# Gene Set Analysis (GSA)

RNA-Seq group, IFB-AVIESAN 2020

2020-10-06

# Purpose of this session

At the end of this session, you will be able to:

1. understand what is gene set analysis
2. identify standard databases
3. tell the differences between GeneID and GeneName

# Bonus

You will also:

- know limitations of gene sets analyses
- question confidence intervals on pathways enrichments

# Take-home methods

You will have protocols to:

- question standard databases such as GO, or KEGG
- perform comparison with gene sets from custom sources

# Definitions

# Gene Set Analysis (GSA)

Unlike differential gene analysis, we do not want to produce results about *single genes*, but rather about **gene sets**.

These networks can be multiple:

# Gene Set Analysis (GSA)

> *" we want to identify pathways (i.e. gene-sets) that are significantly enriched in differentially expressed genes with respect to the background set of genes."*

Example:

In our previous experiment on *A. Thaliana*, we saw that, once treated, the plants did not grow as much as wild type.

Our questions could be: *"Is plant organ morphogenesis involved in the plant growth? What other pathways are impacted by the treatment?"*

# Databases

# Genes annotations: database expectations (1/2)

- **Gene Ontology (GO)**: which hosts a controlled vocabulary (fixed terms) for annotating genes
  - *Molecular Functions*: Molecular-level activities performed by gene products
  - *Cellular Components*: Locations relative to cell compartments and structures
  - *Biological Process*: Larger processes accomplished by multiple molecular activities

`http://geneontology.org/`

# Genes annotations: database expectations (2/2)

- **KEGG**: Kyoto Encyclopedia of Genes and Genomes
  - *Pathways*: Larger processes accomplished by multiple molecular activities
  - ...

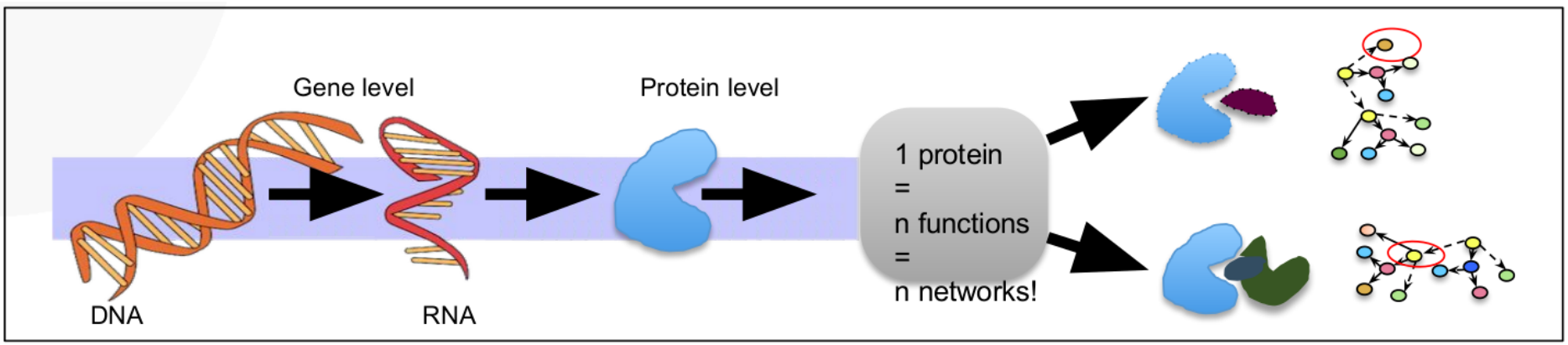
  *https://www.genome.jp/*

- **MSigDB**: Molecular Signatures Database
  - Multiple collections of genes sets (human centered)

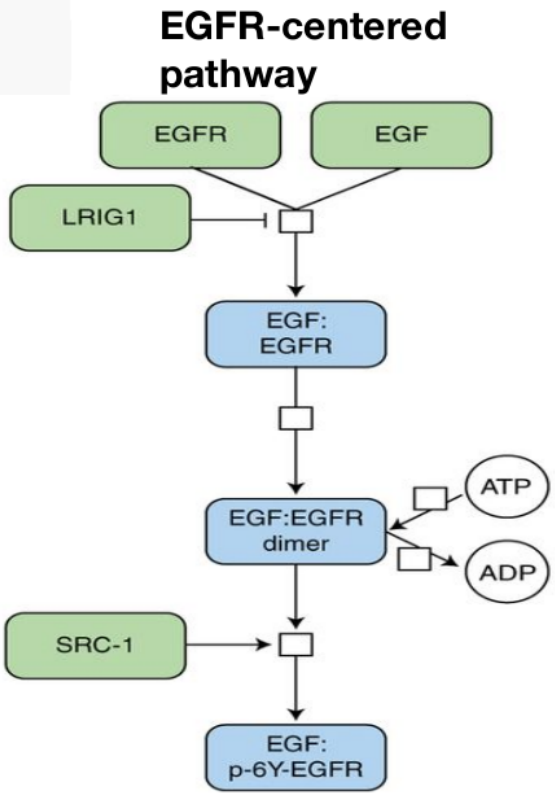
  *http://software.broadinstitute.org/gsea/msigdb/index.jsp*

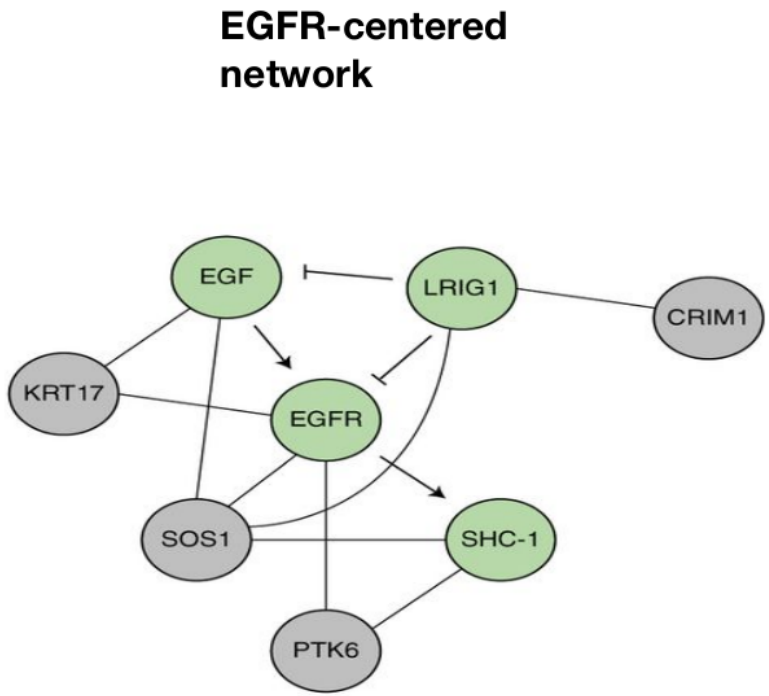
# Protein - Protein Interactions (PPIs)

PPIs are useful for understanding functional relationships between proteins and the biology of the cell



# Pathways vs Network




Adapted from: **Nature Methods**. Pathway and network analysis of cancer genomes (2015)

Institut Pasteur

# Preparation

# Load libraries

To load libraries, we use the function "library", from the package "base". This function takes only one argument: the name of the package.

```
library("clusterProfiler");
library("limma");
library("DOSE");
library("enrichplot");
library("pathview");
library("org.At.tair.db");
```

It is a good practice to keep package names in order to avoid name-space collision. However, during this presentation, for the sake of readability, we shall omit them.

