



Curso de Bioinformática:

Análisis funcional de expresión genética

Ricardo A. Verdugo, Ph.D.

Programa de Genética Humana, ICBM

Facultad de Medicina, U. de Chile

Data analysis workflow

1. Data importation
2. Quality Control (QC)
3. Probe filtering
4. DE testing
5. Clustering
6. Gene set enrichment

3. Differentially expressed genes

- Gene by gene ANOVA

$$y_{ij} = \mu + T_i + \varepsilon_{ij}$$

where,

y_{ij} general logarithm of the gene expression
in i^{th} treatment group of the j^{th} replicate

μ mean

T_i effect of the i^{th} treatment ($i=1 \rightarrow 5$)

ε_{ij} residual effect

2. List of selected genes

- Significant differences between treatments
- FDR transformation of p-values
- 340 selected genes

Table 2. Number of genes significantly affected by the treatments with a global 5% false discovery rate (FDR) accepted.

Treatment	Number of genes	Down regulated		Up regulated	
		Count	Percent	Count	Percent
Arsenic	274	139	50.7	135	49.3
Cadmium	260	143	55.0	117	54.0
Chromate	173	69	39.9	104	60.1
Vanadate	208	93	44.7	115	55.3

Measures of distance

- Manhattan

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

(from Kaufman and Rousseeuw, 1990. pp 11)

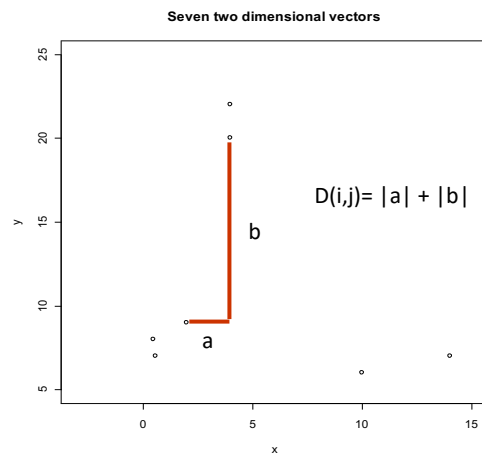
Measures of distance

- Euclidean

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

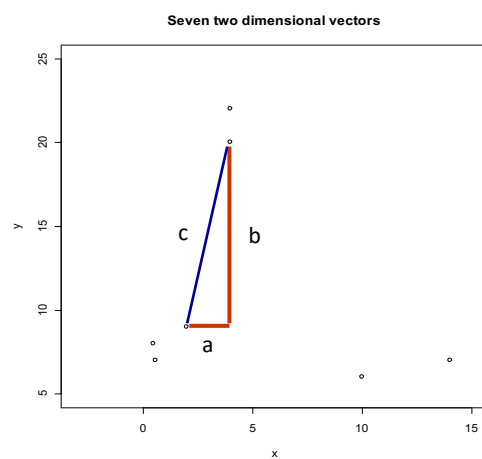
(from Kaufman and Rousseeuw, 1990. pp 11)

Manhattan or city block



Euclidean

$$D(i,j) = \sqrt{a^2 + b^2} = c$$



Measures of distance

- Minkowsky

$$d_m(i,j) = \{ |x_{i1} - x_{j1}|^m + |x_{i2} - x_{j2}|^m + \dots + |x_{ip} - x_{jp}|^m \}^{1/m}$$

(from Drăghici, 2003. pp 265-276)

Measures of distance

- Chebyshev

$$d(i,j) = \max_k |x_{ik} - x_{jk}|$$

(from Drăghici, 2003. pp 265-276)

