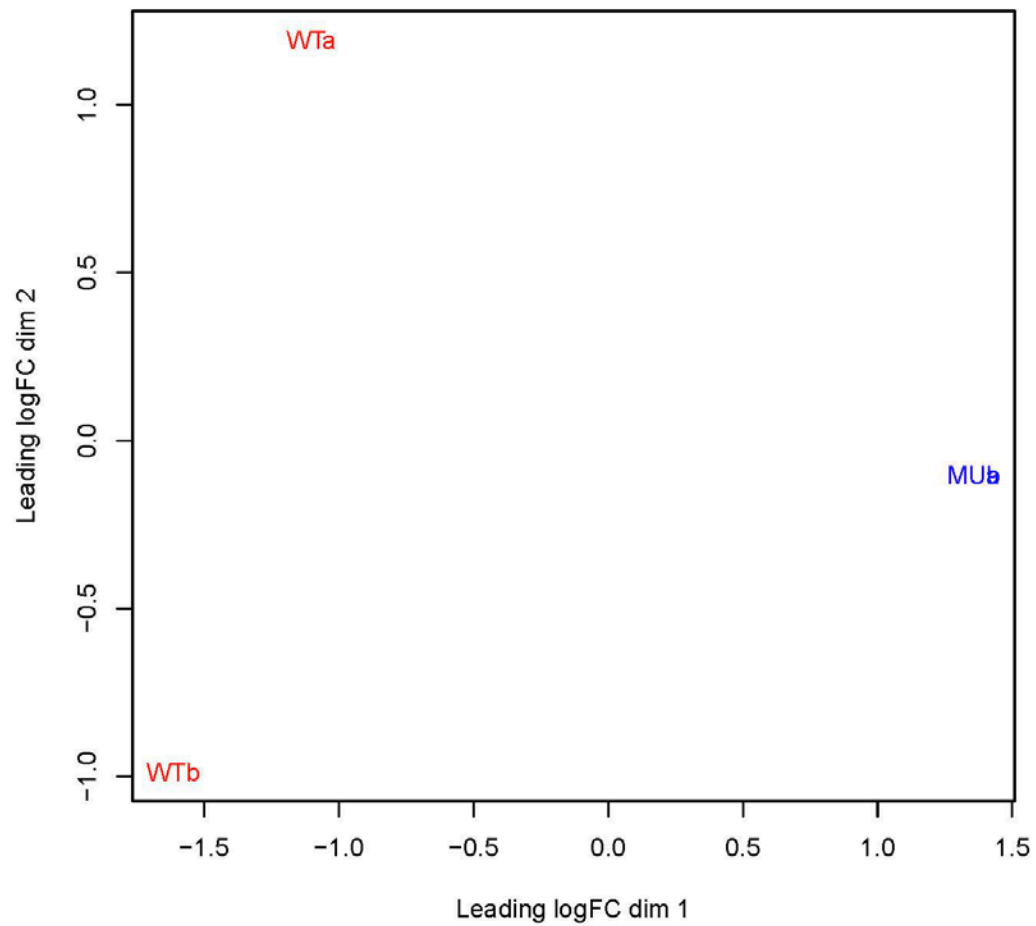


MDS Plot: the two groups are separated well at dim 1



Continue exercise 2

Using EdgeR for DE gene detection

RNA-seq workflow:

<http://cbsu.tc.cornell.edu/lab/userguide.aspx>

```
library("edgeR")
x <- read.delim("edgeR_count.xls", row.names='Gene')
x <- round(x, 0)
group <- factor(c(1,1,1,2,2,2,3,3,3))
y <- DGEList(counts=x,group=group)
# only keep genes with cpm value greater than 1 in at least 3 samples
keep <- rowSums(cpm(y)>=1) >=3
y<-y[keep,]
y <- calcNormFactors(y)
design<-model.matrix(~0+group)
y <- estimateGLMCommonDisp(y,design)
y <- estimateGLMTrendedDisp(y,design)
y <- estimateGLMTagwiseDisp(y,design)
fit<-glmFit(y,design)
```

To compare 2 vs 1

```
lrt.2v1<-glmLRT(fit,contrast=c(1,-1,0))  
top2v1 <- topTags(lrt.2v1, n=2000)  
write.table(top2v1, "diff2-1.txt", sep="\t")
```

To compare 3 vs 1

```
lrt.3v1<-glmLRT(fit,contrast=c(1,0,-1))  
top3v1 <- topTags(lrt.3v1, n=2000)  
write.table(top3v1, "diff3-1.txt", sep="\t")
```

To compare 3 vs 2

```
lrt.3vs2<-glmLRT(fit,contrast=c(0,-1,1))  
top3v2 <- topTags(lrt.3v2, n=2000)  
write.table(top3v2, "diff3-2.txt", sep="\t")
```

```
lrt.2v1<-glmLRT(fit,contrast=c(1,-1,0))
```

Use makeContrast:

```
lrt.2v1<-glmLRT(fit, contrast=makeContrasts(Drug.1h-  
Drug.0h))
```

Use makeContrast in multi-factor analysis:

```
lrt.2v1<-glmLRT(fit, contrast=makeContrasts((Drug.1h-  
Drug.0h)-(Placebo.1h-Placebo.0h)))
```

Connection between RNA-seq results and Biology

- **RNA-seq results showed that ~300 genes were differentially expressed between condition A and B;**
- **What to do next?**

Public and Commercial Resources of Pathway/Function analysis

- **Public resource:**

- DAVID Bioinformatics Resources

- (<http://david.abcc.ncifcrf.gov/>)

- **Commercial Resource:**

- BLAST2GO: Bioinformatics Facility has license

- Ingenuity:

- (License information

- <http://www.biotech.cornell.edu/node/137>)

What is Gene Ontology -1

How to describe the function of a gene?

- Gene description line

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

- **Pathway (KEGG)**
- **Controlled vocabulary (Gene Ontology)**

What is Gene Ontology -1

How to describe the function of a gene?

- Gene description line
- Pathway (KEGG)
- Controlled vocabulary (Gene Ontology)

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

How to Get Gene Ontology Data (1)

Publicly available Reference genome

Ensembl BioMart: <http://www.ensembl.org>

The screenshot shows the Ensembl BioMart interface. At the top, there is a navigation bar with links for BLAST/BLAT, BioMart, Tools, Downloads, and Help & Documentation. Below this, there are buttons for 'New', 'Count', and 'Results'. The main content area is titled 'Dataset' and shows '[None selected]'. A dropdown menu is open, displaying the following options: '- CHOOSE DATABASE -', '- CHOOSE DATABASE -', 'Ensembl Genes 87', 'Mouse strains 87', 'Ensembl Variation 87', 'Ensembl Regulation 87', and 'Vega 67'. The 'Ensembl Genes 87' option is highlighted in blue.