# Load SARTools results

We use the function "read.table" from the package "utils". This is installed by default in your R environments.

```
de_res <- read.table(
  file   = "tables/KOvsWT.complete.txt",   # Path to results
  header = TRUE,                # There are column names
  sep    = "\t",                # This is a tabulation
  stringsAsFactors = FALSE # Colnames are not factors
);
```

# Visualize SARTools' results

Our table is pretty large, and we do not need all of it. With the generic function "head", we can see the first lines of our table.

```
head(de_res);
```

With the generic function "colnames", we can have ... The column names!

```
colnames(de_res);
```

# Select columns from SARTools' results

We want to keep the following columns **only**: `"Id"`, `"padj"`, `"log2FoldChange"`.

```
# Keep selected columns
de_res <- de_res[ ,c("Id", "padj", "log2FoldChange")];

# Check dimensions
dim(de_res);
```

```
[1] 27655     3
```

# Filter SARTools' results

And now, we want to filter out the values with a `padj` above 0.001.

```
# List all values that are to be kept
threshold <- de_res[, "padj"] <= 0.001;

# Keep only non-null and significant values
de_res <- de_res[which(threshold), ];

# Check dimensions
dim(de_res);
```

```
[1] 1807    3
```

# Sort SARTools results

The order of the genes within a list of genes is **very** important. We need to sort our genes. Here, we sort our differentially expressed genes by their `log2FoldChange` value:

```
ordered_lines_number <- order(
  de_res$log2FoldChange,      # Select the column
  decreasing = TRUE             # Decreasing sort
);

de_res <- de_res[ordered_lines_number, ];
```

# Genes identifiers

# Fix gene identifiers

In our table, the genes identifiers begin with "gene:". This going to break further analysis! For a computer: "`gene:AT1G01010`" is not "`AT1G01010`"

We need a raw gene identifier:

```
# Replace the names in the ID column
de_res[, "Id"] <- sub("gene:", "", de_res[, "Id"]);
```

And we can check our genes identifiers with the function "`head`":

```
head(de_res);
```

# From gene identifiers to gene names

TAIR identifiers come from the official TAIR project. Entrez ID are only composed of numbers, they are not designed to be used by humans, however, they are unique. Genes symbols are the human-readable names, they are not unique and contains many aliases.

In order to get theses identifiers, we use the function "bitr" from the package "clusterProfiler":

```
annotation <- bitr(
  geneID   = de_res$Id,                    # Our gene list
  fromType = "TAIR",                       # We have TAIR ID
  toType   = c("ENTREZID", "SYMBOL"),# Other ID list
  OrgDb    = org.At.tair.db                # Our annotation
);
```

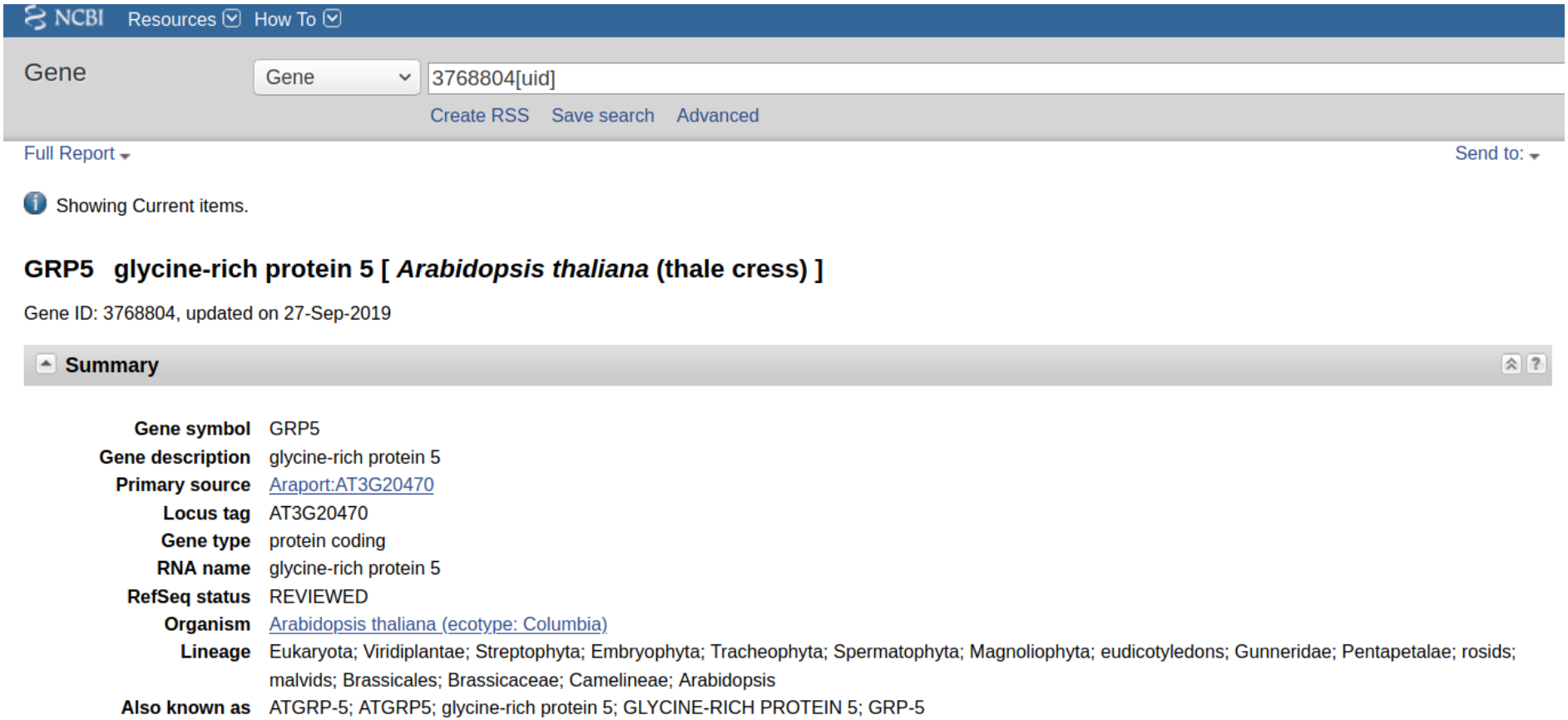
# ENTREZ Identifiers

One single ID refers to one single genomic location in one single organism. It is unique, but not human readable.

```
entrez <- data.frame(table(annotation$ENTREZID));
head(entrez[entrez$Freq > 1, ]);
```

```
      Var1 Freq
8  2745963    2
16 3768391    2
20 3768804    4
22 3770591    2
37  814670    2
40  814714    5
```

# ENTREZ Identifiers



# Genes' SYMBOLS

They are repeated, and still not human readable, because you might confuse them, use versatile aliases, etc.

```
symbol <- data.frame(table(annotation$SYMBOL));
head(symbol[symbol$Freq > 1, ]);
```

```
      Var1 Freq
33     ABR    2
149   ARP2    2
386 ATHEXP    6
979    ELP    2
1096   FMO    2
1234  HB-8    2
```

# Gene's SYMBOLS

To whom does ARP2 refer? Both genes down here does not belong to the same genomic location: Chr1:3A22720431-22723281 or Chr3:3A9952305-9956158

| | Locus | Description |
|---|---|---|
| 1 | AT1G61580 | Other names: ARABIDOPSIS RIBOSOMAL PROTEIN 2, ARP2, R-PROTEIN L3 B, RPL3B<br>R-protein L3 B;(source:Araport11) |
| 2 | AT3G27000 | Other names: ACTIN RELATED PROTEIN 2, ARP2, ATARP2, WRM, WURM<br>encodes a protein whose sequence is similar to actin-related proteins (ARPs) in other organisms. its transcript level is down regulated by light and is expressed in very low levels in all organs |

# Gene's SYMBOLS

ARP2 is not even related to A. Thaliana!