Measures of distance

- Correlation complement

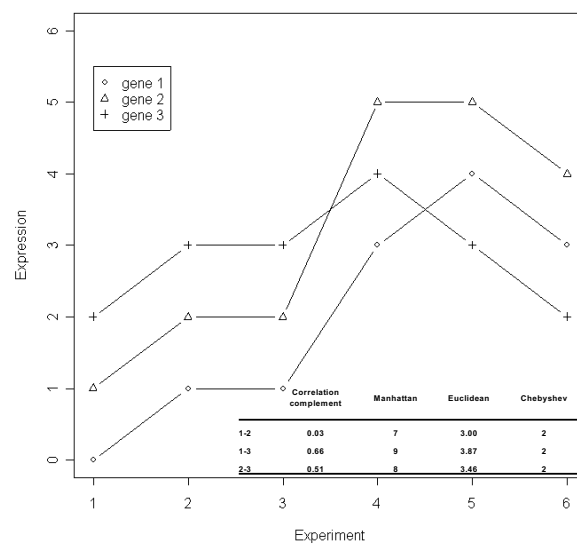
$$d(i,j) = 1 - r$$

Where

r Pearson Correlation

$$r_{xy} = \frac{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{n-1}}{\sqrt{\frac{\sum (X - \bar{X})^2}{n-1} * \frac{\sum (Y - \bar{Y})^2}{n-1}}}$$

