

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(~group)
y <- estimateGLMCommonDisp(y,design)
y <- estimateGLMTrendedDisp(y,design)
y <- estimateGLMTagwiseDisp(y,design)
fit<-glmFit(y,design)
```

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")
```

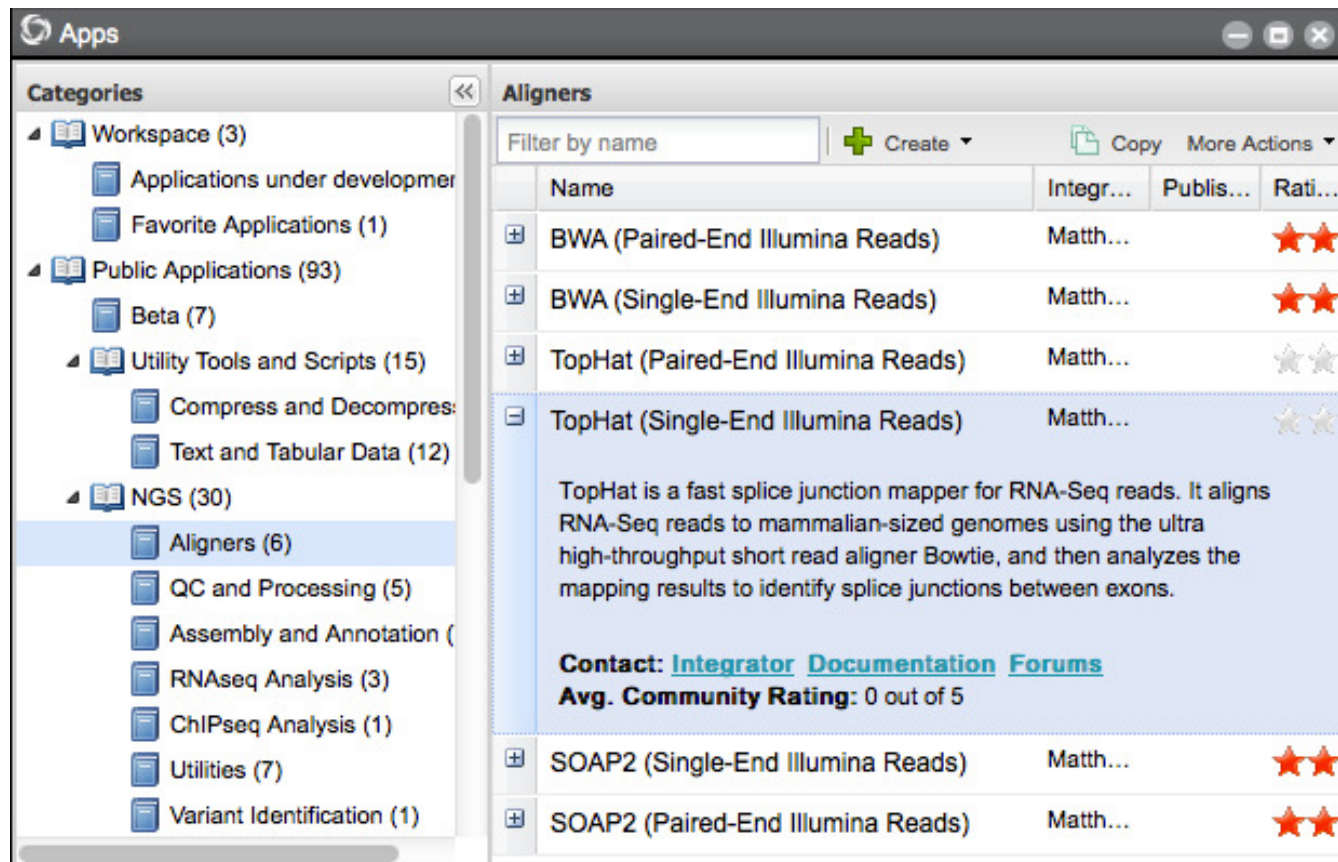
Other computational resources

- **iPlant Discovery Environment**

Tutorials:

General: <http://www.iplantcollaborative.org/learning-center/all-tutorials>

RNA-seq: <http://www.iplantcollaborative.org/learning-center/discovery-environment/de-003-characterizing-differential-expression-rna-seq>



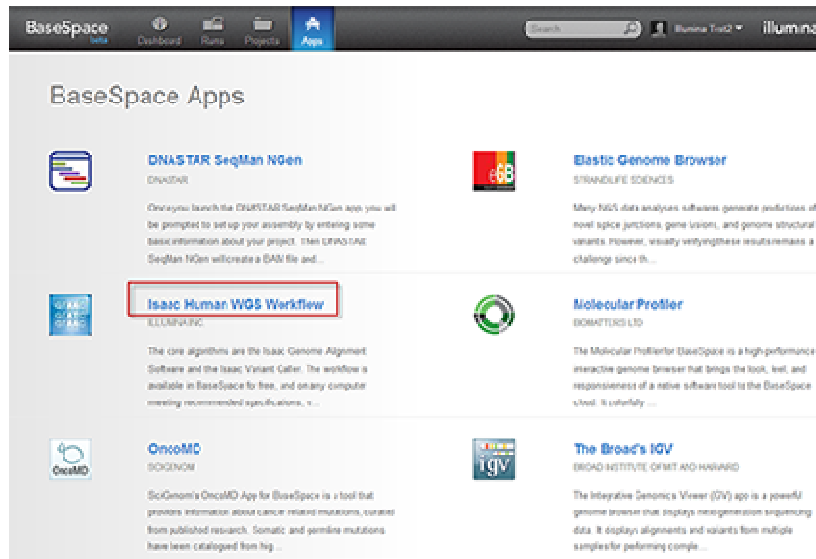
The screenshot shows the 'Apps' window in the iPlant Discovery Environment. The left sidebar displays a tree view of application categories, with 'Aligners (6)' selected under the 'NGS (30)' category. The main panel shows a list of aligners with a search filter, a 'Create' button, and a 'More Actions' dropdown. The list includes BWA and TopHat for both Paired-End and Single-End Illumina Reads. The 'TopHat (Single-End Illumina Reads)' application is expanded, showing a detailed description and contact information.