**Search results**

**Items: 1 to 20 of 6078**

<< First < Prev Page 1 of 304 Next > Last >>

See also 195 discontinued or replaced items.

| Name/Gene ID | Description | Location | Aliases | MIM |
|---|---|---|---|---|
| ARP2 ID: 851532 | actin-related protein 2 [*Saccharomyces cerevisiae S288C*] | Chromosome IV, NC_001136.10 (399340..400638) | YDL029W, ACT2 | |
| Arp2 ID: 32623 | Actin-related protein 2 [*Drosophila melanogaster* (fruit fly)] | Chromosome X, NC_004354.4 (16548290..16553968, complement) | Dmel_CG9901, ARP14D, ARP2, Actr14D, Arp14D, CG9901, Dmel\CG9901, arp2 | |
| arp2 ID: 5802965 | ARP2/3 actin-organizing complex subunit Arp2 [*Schizosaccharomyces pombe* (fission yeast)] | Chromosome I, NC_003424.3 (4783007..4784765) | SPAC11H11.06, SPAC22F8.01 | |
| ARP2 ID: 822317 | actin related protein 2 [*Arabidopsis thaliana* (thale cress)] | Chromosome 3, NC_003074.8 (9952479..9955982, complement) | AT3G27000, ACTIN RELATED PROTEIN 2, ATARP2, WRM, WURM, actin related protein 2 | |
| ARP2 ID: 30037080 | actin-related protein 2 [*Sugiyamaella lignohabitans*] | Chromosome D, NC_031673.1 (877948..878958) | AWJ20_4899 | |
| arp2 ID: 9626912 | actin-related protein Arp2 [*Volvox carteri f. nagariensis*] | | VOLCADRAFT_107669 | |

# We also lost some genes in the process

Read the warnings:

```
In bitr(geneID = sorted$Id, fromType = "TAIR"...  :
  1.44% of input gene IDs are fail to map...
```

# Have a look at our result

We finally have our annotated ranked list of genes! Yet, some of them do not have human readable names.

```
head(annotation);
```

| TAIR | ENTREZID | SYMBOL |
|---|---|---|
| AT3G09160 | 820071 | NA |
| AT1G20120 | 838601 | NA |
| AT5G24240 | 832491 | AtPI4Kgamma3 |
| AT5G24240 | 832491 | MOP9.5 |
| AT1G29090 | 839784 | NA |
| AT5G60830 | 836204 | AtbZIP70 |

# Conclusion on Gene names and identifiers

1. One gene has multiple names from multiple sources. They also have unique identifiers used by computers to refer to them as precisely as possible.
2. Do not use human readable gene names, they do not always refer to what you think.
3. Use ENTREZ identifiers when you can, then translate back to human readable names for your graphs.

# ORA Analysis

# Contingency table (1/3)

Under confidence interval, we are in the following case:

- We have 1807 differentially expressed genes with padj $< 0.05$
- We have a total of 26 698 not differentially expressed genes

ORA stands for Over-representation analysis. With ORA, we can know whether a pathway is enriched in differentially expressed genes or not.

# Contingency table (2/3)

We are interested in the "Root morphogenesis". There are 45 genes in this pathway. 7 of them are differentially expressed.

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

# Contingency table (3/3)

We can build the following associative table:

| | Differentially Expressed | Not differentially Expressed |
|---|---|---|
| In pathway | 7 | 38 |
| not in pathway | 1 800 | 28 483 |

# Fisher test

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

Let's use a fisher tests for example:

```
fisher.test(contingency)
```

```
    Fisher's Exact Test for Count Data

data:  contingency
p-value = 0.01635
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.096784 6.619033
sample estimates:
odds ratio
  2.914753
```

# Anatomy of a Statistical test

1. Have a `null hypothesis` (also called `H0` (H-zero))
2. Compute events that are very unlikely to happen **under H0**
3. Compare your observations with 2.
4. If your observations matches with 2., reject H0
5. If your observations does not mach 2., do not reject H0

# The null hypothesis (1/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

How many genes does it take to make the pathway "Root morphogenesis" *enriched in differentially expressed genes* ?

- 1 gene?
- 10 genes?
- 10% of the overall genes?
- 20% of the genes in the pathway?

# The null hypothesis (2/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

On the other side:

1. We know the number of differentially expressed genes in pathways in general.

# The null hypothesis (2/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

On the other side:

1. We know the number of differentially expressed genes in pathways in general.
2. We can guess an interval of likely and unlikely values

# The null hypothesis (2/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

On the other side:

1. We know the number of differentially expressed genes in pathways in general.
2. We can guess an interval of likely and unlikely values
3. We can compare this number of differentially expressed genes with the one in "Root morphogenesis".

# The null hypothesis (2/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

On the other side:

1. We know the number of differentially expressed genes in pathways in general.
2. We can guess an interval of likely and unlikely values
3. We can compare this number of differentially expressed genes with the one in "Root morphogenesis".
4. Then, we can say weather the number of higher or lesser.

# The null hypothesis (2/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

On the other side:

1. We know the number of differentially expressed genes in pathways in general.
2. We can guess an interval of likely and unlikely values
3. We can compare this number of differentially expressed genes with the one in "Root morphogenesis".
4. Then, we can say weather the number of higher or lesser.
5. We can tell if this number if equal.

# The null hypothesis (3/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

1. Our `null hypothesis` is: The pathway "Root morphogenesis" has the same number of differentially expressed genes as any other pathway.

# The null hypothesis (3/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

1. Our `null hypothesis` is: The pathway "Root morphogenesis" has the same number of differentially expressed genes as any other pathway.
2. We compute a confidence interval around the distribution of the number of differentially expressed genes in a pathway

# The null hypothesis (3/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

1. Our `null hypothesis` is: The pathway "Root morphogenesis" has the same number of differentially expressed genes as any other pathway.
2. We compute a confidence interval around the distribution of the number of differentially expressed genes in a pathway
3. We compare the observed number of differentially expressed genes in "Root morphogenesis" and the confidence interval in 2.

# The null hypothesis (3/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

1. Our `null hypothesis` is: The pathway "Root morphogenesis" has the same number of differentially expressed genes as any other pathway.
2. We compute a confidence interval around the distribution of the number of differentially expressed genes in a pathway
3. We compare the observed number of differentially expressed genes in "Root morphogenesis" and the confidence interval in 2.
4. If our observation is very unlikely, then we reject the null hypothesis

# The null hypothesis (3/3)

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

1. Our `null hypothesis` is: The pathway "Root morphogenesis" has the same number of differentially expressed genes as any other pathway.
2. We compute a confidence interval around the distribution of the number of differentially expressed genes in a pathway
3. We compare the observed number of differentially expressed genes in "Root morphogenesis" and the confidence interval in 2.
4. If our observation is very unlikely, then we reject the null hypothesis
5. If our observations are within the confidence interval, we do not reject the null hypothesis

# Fisher test

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

Let's use a fisher tests for example:

```
fisher.test(contingency)
```

```
    Fisher's Exact Test for Count Data

data:  contingency
p-value = 0.01635
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.096784 6.619033
sample estimates:
odds ratio
  2.914753
```

# Odds ratio

We have an `odds ratio` of 2.914753.

This mean that a gene, within the pathway "Root morphogenesis" is 2.914753% more likely to be differentially expressed than any other random gene.

# Fisher test

**Our question**: Is the pathway "Root morphogenesis" significantly enriched in differentially expressed genes?

Let's use a fisher tests for example:

```
fisher.test(contingency)
```

```
    Fisher's Exact Test for Count Data

data:  contingency
p-value = 0.01635
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.096784 6.619033
sample estimates:
odds ratio
  2.914753
```

# Conclusion on ORA (1/2)

> *With ORA, you do not test if a pathway is affected or not by your experience.*

If I want to know if "root morphogenesis" is changed during the experience, I grow *A. Thaliana* fields, cut their roots, compare their roots (weight, length, diameter, density ...). I do not need any RNA-Seq nor ORA.

If your favorite pathway does not *"show up"* in ORA, it does not mean that pathway is not affected.

# Conclusions on ORA (2/2)

> *With ORA, the ultimate p-value measure weather a pathway is enriched in differentially expressed genes.*

A single gene can be a staple of your pathway. A single gene can block a whole pathway. For ORA, all genes are equals among a pathway.

If you *know* a gene is important, whatever the ORA p-value, then this gene matters.

