Clustering
&
Function Enrichment Analysis

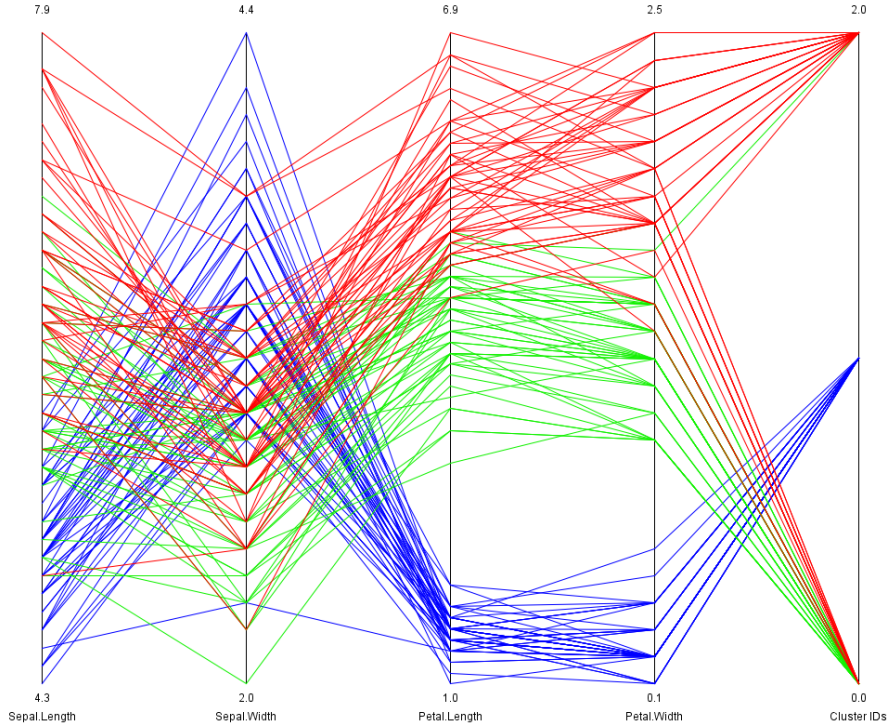
Qi Sun

Bioinformatics Facility

Cornell University

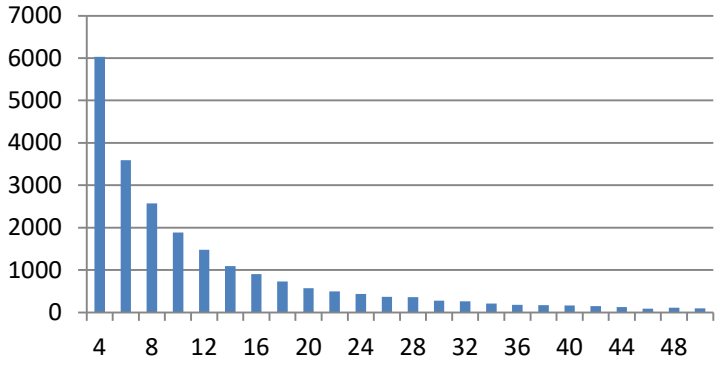
Clustering analysis

1. Hierarchical
2. K-means
3. Co-expression network

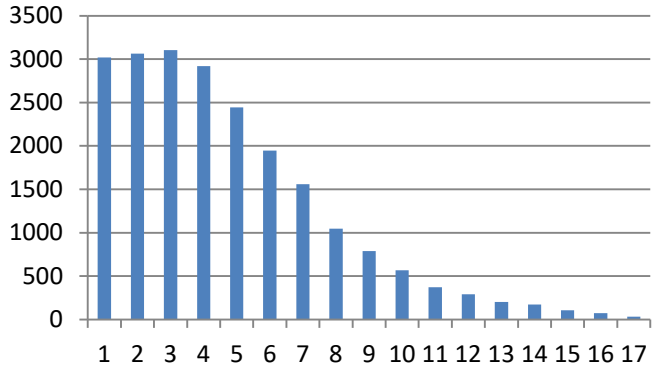


Prepare data for clustering

Step 1. LOG transformation of CPM value to improve the distribution



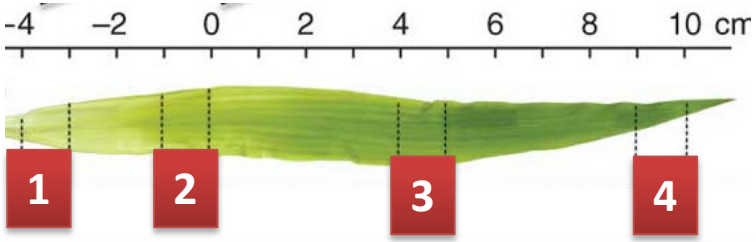
CPM

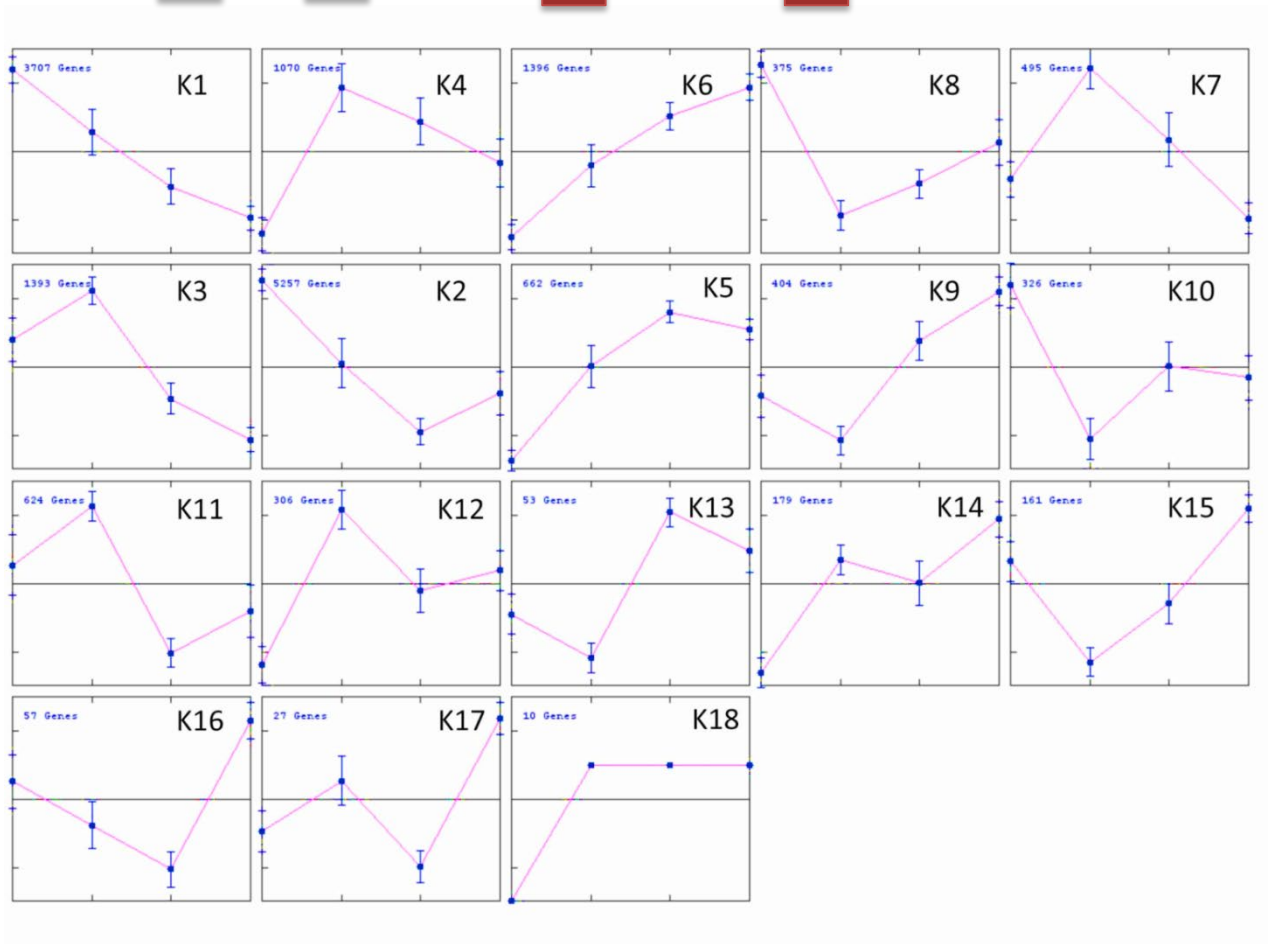


Log2(CPM)

