stage/ **Directory structure**

```
morlon@vega:~/stage$ tree -L 2
.
├── Go enrichment for non-model species.md
├── data
│   ├── final
│   ├── intermediate
│   └── raw
├── go_enrichment_analysis.md
├── images
│   └── FULL5heatmapEuclideanComplete.svg
├── results
│   ├── compare_cluster
│   ├── go_enrichment_analysis
│   ├── go_enrichment_analysis_v2
│   ├── go_enrichment_analysis_v2_dotplot
│   └── topgo_go_enrichment_analysis
├── script
│   ├── clusterprofiler_go_enrichment_analysis.R
│   ├── clusterprofiler_go_enrichment_analysis_v2.R
│   ├── clusterprofiler_go_enrichment_analysis_v2_dotplot.R
│   ├── clusters_pvalue_distribution.R
│   ├── compare_cluster.R
│   ├── orthologs_analysis_test.R
│   ├── topgo_go_enrichement_analysis_v2.R
│   └── topgo_go_enrichment_analysis.R
└── scripts.md
```

- images/: heatmap (stage martin)
- data/raw: b2g.reformatted.annot file + clusters.txt (stage martin) – see below
- script/ *script_name.R*
- scripts.md: description of scripts written so far
- results/*script_name.R*: results of a given script.R
- + various .md which go over the work done

```
strigamia-acuminata_seq1  GO:0003677  UniRef50_T1IWD7PWWP domain-containing protein n=1 Tax=Strigamia maritima TaxID=126957 RepID=T1IWD7_STRMM
strigamia-acuminata_seq3  GO:0005515  UniRef50_T1IWD5LRRCT domain-containing protein n=1 Tax=Strigamia maritima TaxID=126957 RepID=T1IWD5_STRMM
strigamia-acuminata_seq8  GO:0017065  UniRef50_A0A158NRI9Uracil-DNA glycosylase-like domain-containing protein n=23 Tax=Formicidae TaxID=36668 RepID=A0A158NRI9_ATTCE
strigamia-acuminata_seq8  GO:0006284
strigamia-acuminata_seq8  EC:3.2.2.27
strigamia-acuminata_seq13 GO:0008324  UniRef50_A0A4S2KDE5Solute carrier family 41 member 2 n=2 Tax=Endopterygota TaxID=33392 RepID=A0A4S2KDE5_9HYME
strigamia-acuminata_seq13 GO:0006812
strigamia-acuminata_seq15 GO:0005509  UniRef50_T1IWC6EGF-like domain-containing protein n=1 Tax=Strigamia maritima TaxID=126957 RepID=T1IWC6_STRMM
strigamia-acuminata_seq16 GO:0031267  UniRef50_O43592Exportin-T n=1481 Tax=Eumetazoa TaxID=6072 RepID=XPOT_HUMAN
```

```
"x"
"strigamia-acuminata_seq2"
"strigamia-acuminata_seq16"
"strigamia-acuminata_seq17"
"strigamia-acuminata_seq23"
"strigamia-acuminata_seq30"
"strigamia-acuminata_seq40"
"strigamia-acuminata_seq42"
"strigamia-acuminata_seq50"
"strigamia-acuminata_seq65"
"strigamia-acuminata_seq67"
```

Non-model org GO analysis - **which options ?**

Might be possible ?

- Difficult to implement in a script
- Need to generate a PANTHER generic mapping file (I think the genes name need to be standard)

BiNGO

Cytoscape extension:

- Allow use of custom dataset
- Necessitate Cytoscape
- Not well maintained

clusterProfiler (BioC package)

enricher() function:

- Allow use of custom dataset
- Many visualization option

enrichGO() function:

- If an OrgDb database is created for the universe dataset (possible ?)
- Allow to choose sub-ontology

topGO (BioC package)

Takes GO topology into account:

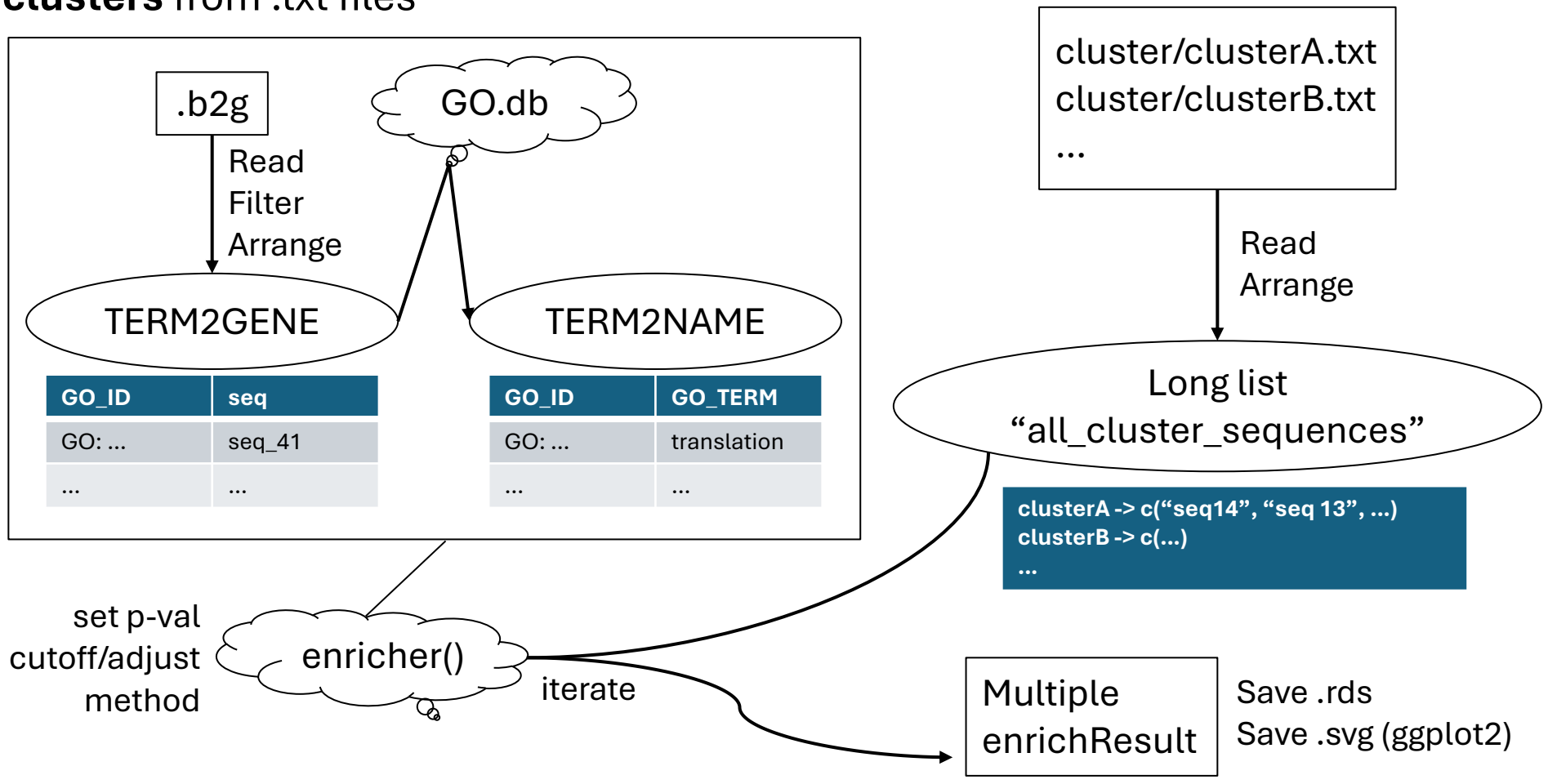
- Choose sub-ontology
- Less visualization options
- Multiple algorithms and test statistics available

Done

Not done/possible

GO analysis using **clusterProfiler**

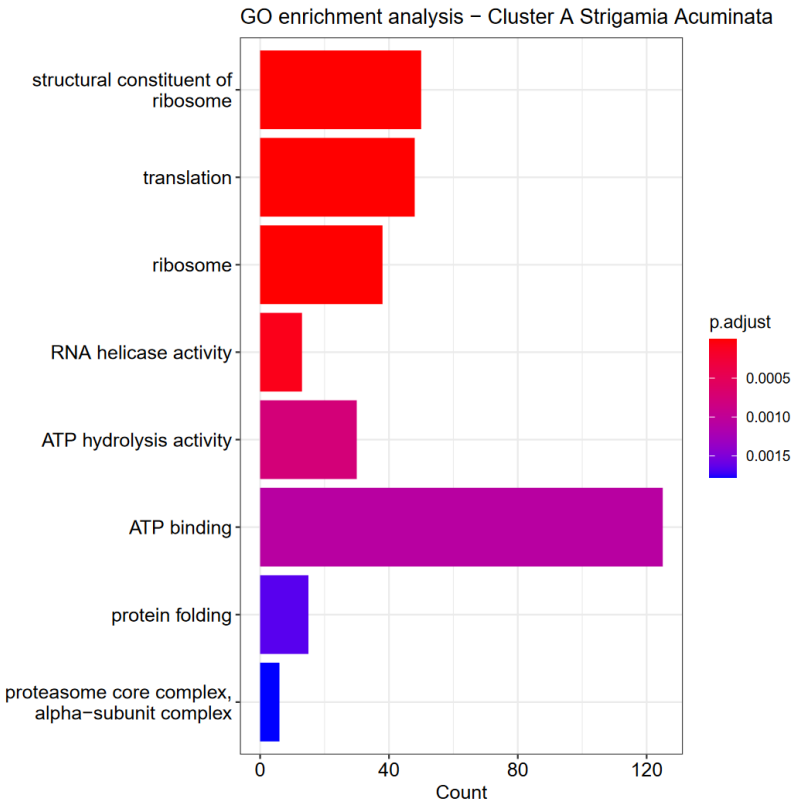
1. Create **TERM2GENE** (GO_ID – gene) dataframe (using .b2g file)
2. Using TERM2GENE + GO.db, create **TERM2NAME** (GO_ID – GO_term)
3. Create **large list of clusters** from .txt files
4. Execute **analysis**
   - Cut-off p-val
5. **Visualize**

