Project 6

Tags

The work was performed by Maria Uzun and Alisa Fedorenko.

1. Input data

yeast reads:

SRR941816: fermentation 0 minutes replicate 1

ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941816/SRR941816.fastq.gz (413 Mb)

SRR941817: fermentation 0 minutes replicate 2

ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941817/SRR941817.fastq.gz (455 Mb)

SRR941818: fermentation 30 minutes replicate 1 ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941818/SRR941818.fastq.gz (79.3 Mb)

SRR941819: fermentation 30 minutes replicate 2 ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941819/SRR941819.fastq.gz (282 Mb)

As a reference genome we will use *Saccharomyces cerevisiae*, in the genome database at NCBI. Make sure you have strain S288c and assembly R64.

reference genome file:

ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.fna.gz

annotation file:

ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.gff.gz

Fasta cDNA dumps:

https://ftp.ensembl.org/pub/release-111/fasta/saccharomyces_cerevisiae/cdna/Saccharomyces_cerevisiae.R64-1-1.cdna.all.fa.gz

The data were downloaded by SRA-tools:

```
fastq-dump --gzip SRR941816 SRR941817 SRR941818 SRR941819
```

```
# cDNA file
wget https://ftp.ensembl.org/pub/release-111/fasta/saccharomyces_cerevisiae/cdna/Saccharomyces_cerevisiae.R64-1-1.cdna.all.fa.gz

# annotation file
wget http://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.gff.gz
```

2. Data preprocessing

a) Salmon

Because we will explore how RNA expression levels we decided to use Salmon pipeline to work with RNA-reads. We have 4 single-end reads with good quality control.

Firstly, we suggest to prepare file environment.yml

```
nano environment.yml
```

```
name: salmon
channels:
  - bioconda
  - conda-forge
  - defaults
dependencies:
  - libgcc-ng=9.1.0=hdf63c60_0
  - tbb=2020.2=hc9558a2_0
  - salmon=1.4.0=hf69c8f4_0
```

```
# to install and activate salmon environment
conda env update --file environment.yml
conda activate salmon
```

Then, we prepared the transcriptome index:

```
salmon index -t Saccharomyces_cerevisiae.R64-1-1.cdna.all.fa.gz -i salmon_index
```

Prepare the list of our sample names as a samples.txt file

```
ls -1 /path/to/your/folder/*.fastq.gz | sed 's/^\.\///; s/\.fastq\.gz$//' > samples.txt
```

Let’s launch our Salmon pipeline based on *pseudoalignment* algorithm:

```
#!/usr/bin/env bash
"""
Salmon_mapping.sh script processes single-end RNA-seq reads using fastp for quality control a
nd filtering,
and then maps the filtered reads to a transcriptome using Salmon quant.

Usage:
  ./Salmon_mapping.sh <samples.txt>

Arguments:
  <samples.txt>    A text file containing the list of sample names.

Each sample name in the <samples.txt> file should correspond to the base name of the FASTQ fi
le (without the .fastq.gz extension).

Example:
  ./Salmon_mapping.sh samples.txt
"""

# Полный путь к директории с файлом samples.txt
SAMPLES_DIR="/path/to/your/samples/directory"

while read SAMPLE; do
        echo "Running sample ${SAMPLE}"

        FASTQ="${SAMPLES_DIR}/${SAMPLE}.fastq.gz"    # полный путь к FASTQ файлу

        ########################################
```

```
        ##     Map samples with Salmon     ##
        #########################################

        SALMON_TRANSCRIPTOME_INDEX_DIR="/path/to/salmon/transcriptome/index/dir" # <- add dir
ectory with transcriptome index
        SALMON_OUT_DIR="/path/to/salmon/output/dir/${SAMPLE}"                # <- add direct
ory to store salmon output

        salmon quant -i ${SALMON_TRANSCRIPTOME_INDEX_DIR} \
                -l A \
                -r ${FASTQ} \
                -p 2 \
                -o ${SALMON_OUT_DIR} \
                --useVBOpt \
                --seqBias \
                --validateMappings

done < "${SAMPLES_DIR}/samples.txt"
```

More than 80% of the analyzed reads of each sample were mapped to the reference transcriptome.

As the output we should get 4 quant.sf files with counts. These files we will take for deseq2 analysis.

b) Find differentially expressed genes with Deseq2

Deseq2 provides analysis of count data from RNA-seq is the detection of differentially expressed genes. We performed this analysis in R. To find more information about deseq2 follow the link.

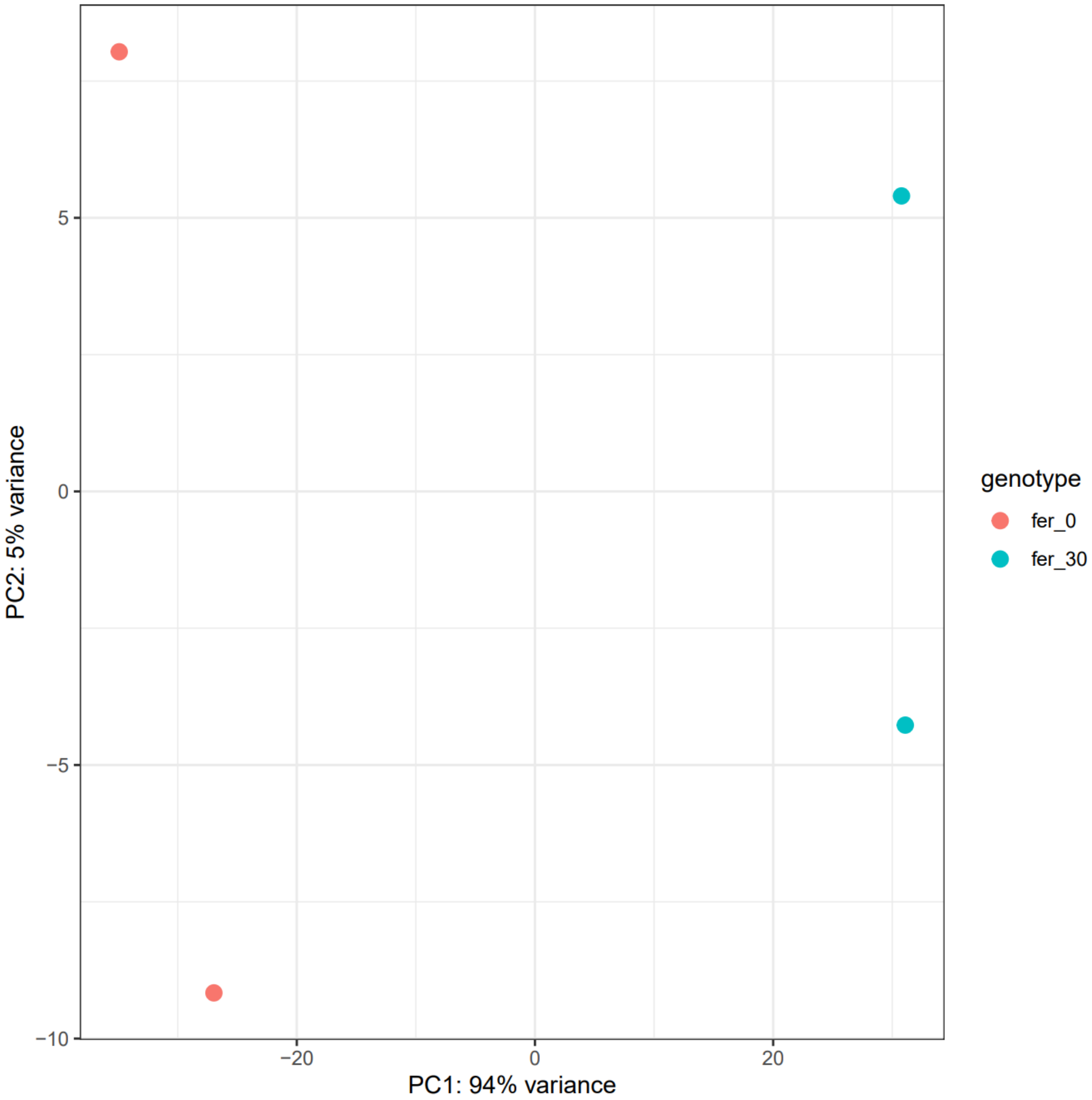
Steps of deseq2 analysis:

1. Annotation (Return the Ensembl EnsDb information for Saccharomyces cerevisiae)
2. Tximport (to import transcript abundances and construct a gene-level DESeqDataSet object from Salmon quant.sf)
3. Count normalization
4. Quality control
5. Statistical testing (counting up- and down-regulated genes)
6. Visualizing the DE analysis results (heatmap, volcano plot)
7. Functional analysis (GO)

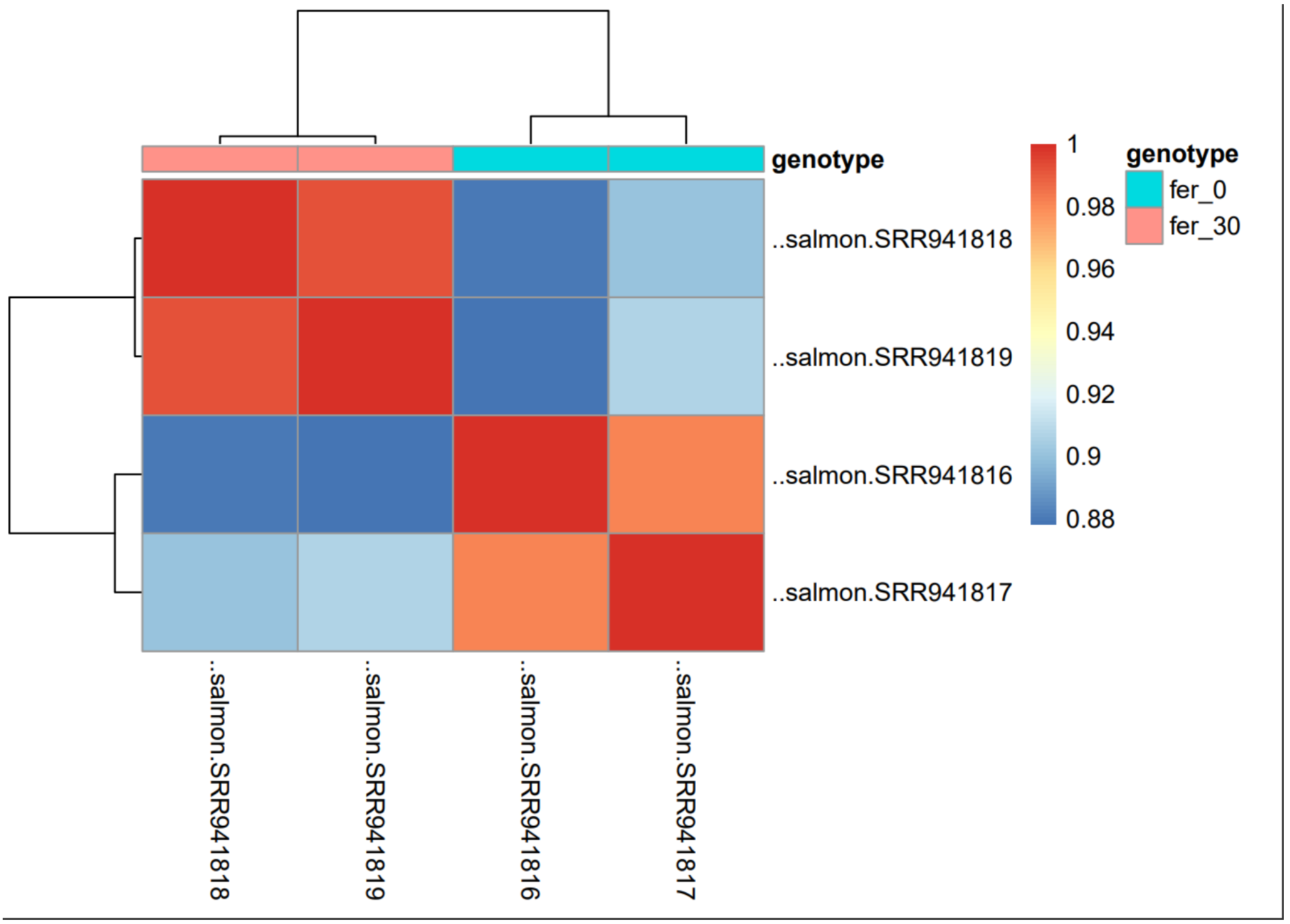
Quality control

Principal component analysis (PCA) can be used to visualize variation between expression analysis samples. As we can see that samples were produced under two experimental conditions (e.g. 0 minutes of fermentation vs. 30 minutes of fermentation). They are clustered together.