The screenshot shows the Ensembl BioMart interface with the 'Ensembl Genes 87' dataset selected. The 'Attributes' section is expanded, showing a list of attributes with checkboxes. The 'Ensembl' section includes: Gene ID (checked), Transcript ID, Protein ID, Exon ID, Description, Chromosome/scaffold name, Gene Start (bp), Gene End (bp), Strand, Band, Transcript Start (bp), Transcript End (bp), Transcription Start Site (TSS), and Transcript length (including UTRs and CDS). The 'EXTERNAL' section includes: GO Term Accession (checked), GO Term Name, and GO Term Definition. On the right side, there are additional attributes: Associated Gene Name, Associated Gene Source, Associated Transcript Name, Associated Transcript Source, Transcript count, % GC content, Gene type, Transcript type, Source (gene), Source (transcript), Status (gene), Status (transcript), Version (gene), and Version (transcript). The 'GO' section includes: GO Term Evidence Code and GO domain.

How to Get Gene Ontology Data (2)

Your own reference genome

BLAST2GO on BioHPC Lab

Details for **blast2go** ([hide](#))

Name:	blast2go
Version:	DB: Mar.2016; Software: v1.2.1
OS:	Linux
About:	Gene Ontology annotation and function enrichment analysis.
Added:	4/15/2013 5:20:07 PM
Updated:	4/25/2016 12:13:57 PM
Link:	https://www.blast2go.com/
Manual:	https://www.blast2go.com/images/b2g_pdfs/blast2go_cli_manual.pdf
Download:	https://www.blast2go.com/blast2go-pro/b2g-register-basic

Notes:

```
#####  
  
### Run BLAST on any BioHPC computer #####  
#####  
#you can run blast on any of the biohpc computers, adjust the num_threads based on computer you are  
using:general machine: 8; medium memory:24; large memory: 64  
# you have an option to use swissprot, refseq or nr for blast database. In most cases swissprot is fast and good  
enough. However, if a large percentage of your genes have no blast hits to swissprot, you can try refseq. The nr  
database is too big, the blast run would take very long time.  
#replace test.fa with your own fasta file. Make sure you are using the right blast software (blastx or blastp). To  
save time, it is preferable to use blastp on protein queries. We recommend to use TransDecoder software to  
identify protein coding sequences from cDNA sequences.  
#replace swissprot with nr if you want to blast against nr database  
#adjust the blast parameters in blast command  
# BLAST might take hours to finish. With nr, it might take days  
  
#commands (use swissprot as an example. To use refseq, replate swissprot with refseq_protein)  
  
cd /workdir/myUserName  
cp /shared_data/genome_db/BLAST_NCBI/swissprot* ./  
  
blastp -num_threads 24 -query test.fa -db swissprot -out blastresults.xml -max_target_seqs 20 -evaluate 1e-5 -  
outfmt 5 -culling_limit 10 >& blastlogfile &  
  
After this step, the blast result file blastresults.xml will be created. Copy this file to your home directory.  
  
#####  
### Optional: Run Interproscan on any BioHPC computer #####  
#####  
#you can run interproscan on any of the biohpc computers,  
  
Follow the instruction to run interproscan on BioHPC lab
```

How to do GO analysis?

Using Fisher's Exact Test to identify over represented genes in a pathway or function category

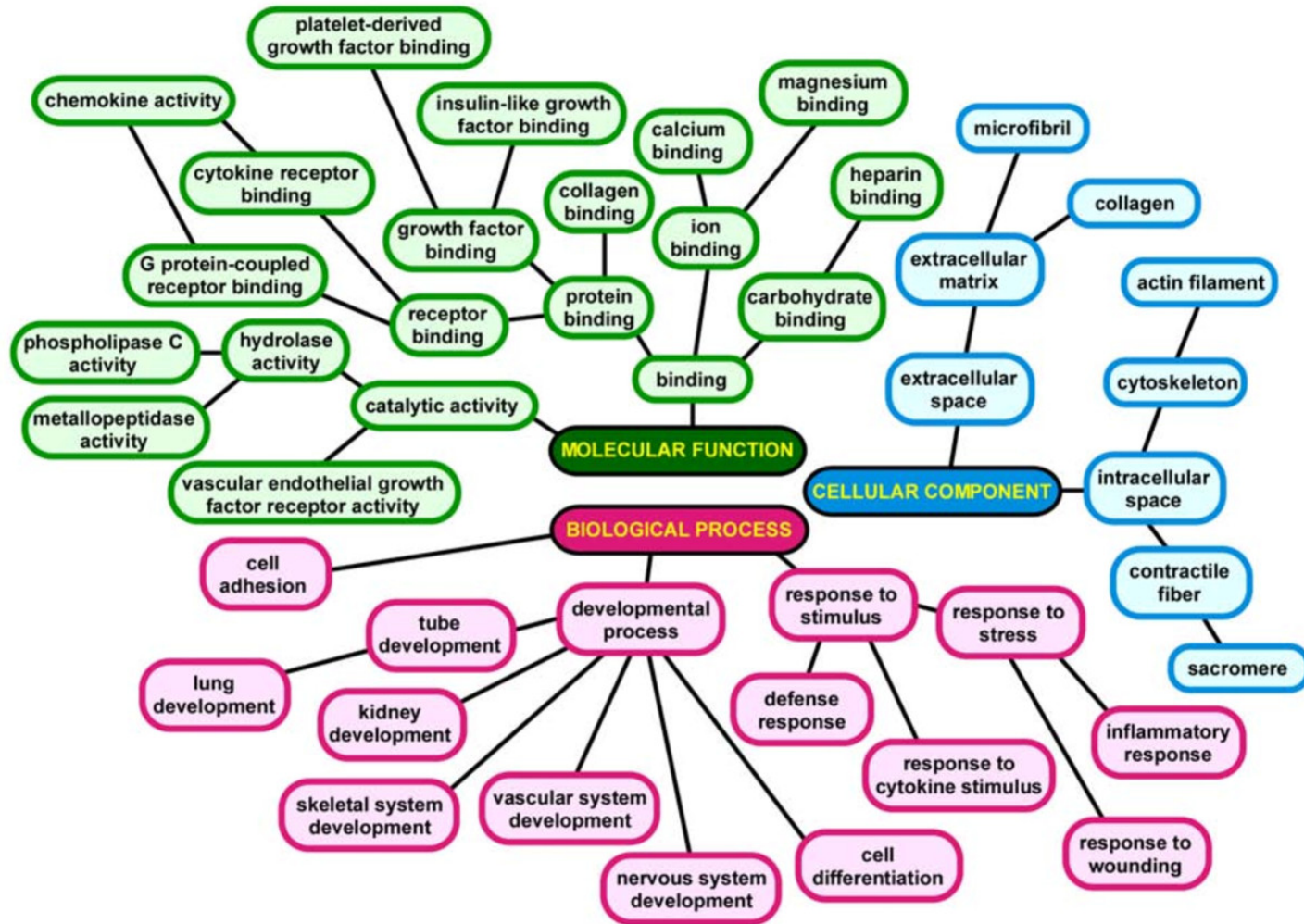
	Genes in the genome	DE genes in a experiment
P53 Pathway	40	3 -1
Not P53 Pathway	29960	297

Standard Fisher's exact test: P value= 0.008

EASE Score (in red): P value=0.06

http://david.abcc.ncifcrf.gov/content.jsp?file=functional_annotation.html

Hierarchical structure of gene ontology?