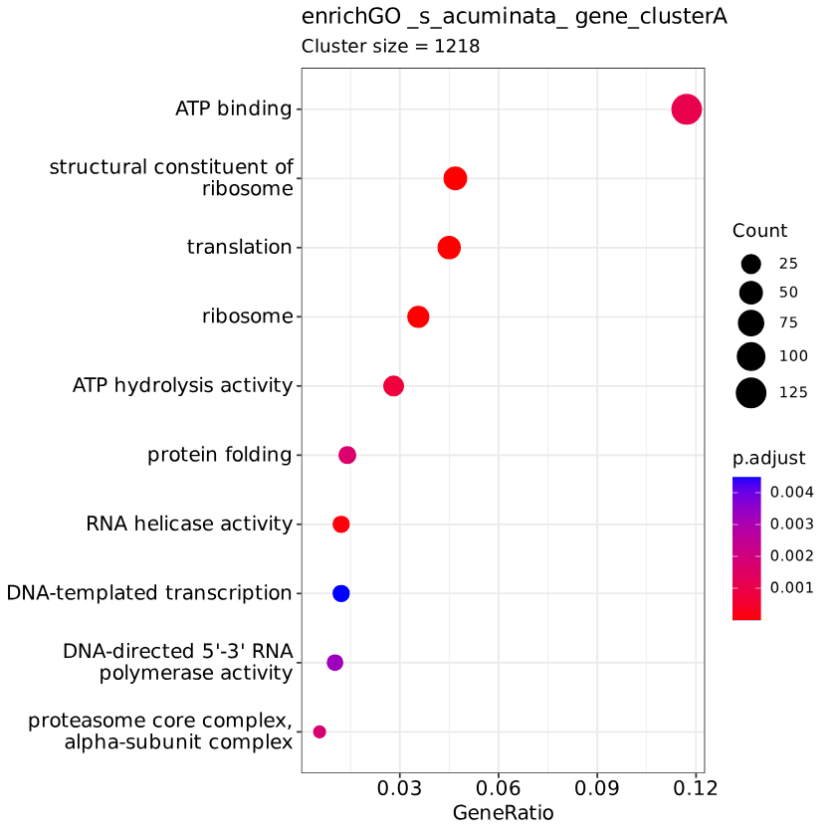


GO analysis using **clusterProfiler**

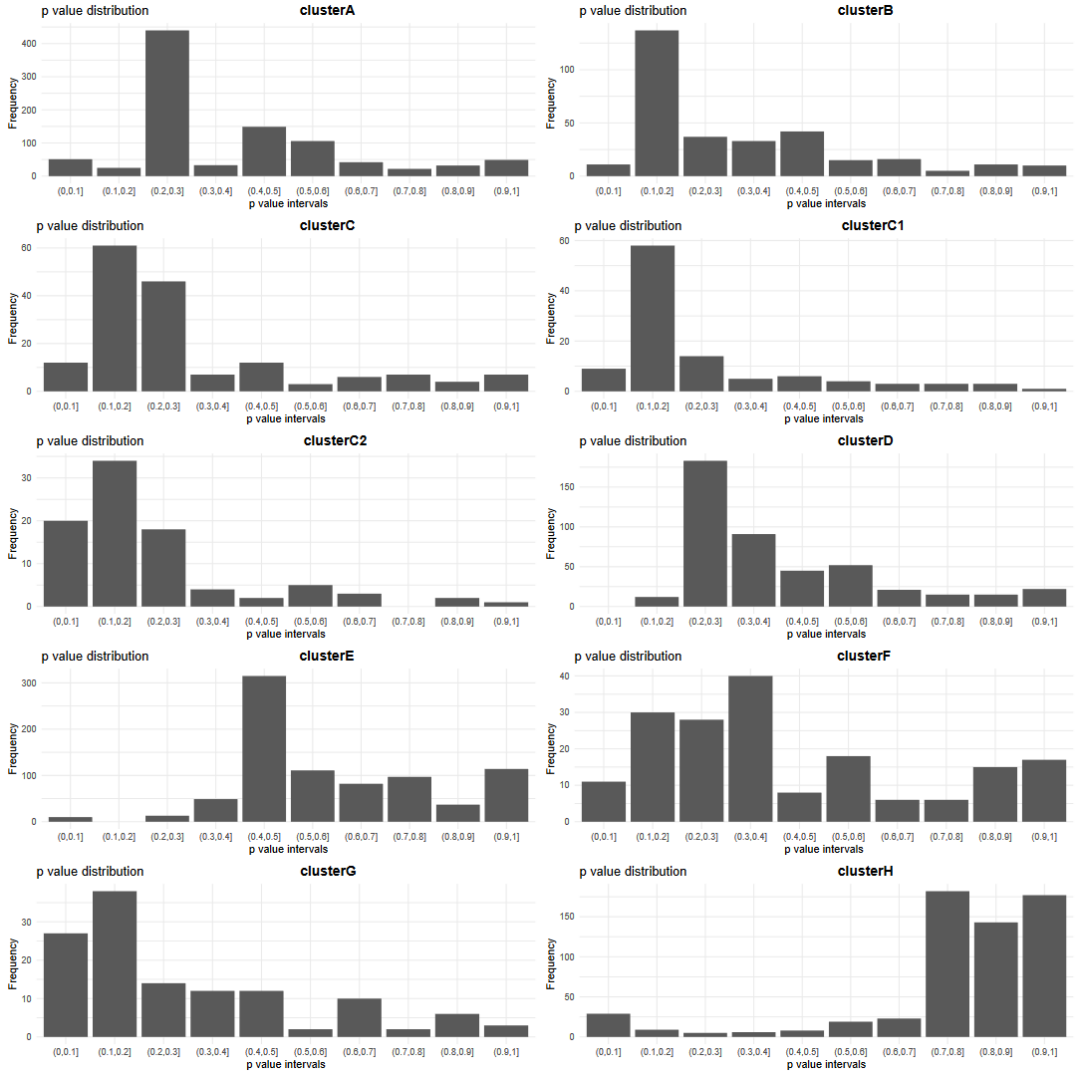
1. Create **TERM2GENE** (GO_ID – gene) dataframe (using .b2g file)
2. Using TERM2GENE + GO.db, create **TERM2NAME** (GO_ID – GO_term)
3. Create **large list of clusters** from .txt files
4. Execute **analysis**
   - Cut-off p-val
5. **Visualize**