# Cluster Profiler Enrichment on GO: Cellular Components

We would like to perform Gene Set Enrichment analysis against the Gene Ontology's Cellular Components:

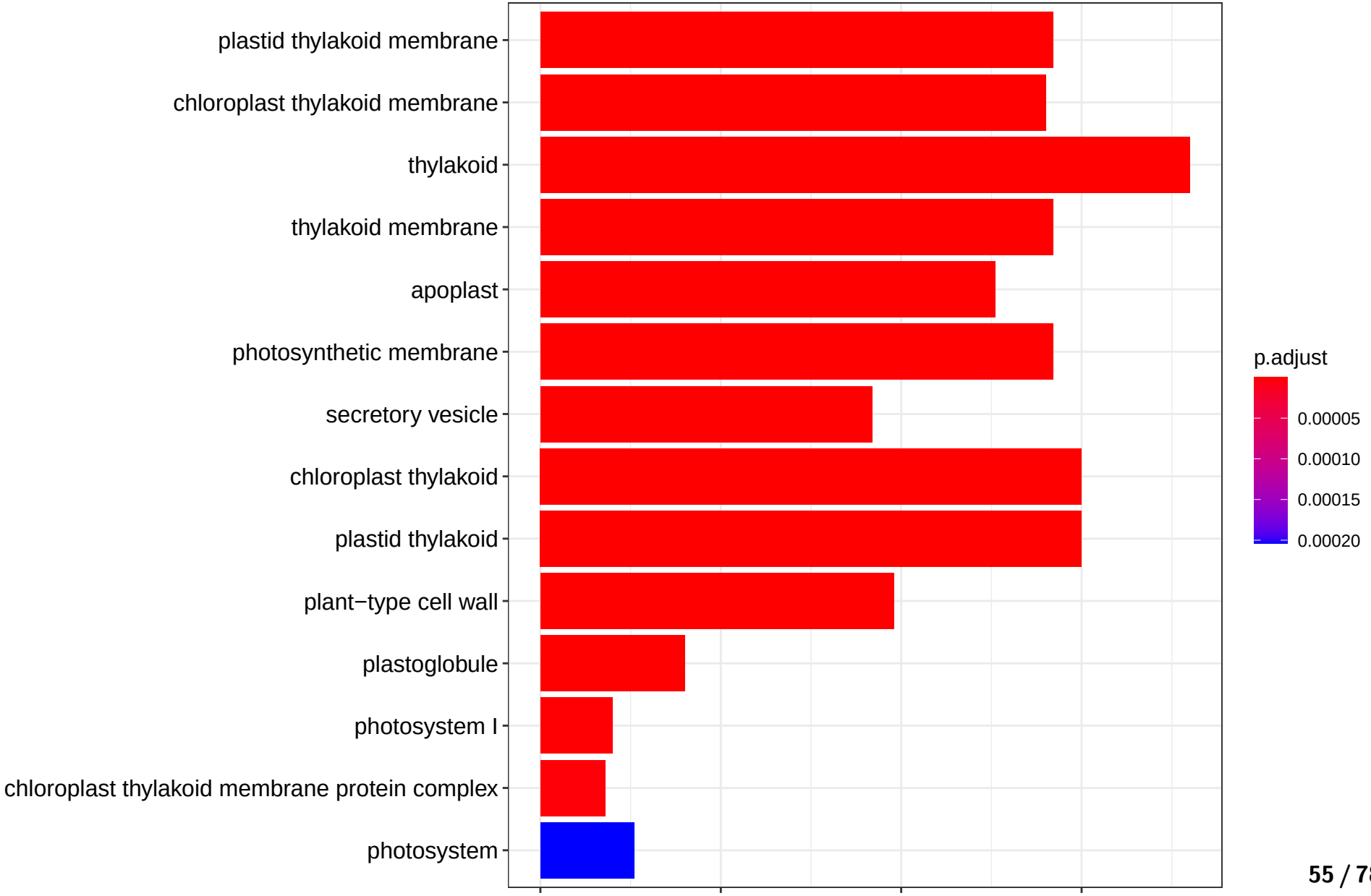
```
ego <- enrichGO(
  gene     = annotation$ENTREZID,# Ranked gene list
  OrgDb    = org.At.tair.db,     # Annotation
  keyType = "ENTREZID",          # The genes ID
  ont      = "CC",               # Cellular Components
  pvalueCutoff = 0.001,          # Significance Threshold
  pAdjustMethod = "BH"           # Adjustment method
);
```

# Cluster Profiler: Plots (1/3)

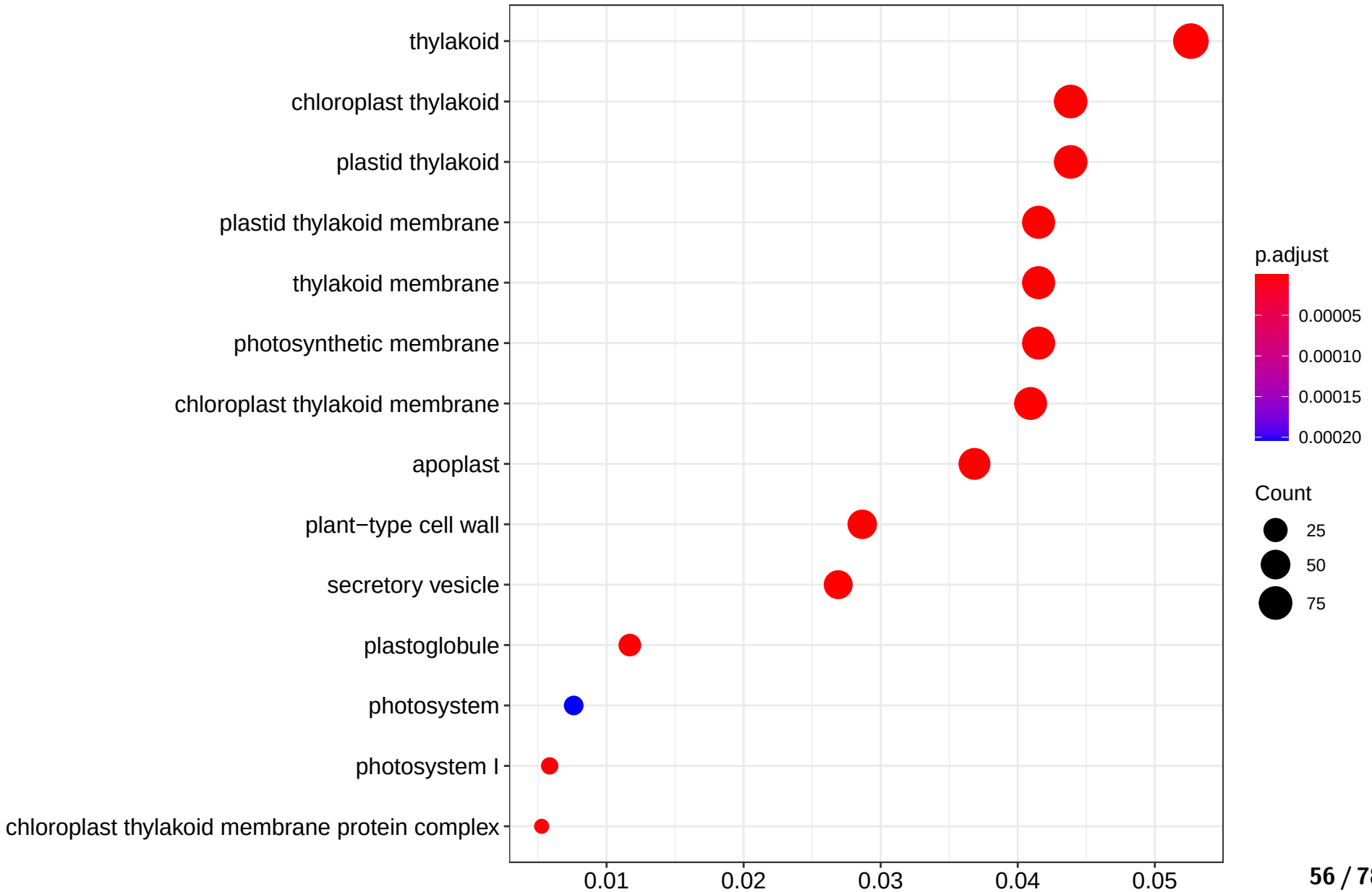
We want to visualize these results. To do so, let's use the function `"barplot"` and the function `"dotplot"` from the `"enrichplot"` package:

```
barplot(ego, showCategory=20);
dotplot(object = ego, showCategory=20);
```