Tools for function Enrichment analysis

- DAVID

- Web based (<http://david.abcc.ncifcrf.gov/>)
- Recognized Gene IDs are limited

Functional Annotation Chart
 Current Gene List: demolist1
 Current Background: Homo sapiens
 171 DAVID IDs

Options
 Count Threshold: 2
 EASE Threshold: 0.1
 # of Records Displayed: 1000

Download File

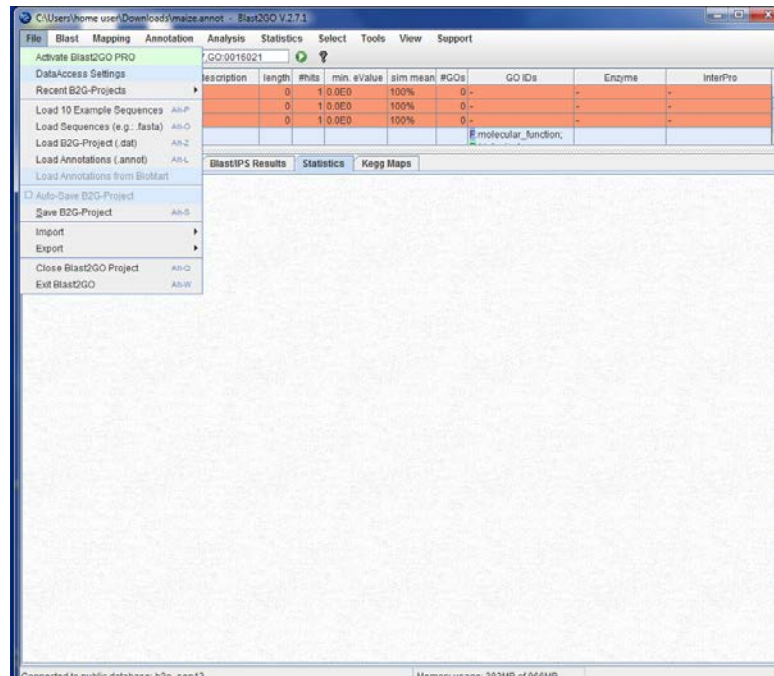
Sublist	Category	Term	RT	Genes	Count	%	P-Value
<input type="checkbox"/>	SP_PIR_KEYWORDS	signal	RT		47	27.5%	3.0E-10
<input type="checkbox"/>	SP_PIR_KEYWORDS	glycoprotein	RT		51	29.8%	4.9E-8
<input type="checkbox"/>	GOTERM_CC_ALL	extracellular region	RT		32	18.7%	1.1E-7
<input type="checkbox"/>	SP_PIR_KEYWORDS	alternative splicing	RT		49	28.7%	6.4E-6
<input type="checkbox"/>	SP_PIR_KEYWORDS	chromoprotein	RT		7	4.1%	1.1E-5
<input type="checkbox"/>	SP_PIR_KEYWORDS	direct protein sequencing	RT		33	19.3%	1.2E-5
<input type="checkbox"/>	SP_PIR_KEYWORDS	phosphorylation	RT		31	18.1%	1.6E-5
<input type="checkbox"/>	UP_SEQ_FEATURE	signal peptide	RT		47	27.5%	3.7E-5
<input type="checkbox"/>	SP_PIR_KEYWORDS	metalloprotein	RT		8	4.7%	4.7E-5
<input type="checkbox"/>	GOTERM_BP_ALL	response to chemical stimulus	RT		14	8.2%	6.1E-5

Annotations:

- Gene list and population background being analyzed (points to Current Gene List and Current Background)
- Minimum number of genes for the corresponding term (points to Count Threshold)
- Maximum EASE Score/P-Value (points to EASE Threshold)
- Maximum number of record per page (points to # of Records Displayed)
- Original database/resource where the terms orient (points to SP_PIR_KEYWORDS, GOTERM_CC_ALL, UP_SEQ_FEATURE, GOTERM_BP_ALL)
- Enriched terms associated with your gene list (points to Term column)
- Related Term Search (points to RT column)
- Genes involved in the term (points to Genes column)
- Percentage, e.g. 14/171=8.2% (involved genes/total genes) (points to % column)
- Modified Fisher Exact P-Value, EASE Score. The smaller, the more enriched. (points to P-Value column)

Function Enrichment analysis

- BLAST2GO
 - Flexible input file for reference genome, can do sequence based function annotation
 - Input file: Sequence FASTA, BLAST results, GO annotation file
 - Do Fisher's Exact test with a graphic user interface



Fisher's Exact Test with BLAST2GO

Fisher's Exact Test

Select Test-Set:

Select Reference (optional):

Term Filter Value:

Term Filter Mode: **FDR**

Two-Tailed:

Create GO->IDs List:

Remove double IDs:

Genes in test set

Genes in reference set
(filtered gene list)