Plot PCA graph:



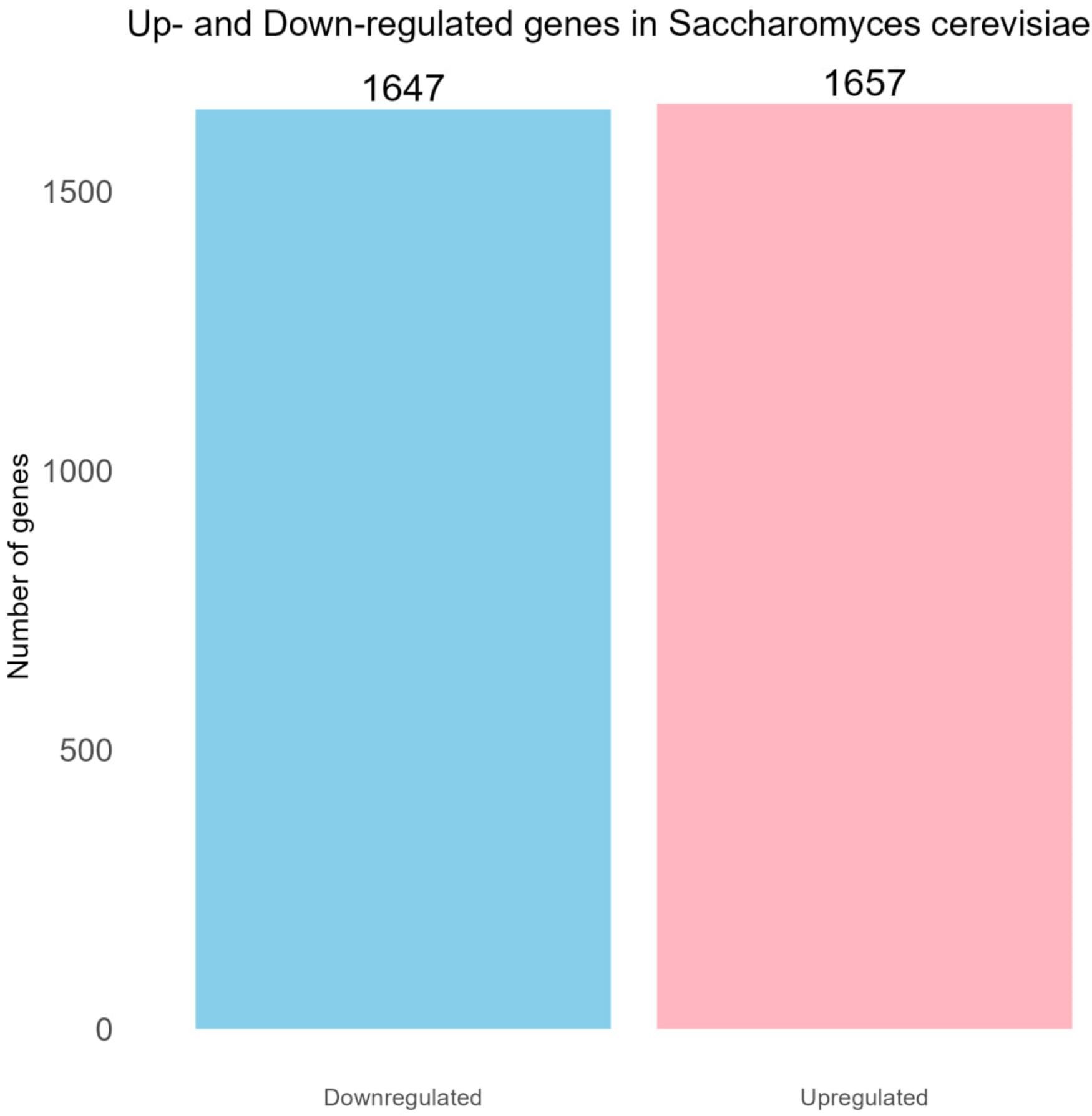
Samples Correlation Heatmap:



Statistical testing

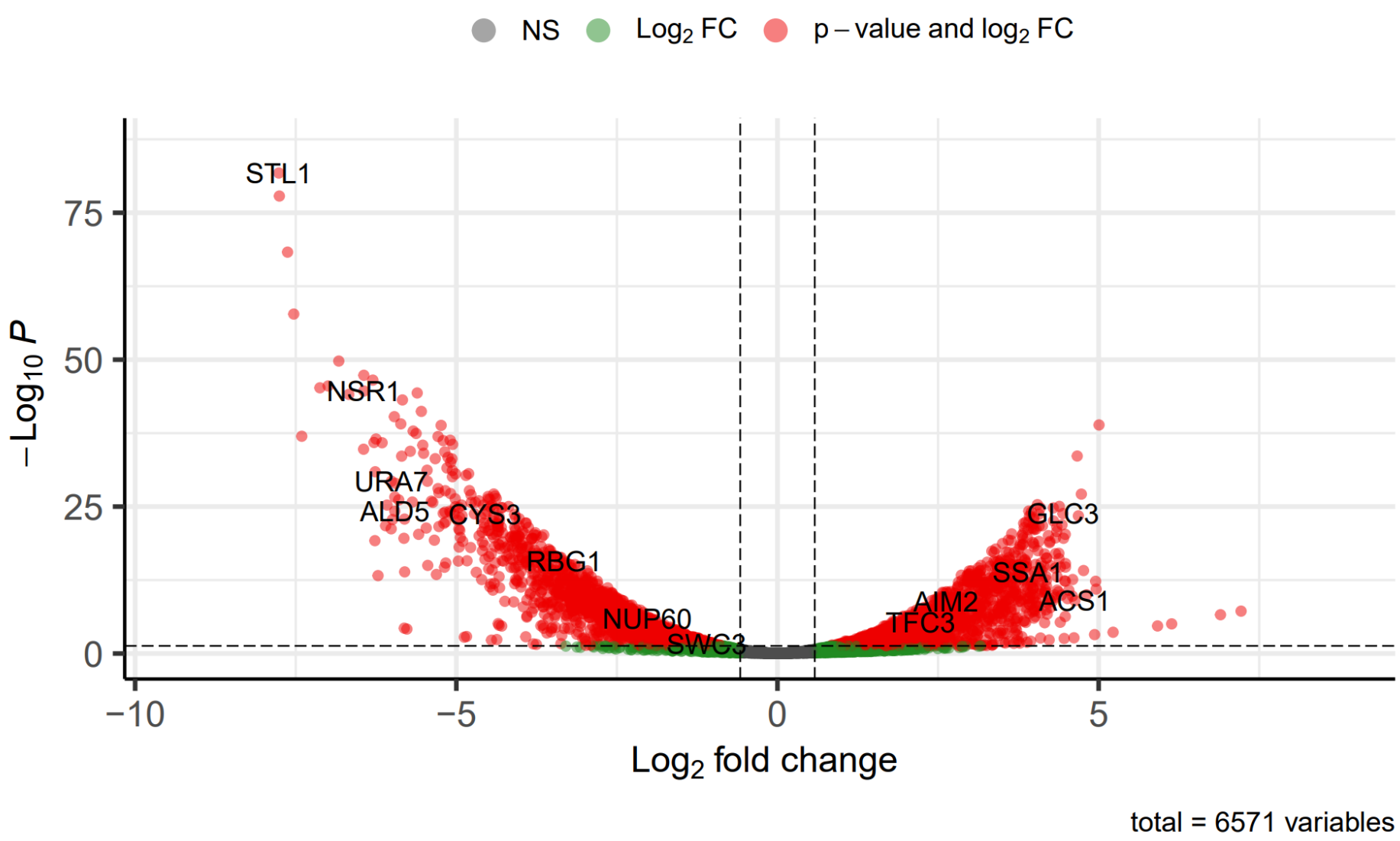
We choose adjusted p-value < 0.05 (padj.cutoff ← 0.05) and log2 fold changes so this translates to an actual fold change of 1.5 (lfc.cutoff ← 0.58) to select up- and down-regulated genes.

```
out of 6443 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 1657, 26%
LFC < 0 (down)     : 1647, 26%
outliers [1]       : 0, 0%
low counts [2]     : 125, 1.9%
(mean count < 1)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

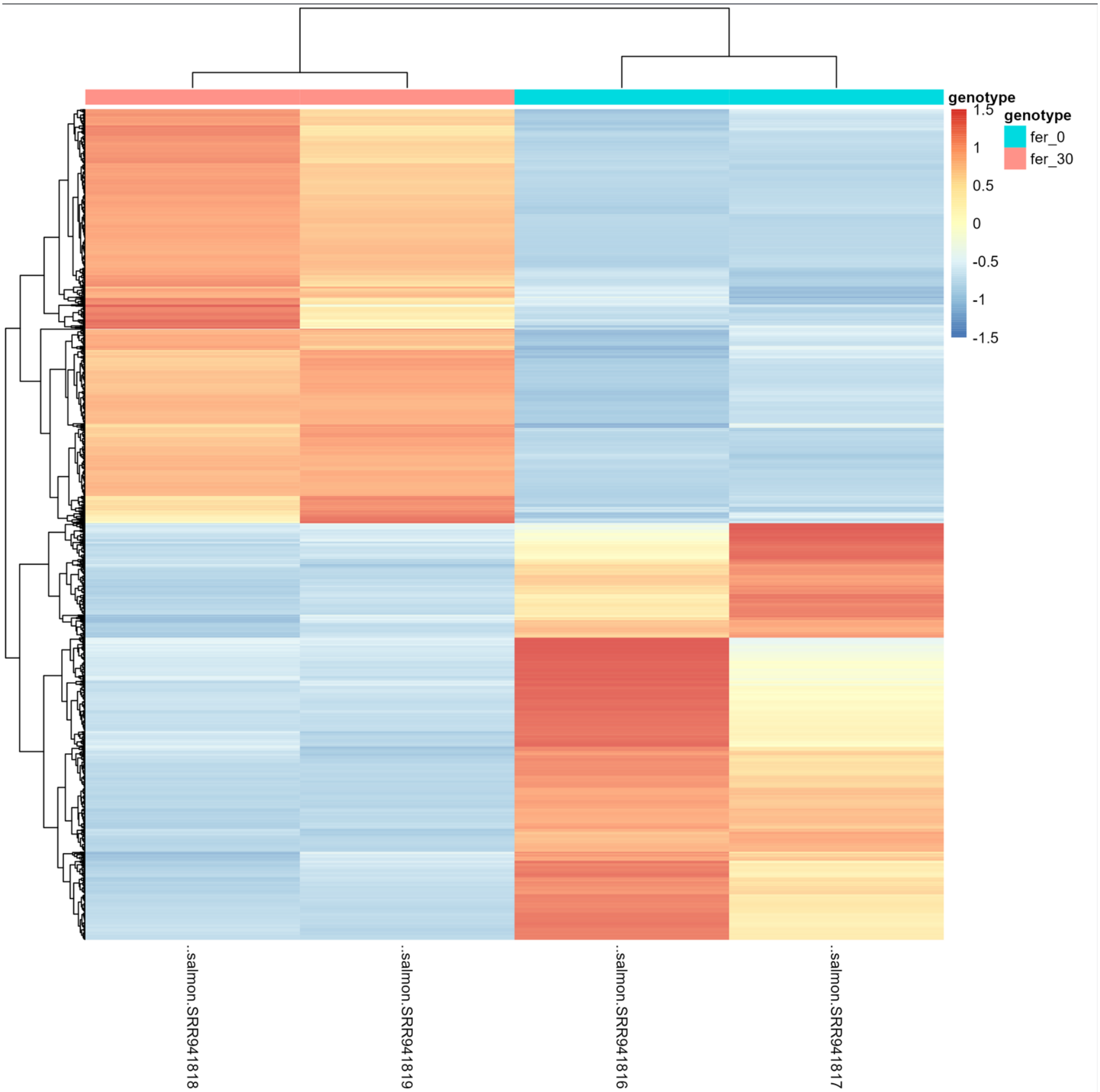


Visualizing the DE analysis results

Volcano plot is the graphical representation of a differental expression analysis. Volcano plots indicate the fold change (either positive or negative) in the x axis and