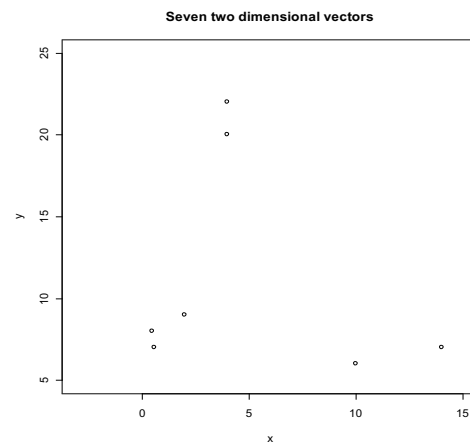
Distance measures



Clustering analysis

- Hierarchical clustering

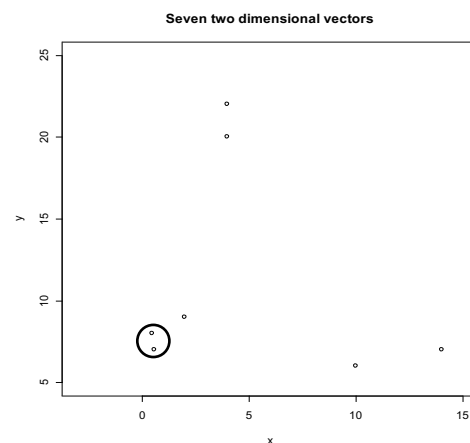
Single linkage



Clustering analysis

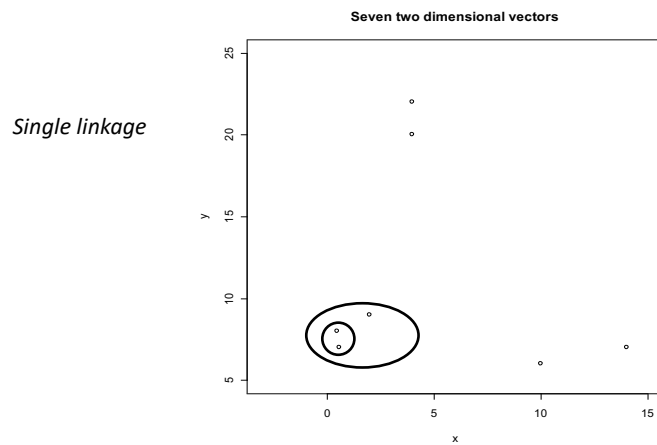
- Hierarchical clustering

Single linkage



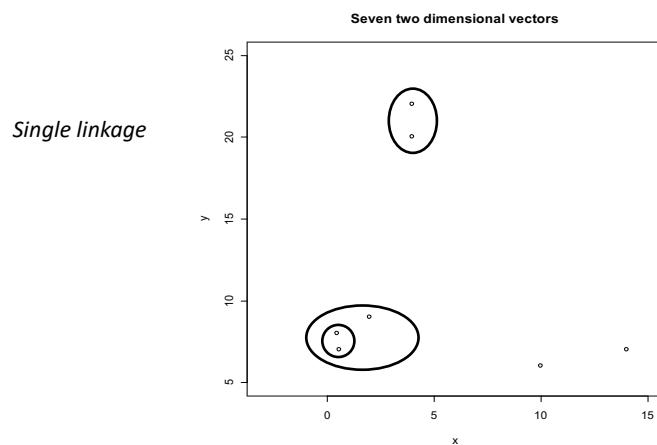
Clustering analysis

- Hierarchical clustering



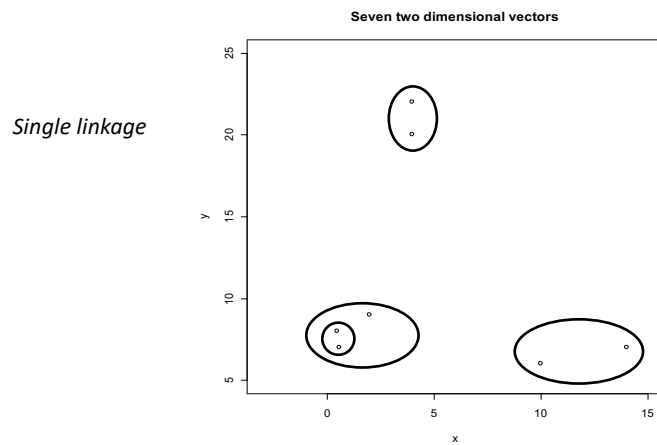
Clustering analysis