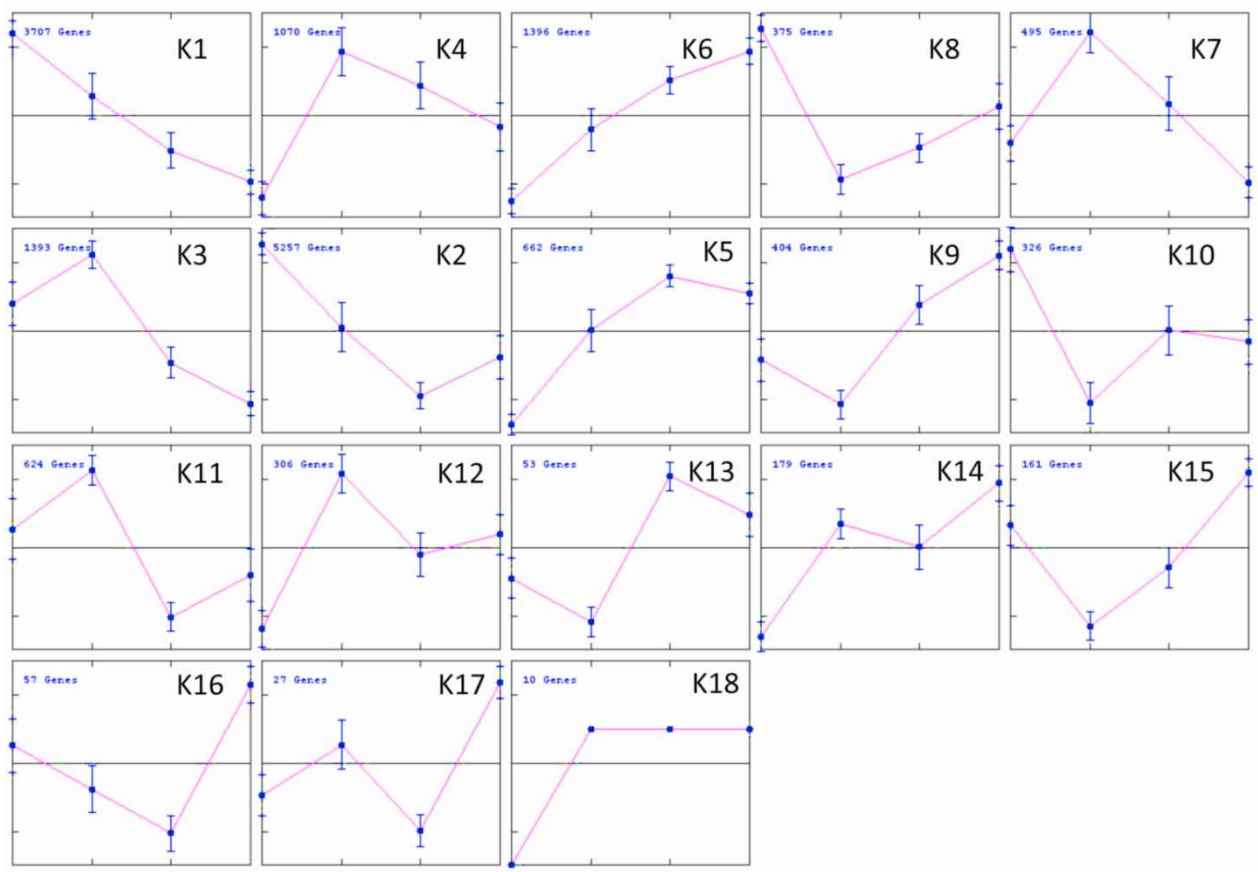
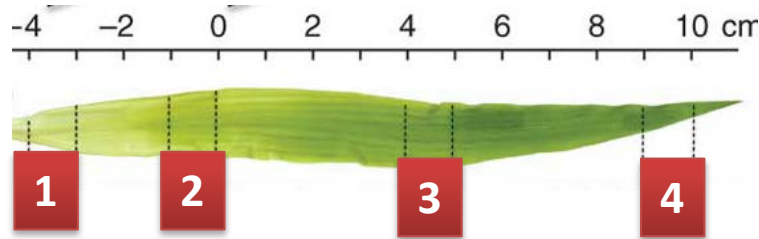
Gossip Fisher's Exact Test Results: testset_example.txt

GOSSIP
Test-Set: testset_example.txt
Tests for all Gene Ontology terms if they are enriched in a test group when compared to a reference group using Fisher's exact test with multiple testing correction.
[Pub: Biological Profiling of Gene Groups utilizing Gene Ontology A Statistical Framework](#)
[Poster: GOSSIP: Biological Profiling of Gene Groups utilizing Gene Ontology](#)
by Nils Blthgen, Karsten Brand, Hanspeter Herzel, Dieter Beule

GO Term	Name	FDR	FWER	single test p-Value	# in test group	# in reference group	# non annot test	# non annot reference group	Over/Under
GO:0044464	cell part	5.85654E-4	2.92787E-4	1.53838E-4	29	166	32	60	under
GO:0005623	cell	5.85654E-4	2.92787E-4	1.53838E-4	29	166	32	60	under
GO:0003824	catalytic activity	0.0067865	0.0050773	9.6063E-4	18	119	43	107	under
GO:0006790	sulfur metabolic process	0.0097901	0.00258308	8.84665E-5	8	2	53	224	over
GO:0004364	glutathione transferase activity	0.0097901	0.0152647	3.79698E-4	5	0	56	226	over
GO:0042221	response to chemical stimulus	0.0097901	0.0156898	3.91899E-4	17	21	44	205	over
GO:0006749	glutathione metabolic process	0.0097901	0.0187977	4.39258E-4	6	1	55	225	over

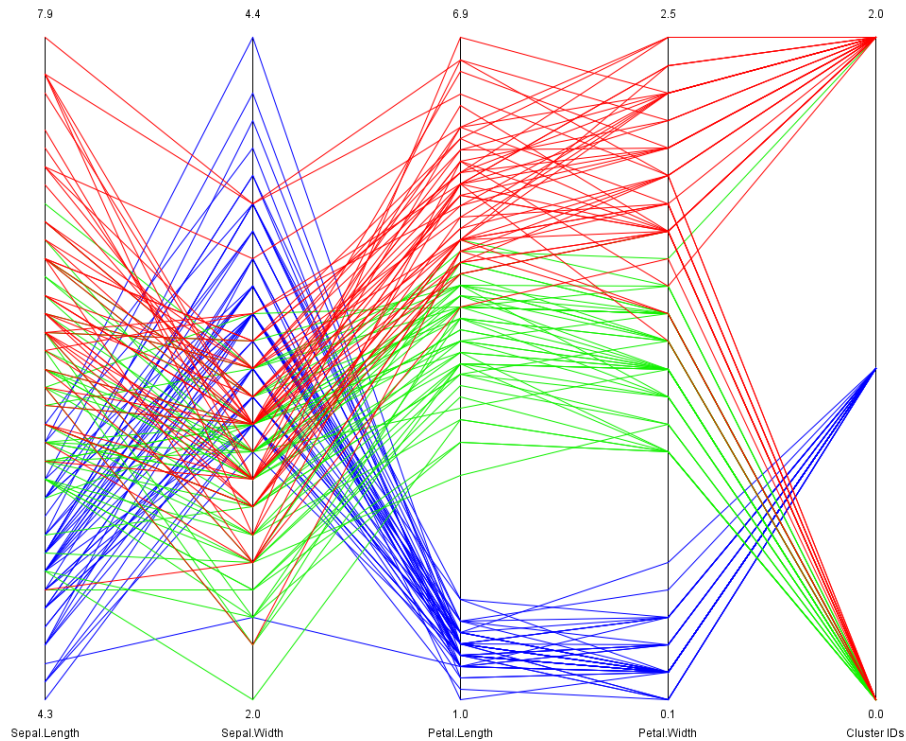
https://www.blast2go.com/images/b2g_pdfs/b2g_user_manual.pdf