Step 2. Remove genes with no variation across samples

Clustering analysis on multiple conditions of RNA-seq data

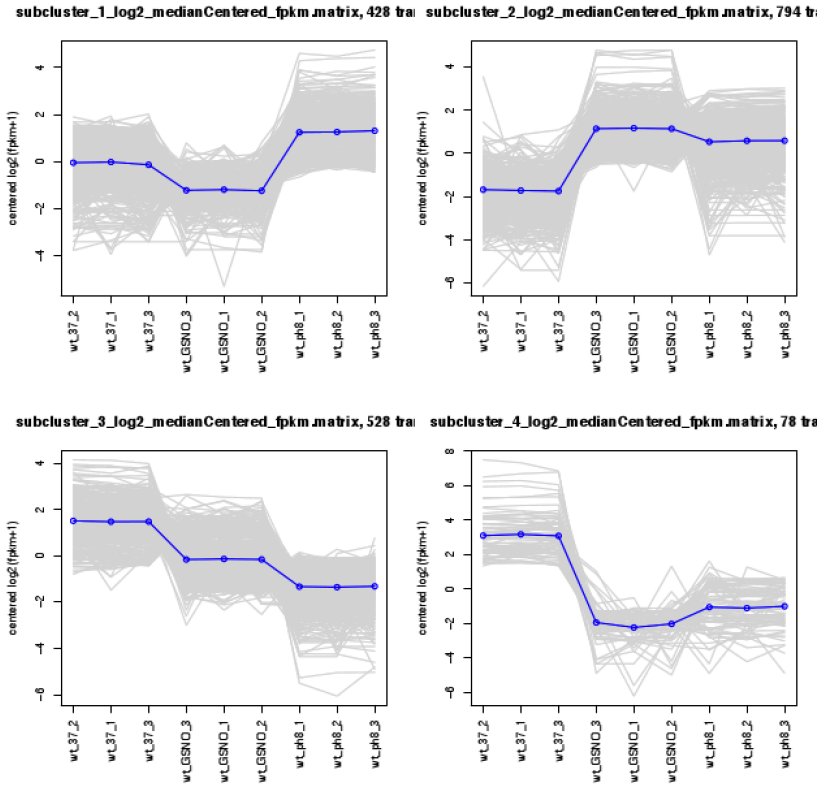




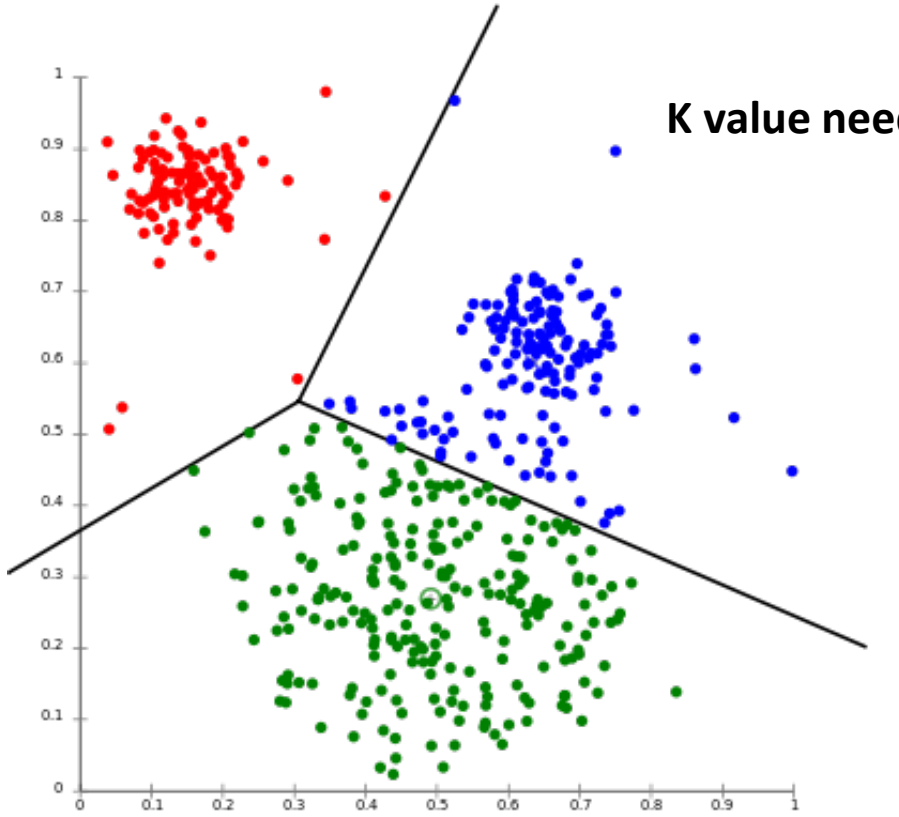
Hierarchical clustering

K-means clustering

```
$TRINITY_HOME/Analysis/DifferentialExpression/
define_clusters_by_cutting_tree.pl  -R
diffExpr.P0.001_C2.matrix.RData -K 18
```



K-means clustering

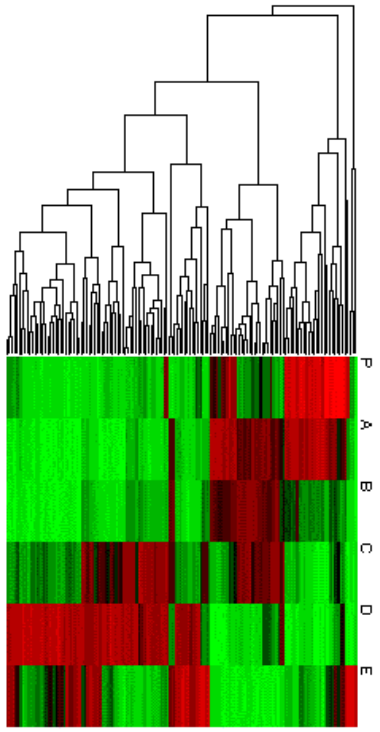


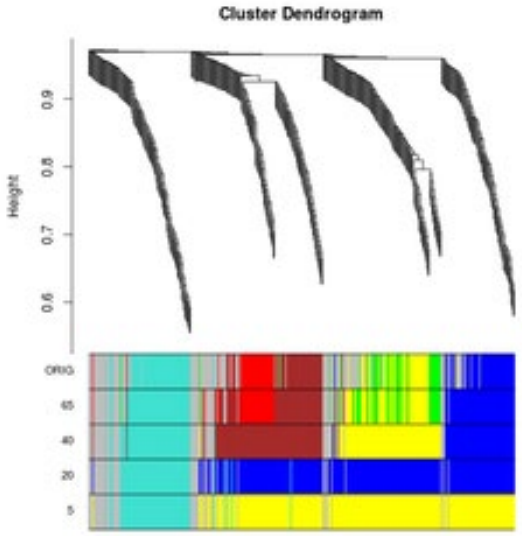
K value need to be pre-specified

Co-expression network modules

WGCNA (weighted correlation network analysis)

- transform the initial distance matrix into Topological Overlap Matrix





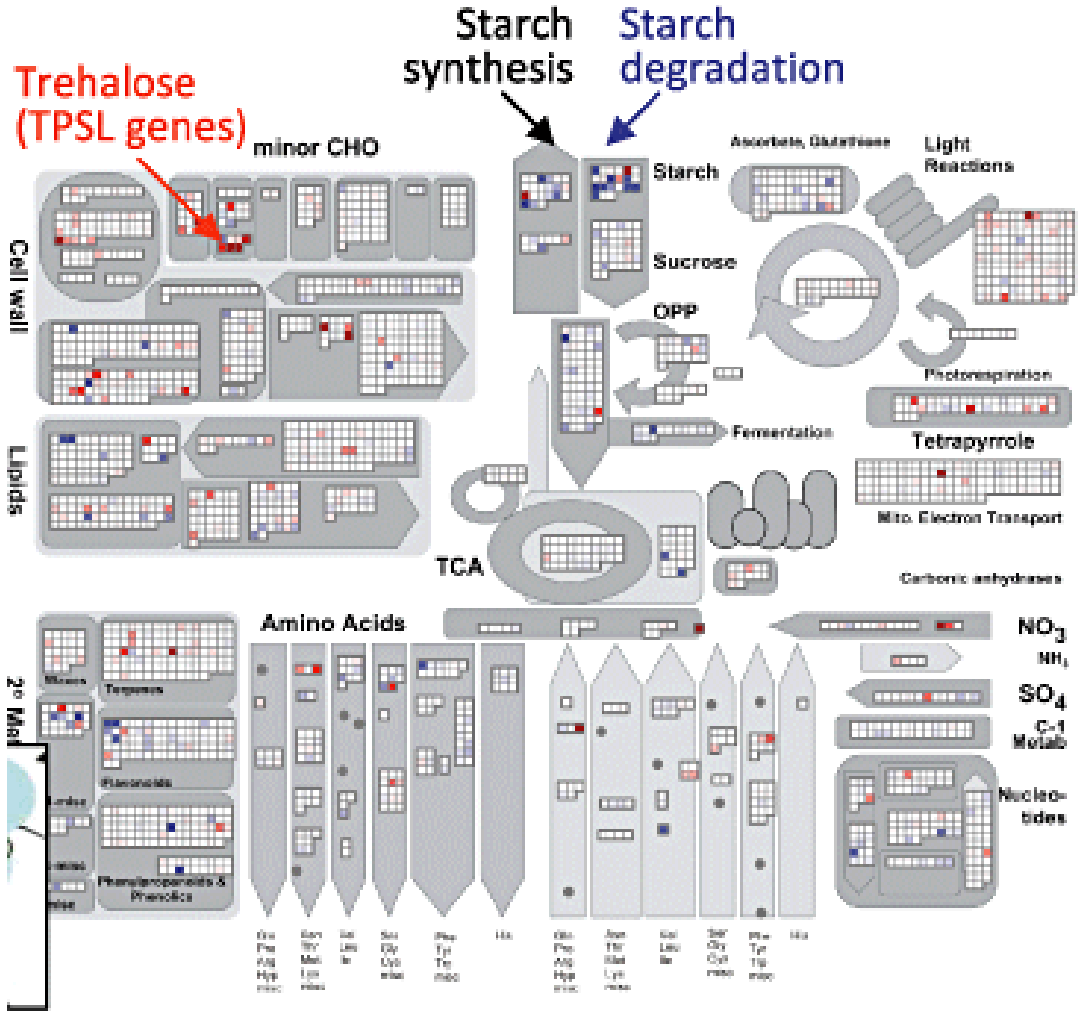
http://rgm3.lab.nig.ac.jp/RGM/R_image_list?package=WGCNA&init=true

Connect RNA-seq results to biological pathways

RNA-seq results

169 genes up-regulated

340 genes down-regulated



Mapman pathway viewer

https://doi.org/10.1111/j.1365-3040.2009.01978.x

From a biological question to a statistical problem

In the whole genome

Number of genes:	29960
Genes in P53 pathway:	40

RNA-seq results

Number of DE genes:	297
DE genes in P53:	3

If the P53 genes over-represented in the DE genes?

How to tag a gene with function/pathway categories?

- **Free text description**

Gene ID	Gene description
GRMZM2G002950	Putative leucine-rich repeat receptor-like protein kinase family
GRMZM2G006470	Uncharacterized protein
GRMZM2G014376	Shikimate dehydrogenase; Uncharacterized protein
GRMZM2G015238	Prolyl endopeptidase
GRMZM2G022283	Uncharacterized protein

- **Controlled vocabulary (Gene Ontology)**

Gene Ontology (GO)

A controlled-vocabulary system for gene function/pathways

Gene ID	GO
GRMZM5G888620	GO:0003674
GRMZM5G888620	GO:0008150
GRMZM5G888620	GO:0008152
GRMZM5G888620	GO:0016757
GRMZM5G888620	GO:0016758
GRMZM2G133073	GO:0003674
GRMZM2G133073	GO:0016746