Name	Integr...	Publis...	Rati...
BWA (Paired-End Illumina Reads)	Matth...		★★★
BWA (Single-End Illumina Reads)	Matth...		★★★
TopHat (Paired-End Illumina Reads)	Matth...		★★★
TopHat (Single-End Illumina Reads)	Matth...		★★★
<p>TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.</p> <p>Contact: Integrator Documentation Forums Avg. Community Rating: 0 out of 5</p>			
SOAP2 (Single-End Illumina Reads)	Matth...		★★★
SOAP2 (Paired-End Illumina Reads)	Matth...		★★★

Other computational resources

- **Illumina BaseSpace**

<https://basespace.illumina.com/home/index>



- **Galaxy**

<https://usegalaxy.org/>

Commercial Software @ Cornell

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

- **LaserGene (Ngen)**
- **Geneous**
- **Ingenuity Pathways Analysis**

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?**

What is Gene Ontology -1

How to describe the function of a gene?

- Gene description line

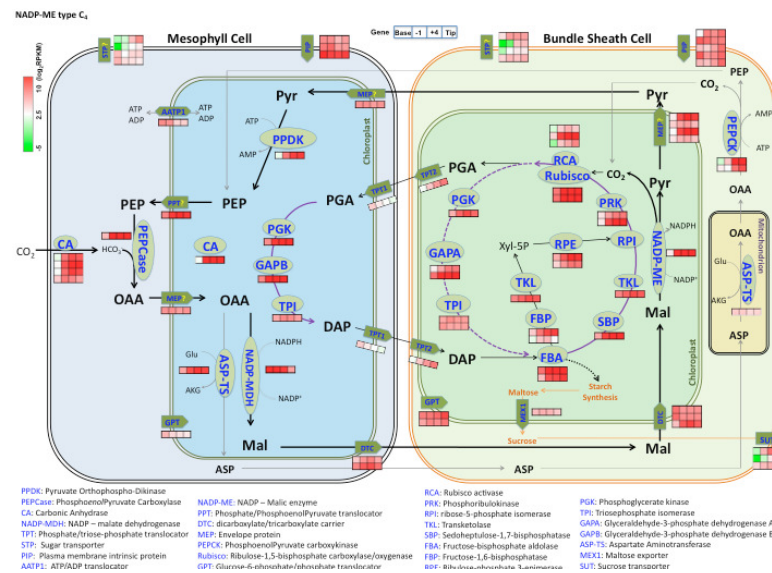
GRMZM2G002950	Putative leucine-rich repeat receptor-like protein kinase family protein
GRMZM2G006470	Uncharacterized protein
GRMZM2G014376	Shikimate dehydrogenase; Uncharacterized protein
GRMZM2G015238	Prolyl endopeptidase
GRMZM2G022283	Uncharacterized protein

- **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)

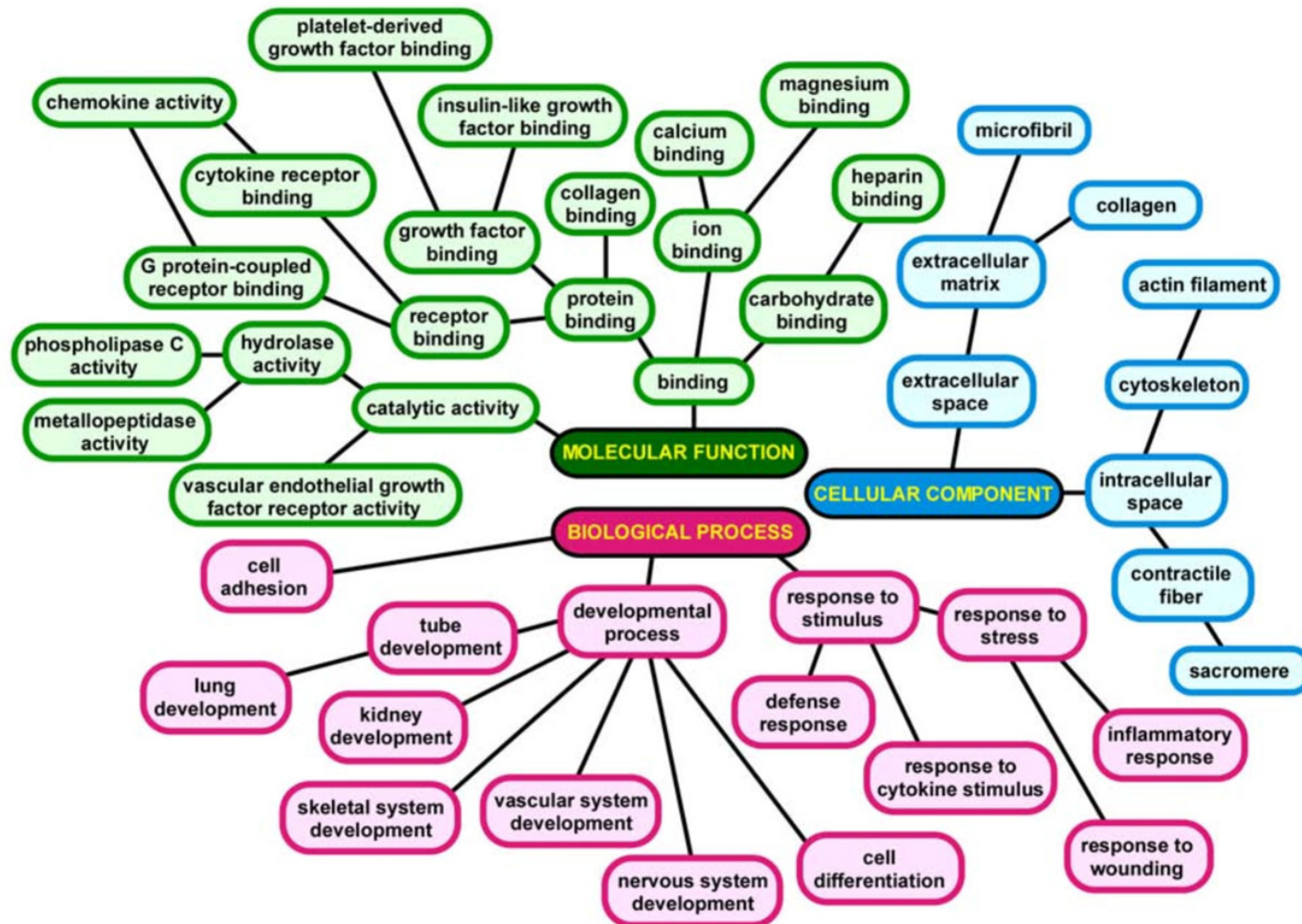
What is Gene Ontology -1

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

Hierarchical structure of gene ontology?