Clustering analysis on multiple conditions of RNA-seq data



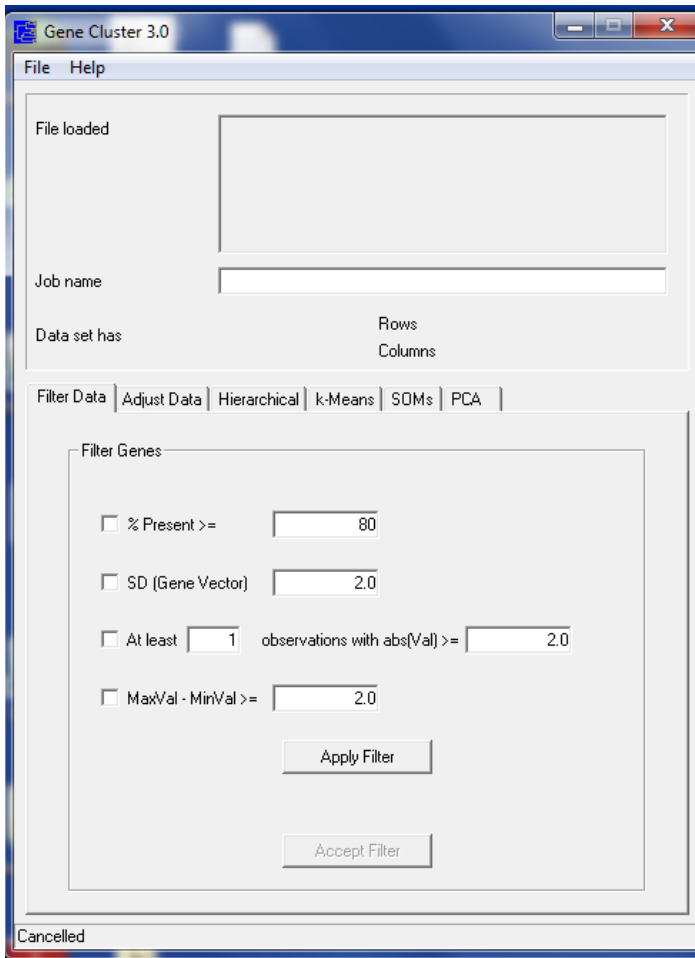
Clustering analysis

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



Using free software Cluster 3.0 for hierarchical and k-means clustering

<http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm>



tracking_id	s1_FPKM	s2_FPKM	s3_FPKM	s4_FPKM
AC14815 2.3_FG00 1	• 1	• 1	• 1.085823	• 1.237447
AC14815 2.3_FG00 2	• 1	• 1	• 1	• 1
AC14815 2.3_FG00 5	• 1.054317	• 6.65432	• 1.089866	• 1
AC14815 2.3_FG00 6	• 1.044314	• 1.223353	• 1	• 1
AC14815 2.3_FG00 7	• 1	• 1	• 1	• 1
AC14815 2.3_FG00 8	• 3.13339	• 20.1778	• 68.1838	• 88.5417
AC14816 7.6_FG00 1	• 17.603	• 43.4081	• 54.7869	• 37.5133
AC14947 5.2_FG00 2	• 149.468	• 10.75707	• 14.3301	• 11.8052
AC14947 5.2_FG00 3	• 101.308	• 34.2556	• 30.6524	• 20.2889
AC14947 5.2_FG00 4	• 1.053882	• 1	• 1	• 1

* Add 1 to each FPKM value before loading into Cluster

Alternative software

- **Gene-E**

<http://www.broadinstitute.org/cancer/software/GENE-E/>

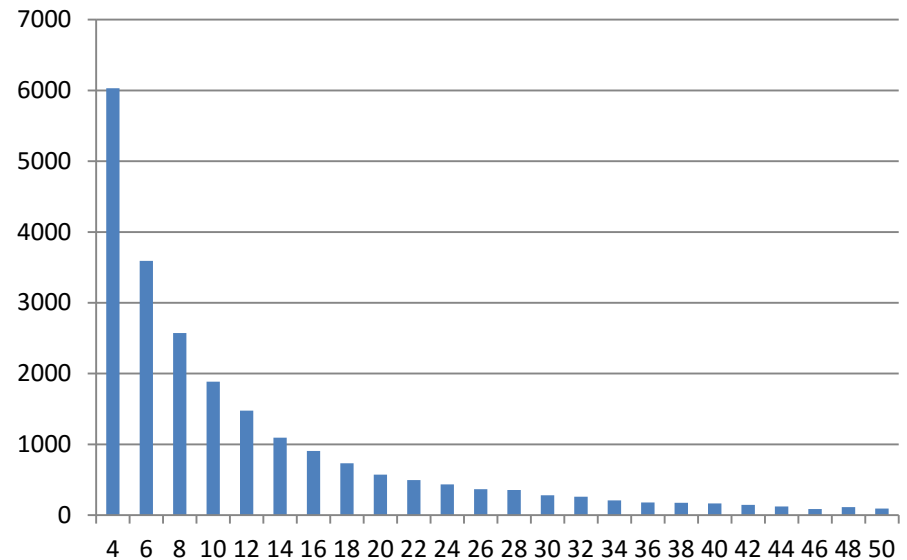
- **Bioconductor: hclust & kmeans**

- Free R package

Prepare data for clustering

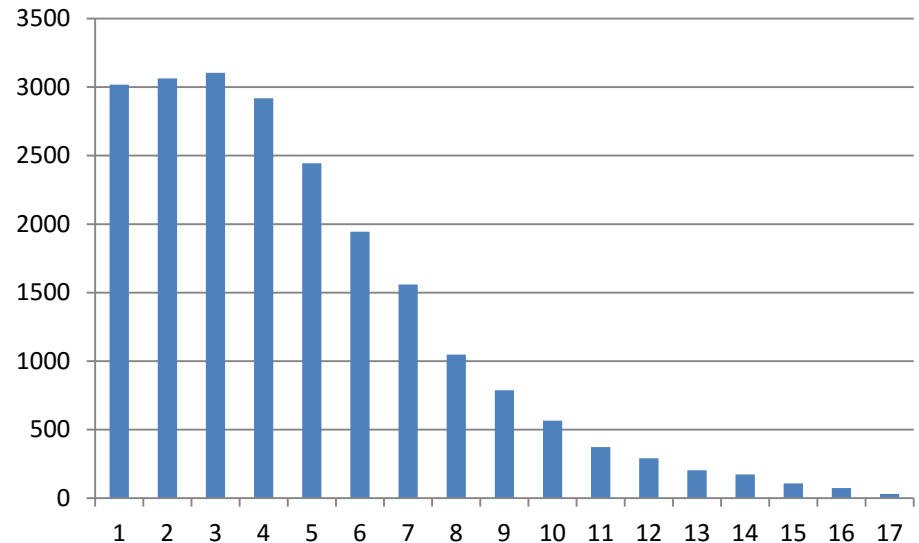
LOG transformation of FPKM (or CPM) value to improve the distribution

FPKM



Log₂(FPKM)

To Avoid log(0), using Excel to add 1 to all FPKM values before loading to Cluster.