e.g. Cluster A (highly conserved genes)





GO analysis using **clusterProfiler**



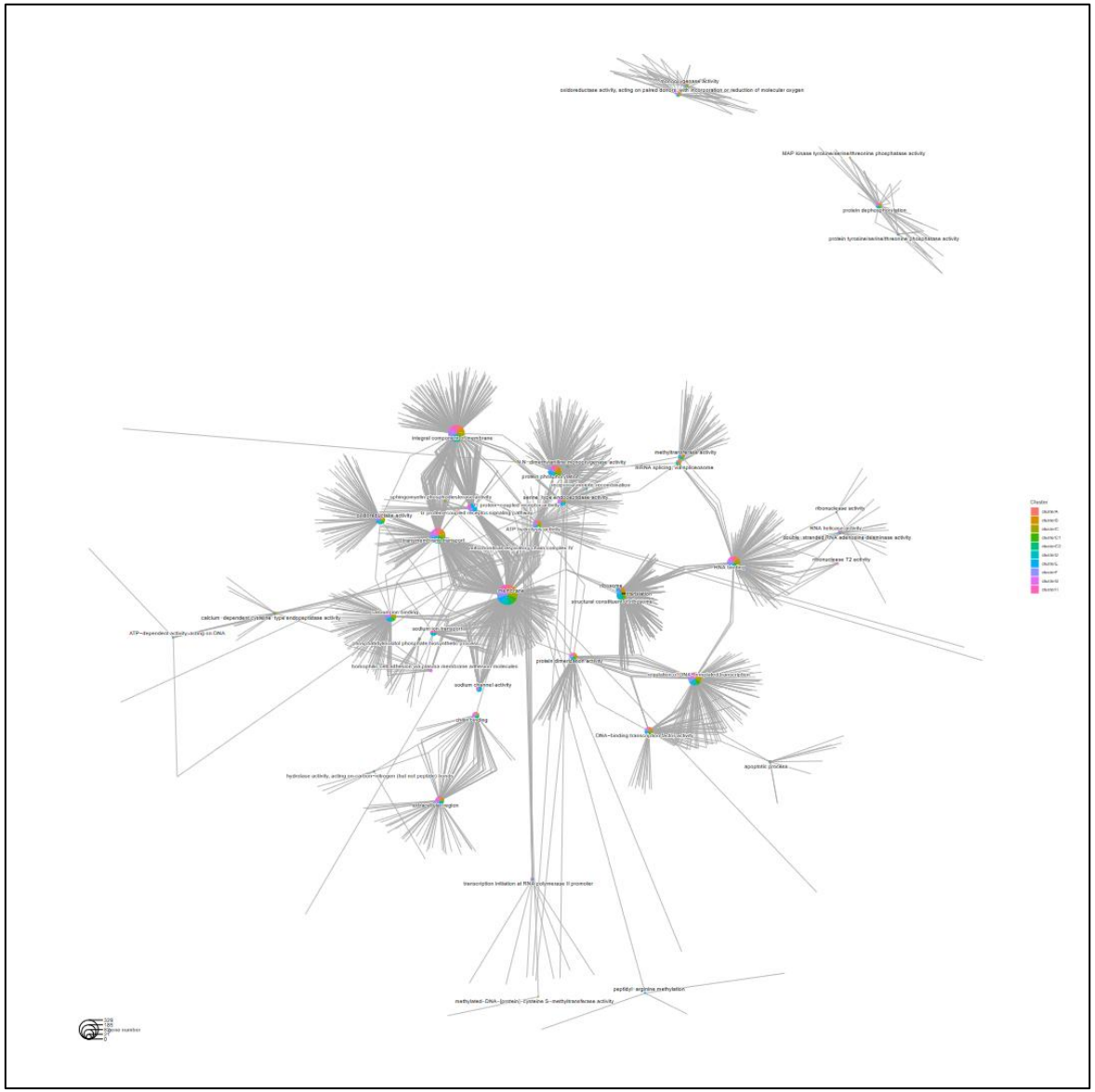
The object resulting from the enricher() analysis allows various plotting options:

- P-value distribution
- Dot-plot w/ all clusters

GO analysis using **clusterProfiler**

The object resulting from the enricher() analysis allows various plotting options:

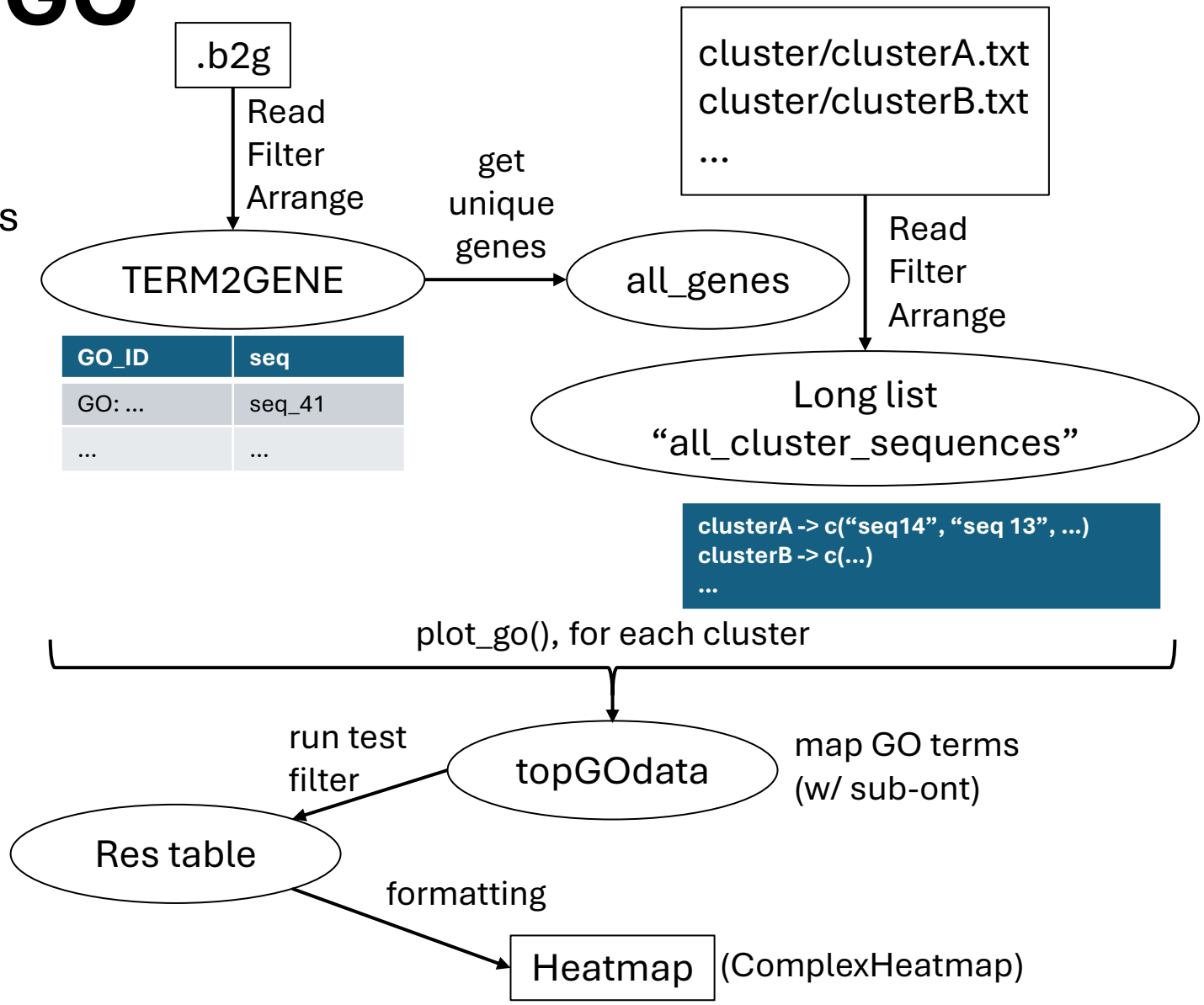
- Gene-concept network



As it is, we cannot choose sub-ontology, which can lead to confusing representations