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

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

Rerun Using Options Create Sublist Download File

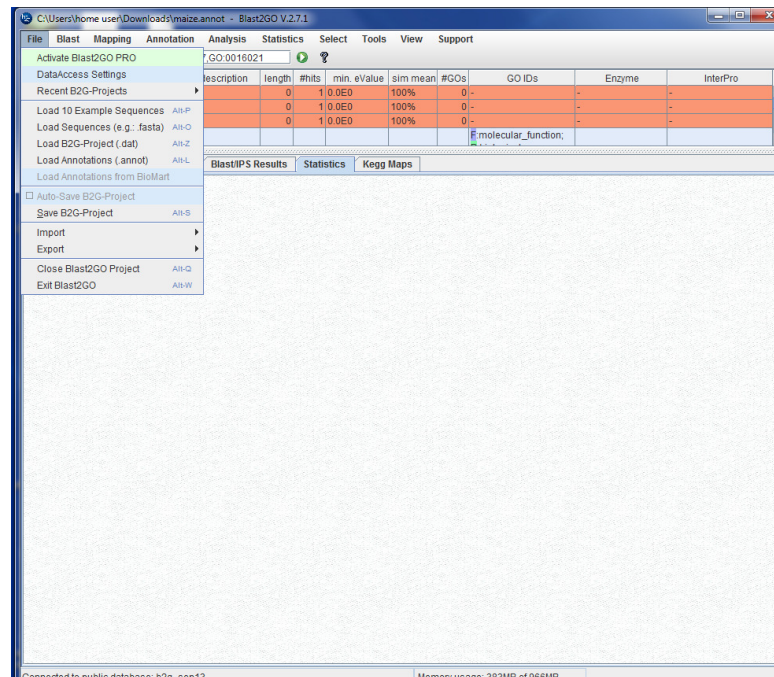
Sublist	Category	Term	RT	Genes	Count	%	P-Value
<input type="checkbox"/>	SP_PIR_KEYWORDS	signal	RT		47	27.5%	3.0E-10
<input type="checkbox"/>	SP_PIR_KEYWORDS	glycoprotein	RT		51	29.8%	4.9E-8
<input type="checkbox"/>	GOTERM_CC_ALL	extracellular region	RT		32	18.7%	1.1E-7
<input type="checkbox"/>	SP_PIR_KEYWORDS	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: demolist1")
- Minimum number of genes for the corresponding term (points to "Count Threshold: 2")
- Maximum EASE Score/P-Value (points to "EASE Threshold: 0.1")
- Maximum number of record per page (points to "# of Records Displayed: 1000")
- Original database/resource where the terms orient (points to "GOTERM_BP_ALL")
- Enriched terms associated with your gene list (points to "response to chemical stimulus")
- 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

http://www.blast2go.com/data/blast2go/b2g_user_manual_22102013.pdf

Public and Commercial Resources

- **Public resource:**
 - DAVID Bioinformatics Resources
(<http://david.abcc.ncifcrf.gov/>)

- **Commercial Resource:**
 - Ingenuity
(License information
<http://www.biotech.cornell.edu/node/137>)

Biological Databases @ Cornell Library

- **KEGG**

- Biological pathway databases

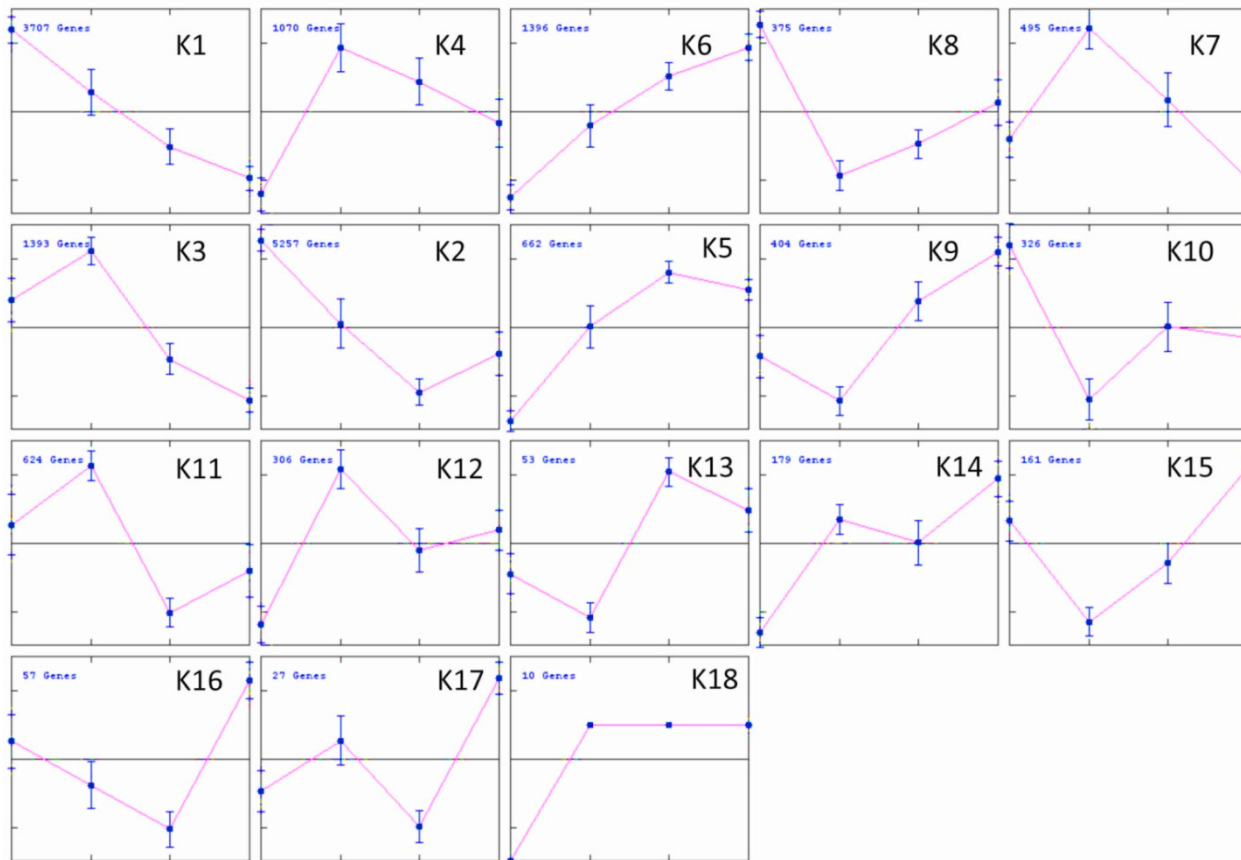
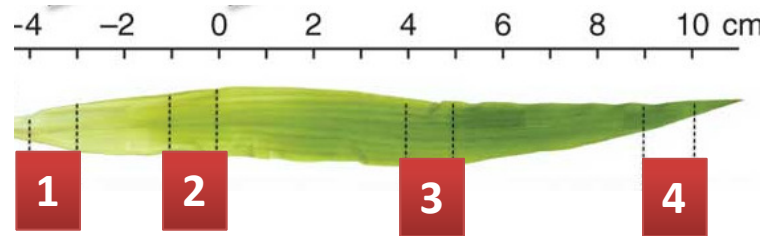
- <https://catalog.library.cornell.edu/cgi-bin/Pwebrecon.cgi?BBID=8327047&DB=local>