Filter data

To make the analysis computational feasible on a desktop computer, pre-filter the data to remove

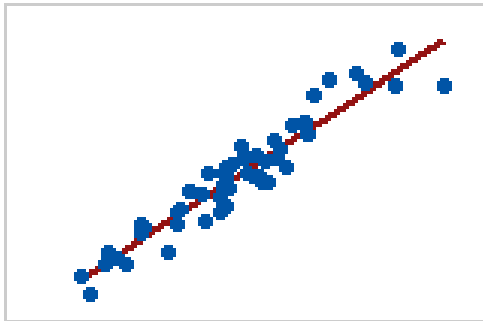
- Low expressed genes;
- Invariant genes.

Construction of pairwise distance matrix of all genes

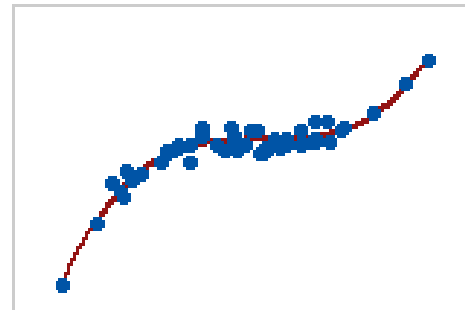
Pearson : Linear correlation (Default)

vs

Spearman: Ranked correlation



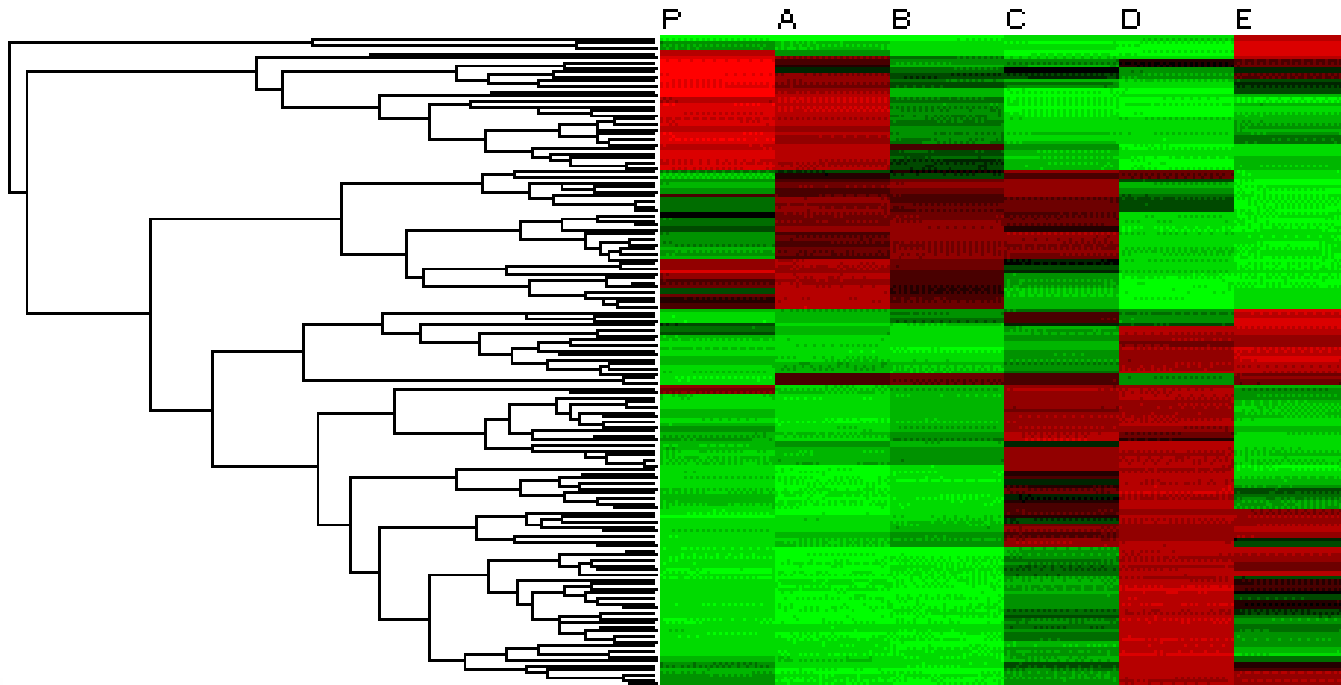
Use Pearson



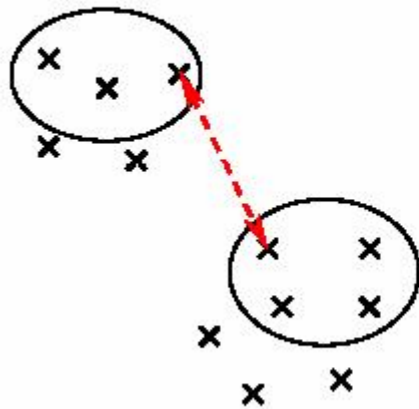
Use Spearman

Hierarchical clustering

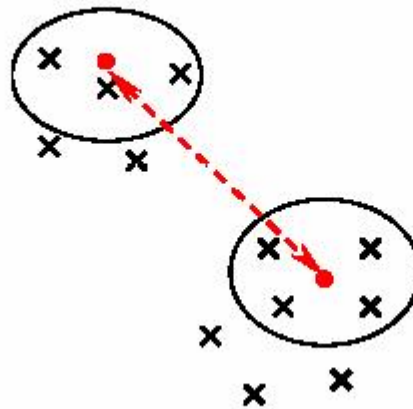
Linkage criteria in hierarchical clustering