# Cluster Profiler: Plots (2/3)



# Cluster Profiler: Plots (3/3)



# GSEA

# Re-load the original data

We need the *complete, unfiltered* list of genes for GSEA:

```
de_res <- read.table(
  file   = "tables/KOvsWT.complete.txt",   # Path to results
  header = TRUE,                # There are column names
  sep    = "\t",                # This is a tabulation
  stringsAsFactors = FALSE # Colnames are not factors
);
de_res <- de_res[ ,c("Id", "stat")];
de_res[, "Id"] <- sub("gene:", "", de_res[, "Id"]);
```

# Prepare Data (1/2)

To perform a Gene Set Enrichment Analysis (GSEA), we need to give a list of weighted ranked genes in order to compute the enrichment score for our table called "`sorted`". We also need the ENTREZ identifiers, from our table called "`annotation`". We need to merge them according the the TAIR identifiers:

```
# Rename the columns
colnames(de_res) <- c("TAIR", "stat");

# Merge the frames
geneFrame <- merge(de_res, annotation, by="TAIR");
geneFrame <- unique(geneFrame); # Drop duplicates
```

# Prepare Data (2/2)

GSEA expects a list of decreasing numeric values. Let's build this structure:

```
# Build a numeric vector
geneList <- as.numeric(geneFrame$stat);

# Get the genes identifiers
names(geneList) <- geneFrame$ENTREZID;

# Sort this list
geneList <- sort(geneList, decreasing = TRUE);
```

# Run analysis

Dear statisticians, please look aside for a minute.

```
gsea <- gseGO(
  geneList = geneList,          # Ranked gene list
  ont      = "BP",              # Biological Process
  OrgDb    = org.At.tair.db,    # Annotation
  keyType  = "ENTREZID",        # Identifiers
  pAdjustMethod = "BH",         # Pvalue Adjustment
  pvalueCutoff = 1              # Significance Threshold
);
```

# GSEA plot (1/4)

Let's see the top 8 of the over-represented genes sets:

```
columns_of_interest <- c(
  "Description",
  "enrichmentScore",
  "p.adjust"
);
head(
  x = gsea[, columns_of_interest], # Pathway ID
  8                                # lines to display
);
```

# GSEA plot (2/4)

Let's see the top 8 of the over-represented genes sets:

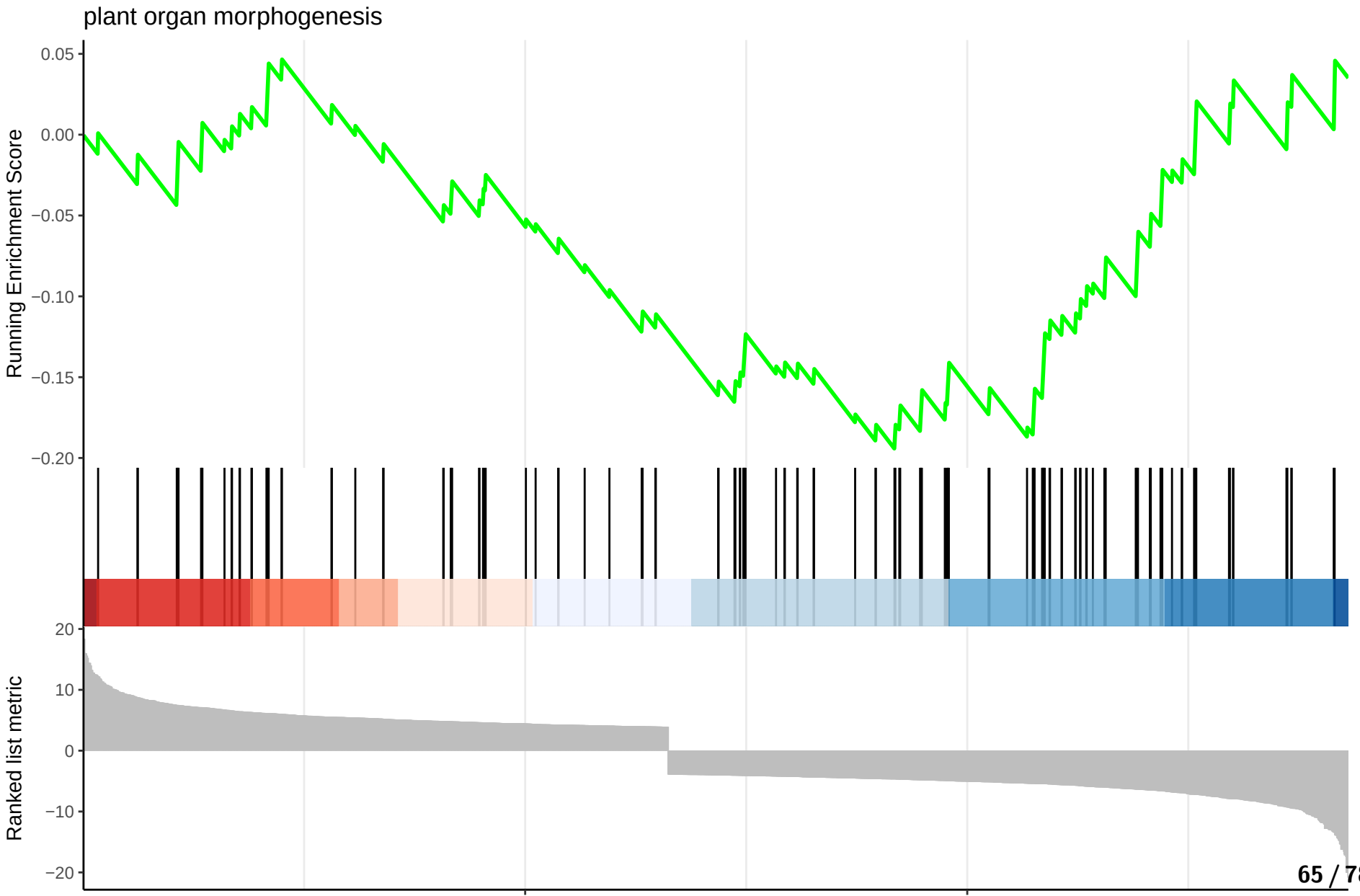
| Description | enrichmentScore | p.ac |
|---|---|---|
| nucleobase-containing compound metabolic process | 0.2925762 | 0e |
| gene expression | 0.2831054 | 0e |
| heterocycle metabolic process | 0.2681485 | 0e |
| nucleic acid metabolic process | 0.3016717 | 0e |
| RNA metabolic process | 0.2941390 | 0e |
| cellular aromatic compound metabolic process | 0.2503229 | 0e |
| mRNA metabolic process | 0.5630310 | 0e |
| organic cyclic compound metabolic process | 0.2367644 | 1 |

# GSEA plot (3/4)

Finally, building the GSEA plot is being done with the function "gseaplot2" from "clusterProfiler":

```
# We need the number of the line
# Containing our pathway of interest
gsea_line <- match(
  "plant organ morphogenesis",
  gsea$Description
);
gseaplot2(
  x          = gsea,                    # Our analysis
  geneSetID = gsea$ID[gea_line],  # Pathway ID
  title      = "plant organ morphogenesis" # Its name
);
```

