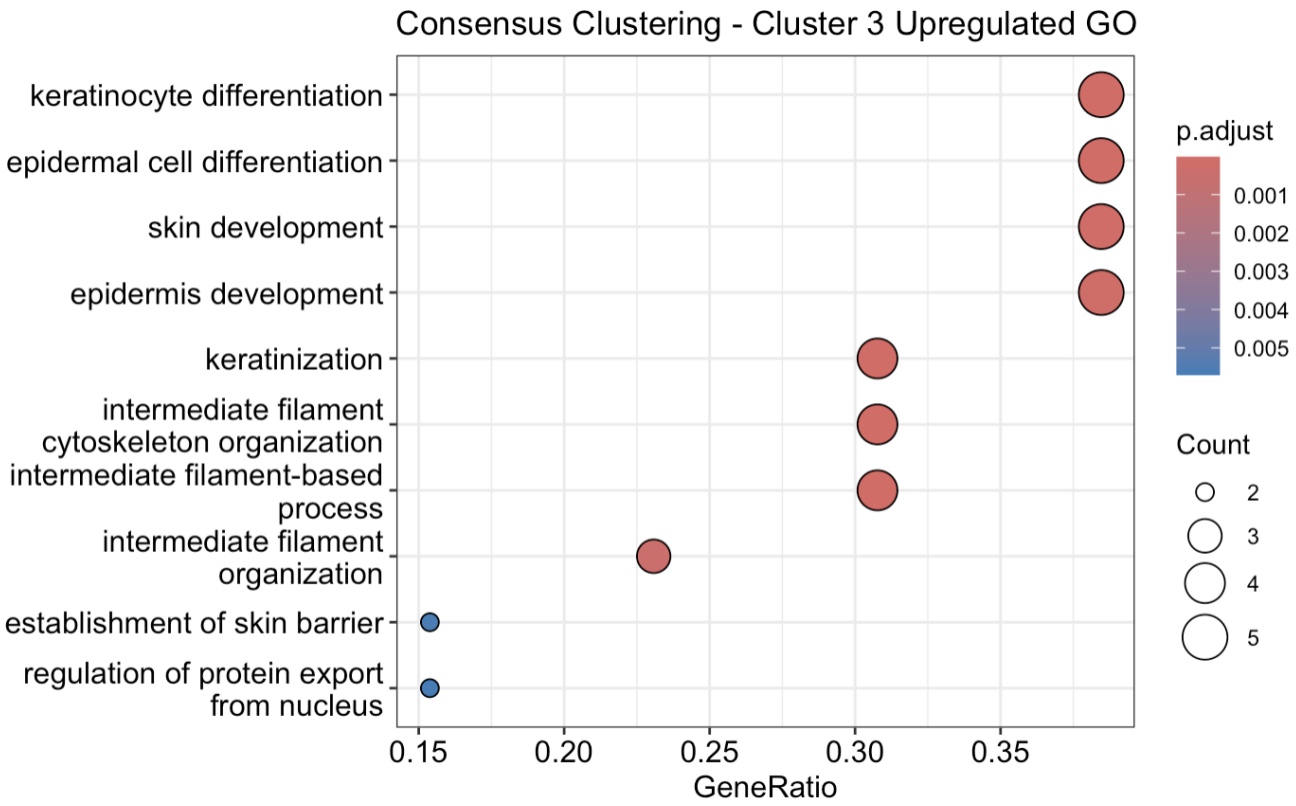

Consensus Clustering - Cluster 1 Downregulated GO
extracellular matrix organization
extracellular structure organization
external encapsulating structure organization
cell-substrate adhesion
ossification
collagen fibril organization
osteoblast differentiation
regulation of cell-substrate adhesion
collagen metabolic process
endodermal cell differentiation
0.06
0.09
0.12
0.15
GeneRatio
p.adjust
1e-09
2e-09
3e-09
Count
20
30
40
50