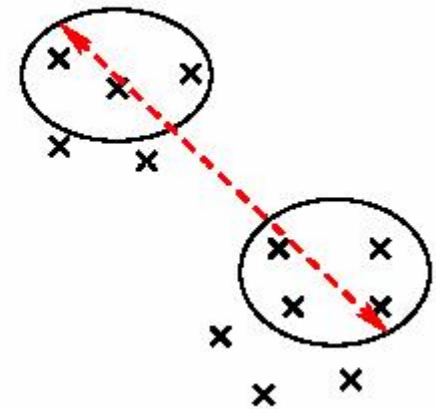
- Simple linkage



- Average linkage



- Complete linkage



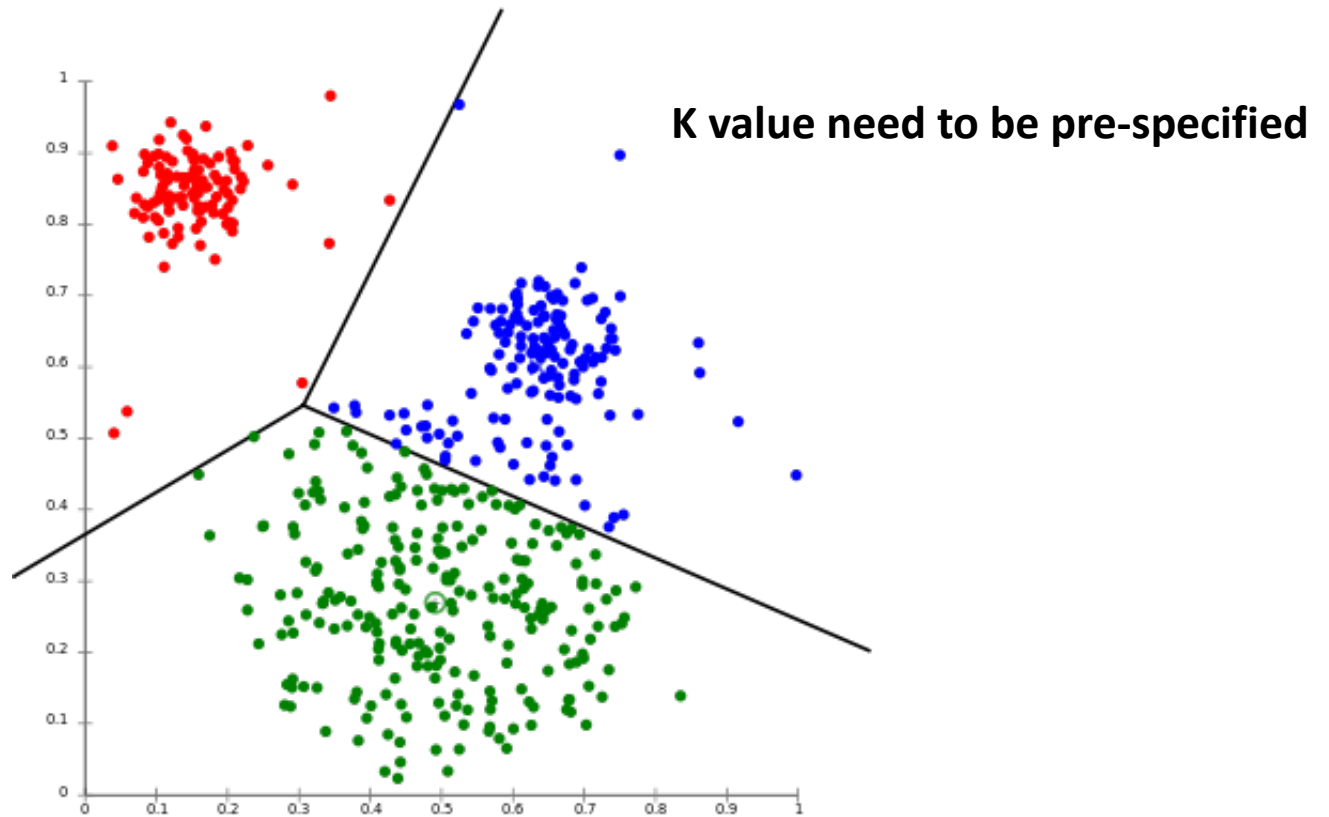
Default

Visualize the clustering results with Treeview



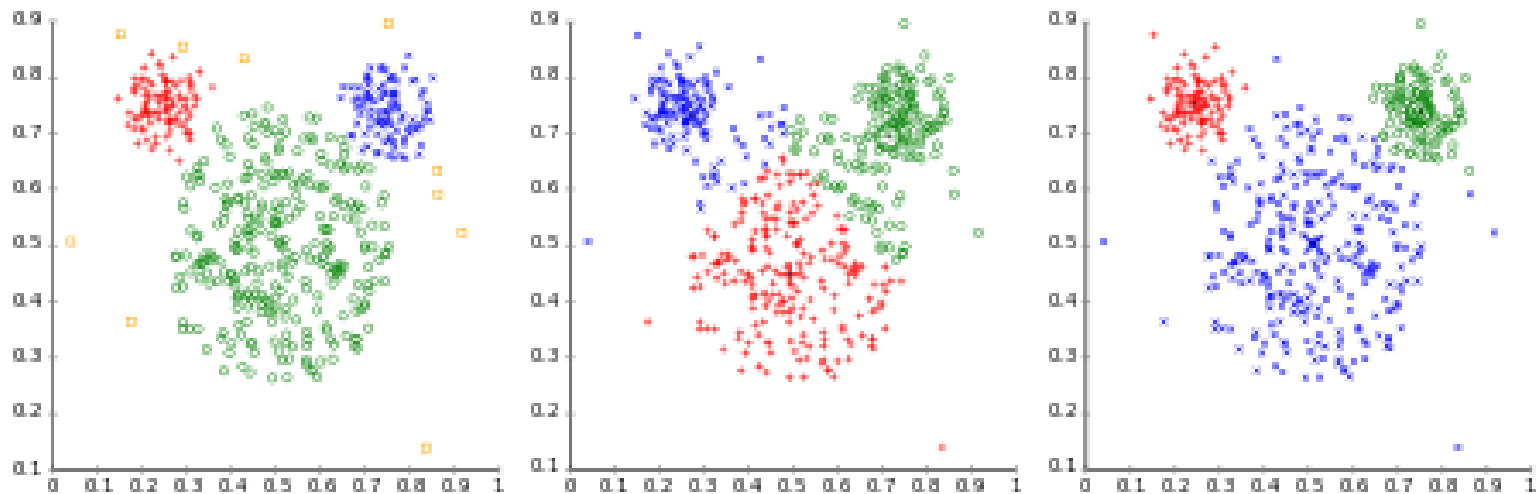
The software has functions to select nodes and export genes in selected node.

K-means clustering



The tendency of k -means to produce equal-sized clusters leads to bad results

Different cluster analysis results on "mouse" data set:
Original Data k -Means Clustering EM Clustering