- Hierarchical clustering



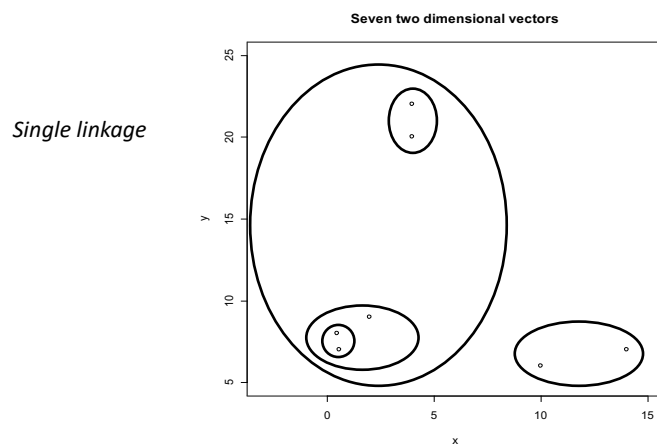
Clustering analysis

- Hierarchical clustering



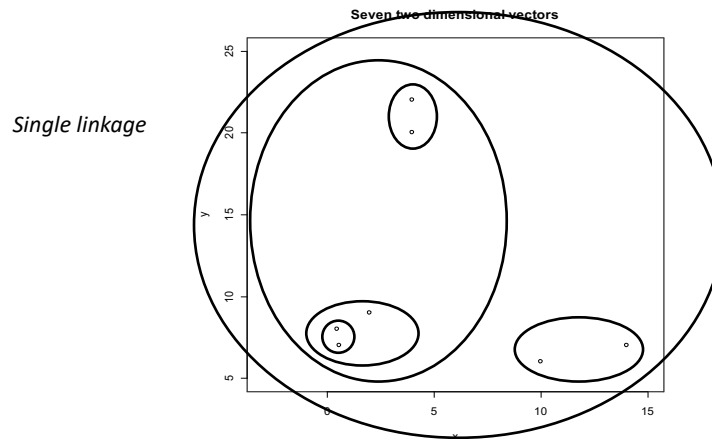
Clustering analysis

- Hierarchical clustering

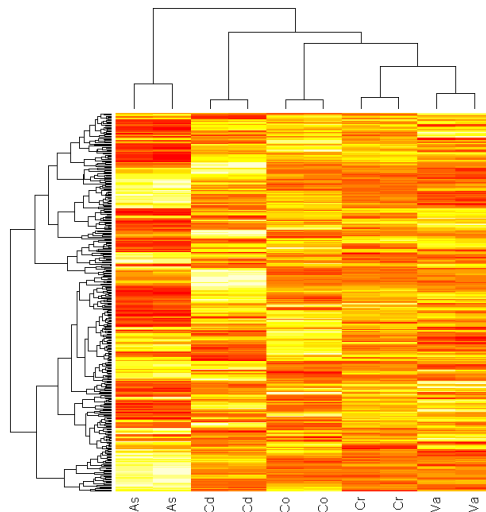


Clustering analysis

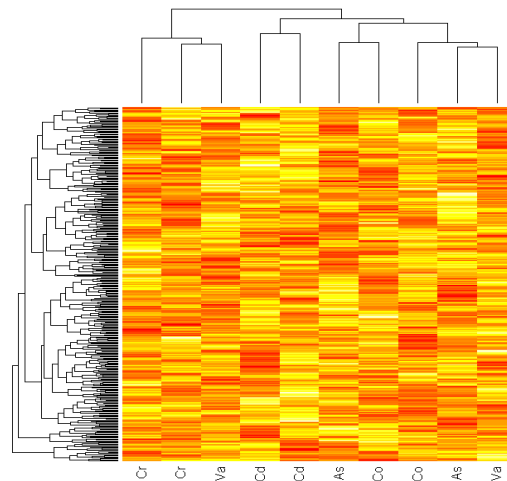
- Hierarchical clustering



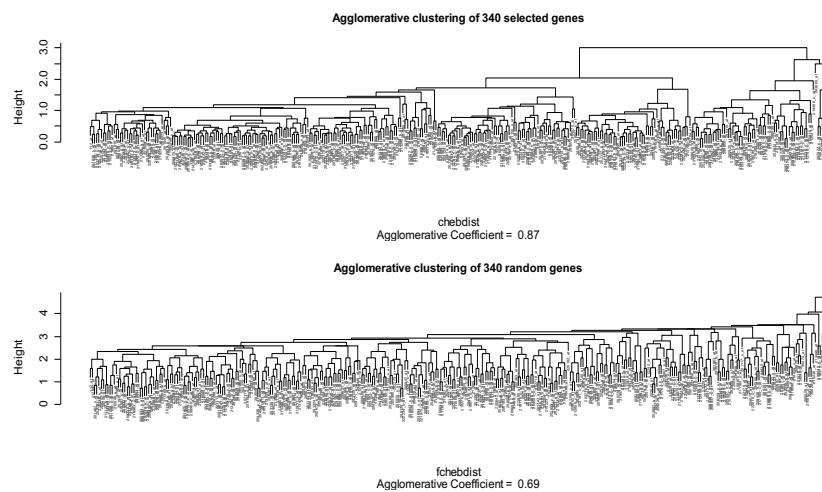
Hierarchical clustering of 340 selected genes