Three Groups of GO Terms

Molecular Function

id: GO:0004396
name: hexokinase activity

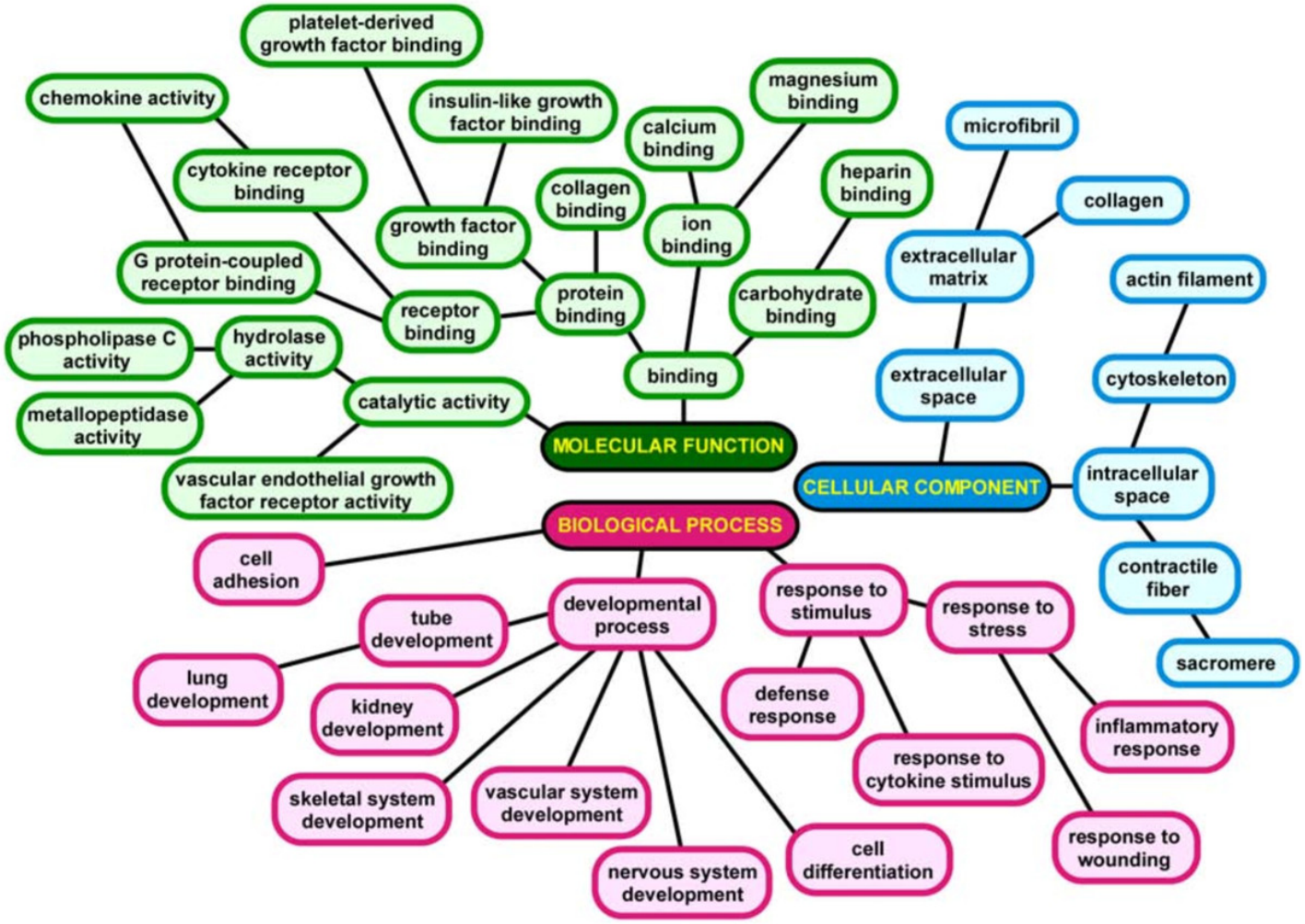
Biological Process

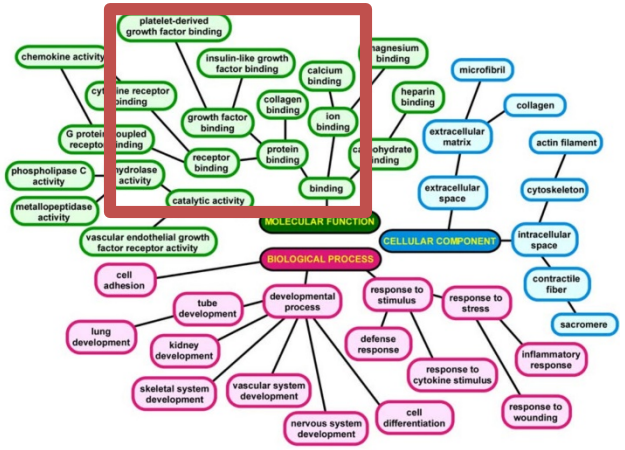
id: GO:0000018
name: regulation of DNA recombination

Cellular Component

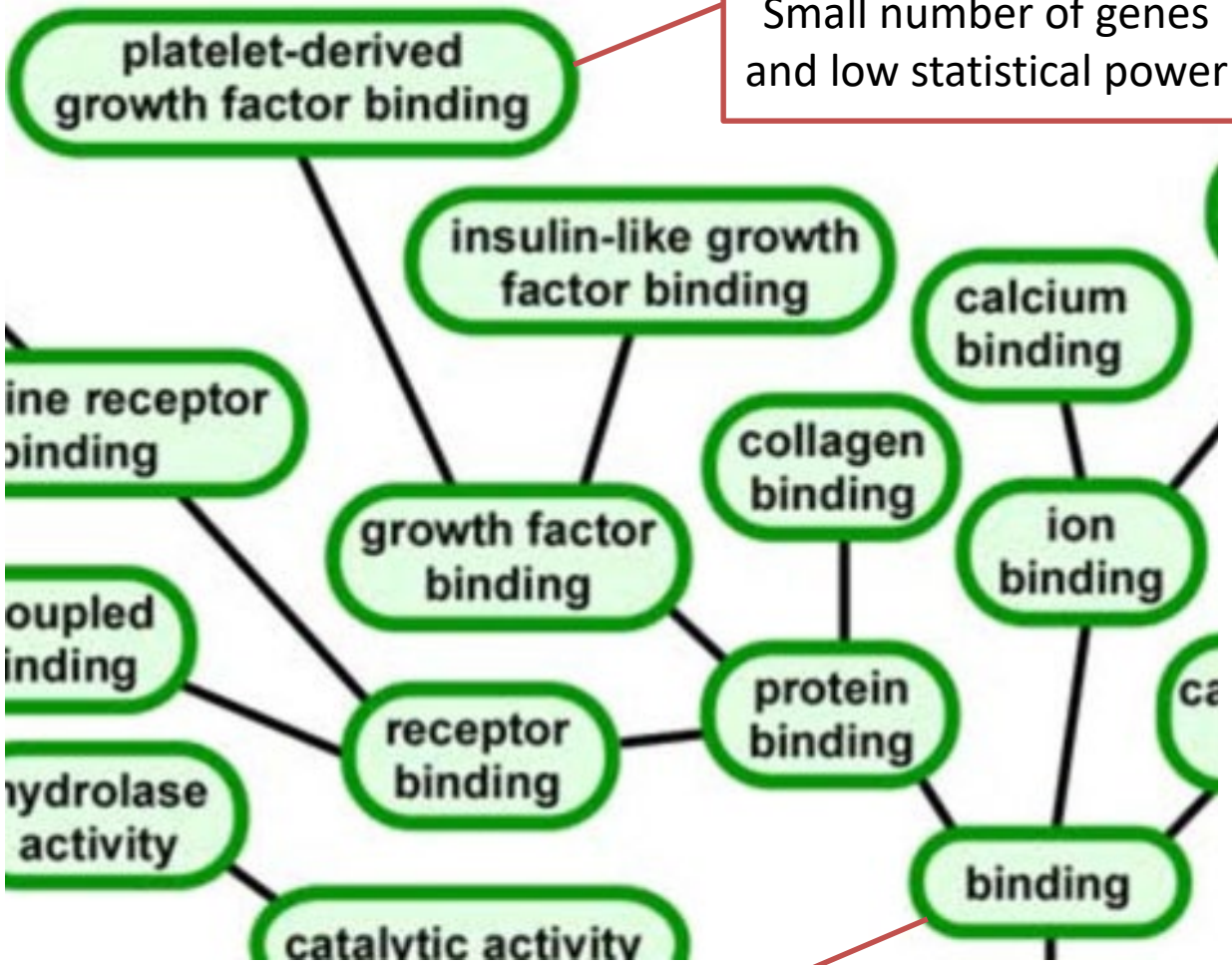
id: GO:0032590
name: dendrite membrane

Hierarchical structure of gene ontology?





Shallow vs deep nodes in GO



Small number of genes and low statistical power

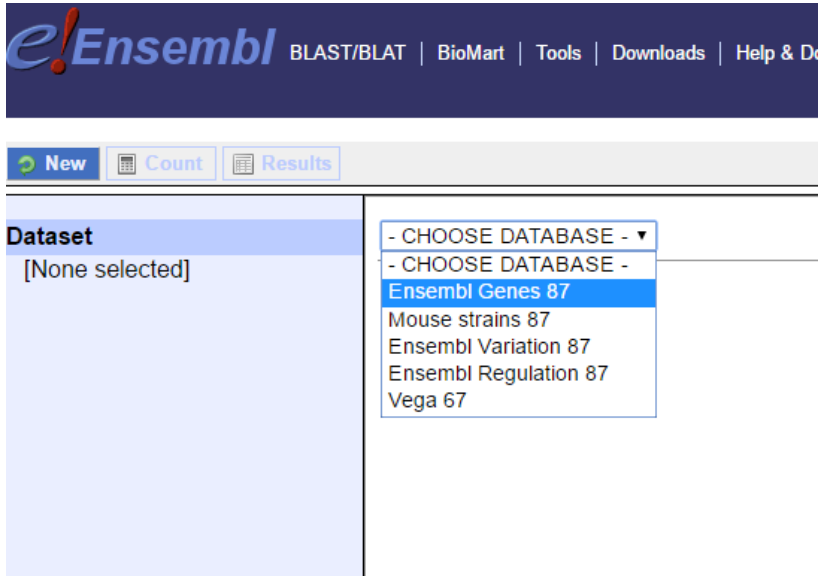
Large number of genes but the term is too general

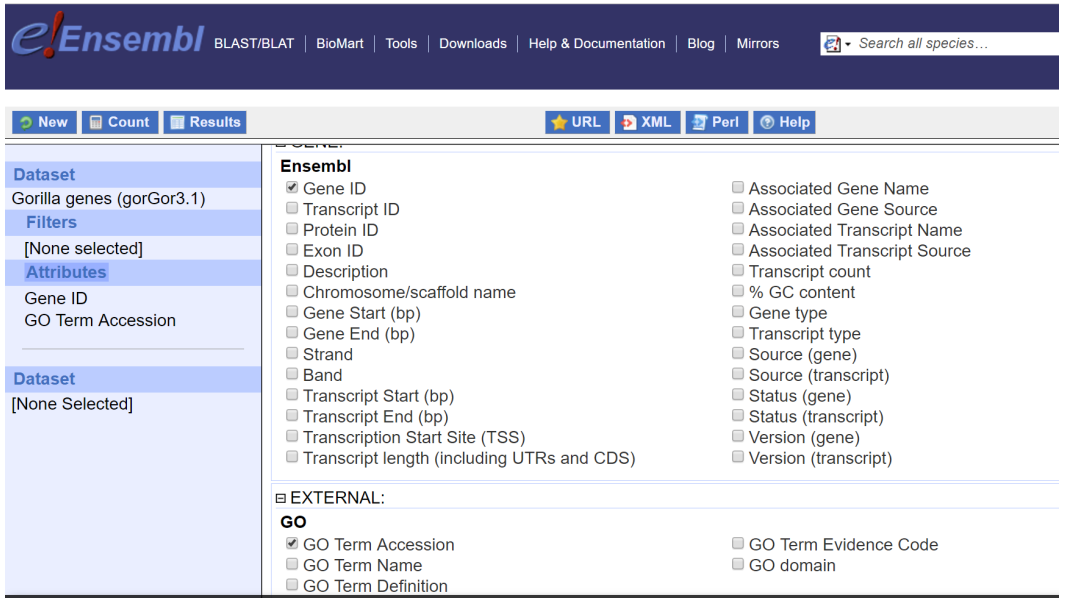
How to get Gene Ontology ?