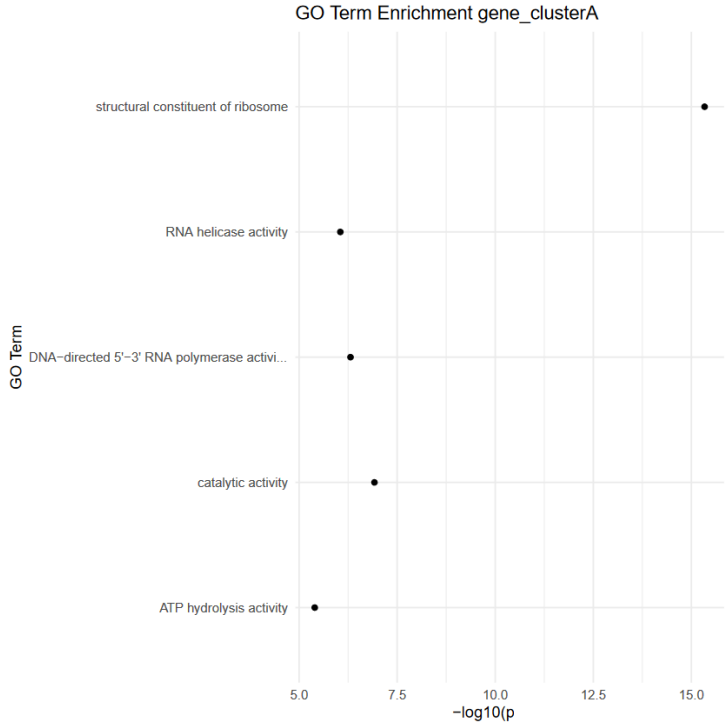
GO analysis using **topGO**

1. Create **TERM2GENE** (GO_ID – gene) dataframe (using .b2g file)
2. Create **large list of clusters** from .txt files
3. Execute **analysis**
   - Choose algorithm
     - **Classic**: no hierarchical order consideration
     - **Elim**: remove gene annotated to a significantly enriched node from all its ancestor (very strict)
     - **Weight & Weight01**: generalize elim idea to weight in the 0-1 interval (reduce false positive + allow to detect locally most significant term)
     - **…**
   - Choose subGO ontology
   - Cut-off p-val
4. **Visualize**



GO analysis using **topGO**

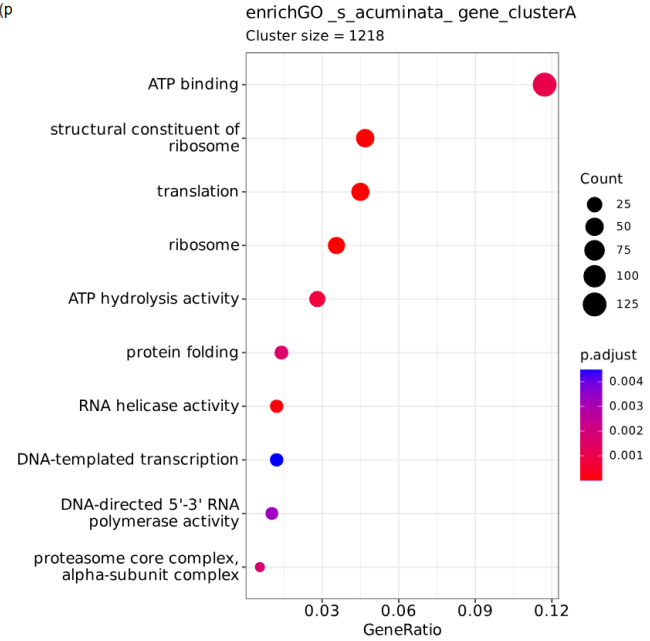
1. Create **TERM2GENE** (GO_ID – gene) dataframe (using .b2g file)
2. Create **large list of clusters** from .txt files
3. Execute **analysis**
   - Choose algorithm
     - **Classic**: no hierarchical order consideration
     - **Elim**: remove gene annotated to a significantly enriched node from all its ancestor (very strict)
     - **Weight & Weight01**: generalize elim idea to weight in the 0-1 interval (reduce false positive + allow to detect locally most significant term)
     - **…**
   - Choose subGO ontology
   - Cut-off p-val
4. **Visualize**