Hierarchical clustering of 340 simulated genes



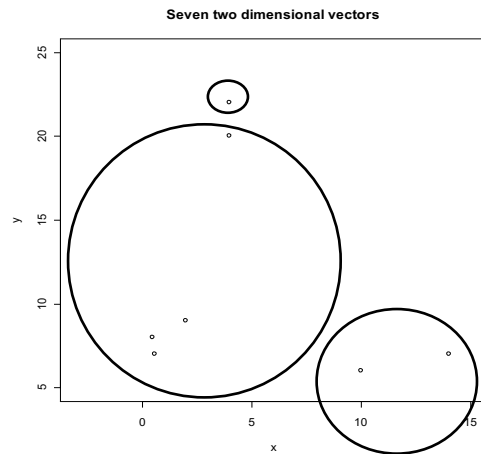
Hierarchical clustering comparison



Partitioning around medoids

Iteration 1

Step 1

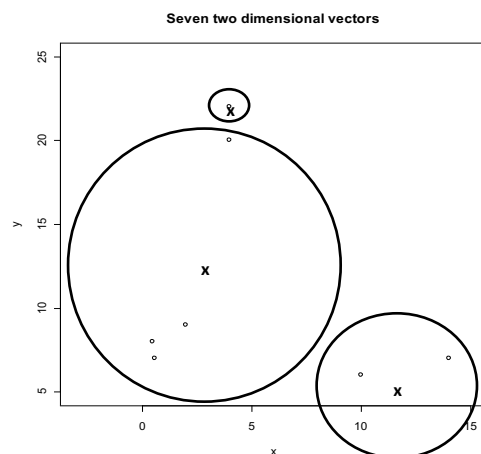


$K=3$

Partitioning around medoids

Iteration 1

Step 2

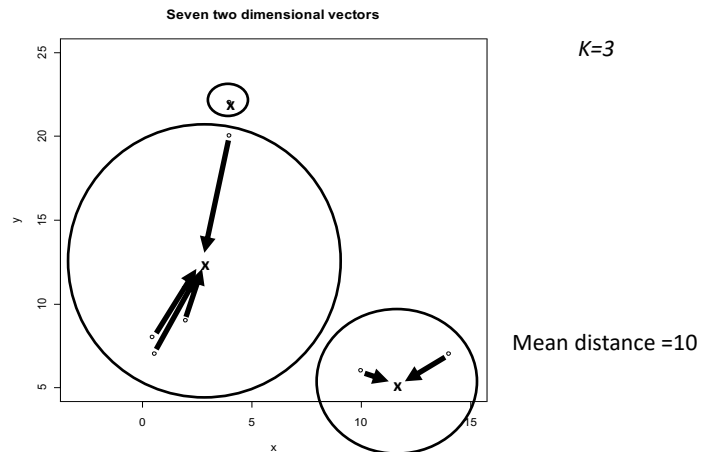


$K=3$

Partitioning around medoids

Iteration 1

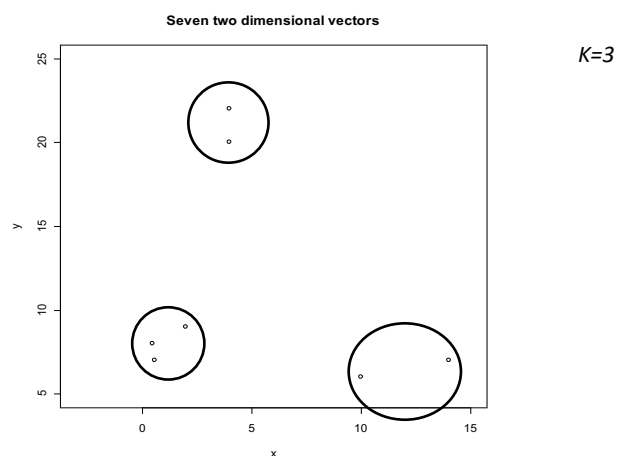
Step 3



Partitioning around medoids

Iteration 2

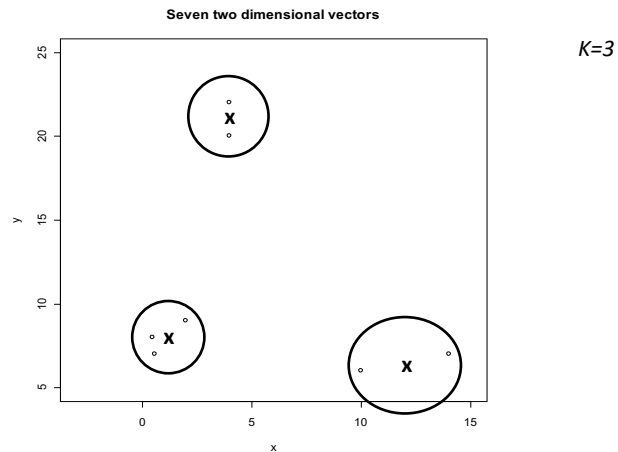
Step 1