GRMZM2G035341	molecular_function	GO:0008270	zinc ion binding
GRMZM2G035341	molecular_function	GO:0046872	metal ion binding
GRMZM2G035341	cellular_component	GO:0005622	intracellular
GRMZM2G035341	cellular_component	GO:0019005	SCF ubiquitin ligase complex
GRMZM2G035341	biological_process	GO:0009733	response to auxin
GRMZM2G047813	molecular_function	GO:0003677	DNA binding
GRMZM2G047813	cellular_component	GO:0005634	nucleus
GRMZM2G047813	cellular_component	GO:0005694	chromosome
GRMZM2G047813	biological_process	GO:0006259	DNA metabolic process
GRMZM2G047813	biological_process	GO:0034641	cellular nitrogen compound metabolic process

Model organisms: Ensembl BioMart:

Animal genomes: http://www.ensembl.org
Plant genomes: http://plants.ensembl.org





Annotate GO by yourself

Public tool: **InterProScan**

Commercial software: **BLAST2GO**

Function enrichment analysis

	ORA (Over Representation Analysis)	GSEA (Gene Set Enrichment Analysis)
Methods		
Statistics	Fisher Exact	K-S (Kolmogorov-Smirnov)
Input	DE gene list	DE gene list + p-values (or score)

Software

- **Free:**
  - DAVID (online tool http://david.abcc.ncifcrf.gov/) Fisher
  - topGO (R package) Fisher & KS (Kolmogorov-Smirnov)
  - GSEA (Win/Mac/Linux software) KS

- **Commercial:**
  - IPA (Ingenuity Pathway Analysis)

(Cornell license information https://library.weill.cornell.edu/node/1050)

Fisher vs KS

Fisher's exact test:

For each GO category, compares the expected number of significant genes at random to the observed number of significant genes, and get a p-value.

Significant genes: hard cut-off based on p-value or score

The KS test

Compares the distribution of gene p-values expected at random to the observed distribution of the gene p-values to arrive at a probability.

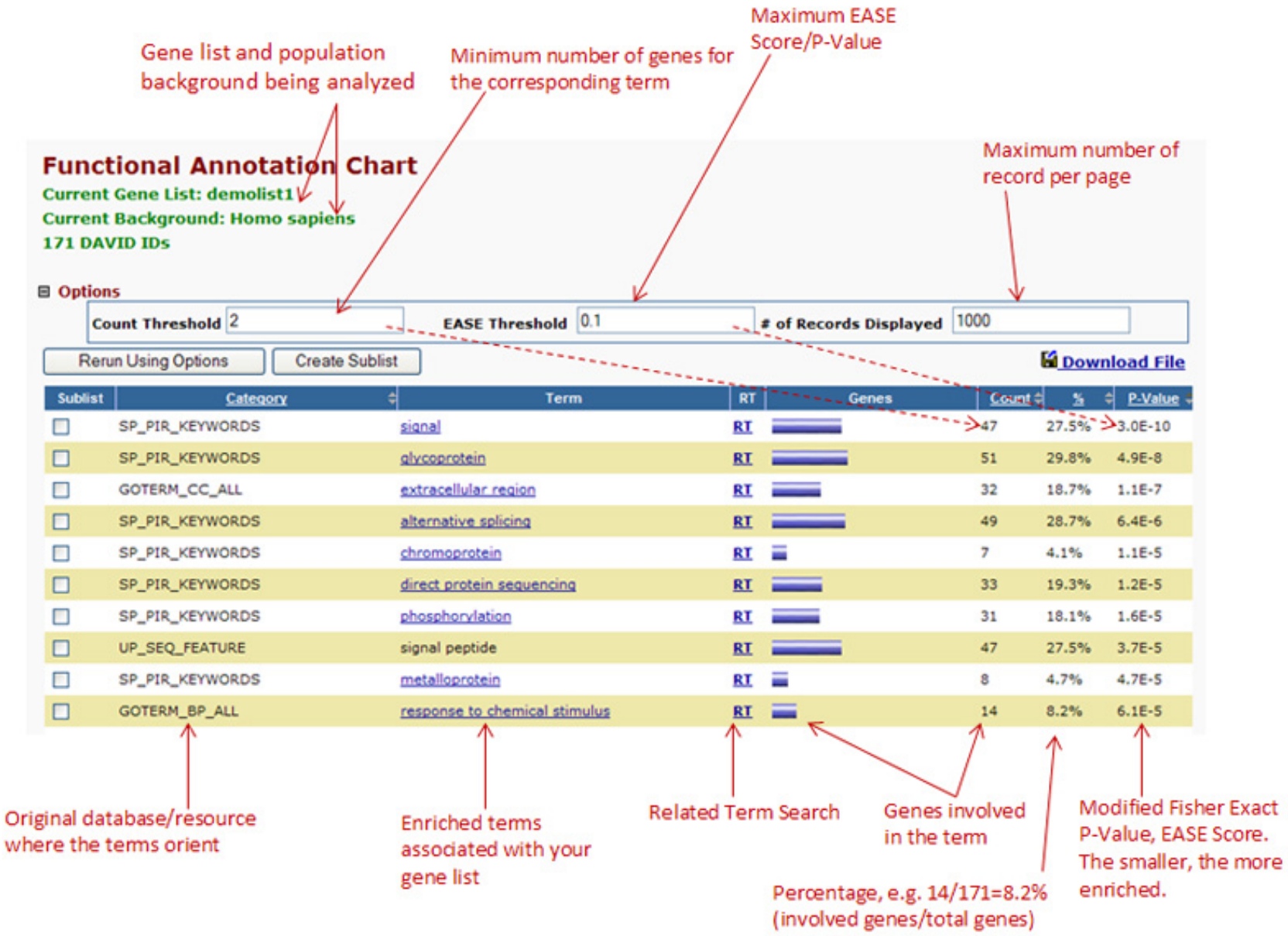
No cutoff is needed. Input file include all genes with p-value and score

Fisher test

	In the genome	Diff. expressed
Total genes	**20,000**	**2,000**
P53 pathway genes	**200**	Expected: **20** Observed: **35**

Online tools

DAVID (http://david.abcc.ncifcrf.gov/



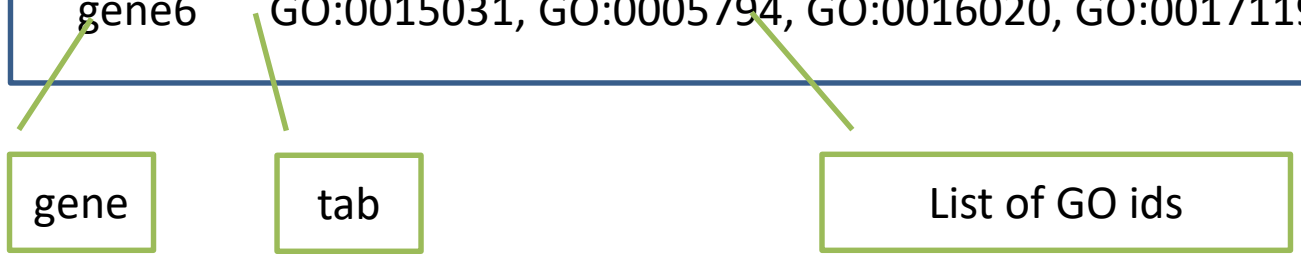
When using a web-based tool or a commercial software, available species are very limited.

- **Option 1: "Humanized" your gene list**

  Convert your gene list to human orthologs using Ensembl BioMart.

- **Option 2: Use custom GO annotation file with topGO**

```
gene1	GO:0005488, GO:0003774, GO:0001539, GO:0006935, GO:0009288
gene2	GO:0005634, GO:0030528, GO:0006355,
gene3	GO:0016787, GO:0017057, GO:0005975, GO:0005783, GO:0005792
gene4	GO:0043565, GO:0000122, GO:0003700, GO:0005634
gene5	GO:0004803, GO:0005634, GO:0008270, GO:0003677
gene6	GO:0015031, GO:0005794, GO:0016020, GO:0017119, GO:0000139
```



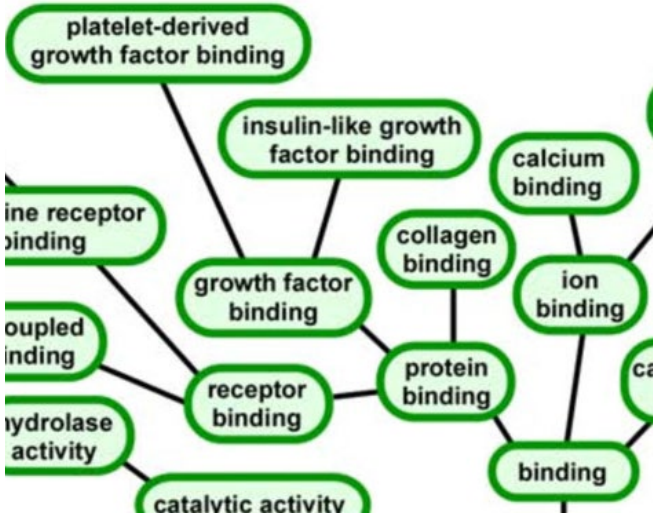
topGo – A Bioconductor Package

Statistics test and algorithm in topGO

	fisher	ks	t	globaltest	sum
classic	✓	✓	✓	✓	✓
elim	✓	✓	✓	✓	✓
weight	✓	–	–	–	–
weight01	✓	✓	✓	✓	✓
lea	✓	✓	✓	✓	✓
parentchild	✓	–	–	–	–

Statistics test: KS vs Fisher

Default in topGO (topgoFisher)



Classic:

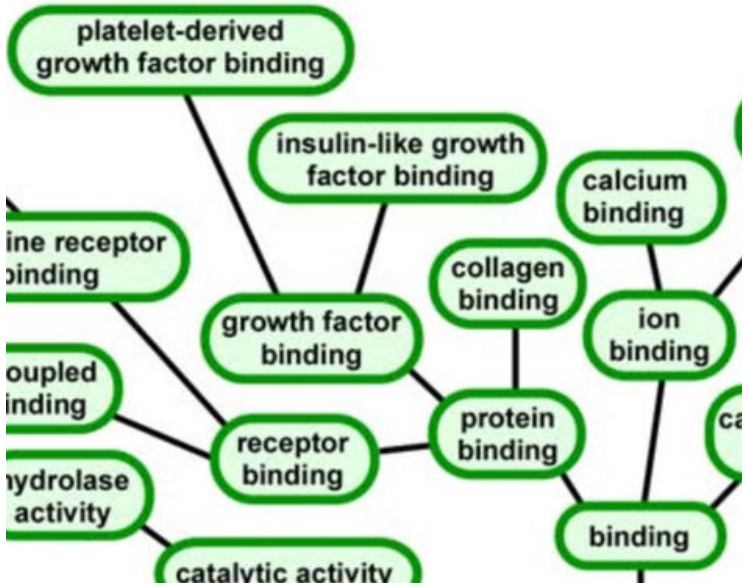
- Each node is tested independently.

Elim:

- Testing from bottom up (starting from most specific node);
- If tested significant in a child node, these genes would not contribute to parental node.

Weight01 (topgoFisher):

- A mixture between classis and elim.



Elim vs Weight algorithm

Default in topGO: Weight01, a mixture between the elim and the weight algorithms

Elim

- Starting from most specific terms to most general terms;
- Iteratively removes the genes mapped to significant GO terms from more general term.

Weight

- A weighting scheme for genes contribution towards neighboring term.