- **TAIR**

- Arabidopsis

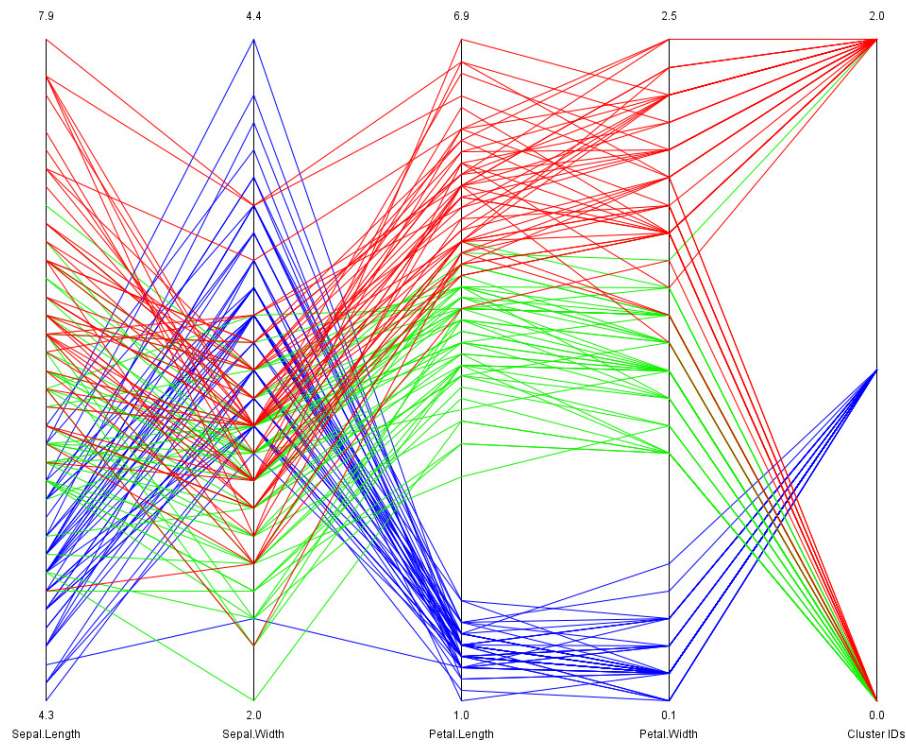
- <https://catalog.library.cornell.edu/cgi-bin/Pwebrecon.cgi?BBID=3924196&DB=local>