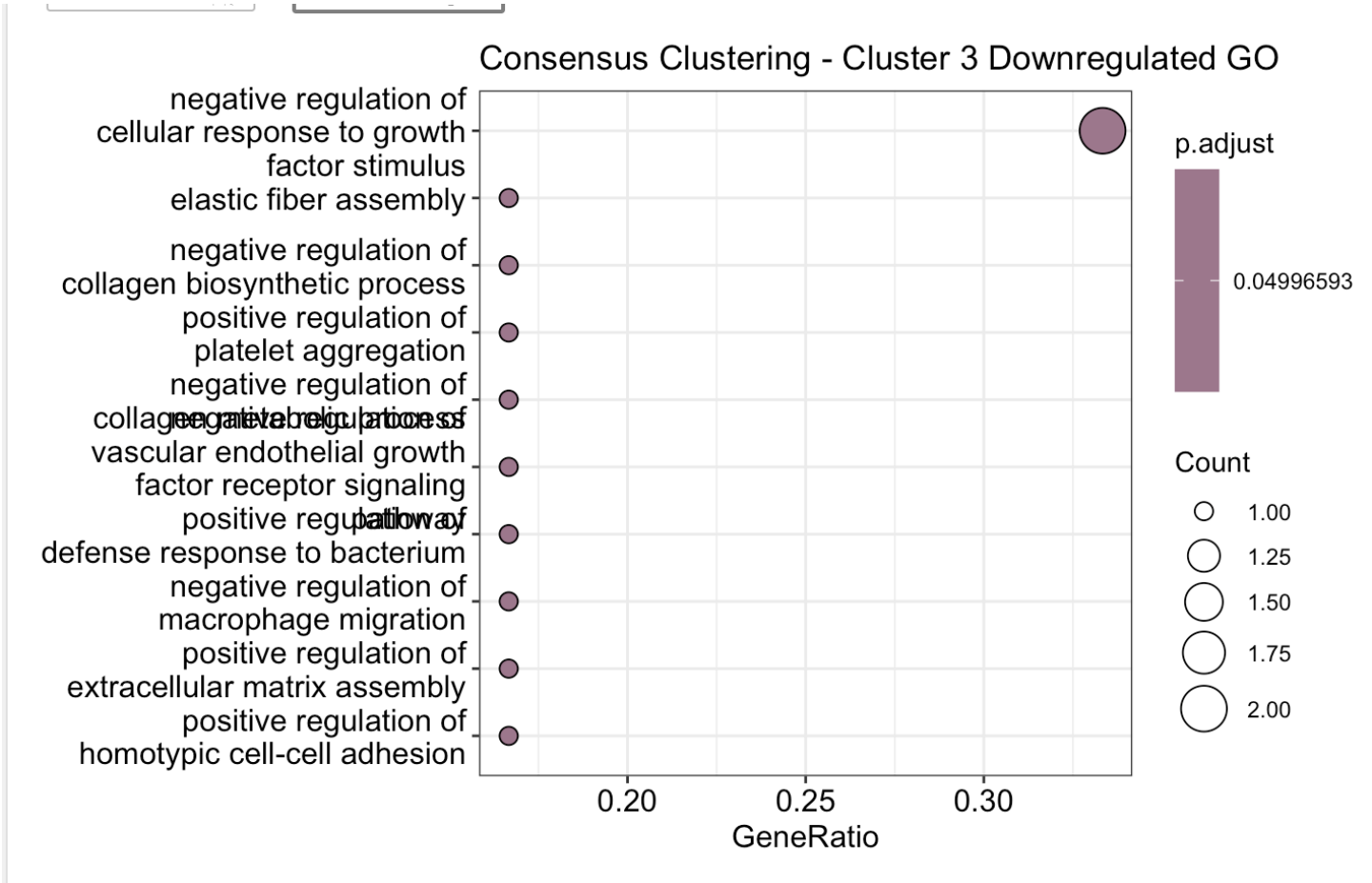
Cluster 2


Consensus Clustering - Cluster 2 Upregulated GO
extracellular matrix organization
extracellular structure organization
external encapsulating structure organization
ossification
cell-substrate adhesion
positive regulation of cell adhesion
cell-matrix adhesion
collagen fibril organization
regulation of cell-substrate adhesion
collagen metabolic process
0.050
0.075
0.100
0.125
GeneRatio
p.adjust
1e-10
2e-10
3e-10
4e-10
Count
20
30
40
50


Consensus Clustering - Cluster 2 Downregulated GO
epidermis development
skin development
epidermal cell differentiation
keratinocyte differentiation
keratinization
defense response to bacterium
antimicrobial humoral response
intermediate filament cytoskeleton organization
peptide cross-linking
establishment of skin barrier
0.05
0.10
0.15
0.20
0.25
GeneRatio
p.adjust
2e-09
4e-09
6e-09
Count
20
30
40
50
60

Cluster 3


Consensus Clustering - Cluster 3 Upregulated GO
keratinocyte differentiation
epidermal cell differentiation
skin development
epidermis development
keratinization
intermediate filament cytoskeleton organization
intermediate filament-based process
intermediate filament organization
establishment of skin barrier
regulation of protein export from nucleus
0.15
0.20
0.25
0.30
0.35
GeneRatio
p.adjust
0.001
0.002
0.003
0.004
0.005
Count
2
3
4
5