Use topGo for Fisher test

```
Rscript topGO.r  go.annot refset testset 0.05 BP myBP
```

Input files

- **go.annot**: Go annotation file
- **Refset**: Reference gene sets (all expressed gene list)
- **Testset**: Test gene set (e.g. DE gene list)

Parameters:

- 0.05: P-value cutoff
- BP: test Biology Process domain (BP CC MF)
- myBP: output file

GO annotation file format

GO annotation from Ensembl

YBR024W GO:0005743
YBR024W GO:0005507
YBR024W GO:0006878
YBR024W GO:0008379
YBR024W GO:0045454
YBR024W GO:0008535
YBR024W GO:0006825
YBR024W GO:0006825
YBR024W GO:0005740
YBR024W GO:0016021
YDL245C GO:0022857
YDL245C GO:0016020
YDL245C GO:0005353
YDL245C GO:0005351
YDL245C GO:0005886
YDL245C GO:0015795
YDL245C GO:0015578
YDL245C GO:0015797

GO annotation required by topGO

YBR024W GO:0005507, GO:0005740, GO:0005743, GO:0006825, …
YDL245C GO:0005351, GO:0005353, GO:0005355, GO:0005886, …

(one gene per line)

Output file: BP.txt

---------- topGOdata object ----------

	GO.ID	Term	Annotated
1	GO:0006189	'de novo' IMP biosynthetic process	9
2	GO:0055114	oxidation-reduction process	363
3	GO:0006730	one-carbon metabolic process	18
4	GO:0019878	lysine biosynthetic process via aminoadi...	11
5	GO:0006799	polyphosphate biosynthetic process	5
6	GO:0000105	histidine biosynthetic process	10
7	GO:0033214	siderophore-dependent iron import into c...	4
8	GO:0005978	glycogen biosynthetic process	22
9	GO:1990961	xenobiotic detoxification by transmembra...	4
10	GO:0031505	fungal-type cell wall organization	152

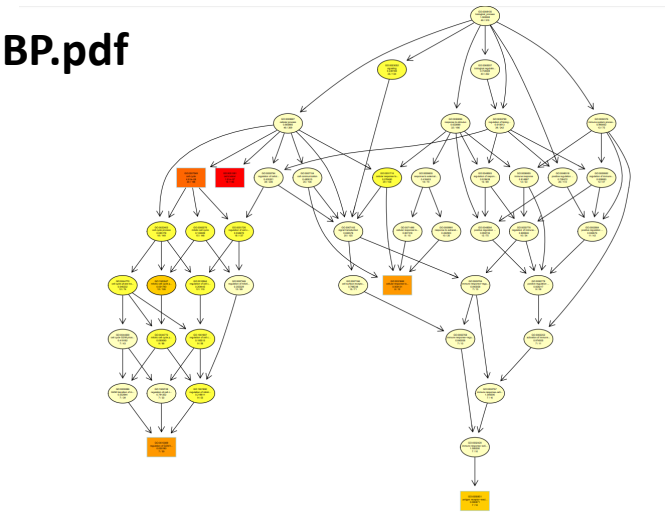
P-values

	Significant	Expected	topgoFisher	classicFisher	elimFisher	parentchildFisher
1	7	0.41	1.30E-08	1.30E-08	1.30E-08	0.35385
2	48	16.71	5.90E-08	5.90E-12	4.70E-09	4.40E-13
3	7	0.83	3.30E-06	8.20E-06	8.20E-06	6.40E-06
4	6	0.51	3.40E-06	3.40E-06	3.40E-06	0.26923
5	4	0.23	2.10E-05	2.10E-05	2.10E-05	2.40E-05
6	5	0.46	4.10E-05	4.10E-05	4.10E-05	7.10E-05
7	4	0.18	9.50E-05	4.40E-06	4.40E-06	0.00824
8	6	1.01	0.00036	0.00036	0.00036	0.06686
9	3	0.18	0.00037	0.00037	0.00037	0.03297
10	15	7	0.00079	0.00396	0.00396	0.62837

Gene List

[1] "Term GO:0006189 genes: YAR015W,YGL234W,YGR061C,YLR359W,YMR120C,YMR300C,YOR128C“
[1] "Term GO:0055114 genes: YAL044C,YAL061W,YBR085W,YBR145W,YBR196C,YBR244W,YCL030C,YDL022W,YDL124W,YDR019C,YDR044W,YER073W,YFL053W,YFR015C,YGR177C,YGR192C,YGR234W,YHL021C,YHR163W,YHR183W,YHR216W,YIL094C,YIL099W,YJL052W,YJL200C,YJR048W,YJR104C,YKL029C,YKL109W,YKL182W,YKR058W,YKR080W,YLR056W,YLR258W,YLR355C,YMR015C,YMR081C,YMR105C,YMR189W,YMR272C,YNL134C,YNR050C,YOL152W,YOR120W,YOR178C,YPL061W,YPR160W,YPR184W“
[1] "Term GO:0006730 genes: YAL044C,YDR019C,YDR502C,YKR080W,YLR058C,YLR180W,YMR189W“
[1] "Term GO:0019878 genes: YBR115C,YDL182W,YIL094C,YIR034C,YJL200C,YNR050C“
[1] "Term GO:0006799 genes: YDR089W,YER072W,YJL012C,YPL019C“
[1] "Term GO:0000105 genes: YBR248C,YCL030C,YER055C,YIL020C,YOR202W“
[1] "Term GO:0033214 genes: YEL065W,YHL040C,YHL047C,YOL158C“
[1] "Term GO:0005978 genes: YFR015C,YKR058W,YLR258W,YMR105C,YOR178C,YPR184W“
[1] "Term GO:1990961 genes: YDR011W,YDR406W,YOR153W“
[1] "Term GO:0031505 genes: YBR067C,YDR055W,YDR077W,YEL040W,YER150W,YGR032W,YJL158C,YJR104C,YKL096W,YKL163W,YLR194C,YLR300W,YOL109W,YOL155C,YPR149W"

Output file: BP.pdf

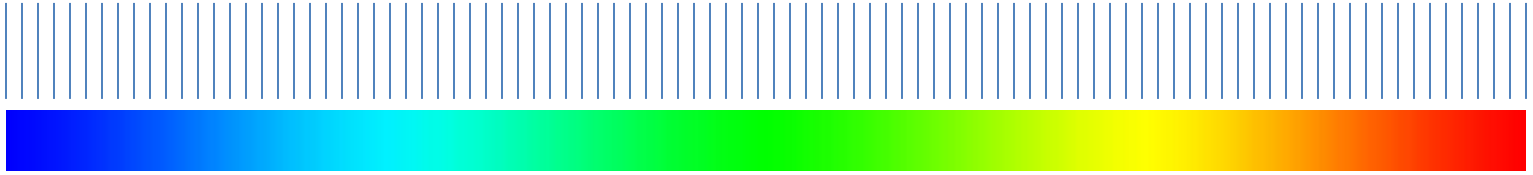


Limitation of Fisher

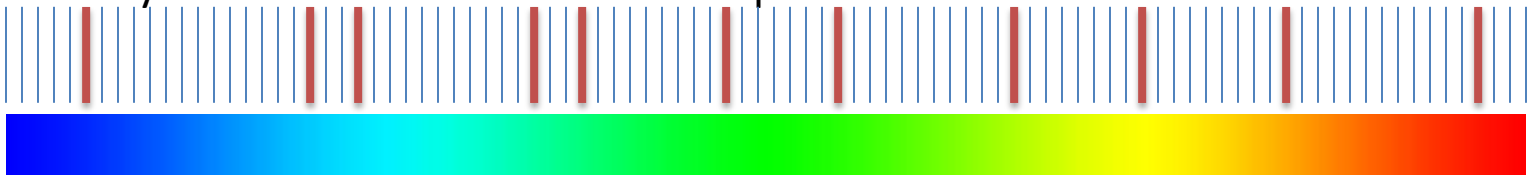
- Arbitrary cutoff for DE gene list;
- Quantitative gene expression information was not used;
- Assume independence among genes;

K-S or Gene Set Enrichment Analysis

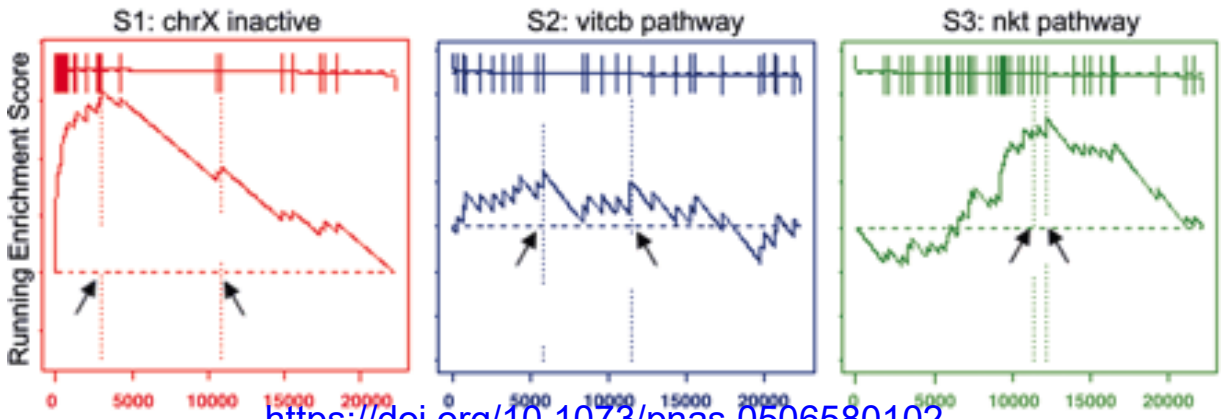
1. Sort all genes based on fold change



2. For each GO term, the null hypothesis is that genes of that GO term should be evenly and randomly distributed across the whole spectrum.