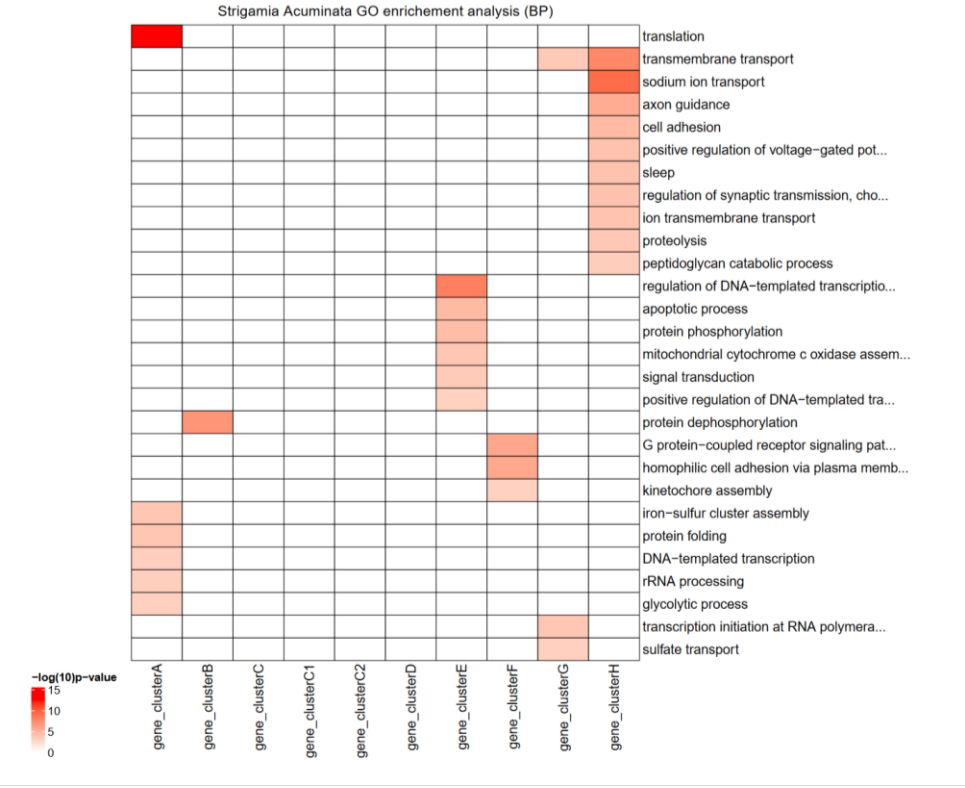
e.g. Cluster A (highly conserved genes)

-

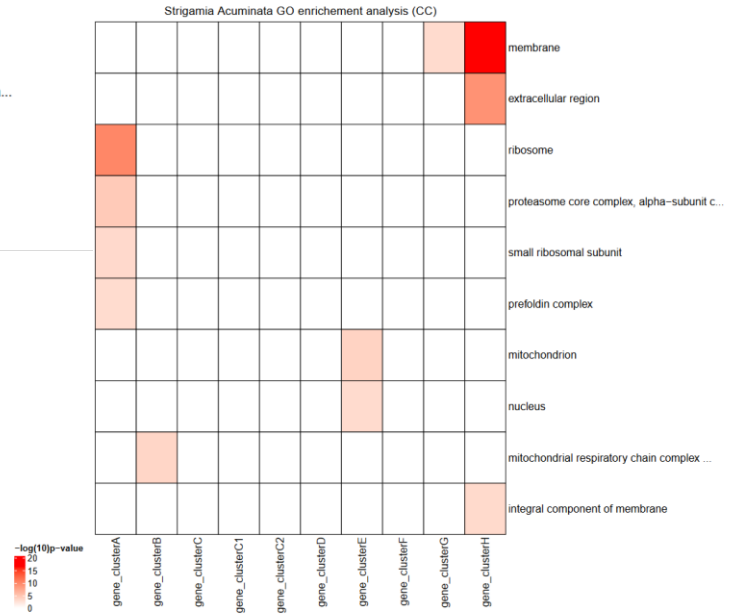
topGO_MF (top)
clusterProfiler (bottom)