Clustering analysis on multiple conditions of RNA-seq data



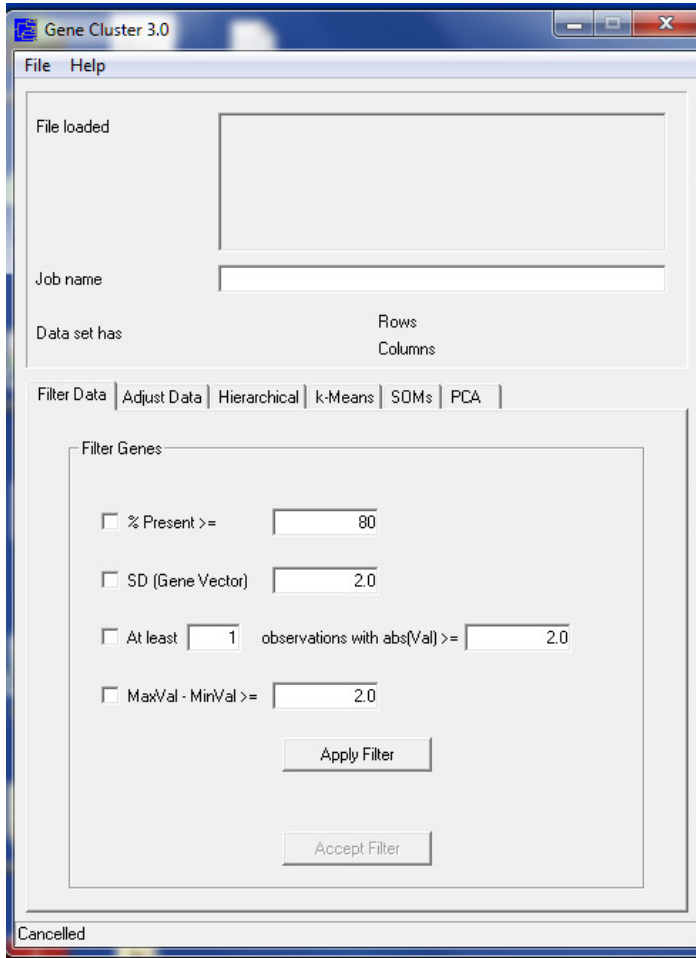
Clustering analysis

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/>

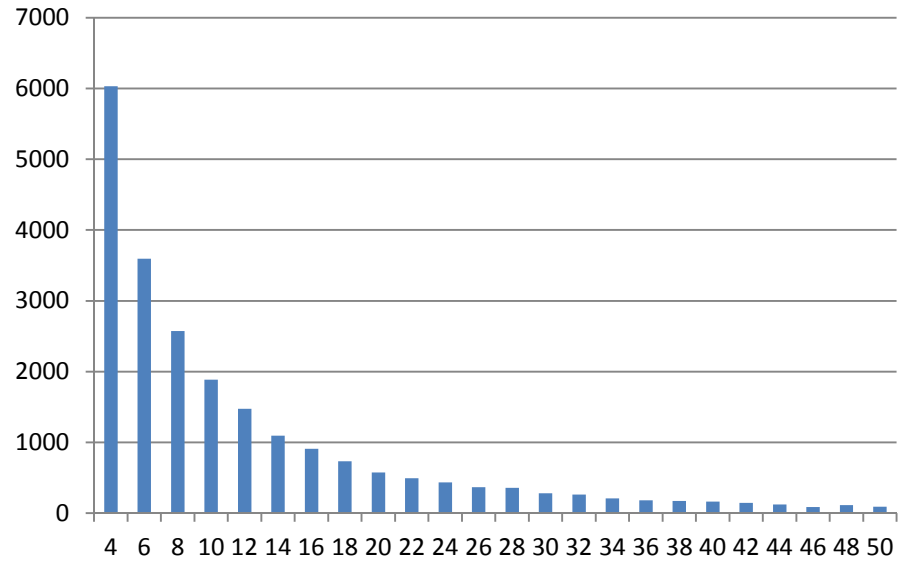
- **Bioconductor: hclust & kmeans**

- Free R package

Prepare data for clustering

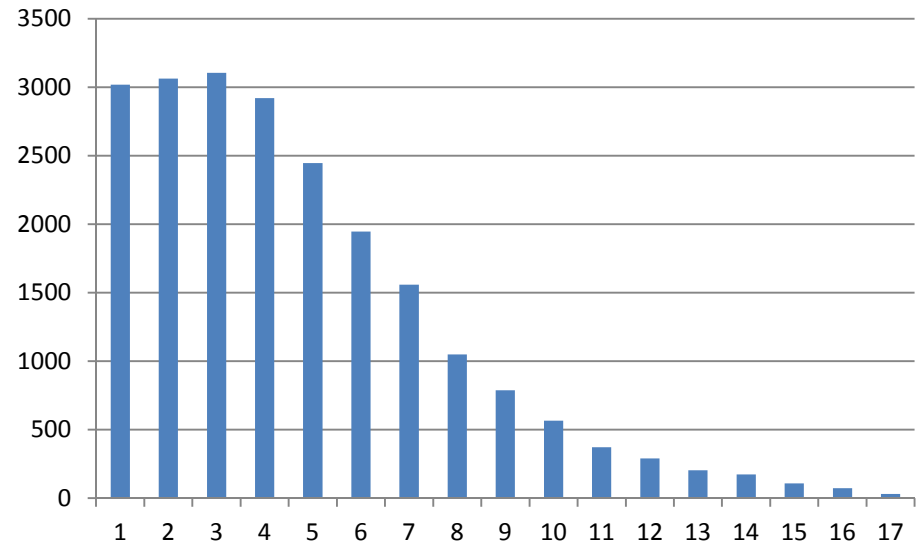
LOG transformation of FPKM (or CPM) value to improve the distribution

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

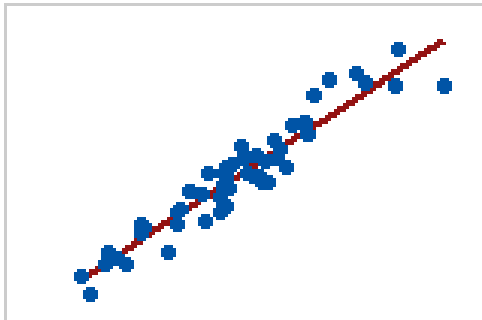
- Low expressed genes;
- Invariant genes.

Construction of pairwise distance matrix of all genes

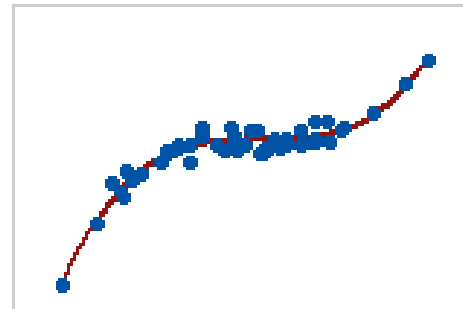
Pearson : **Linear correlation (Default)**