# GSEA plot (4/4)

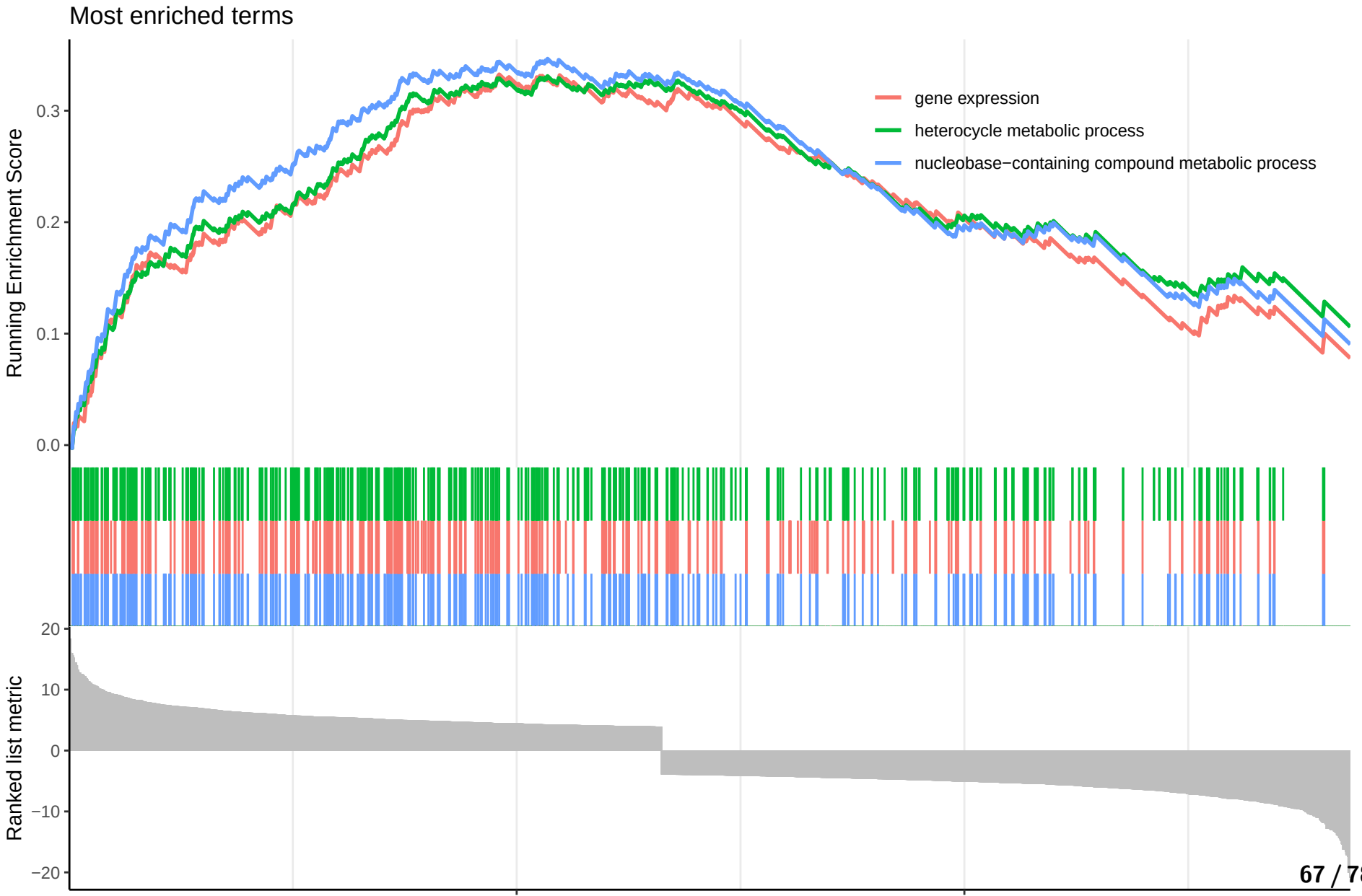

plant organ morphogenesis
Running Enrichment Score
0.05
0.00
−0.05
−0.10
−0.15
−0.20
Ranked list metric
20
10
0
−10
−20

# Multiple GSEA on the same graph (1/2)

... because we can!

```
gseaplot2(
  x = gsea,
  geneSetID = 1:3,
  title = "Most enriched terms"
);
```

# Multiple GSEA on the same graph (2/2)



# Conclusion on GSEA

*With GSEA, you dot not test if a pathway is up or down regulated.*

A pathway contains both enhancers and suppressors genes. An up-regulation of enhancer genes and a down-regulation of suppressor genes will lead to a "bad" enrichment score. However, this will lead to a strong change in your pathway activity!

If your favorite pathway does not have a *"good enrichment score"*, it does not mean that pathway is not affected.