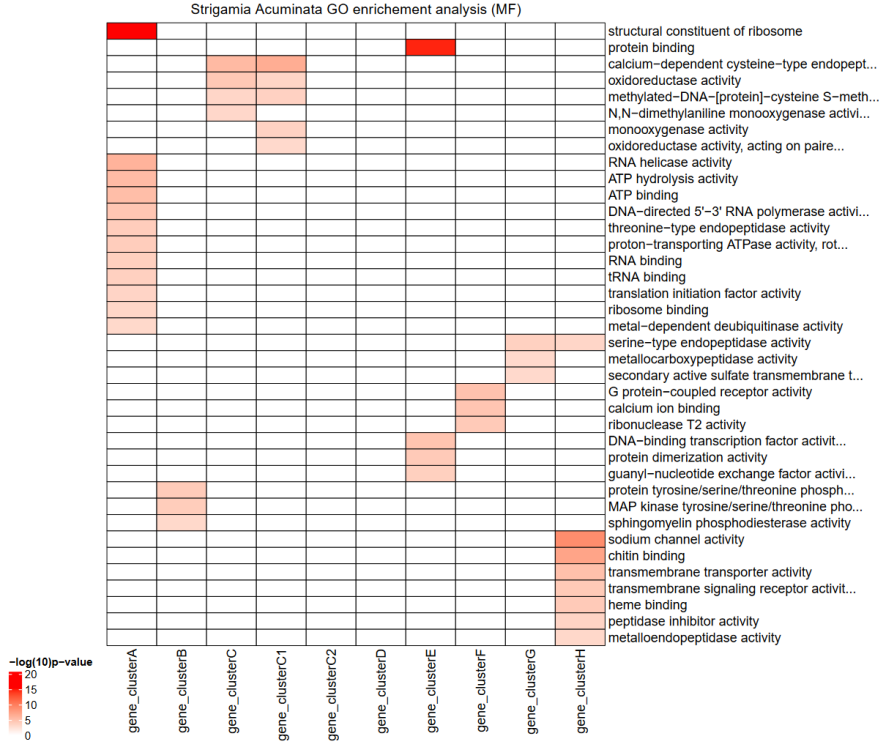


GO analysis using **topGO**



All 3 sub-ontology heatmaps:
p-val = 0.005
algo = weight01
No top term limit





A tool for analyzing and validating clusters

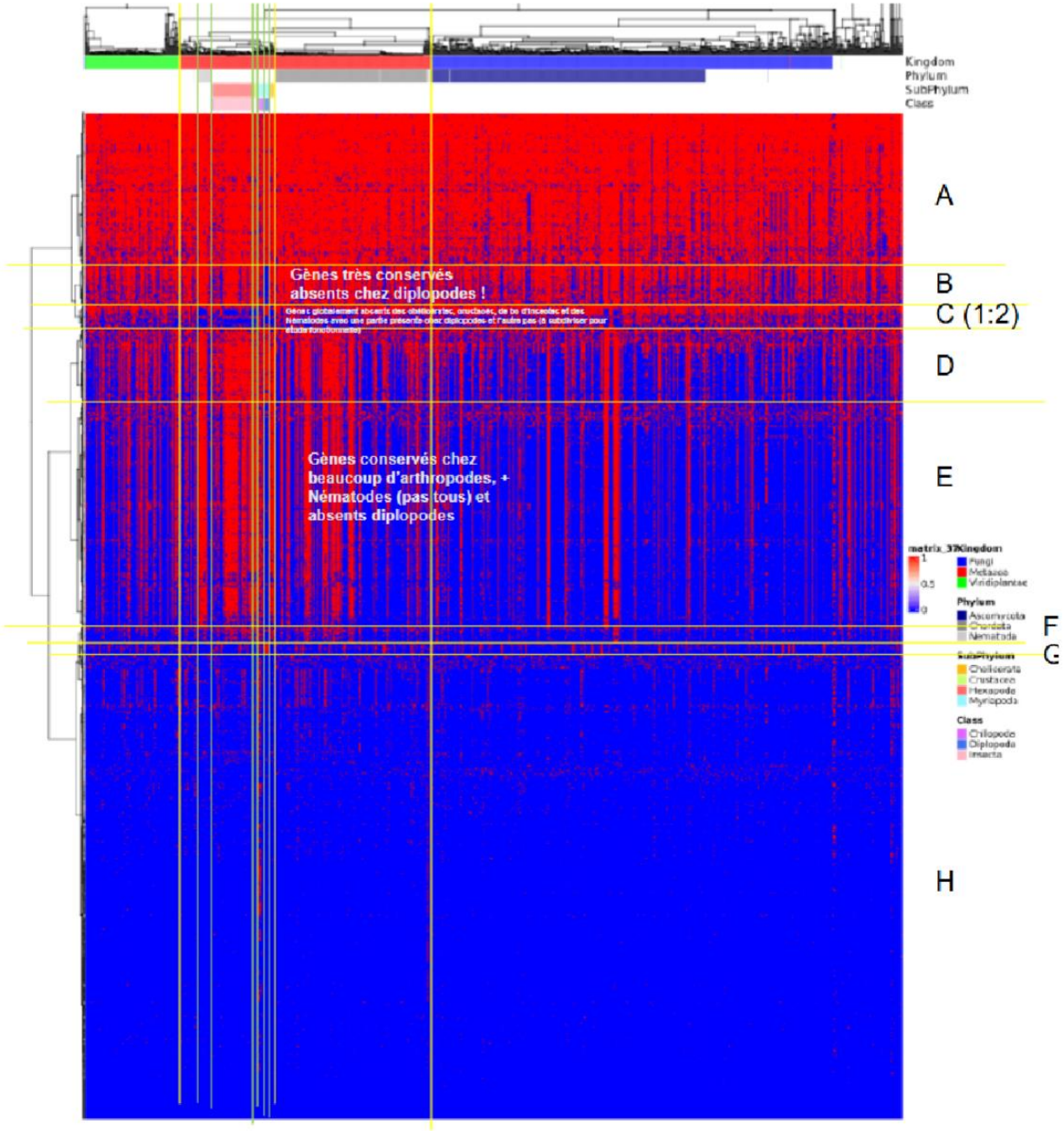
To be done ?

- Incorporate the various compareCluster tools to the clusterProfiler script DONE
- If possible, create OrgDb object for use of enrichGO() function
- Define parameters for topGO analysis – explore visualization options DONE ? Select cutoffpvalue = 0.05 and cutoffclusterterms = 3

Challenges

- Most GO enrichment tools necessitate the organism to be annotated and/or the genes to be named according to UniProt's nomenclature – thus the need for alternative tools

Annexe


Kingdom
Phylum
SubPhylum
Class
A
Gènes très conservés
absents chez diplopodes !
B
C (1:2)
D
Gènes conservés chez
beaucoup d'arthropodes, +
Nématodes (pas tous) et
absents diplopodes
E
F
G
H