Consensus Clustering - Cluster 3 Downregulated GO
negative regulation of cellular response to growth factor stimulus
elastic fiber assembly
negative regulation of collagen biosynthetic process
positive regulation of platelet aggregation
negative regulation of
vascular endothelial growth factor receptor signaling
positive regulation of
defense response to bacterium
negative regulation of macrophage migration
positive regulation of extracellular matrix assembly
positive regulation of homotypic cell-cell adhesion
0.20
0.25
0.30
GeneRatio
p.adjust
0.04996593
Count
1.00
1.25
1.50
1.75
2.00

Cluster 4

```{r}

dotplot(go_up_ccp[[4]], title = "Consensus Clustering - Cluster 4 Upregulated GO")
dotplot(go_down_ccp[[4]], title = "Consensus Clustering - Cluster 4 Downregulated GO")
```

```
Error in ans[ypos] <- rep(yes, length.out = len)[ypos] :
   replacement has length zero
```

Show Traceback

No plot here

conclusion

Cluster	Top 3 Upregulated GO Terms	Top 3 Downregulated GO Terms
Cluster 1	epidermis development skin development epidermal cell differentiation	extracellular matrix organization extracellular structure organization external encapsulating structure organization
Cluster 2	extracellular matrix organization extracellular structure organization external encapsulating structure organization	epidermis development skin development epidermal cell differentiation
Cluster 3	keratinocyte differentiation epidermal cell differentiation skin development	negative regulation of cellular response to growth factor stimulus elastic fiber assembly negative regulation of collagen biosynthetic process
Cluster 4	*No significant GO enrichment*	*No significant GO enrichment*

Q5

up

To compare the clustering results obtained from NMF and Consensus Clustering, we examined both cluster membership and functional enrichment outputs.

The cluster assignments from both methods show considerable overlap. Cross-tabulation of cluster membership revealed near one-to-one correspondence, suggesting high structural consistency across algorithms.

Functionally, GO enrichment analysis of upregulated genes showed strong agreement between corresponding clusters. For example, both methods identified "epidermis development", "skin development", and "epidermal cell differentiation" as the top enriched biological processes in Cluster 1. Similarly, both methods captured extracellular matrix-related signatures in another cluster.

However, discrepancies were observed in Cluster 4 from the Consensus Clustering, which lacked significant upregulated or downregulated GO terms, unlike NMF. This may indicate that Consensus Clustering grouped less coherent or less biologically distinct samples in that cluster.

Overall, the clustering structure and biological interpretations between the two methods were largely consistent, especially for key functional clusters.

down

When comparing the GO enrichment results of downregulated genes across NMF and Consensus Clustering, Clusters 1 and 2 demonstrated strong consistency. Both methods identified processes related to extracellular matrix organization and epidermal development, respectively.

Cluster 3 showed discrepancies: while NMF revealed skin-related pathways, the Consensus Clustering method highlighted immune-related and regulatory processes, possibly reflecting differences in sample partitioning.

Cluster 4 from the Consensus Clustering lacked significantly enriched downregulated pathways, whereas NMF detected nervous system development-related terms. This indicates that NMF may

better capture certain biologically relevant subgroups in this case.

Overall, the GO enrichment results from both methods are largely consistent for the major clusters, with minor differences in specific subgroups.

Q6

To generate the most robust and reliable clusters from gene expression data in an unsupervised setting, the following procedure is recommended:

1. **Feature Selection**: Start with a variance-based filtering strategy, such as MAD (median absolute deviation), to select the most informative genes (e.g., top 1500–3000 genes) and reduce noise.

2. **Apply Multiple Clustering Algorithms**: Use at least two complementary methods, such as:
   - **Non-negative Matrix Factorization (NMF)**: Captures parts-based representation and is particularly suitable for gene expression data.
   - **Consensus Clustering**: Evaluates the stability of clustering results through resampling, useful for determining robust cluster membership.

3. **Cluster Number Estimation**:
   - Use internal metrics (e.g., cophenetic coefficient, silhouette width, residuals) to determine the optimal number of clusters.
   - Avoid choosing `k = 2` unless biologically justified, as it often oversimplifies the heterogeneity.

4. **Validation**:
   - Perform **biological pathway enrichment (GO/KEGG)** to interpret each cluster.
   - Assess whether clusters are biologically meaningful and non-overlapping in terms of functional enrichment.

5. **Cross-Comparison**:

- Compare results from multiple clustering methods.
- Use consensus or overlapping clusters as a confidence measure for stability.

6. **Visualization and Interpretability**:
   - Visualize consensus matrices, silhouette plots, and enrichment bubble plots.
   - Clear visualization helps reveal outliers and cluster boundary ambiguity.

7. **Biological Input and Reproducibility**:
   - Consult domain knowledge for validating clusters.
   - Ensure reproducibility via random seeds, transparent parameter reporting, and standard pipelines.

This integrative and comparative strategy helps identify **robust, reproducible, and biologically relevant clusters** rather than relying on a single method or metric.