Partitioning around medoids

Iteration 2

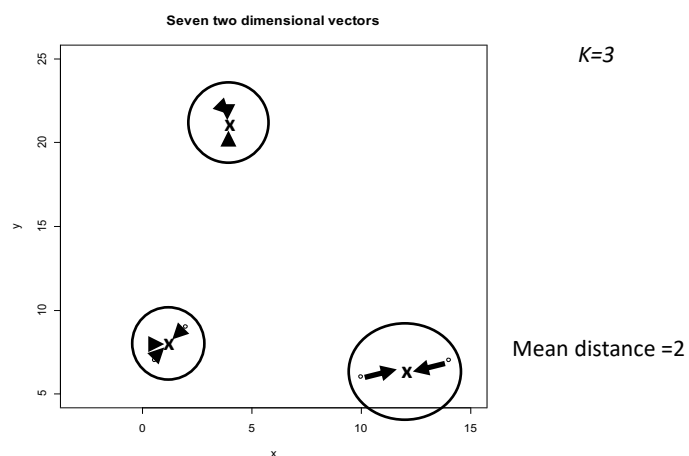
Step 2



Partitioning around medoids

Iteration 2

Step 3



Cluster quality

- Homogeneity

$$H_{ave} = \frac{1}{N_{gene}} \sum_j D(g_j, C(g_j))$$

Chen et al., 2002

g_j is i th gene

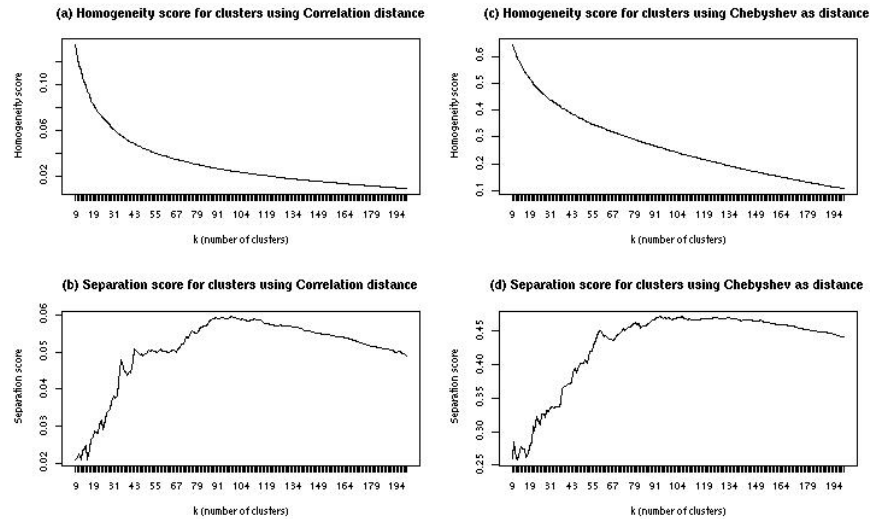
$C(g_j)$ is the center of the cluster that g_j belongs to

Cluster quality

- Separation

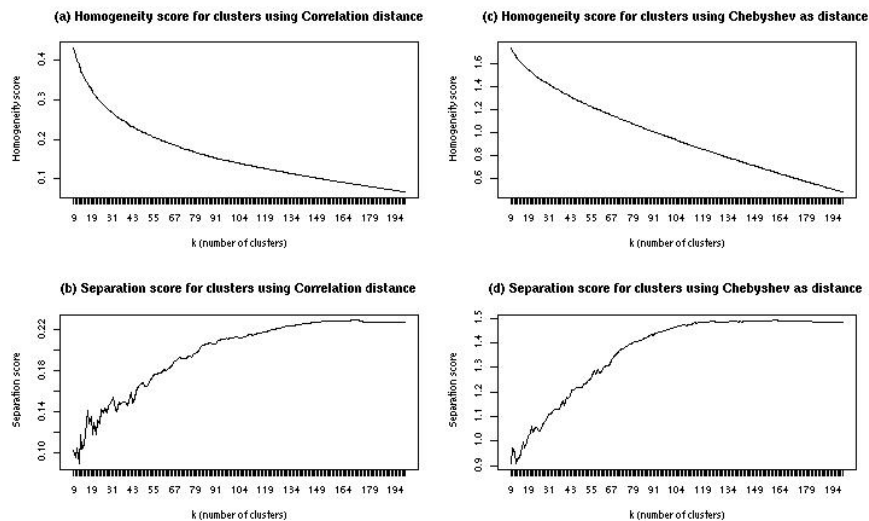
$$S_{min} = \frac{1}{N_{cluster}} \sum_j \min_{i \neq k} (D(g_{ij}, g_{kl}))$$