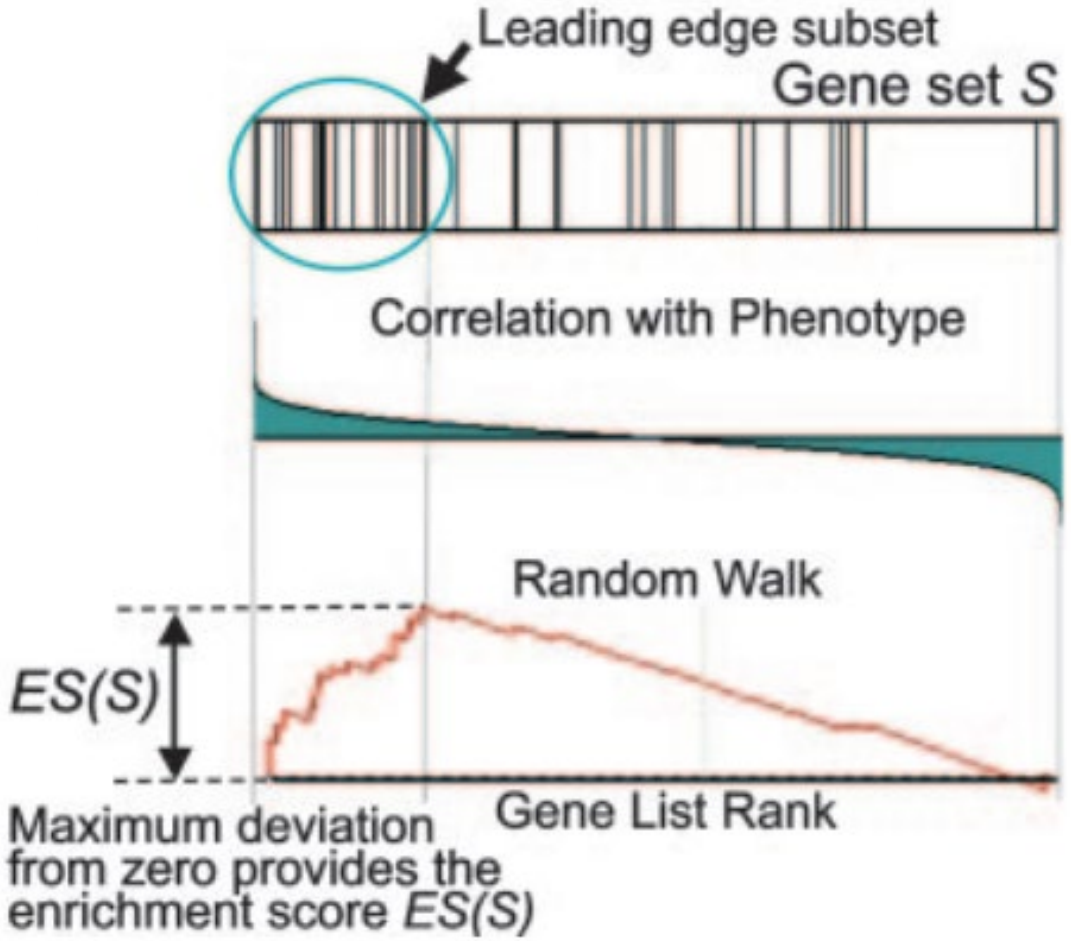
3. Identify GO categories with genes enriched in either one ends of the range.



https://doi.org/10.1073/pnas.0506580102

GSEA - Gene Set Enrichment Analysis

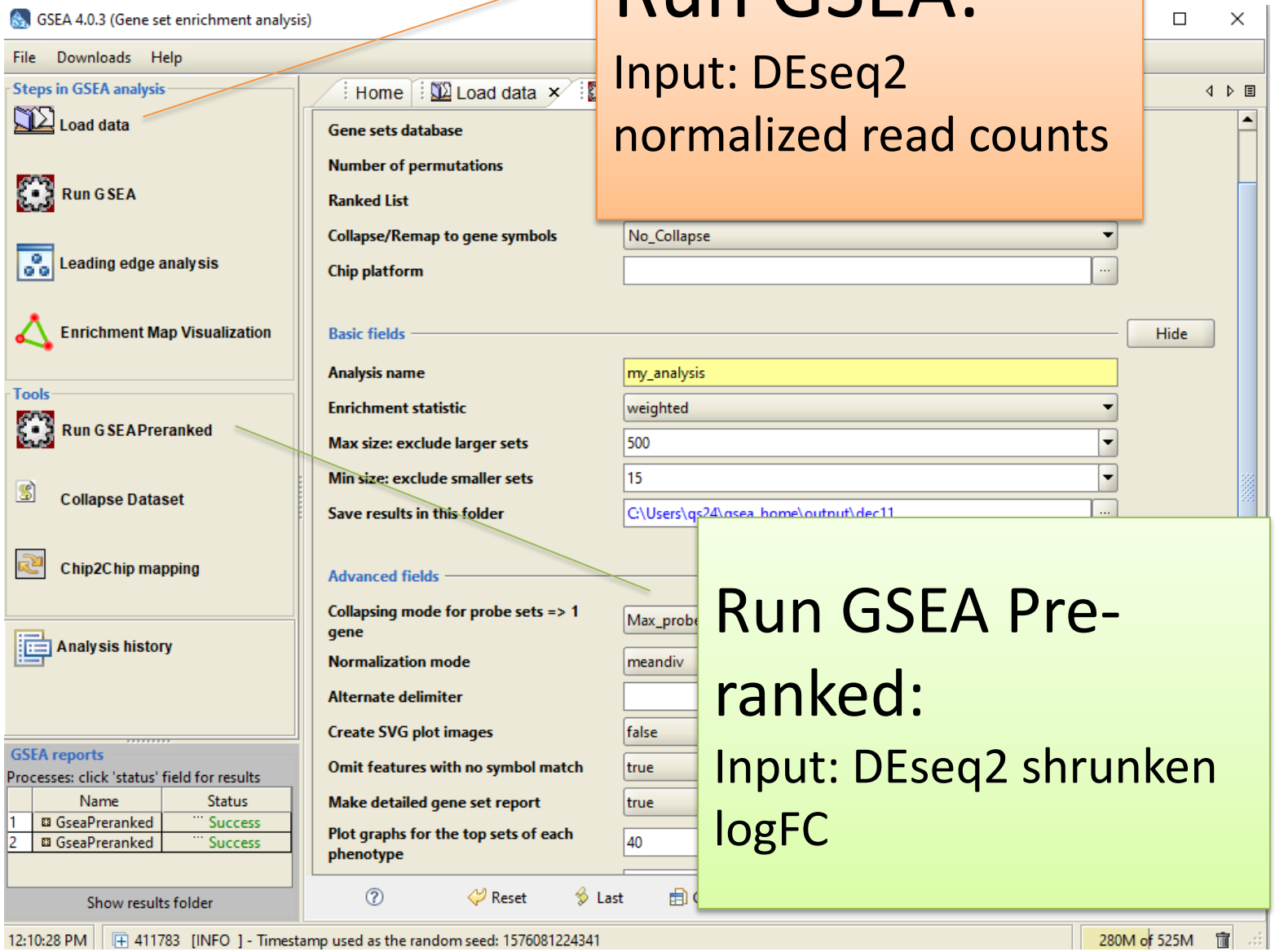
- Rank genes based on shrunken Log_2(Fold_Change) *
- ES score of each gene set (e.g. diabetes related genes)



Two alternative ways to analyze RNA-seq data with GSEA

Run GSEA:

Input: DEseq2 normalized read counts



Run GSEA Pre-ranked:

Input: DEseq2 shrunken logFC

GSEA
Input files

.rnk file
- ranked gene list

Gene	log2(ratio)
YDL248W	0.446508
YDL243C	0.285379
YDL241W	2.006822
YDL240W	-0.87753
YDL239C	-0.00886
YDL238C	0.837298
YDL237W	-0.14496
YDL236W	0.417735
YDL235C	-0.31365
YDL234C	0.832606

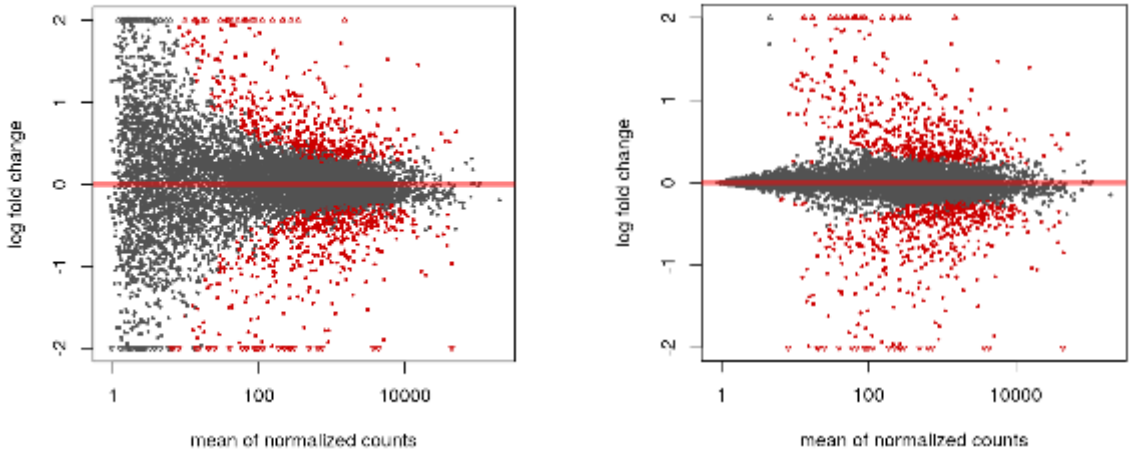
.gmt file
- gene sets

90S_preribosome	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0030686	YBL004W	YBR247C	YCL031C	YCR057C	YDL148C	YDL213C
AP_type_membrane_coat_adaptor	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0030119	YBL037W	YBR288C	YDR358W	YGR261C	YHL019C	YHR108W
ATPase_complex	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:1904949	YAL011W	YAR007C	YBL006C	YBL035C	YBR087W	
COPII_coated_ER_to_Golgi	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0030134	YAL007C	YAL042W	YAR002C-A	YAR033W	YBR210W	YCL001W
COPII_coated_vesicle_budding	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0090114	YCR067C	YDL195W	YFL038C	YGR058W	YHR098C	YIL109C
COPI_coated_vesicle	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0030137	YAR033W	YCL001W	YDL145C	YDR238C		
DASH_complex	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0042729	YBR233W-A	YDR016C	YDR201W	YDR320C-A	YGL061C	YGR113W
RNA_polymerase_II_specific	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0001228	YAL051W	YBL005W	YBR033W	YBR083W	YBR297W	YCR018C

Use shrunken logFC from DESeq2

To shrink the log(Fold-Change) of genes with high noise

MA Plots



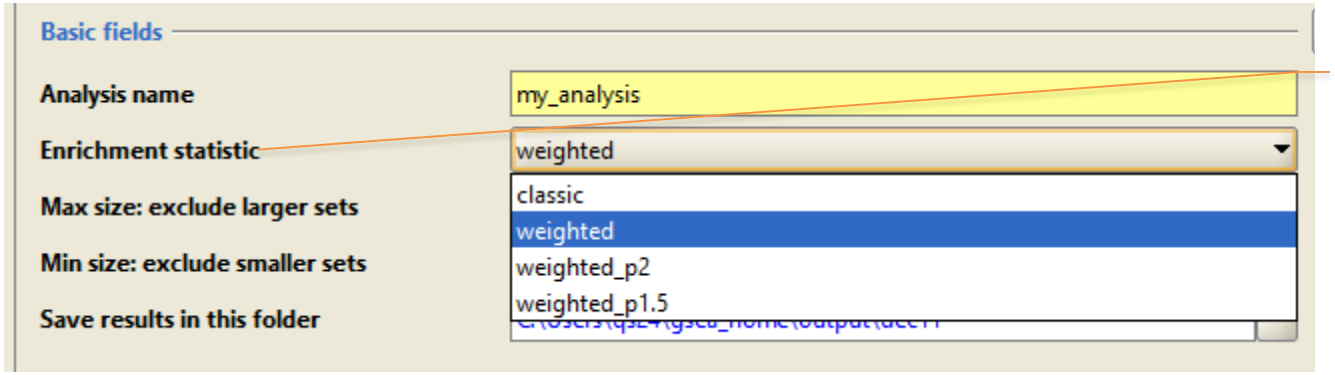
Before shrinkage

After shrinkage

DESeq2 command for shrink logFC

```
resLFC <- lfcShrink(dds,
coef="condition_treated_vs_untreated", type="apeglm")
```