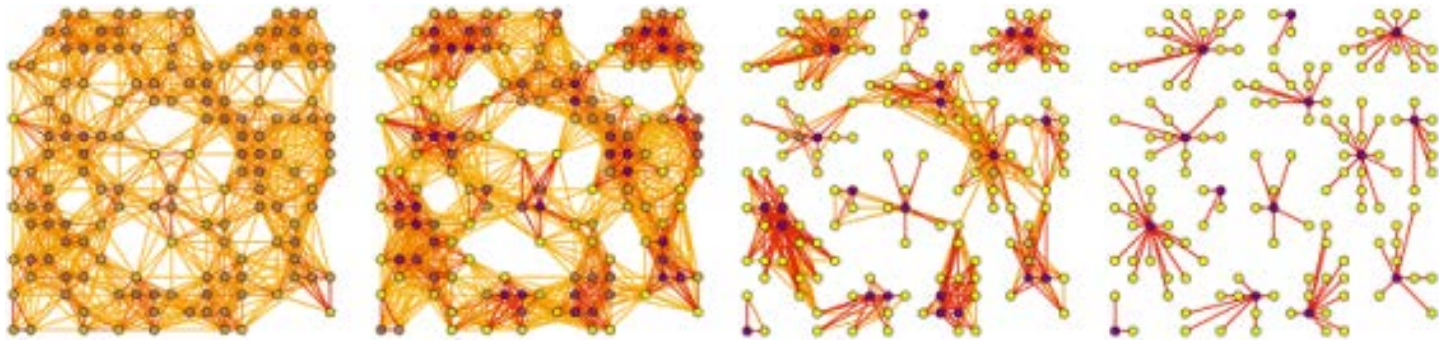


Wikipedia: K-means_clustering

Co-expression network modules

1. MCL (Markov Cluster Algorithm)

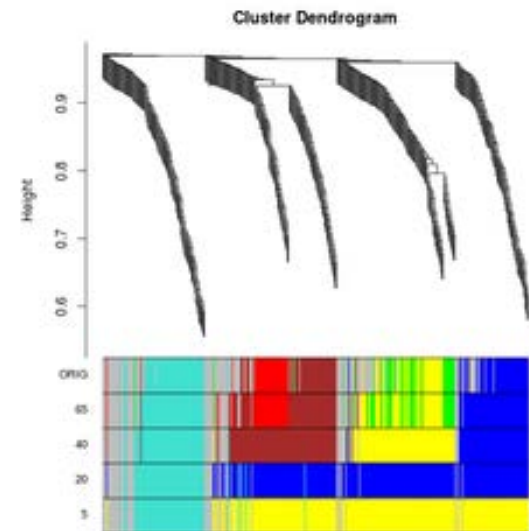
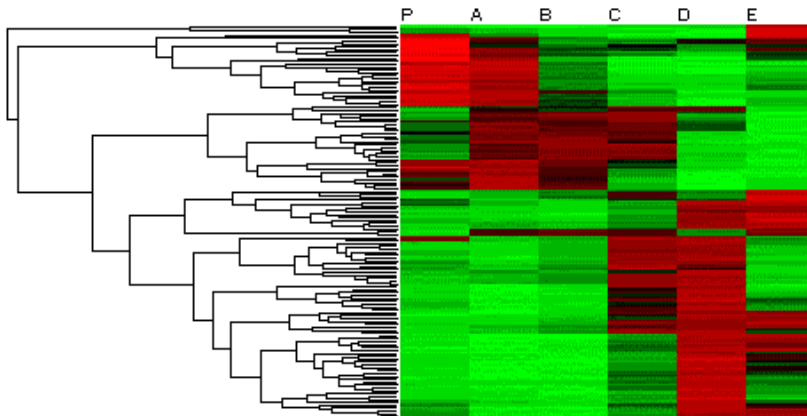
- Easy to use interface: only need a distance matrix and inflation value



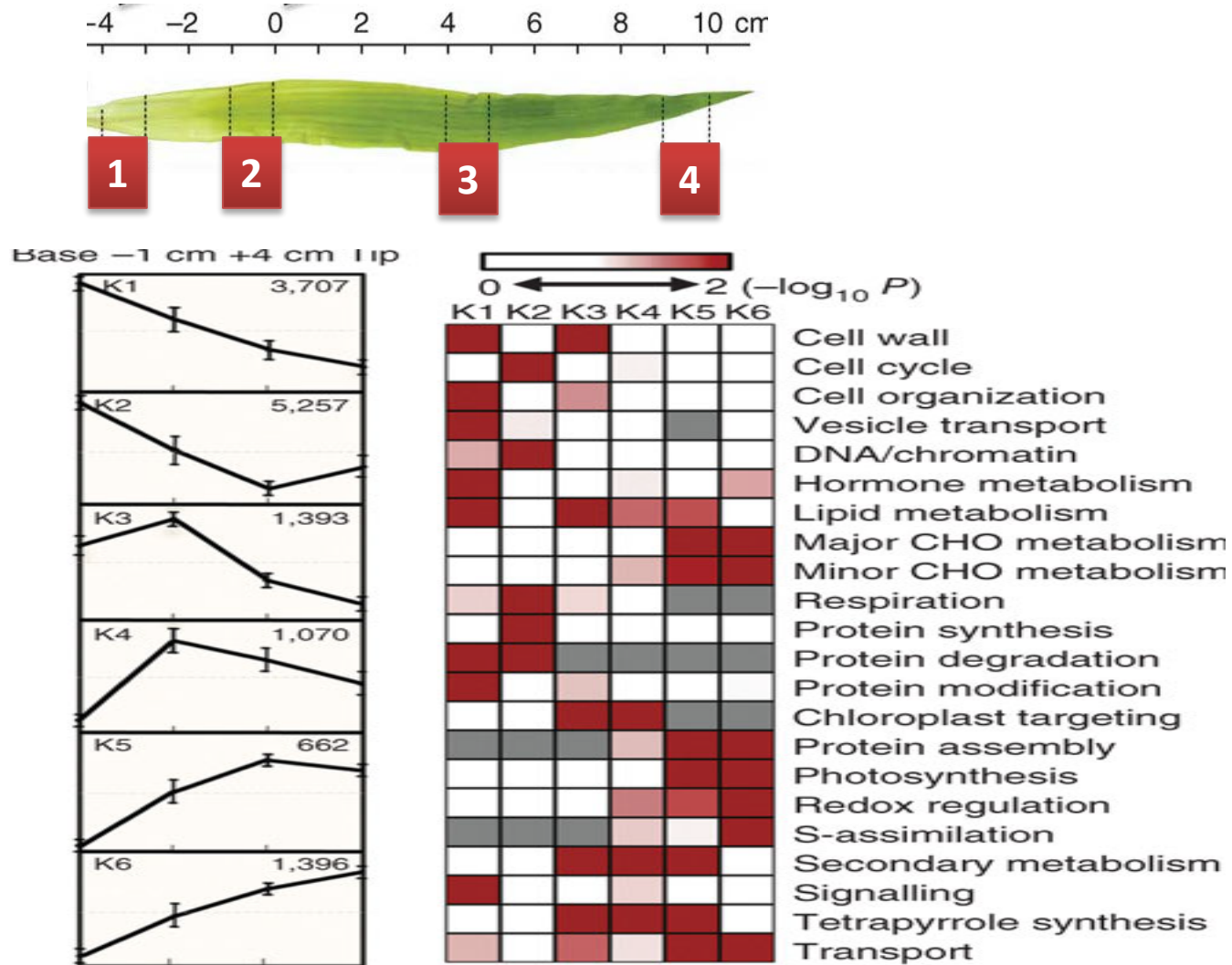
Co-expression network modules

2. WGCNA (weighted correlation network analysis)

- transform the initial distance matrix into Topological Overlap Matrix



Presentation of the results, an example



Homework

- Clustering and Function enrichment analysis.
- Starting file:
 - Normalized Read Count file: genes.txt
 - Rice Gene Ontology annotation file: rice.annot
created with Ensembl BioMart.
- Tasks:
 - Hierarchical clustering
 - K-means clustering
 - Function enrichment analysis with BLAST2GO