a significance value (such as the p-value or the adjusted p-value, i.e. FDR) in the y axis. Points represent individual genes.



Heatmap generated by DESeq summarizing differential expression of the 3304 significantly DE transcripts, at a log-fold change of 1.5 or higher and an FDR of 0.05.



Functional analysis with clusterProfiler

To perform the over-representation analysis, we need a list of background genes and a list of significant genes. For our background dataset we will use all genes tested for differential expression (all genes in our results table). For our significant gene list we will use genes with p-adjusted values less than 0.05 (we could include a fold change threshold too if we have many DE genes).

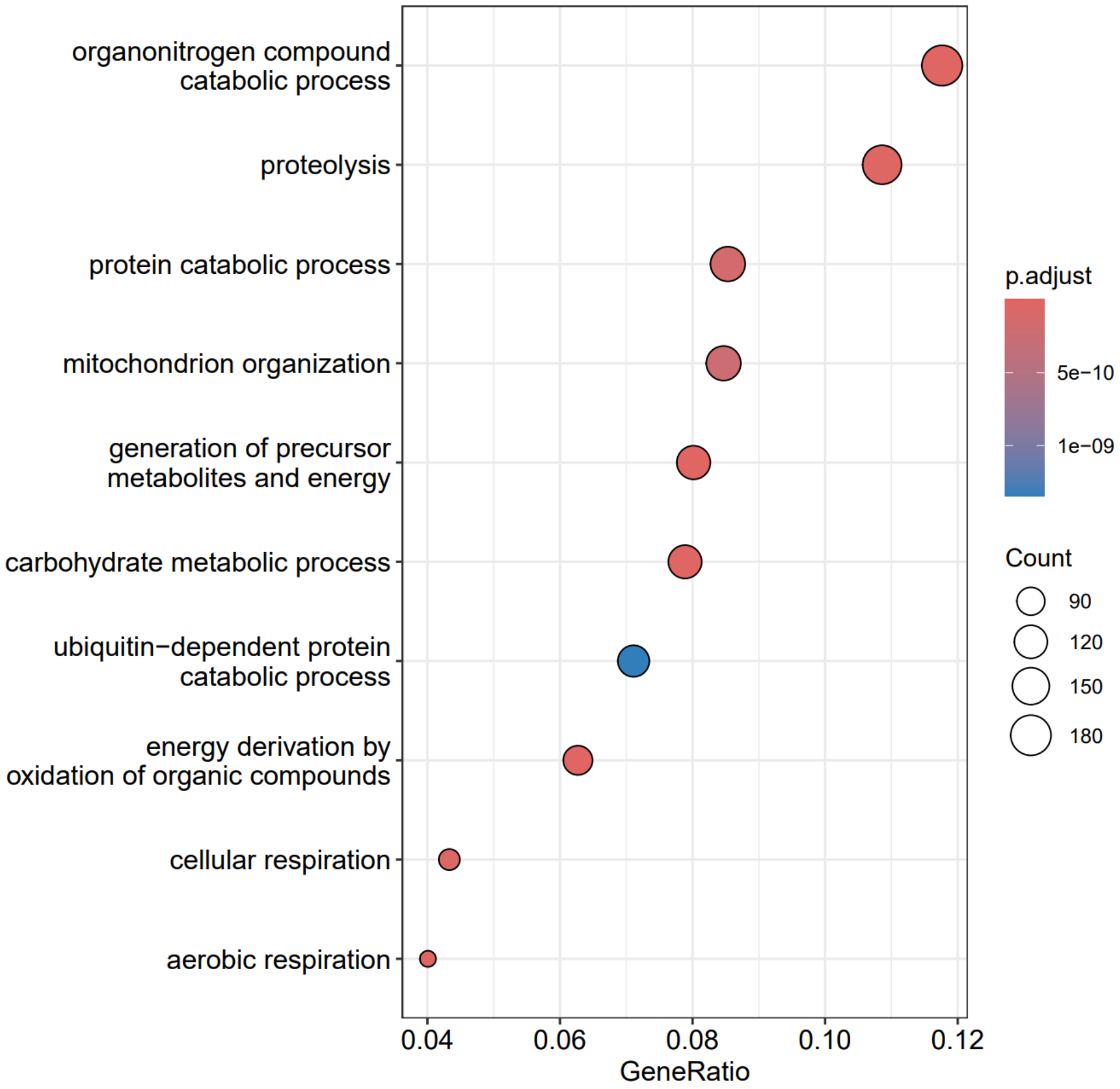
There are 3 types of terms in the gene ontology:

Biological Processes (BP)
Molecular Functions (MF)
Cellular Components (CC)

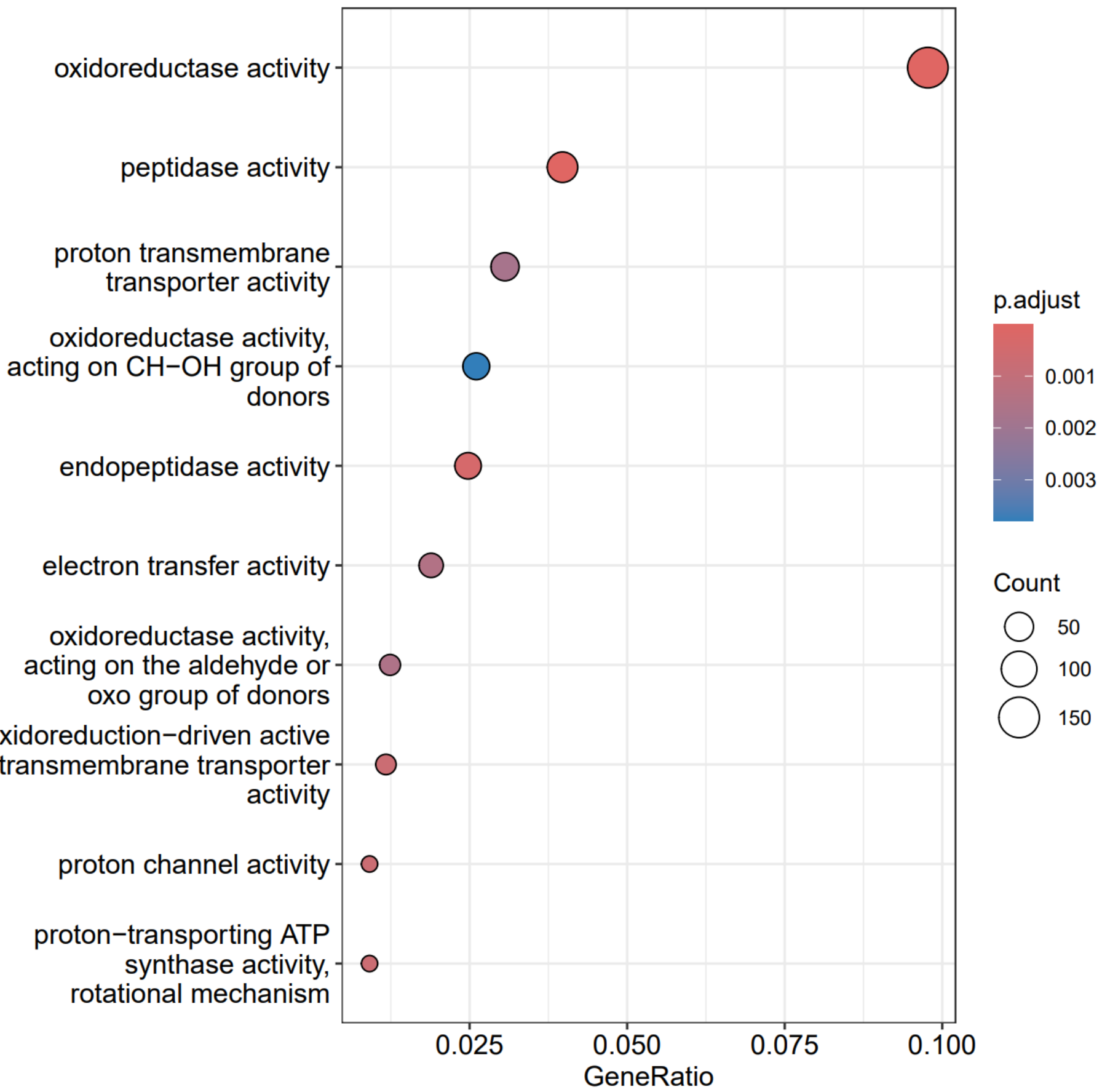
We separated up- and down-regulated genes to see precisely this categories.

For up-regulated genes:

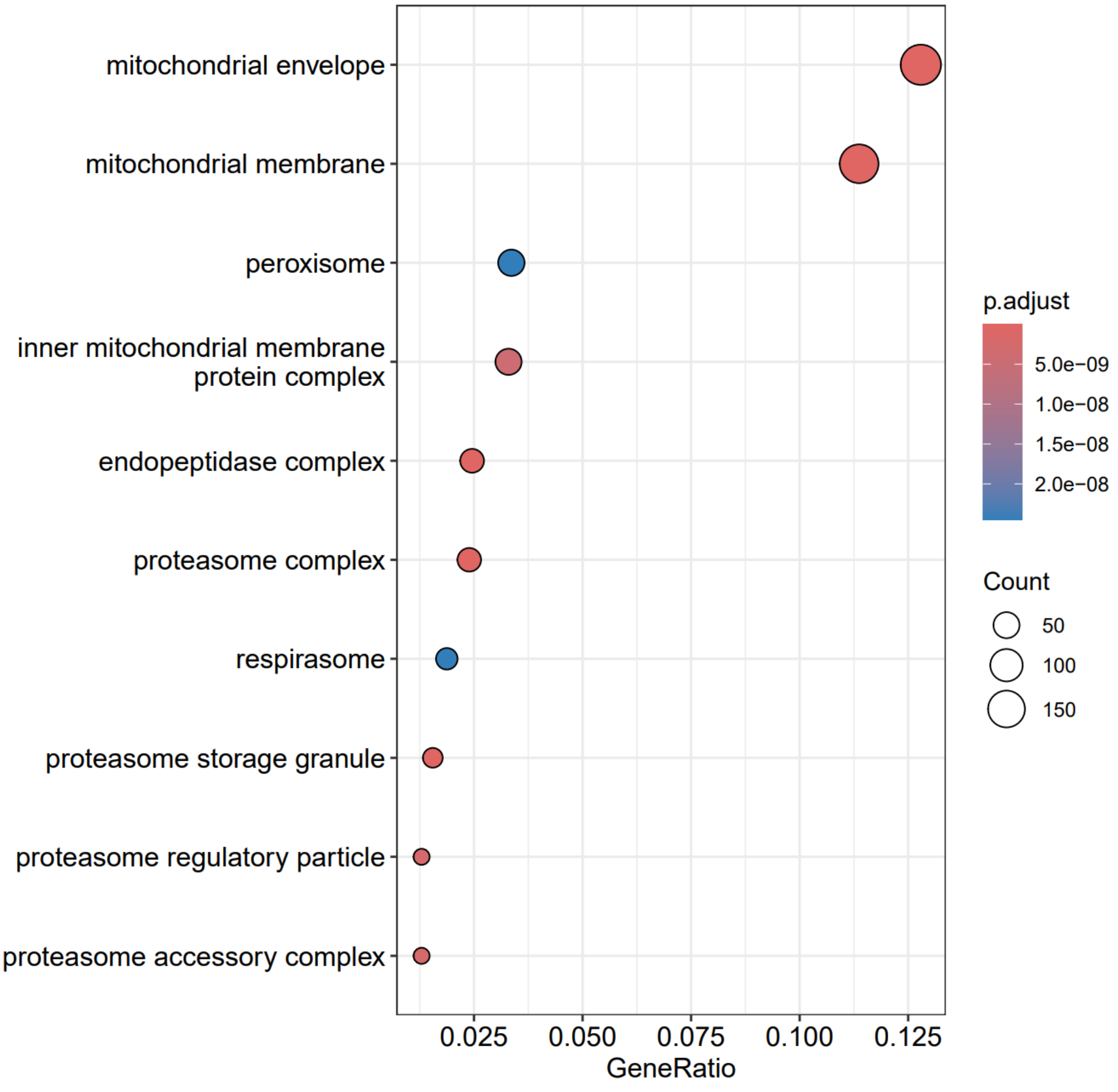
BP:



MF:

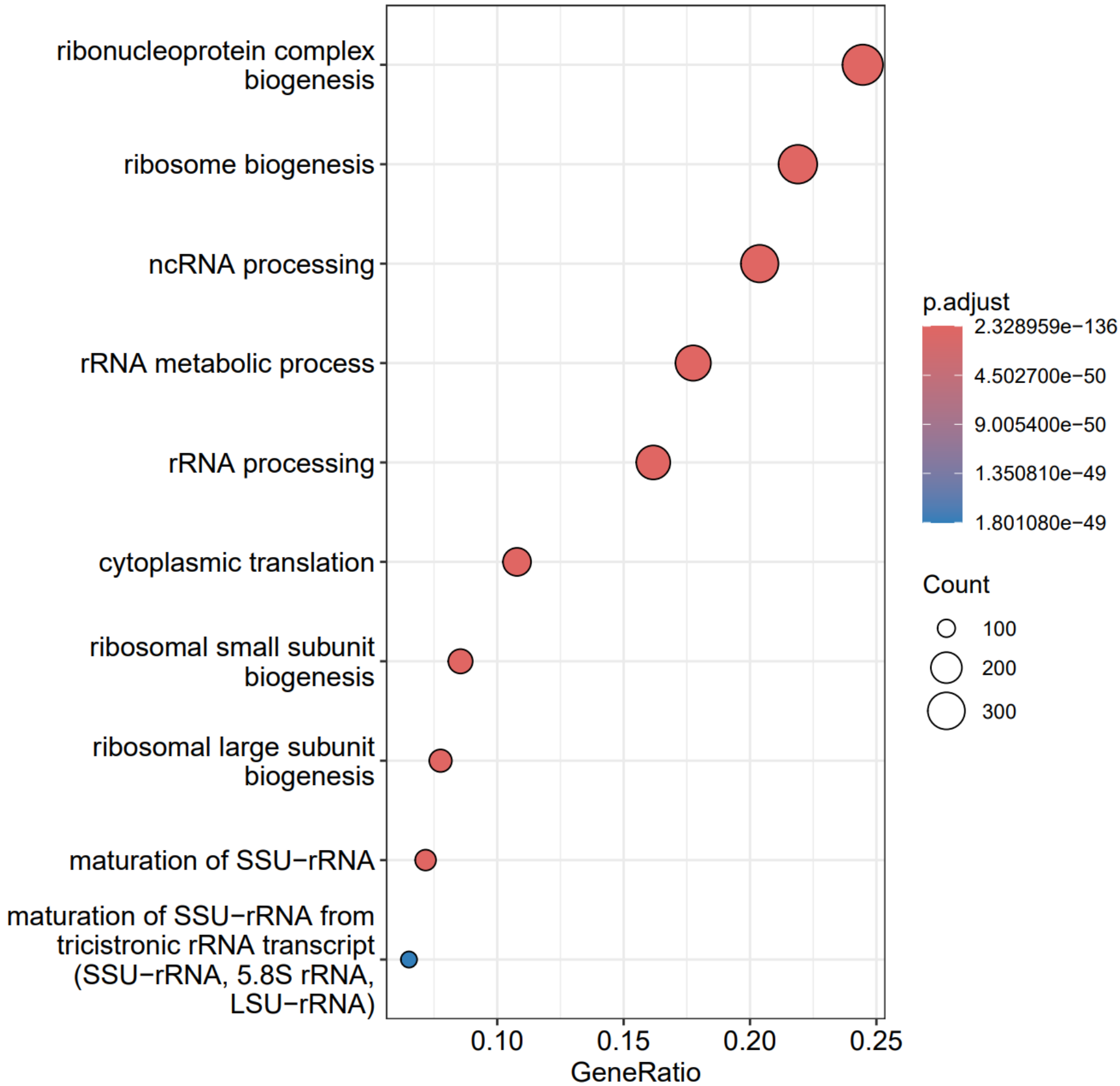

oxidoreductase activity
peptidase activity
proton transmembrane transporter activity
oxidoreductase activity, acting on CH−OH group of donors
endopeptidase activity
electron transfer activity
oxidoreductase activity, acting on the aldehyde or oxo group of donors
oxidoreduction−driven active transmembrane transporter activity
proton channel activity