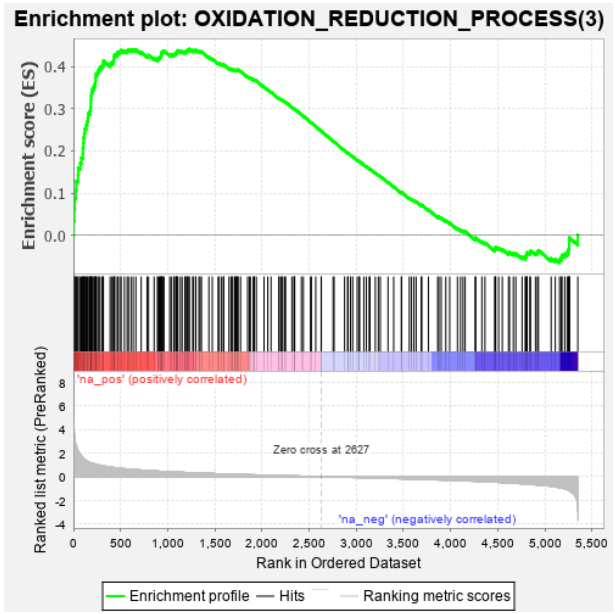
Enrichment statistics



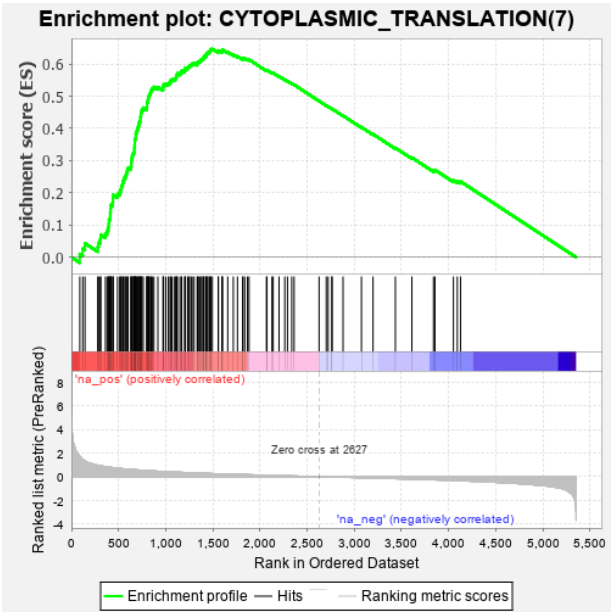
Weighted P-value:
Default: 1

Higher value would enhance the weight of fold change in ES calculation.

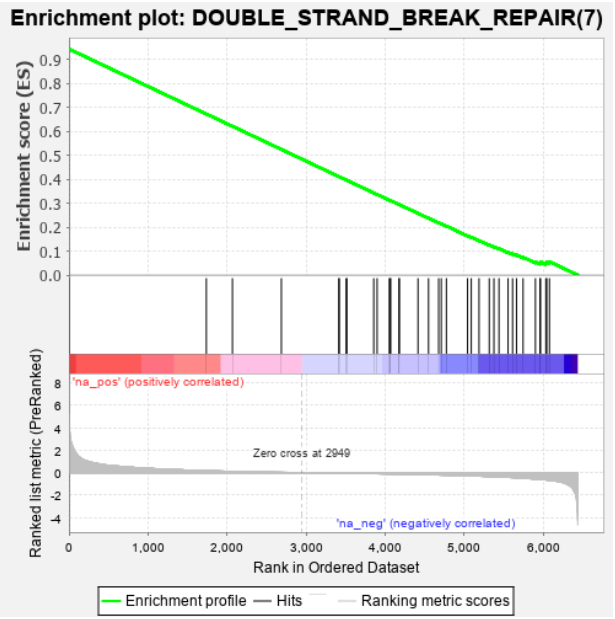
Top hit in ORA



Top hit in GSEA (p=1)



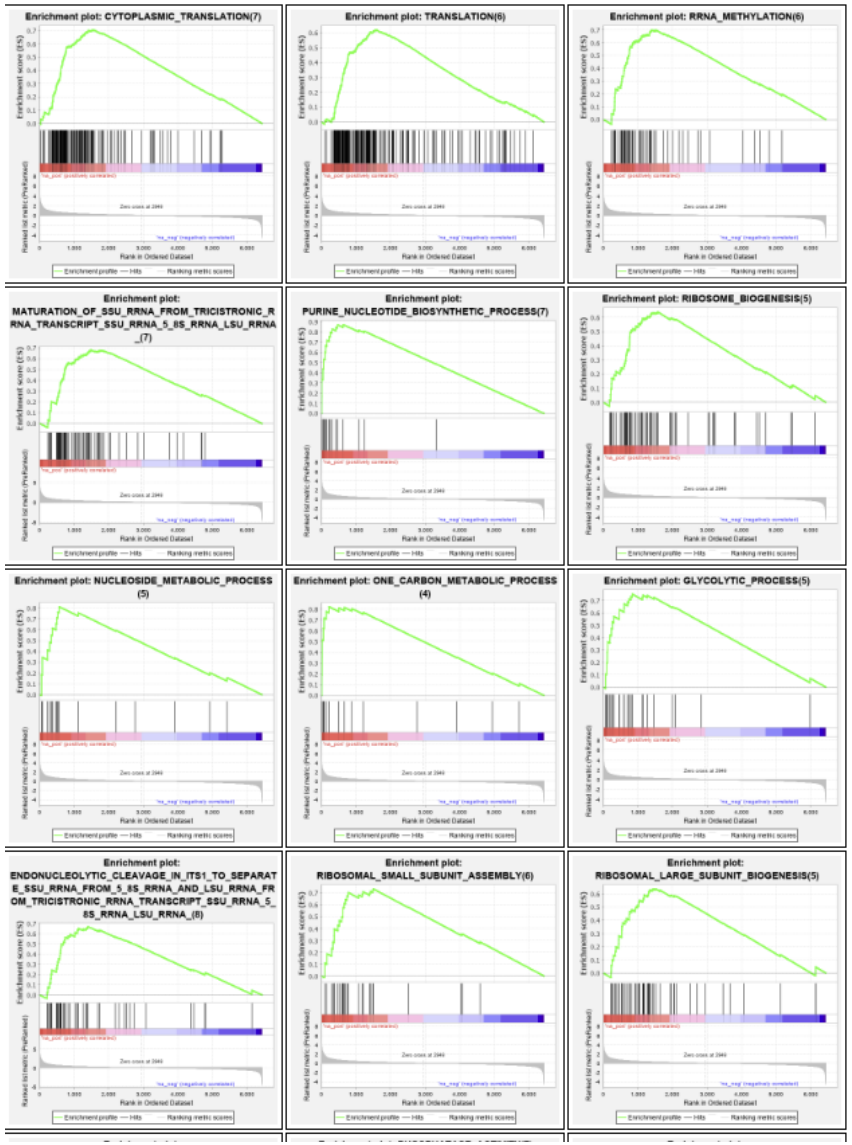
Top hit in GSEA (p=2)



Snap shots of top 12 gene sets with p=1 and p=2

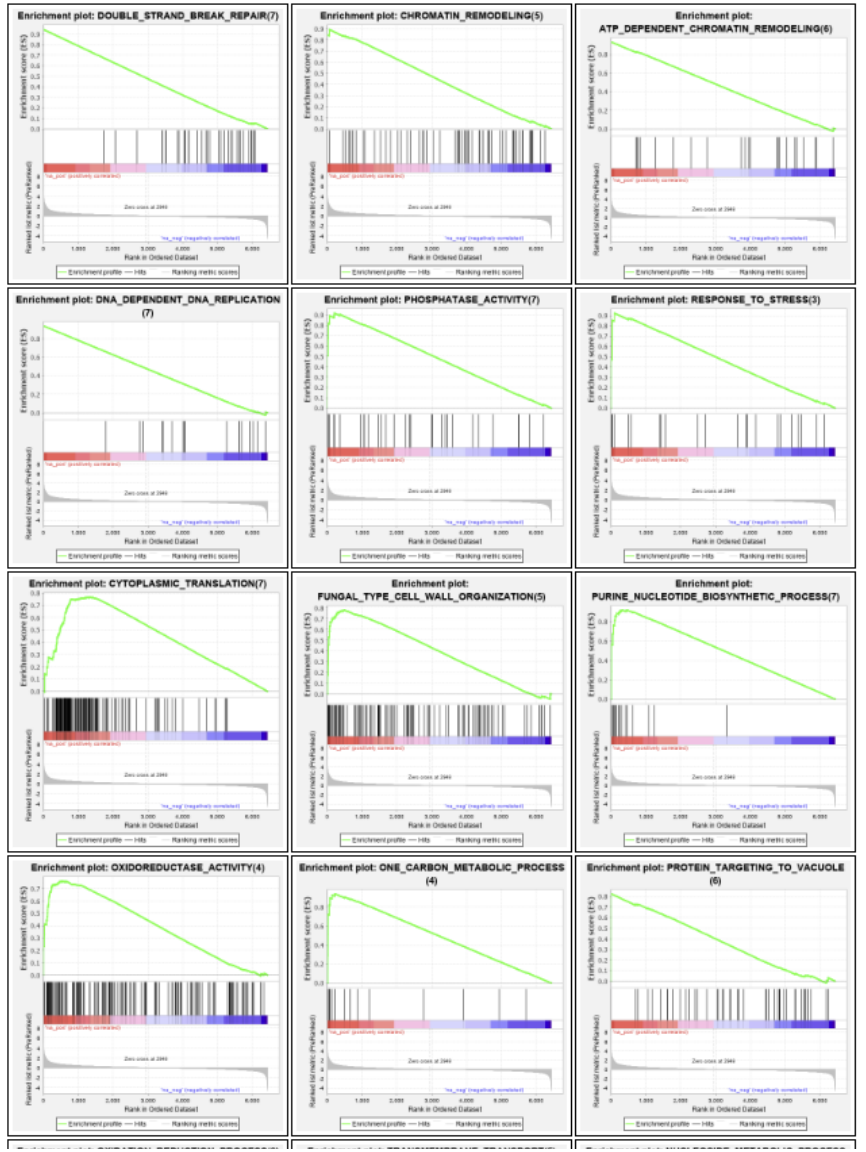
Weighted

Table: Snapshot of enrichment results



Weighted p=2

Table: Snapshot of enrichment results



GSEA Output

Enrichment in phenotype: na

- 199 / 486 gene sets are upregulated in phenotype **na_pos**
- 41 gene sets are significant at FDR < 25%
- 33 gene sets are significantly enriched at nominal pvalue < 1%
- 41 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