vs

Spearman: **Ranked correlation**

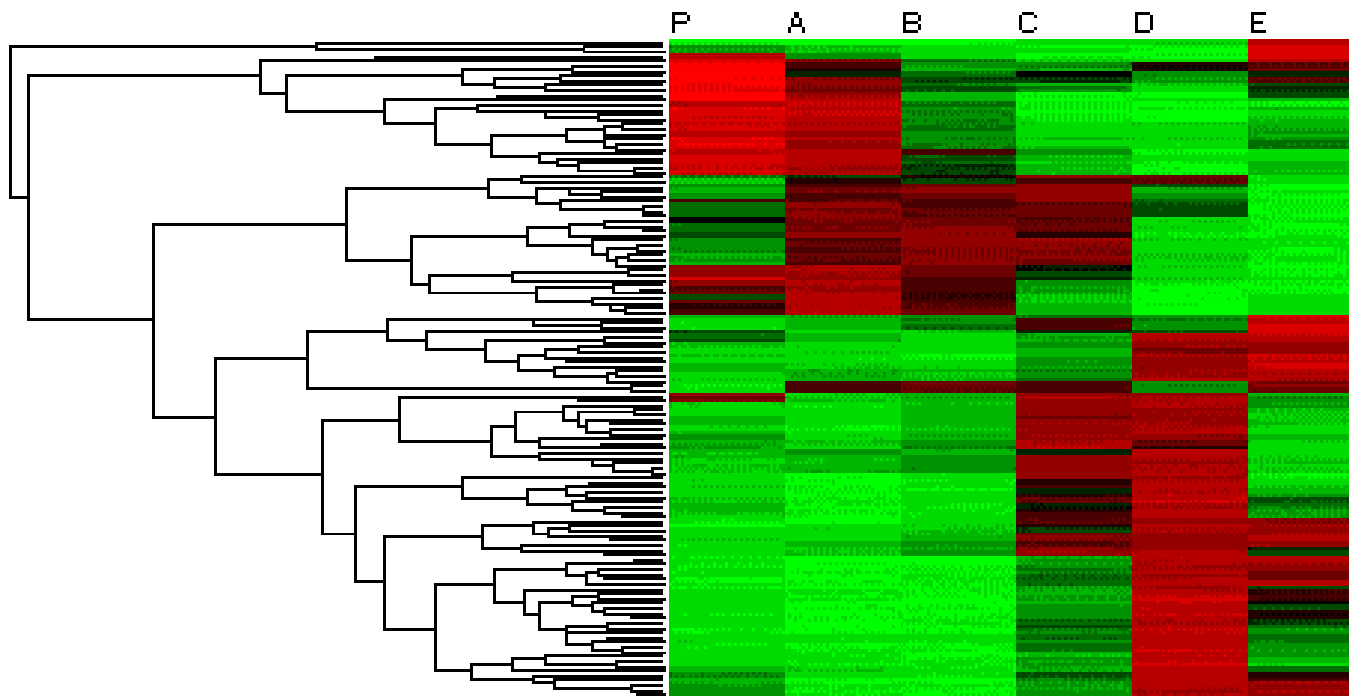


Use Pearson



Use Spearman

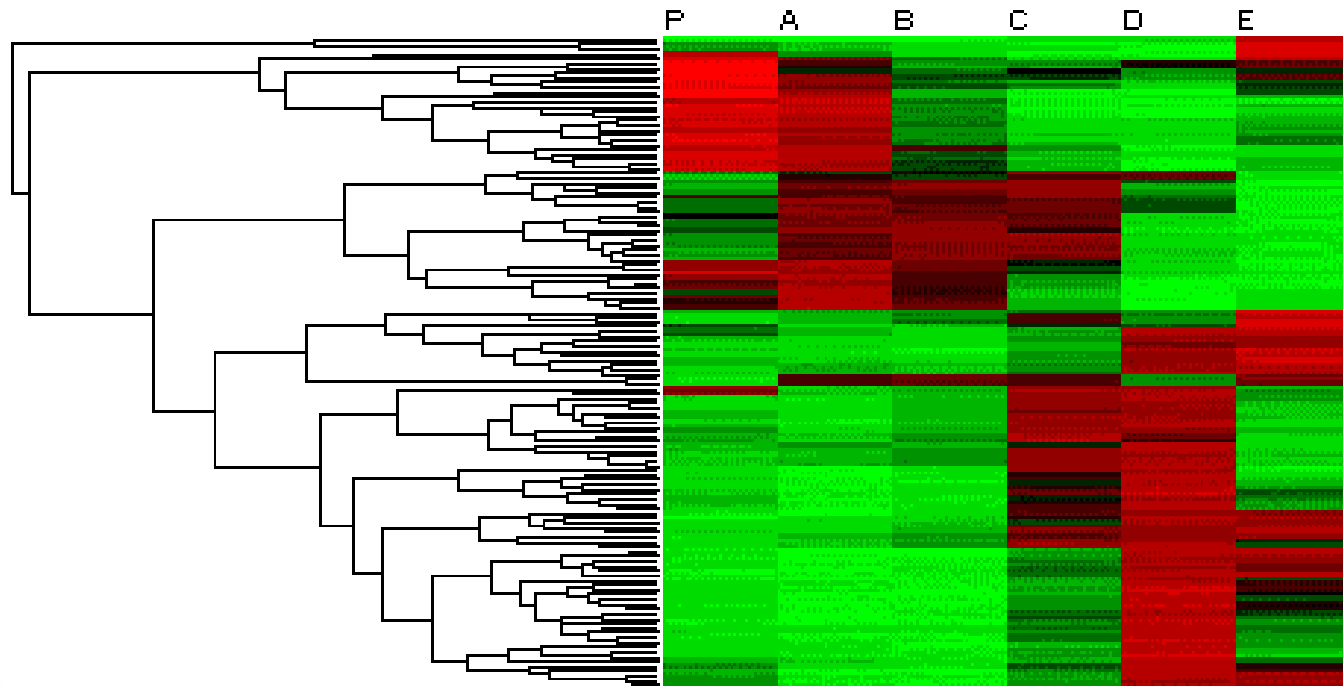
Hierarchical clustering



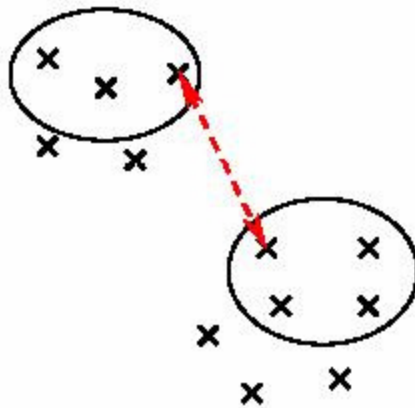
<http://compbio.pbworks.com/w/page/16252903/Microarray%20Clustering%20Methods%20and%20Gene%20Ontology>

Center the data to median value to get green-red color visualization

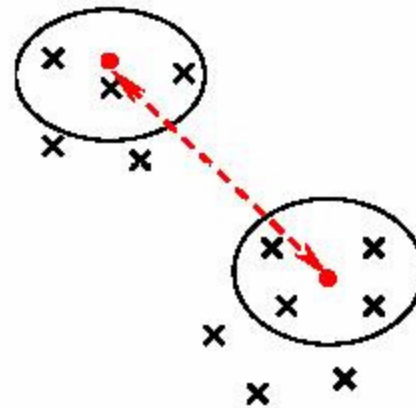
Linkage criteria in hierarchical clustering