proton−transporting ATP synthase activity, rotational mechanism
0.025
0.050
0.075
0.100
GeneRatio
p.adjust
0.001
0.002
0.003
Count
50
100
150

CC:

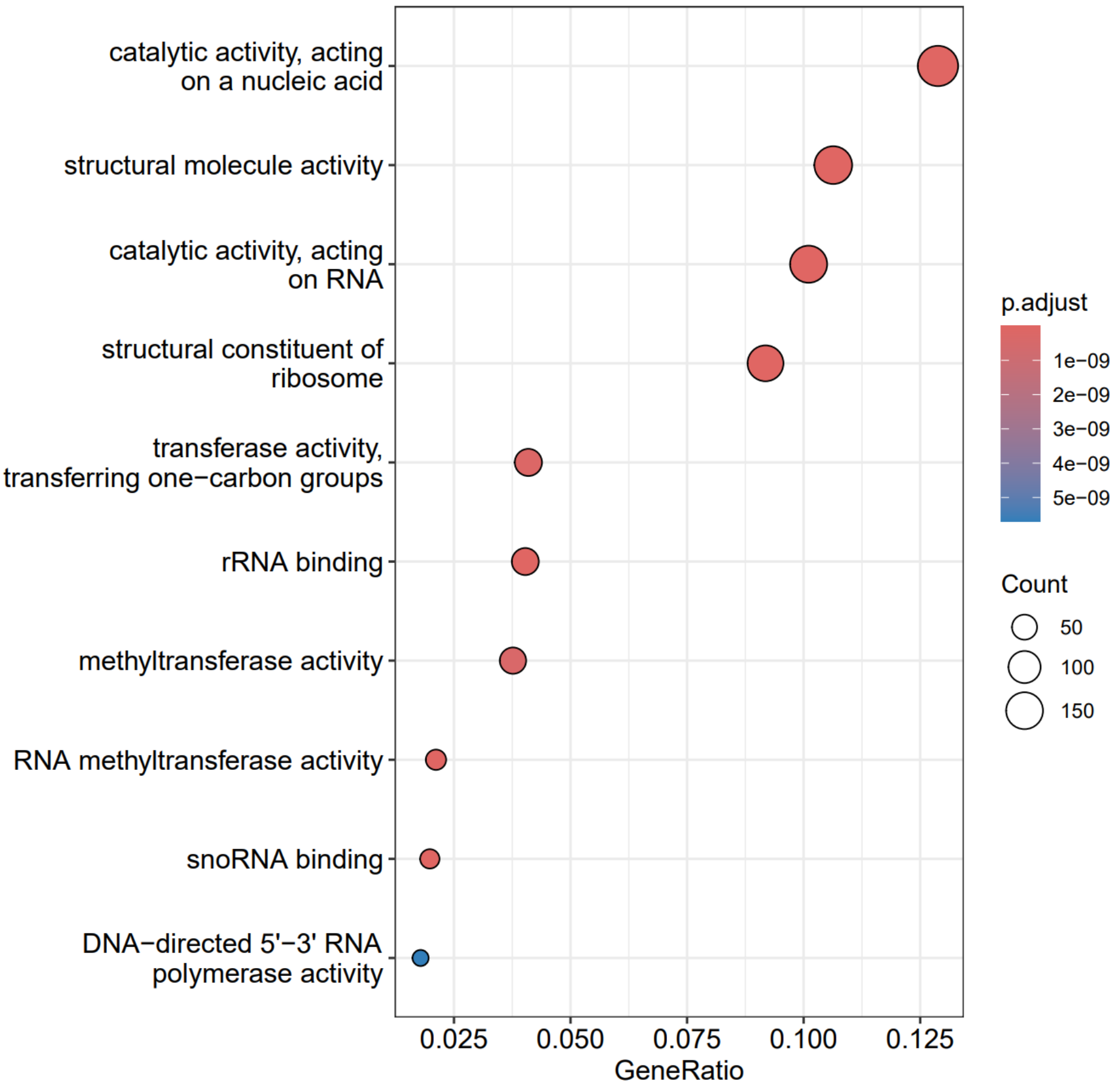


For down-regulated genes:

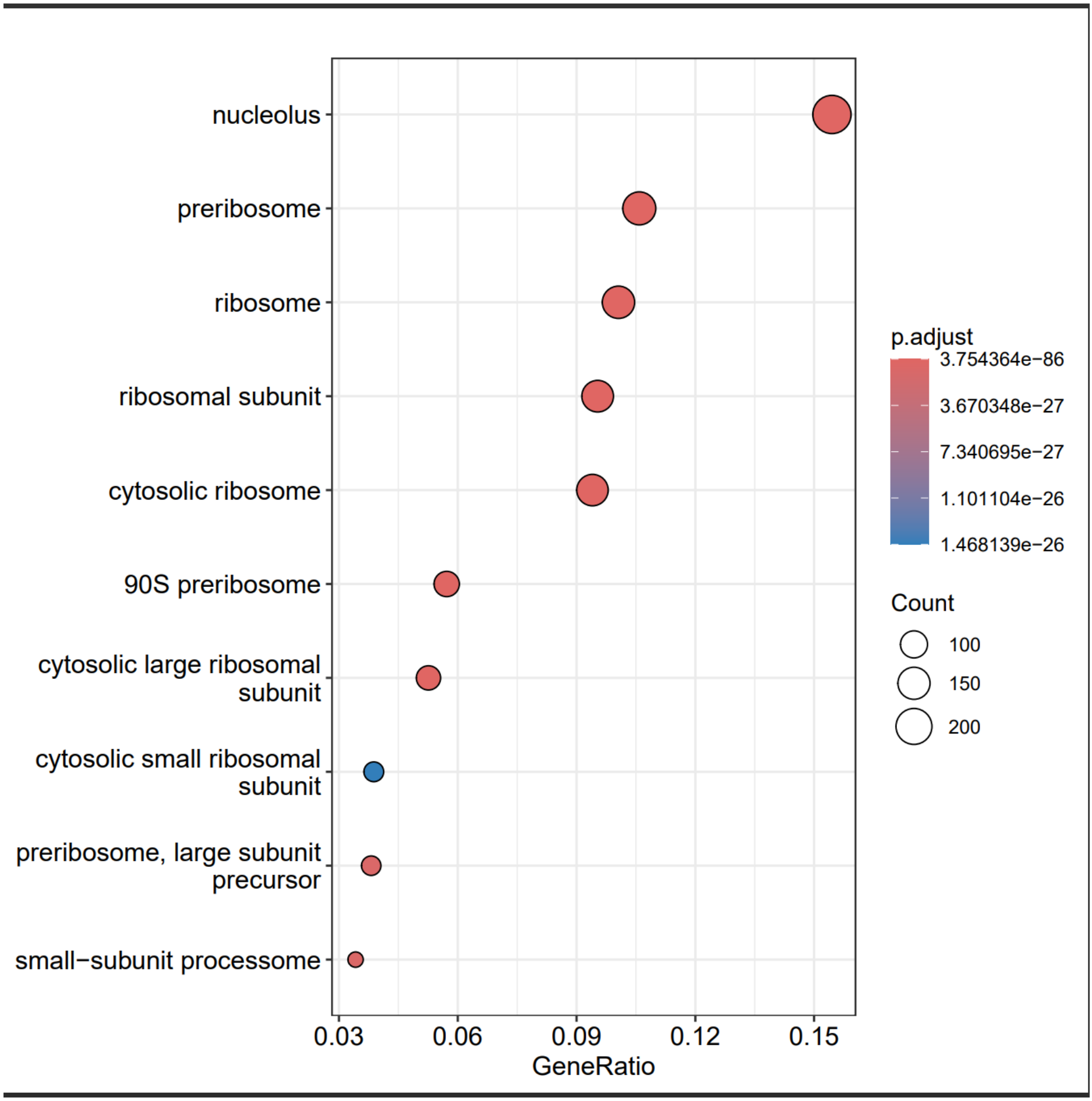
BP:


ribonucleoprotein complex biogenesis
ribosome biogenesis
ncRNA processing
rRNA metabolic process
rRNA processing
cytoplasmic translation
ribosomal small subunit biogenesis
ribosomal large subunit biogenesis
maturation of SSU−rRNA
maturation of SSU−rRNA from tricistronic rRNA transcript (SSU−rRNA, 5.8S rRNA, LSU−rRNA)
0.10
0.15
0.20
0.25
GeneRatio
p.adjust
2.328959e−136
4.502700e−50
9.005400e−50
1.350810e−49
1.801080e−49
Count
100
200
300

MF:


catalytic activity, acting on a nucleic acid
structural molecule activity
catalytic activity, acting on RNA
structural constituent of ribosome
transferase activity, transferring one−carbon groups
rRNA binding
methyltransferase activity
RNA methyltransferase activity
snoRNA binding
DNA−directed 5'−3' RNA polymerase activity
0.025
0.050
0.075
0.100
0.125
GeneRatio
p.adjust
1e−09
2e−09
3e−09
4e−09
5e−09
Count
50
100
150

CC:


nucleolus
preribosome
ribosome
ribosomal subunit
cytosolic ribosome
90S preribosome
cytosolic large ribosomal subunit
cytosolic small ribosomal subunit
preribosome, large subunit precursor
small−subunit processome
0.03
0.06
0.09
0.12
0.15
GeneRatio
p.adjust
3.754364e−86
3.670348e−27
7.340695e−27
1.101104e−26
1.468139e−26
Count
100
150
200