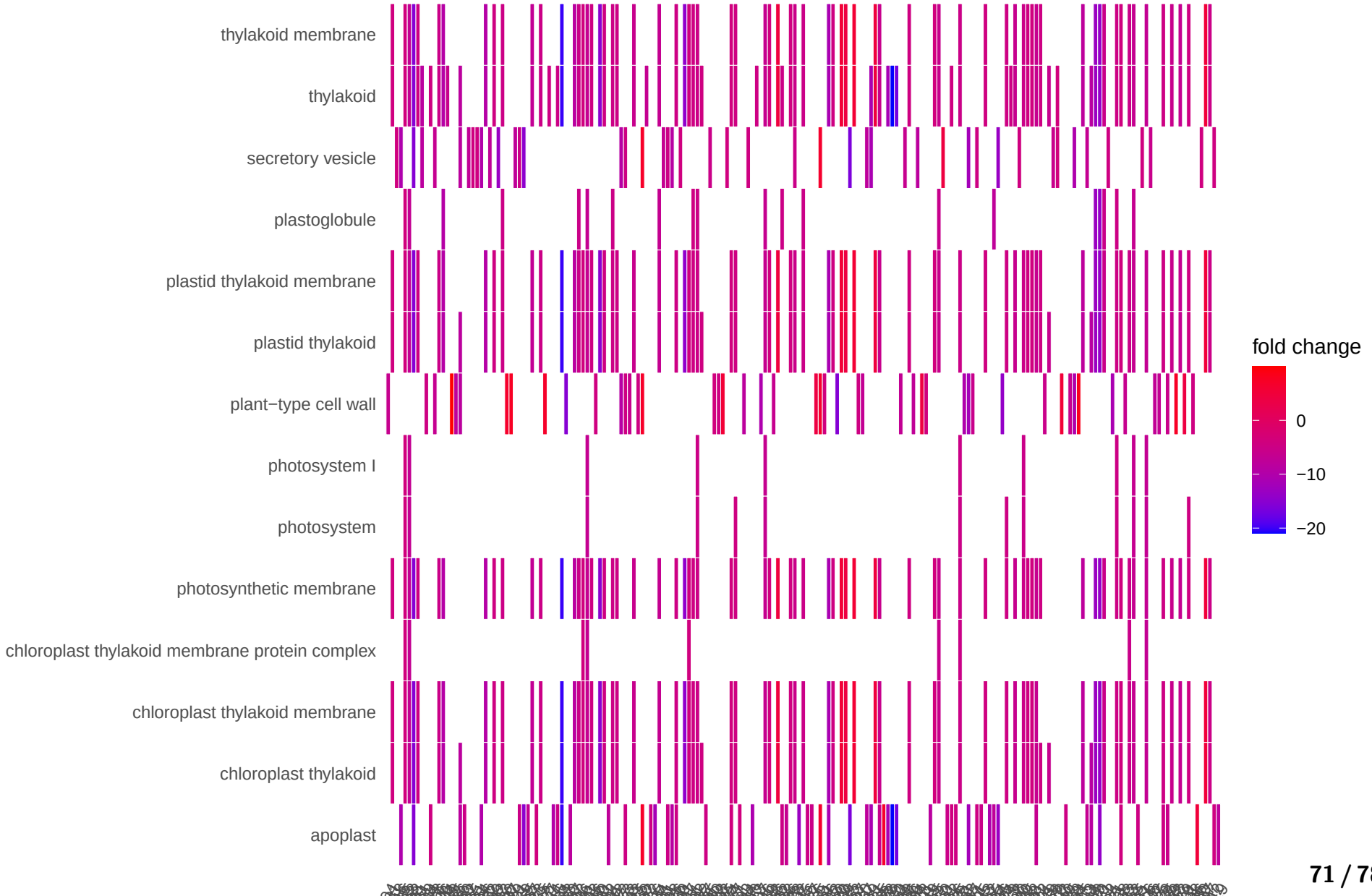
# Set comparison

# Heatmap (1/2)

Very common in publications

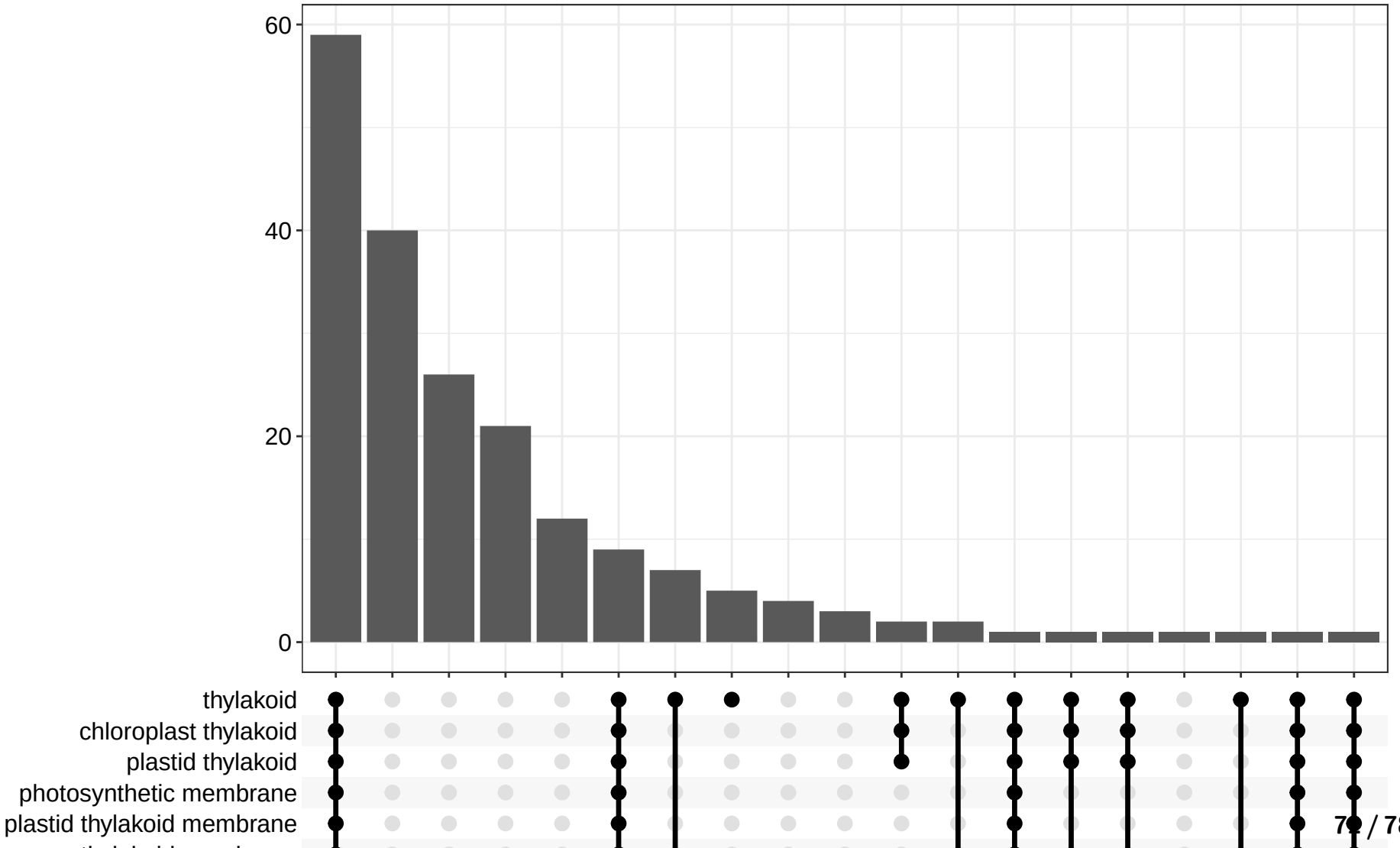
```
heatplot(
  x =  ego,                 # Our enrichment
  showCategory = 15,        # Nb of terms to display
  foldChange = geneList # Our fold changes
);
```

# Heatmap (2/2)


thylakoid membrane
thylakoid
secretory vesicle
plastoglobule
plastid thylakoid membrane
plastid thylakoid
plant−type cell wall
photosystem I
photosystem
photosynthetic membrane
chloroplast thylakoid membrane protein complex
chloroplast thylakoid membrane
chloroplast thylakoid
apoplast
fold change
0
−10
−20

# UpSet plot

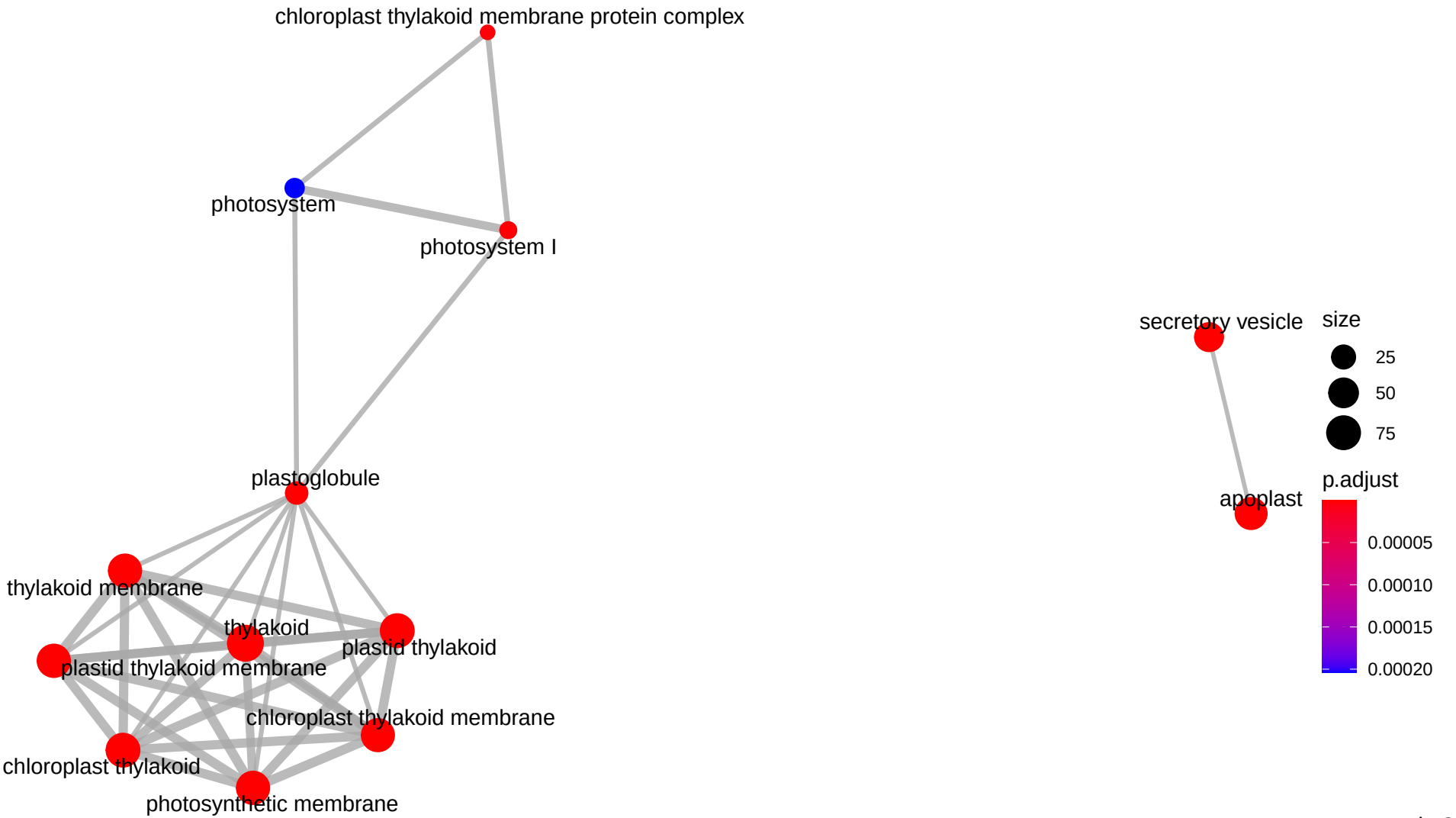
```
upsetplot(x = ego); # From our enrichment analysis
```