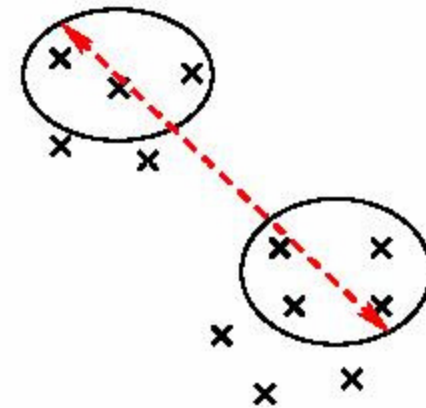
- Simple linkage



- Average linkage



- Complete linkage



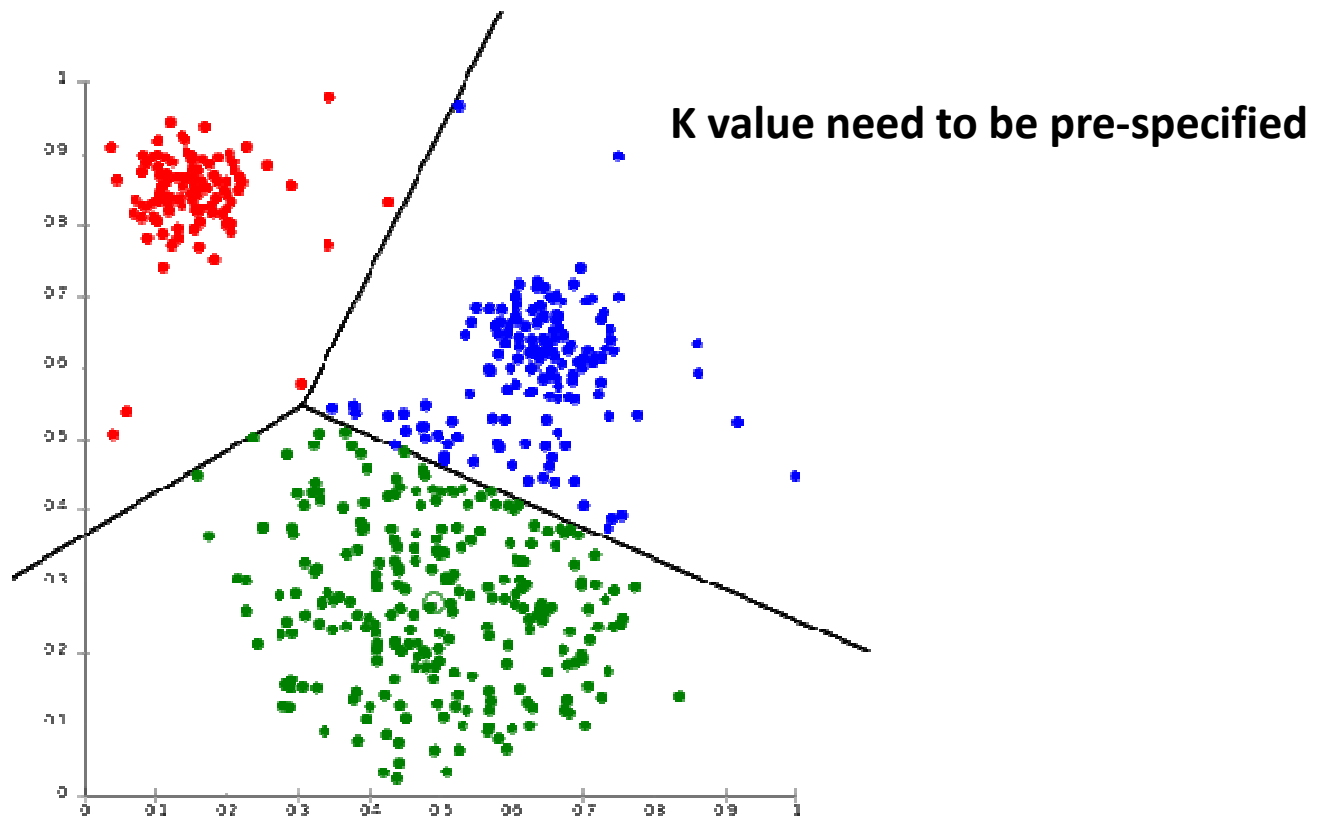
Default

Visualize the clustering results with Treeview



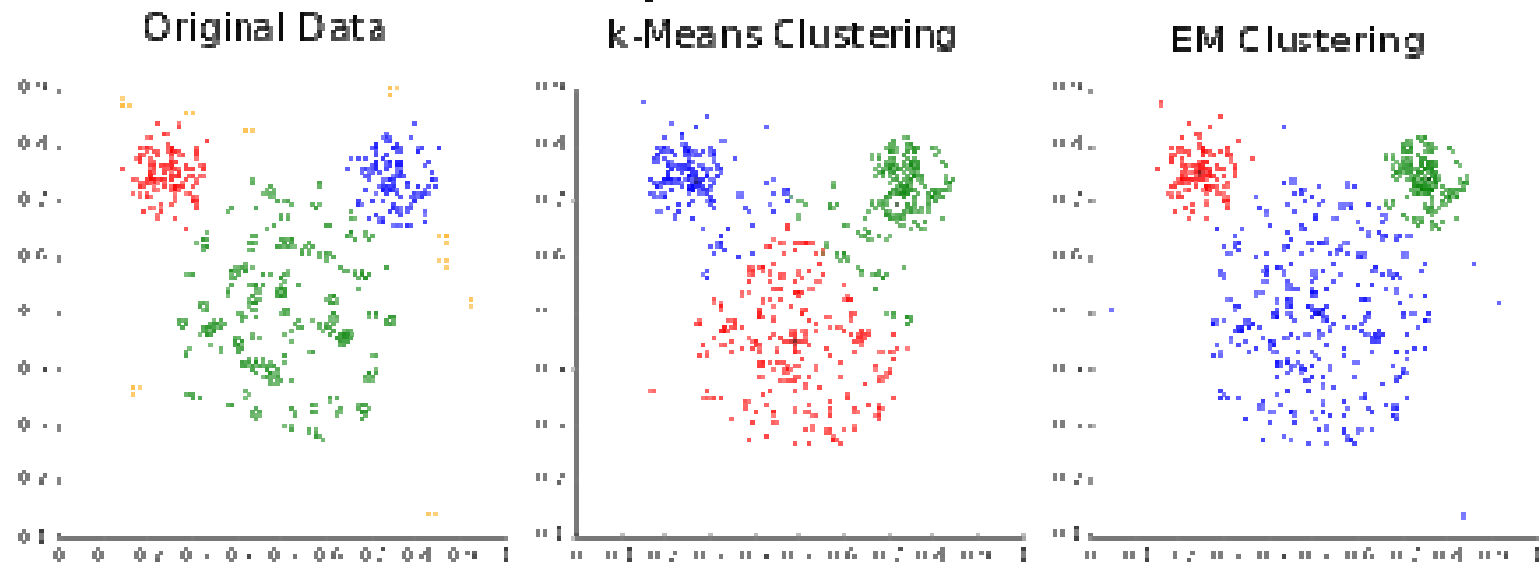
The software has functions to select nodes and export genes in selected node.

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:

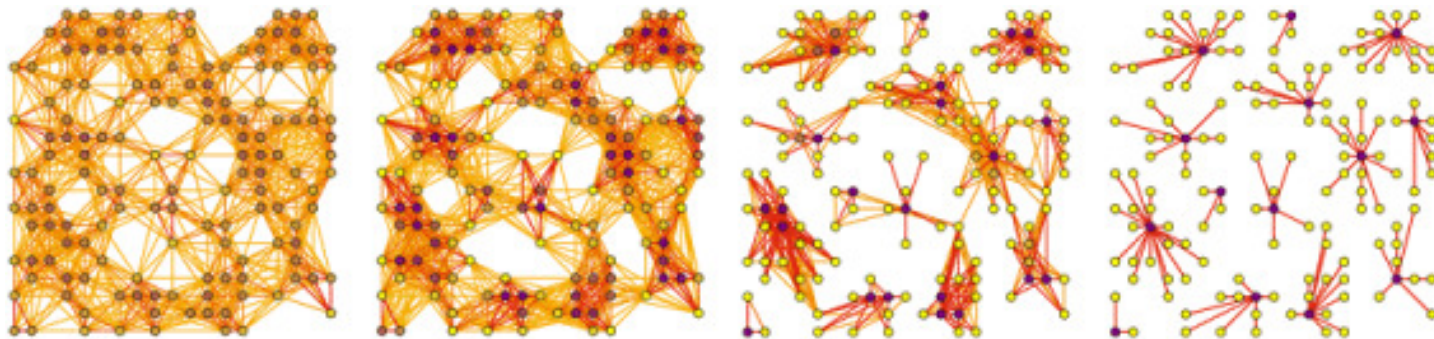


Wikipedia: K-means_clustering

Co-expression network modules

1. MCL (Markov Cluster Algorithm)

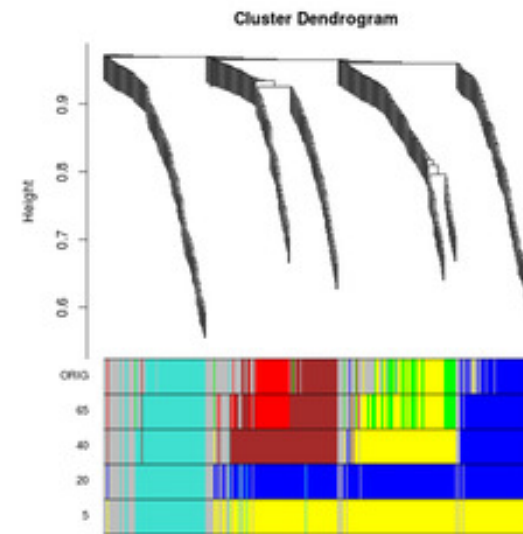
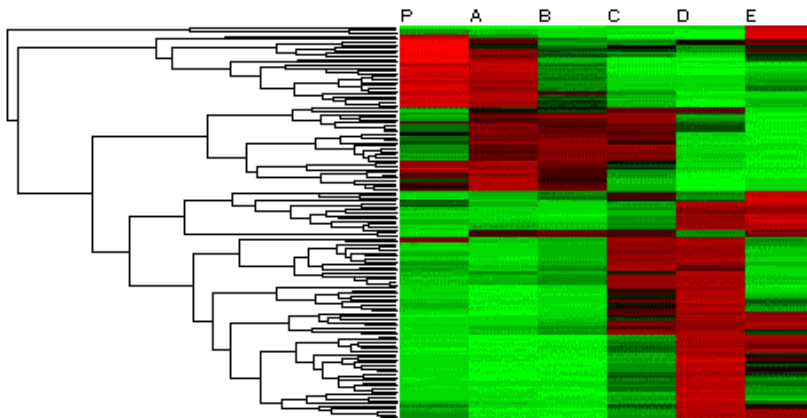
- Easy to use interface: only need a distance matrix and inflation value



Co-expression network modules

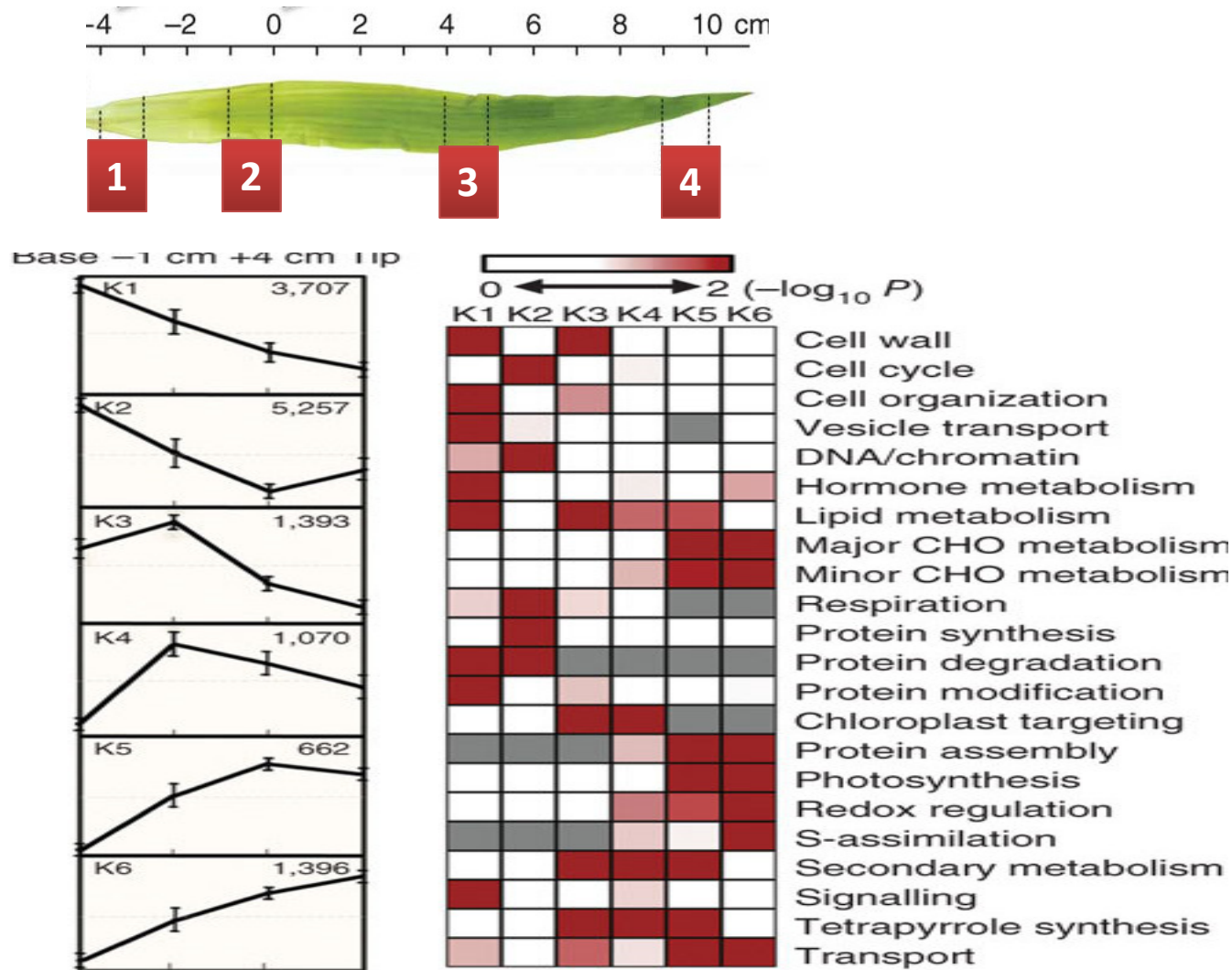
2. WGCNA (weighted correlation network analysis)

- transform the initial distance matrix into Topological Overlap Matrix



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

Presentation of the results, an example



Homework

- Clustering and Function enrichment analysis.
- Starting file:
 - Cuffdiff result: genes.fpk_tracking
 - Rice Gene Ontology annotation file: rice.annot created with Ensembl BioMart.
- Tasks:
 - Hierarchical clustering
 - K-means clustering
 - Function enrichment analysis with BLAST2GO