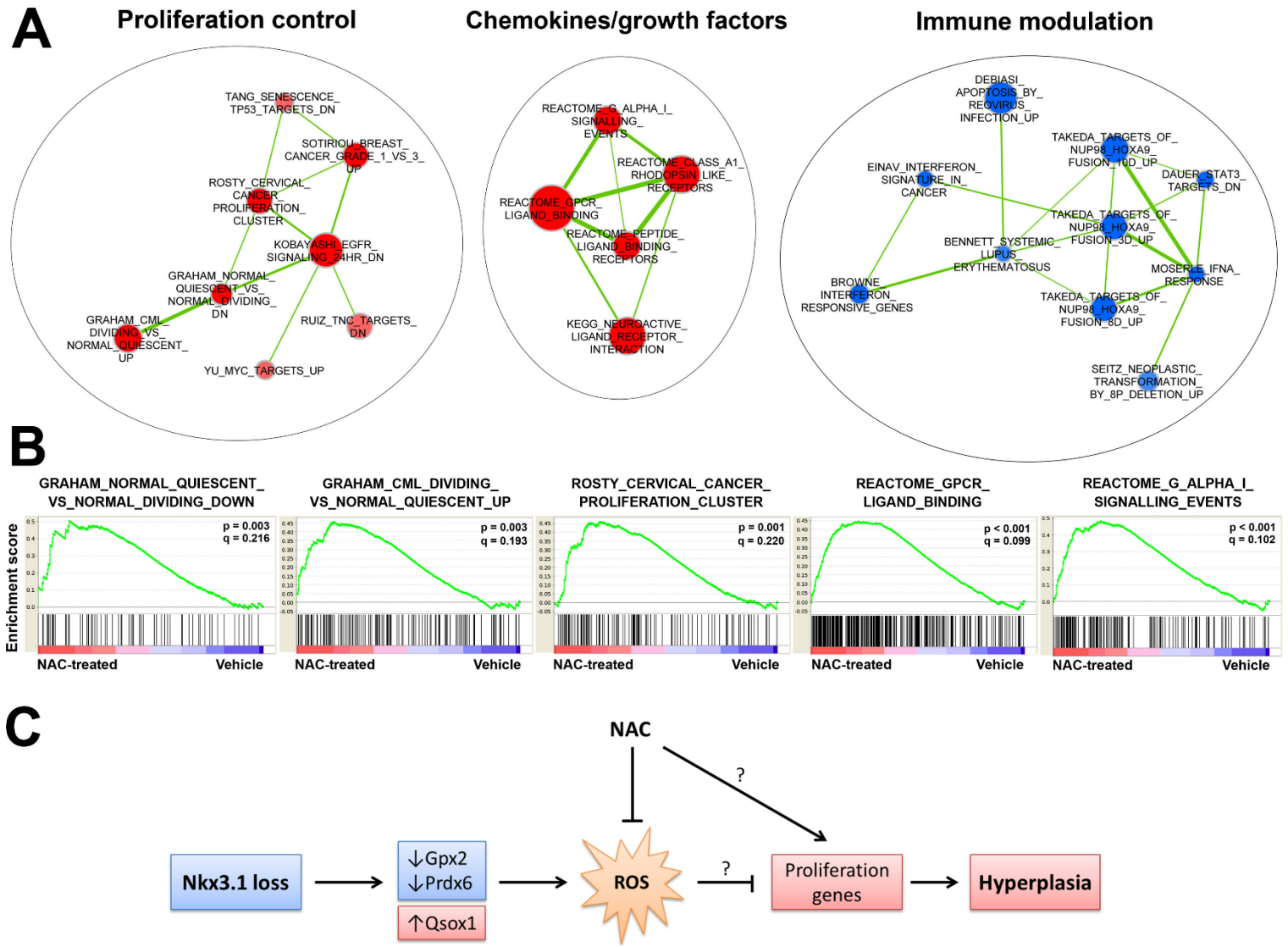
Enrichment in phenotype: na

- 287 / 486 gene sets are upregulated in phenotype **na_neg**
- 70 gene sets are significantly enriched at FDR < 25%
- 55 gene sets are significantly enriched at nominal pvalue < 1%
- 76 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enriched gene sets from GSEA

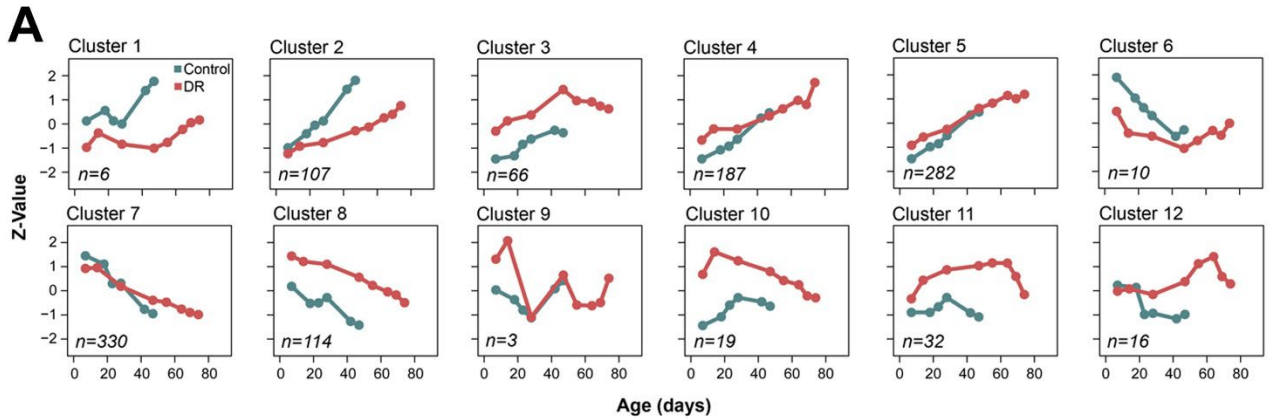
	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
1	CYTOPLASMIC_TRANSLATION(7)	Details ...	151	0.70	2.48	0.000	0.000	0.000
2	TRANSLATION(6)	Details ...	185	0.62	2.22	0.000	0.000	0.000
3	RRNA_METHYLATION(6)	Details ...	57	0.70	2.14	0.000	0.000	0.000
4	MATURATION_OF_SSU_RRNA_FROM_TRICISTRONIC_RRNA_TRANSCRIPT_SSU_RRNA_5_8S_RRNA_LSU_RRNA_(7)	Details ...	70	0.68	2.13	0.000	0.000	0.000
5	PURINE_NUCLEOTIDE_BIOSYNTHETIC_PROCESS(7)	Details ...	16	0.87	2.09	0.000	0.000	0.000
6	RIBOSOME_BIOGENESIS(5)	Details ...	64	0.64	2.00	0.000	0.002	0.017
7	ENDONUCLEOLYTIC_CLEAVAGE_IN_ITS1_TO_SEPARATE_SSU_RRNA_FROM_5_8S_RRNA_AND_LSU_RRNA_FROM_TRICISTRONIC_RRNA_TRANSCRIPT_SSU_RRNA_5_8S_RRNA_LSU_RRNA_(8)	Details ...	43	0.67	1.94	0.000	0.005	0.043
8	RIBOSOMAL_LARGE_SUBUNIT_ASSEMBLY(6)	Details ...	40	0.67	1.94	0.000	0.005	0.044
9	RIBOSOMAL_SMALL_SUBUNIT_ASSEMBLY(6)	Details ...	26	0.74	1.94	0.000	0.004	0.045
10	GLYCOLYTIC_PROCESS(5)	Details ...	24	0.75	1.93	0.002	0.005	0.057
11	NUCLEOSIDE_METABOLIC_PROCESS(5)	Details ...	16	0.81	1.92	0.002	0.006	0.073
12	ONE_CARBON_METABOLIC_PROCESS(4)	Details ...	15	0.82	1.91	0.000	0.006	0.079
13	PHOSPHATASE_ACTIVITY(7)	Details ...	31	0.70	1.91	0.000	0.006	0.086
14	RIBOSOMAL_LARGE_SUBUNIT_BIOGENESIS(5)	Details ...	52	0.64	1.90	0.000	0.006	0.088

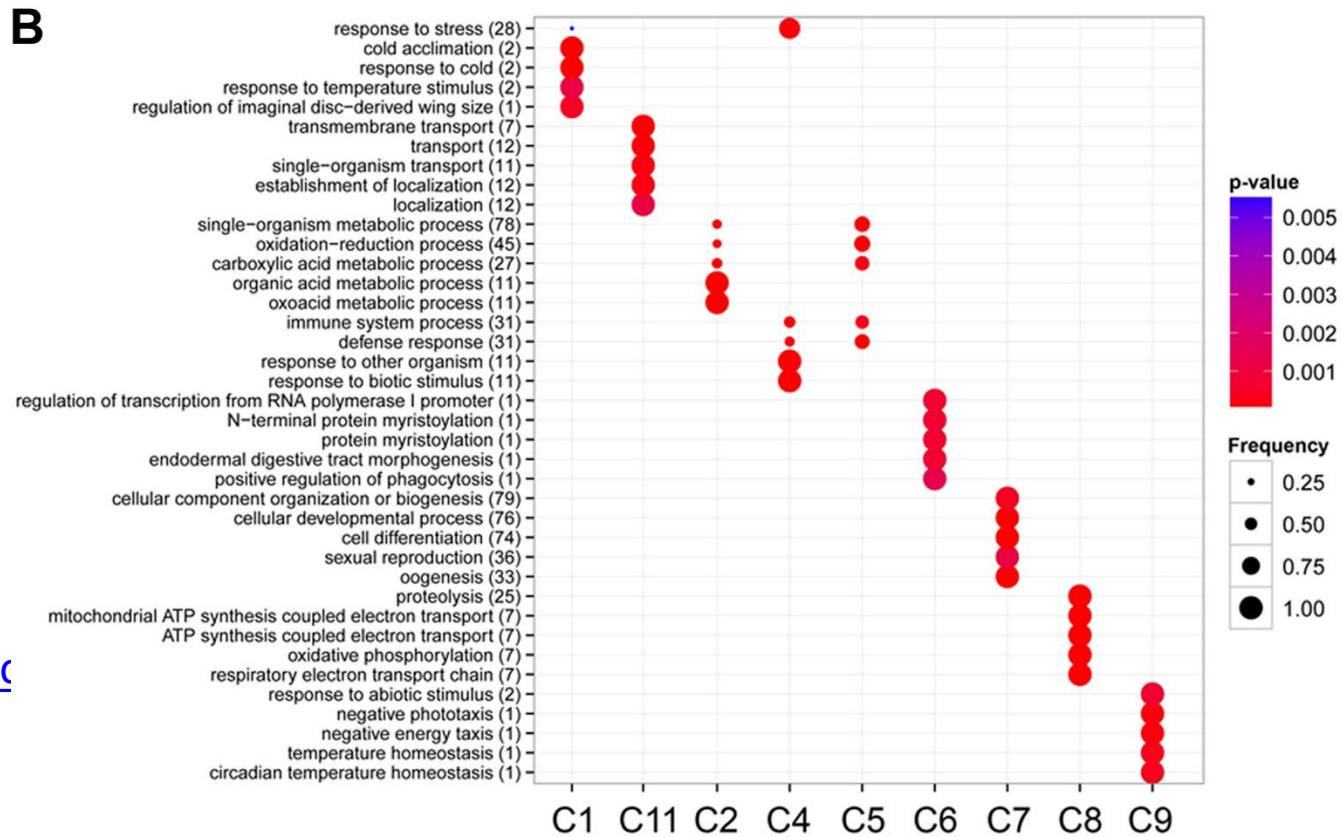
Get network representation of enriched gene sets



https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0046792.g007

clusterProfiler:
enrichment analysis for gene clusters





https://bioconductor.org/packages/release/bio
html/clusterProfiler.html