# Networks

# Enrichment map

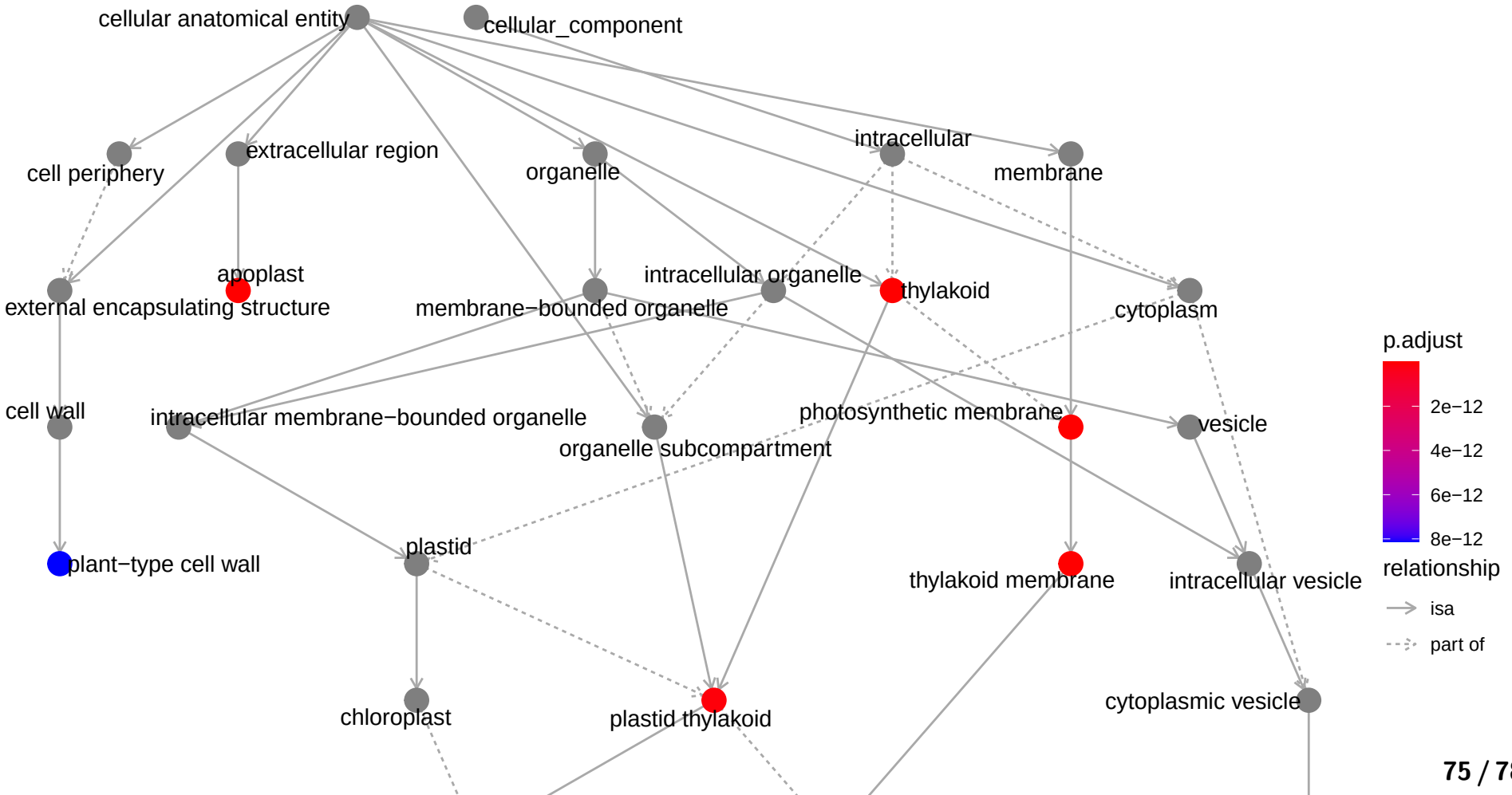
```
emapplot(ego); # From our enrichment analysis
```



# GO plot

Relate enriched terms with each others:

```
goplot(ego); # From our enrichment analysis
```



# Kegg (1/2)

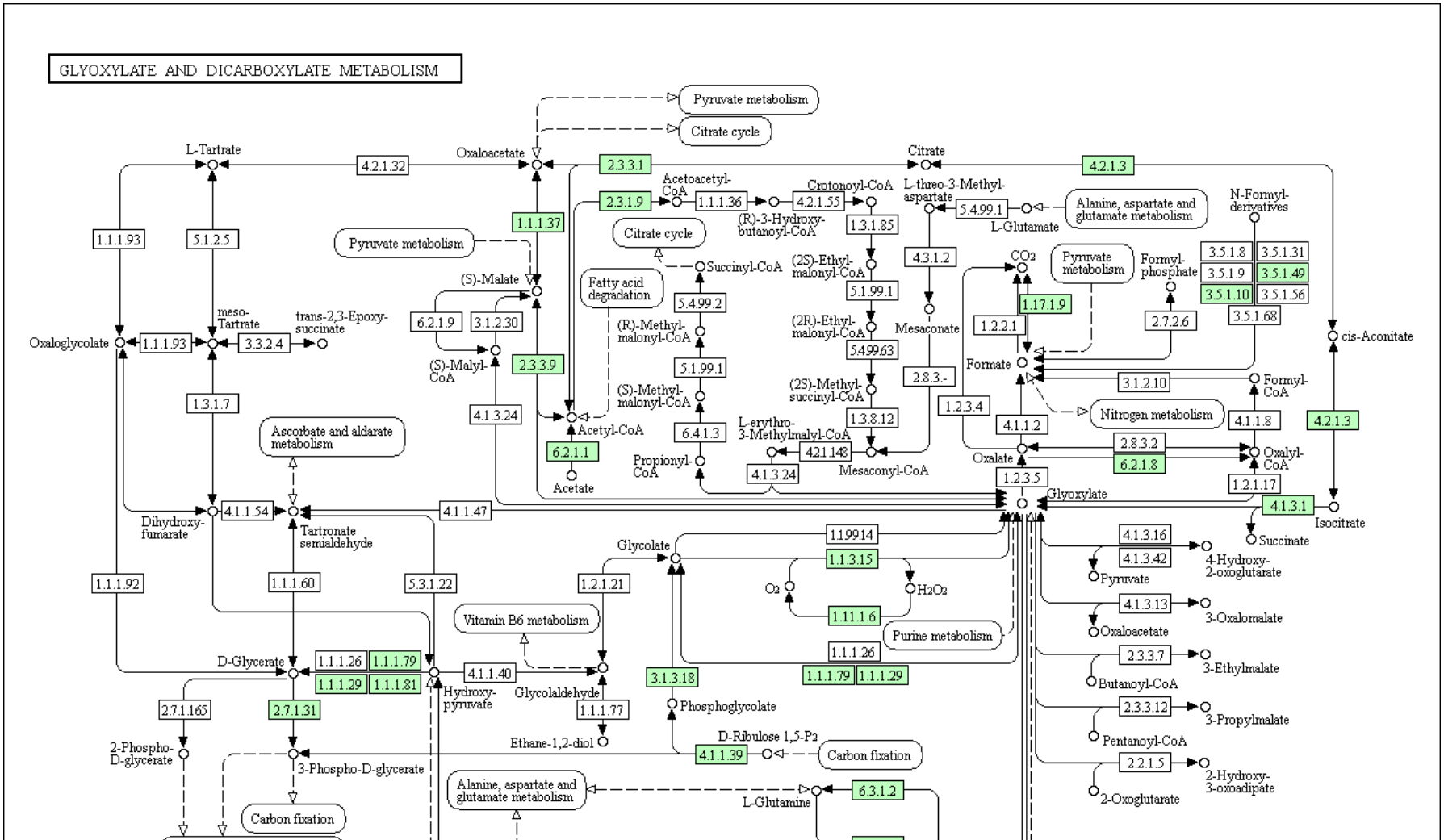
The Kegg analysis is done with the "pathview" package and this eponymous function:

```
names(geneList) <- geneFrame$TAIR;  # Use TAIR id

pv.out <- pathview(
  gene.data = geneList,      # Our gene list
  pathway.id = "ath00630",   # Our pathway
  species = "ath",           # Our organism
  # The color limits
  limit = list(gene=max(abs(geneList))),
  gene.idtype = "TAIR"       # The genes identifiers
);
```

# KEGG (2/2)

There is the representation of our pathway, with differentially expressed genes colored!



# Thanks

Special thanks to Rachel Legendre and Natalia Pietrosemoli, this session was highly based on their work.

Thanks to the rest of the team for their reviews and advises.