Cluster quality evaluation using the Correlation and Chebyshev distance measures
for k between 9 and 200



Cluster analysis on 340 genes using two measures of distance with k from 9 to 200 clusters. Average homogeneity (a) and separation scores (b) between clusters are calculated using Correlation distance between genes. Equivalent parameters using Chebyshev as a measure of similarity are shown in (c) and (d).

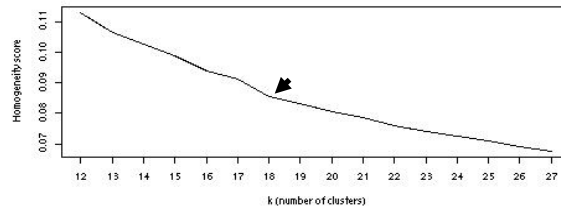
Cluster quality evaluation using the Correlation and Chebyshev distance measures
for k between 9 and 200



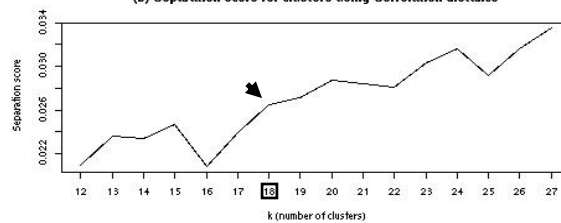
Cluster analysis on 340 genes using two measures of distance with k from 9 to 200 clusters. Average homogeneity (a) and separation scores (b) between clusters are calculated using Correlation distance between genes. Equivalent parameters using Chebyshev as a measure of similarity are shown in (c) and (d).

Cluster quality evaluation using the Correlation distance measure
for k between 12 and 27

(a) Homogeneity score for clusters using Correlation distance



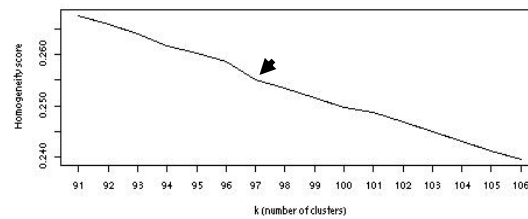
(b) Separation score for clusters using Correlation distance



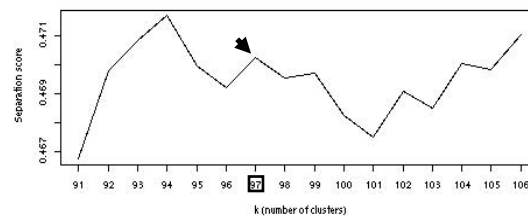
Cluster analysis on 340 genes using the Correlation measure of distance between genes with k from 12 to 27 clusters. Average homogeneity (a) and separation scores (b) between clusters are calculated.

Cluster quality evaluation using the Chebyshev distance measure
for k between 91 and 106

(a) Homogeneity score for clusters using Chebyshev distance



(b) Separation score for clusters using Chebyshev distance



Cluster analysis on 340 genes using the Chebyshev measure of distance between genes with k from 91 to 106 clusters. Average homogeneity (a) and separation scores (b) between clusters are calculated.

Functional Analysis

- We would like to know:
 1. Are these clusters meaningful?
 2. What do these genes have in common?

Functional Analysis

Are these clusters meaningful?

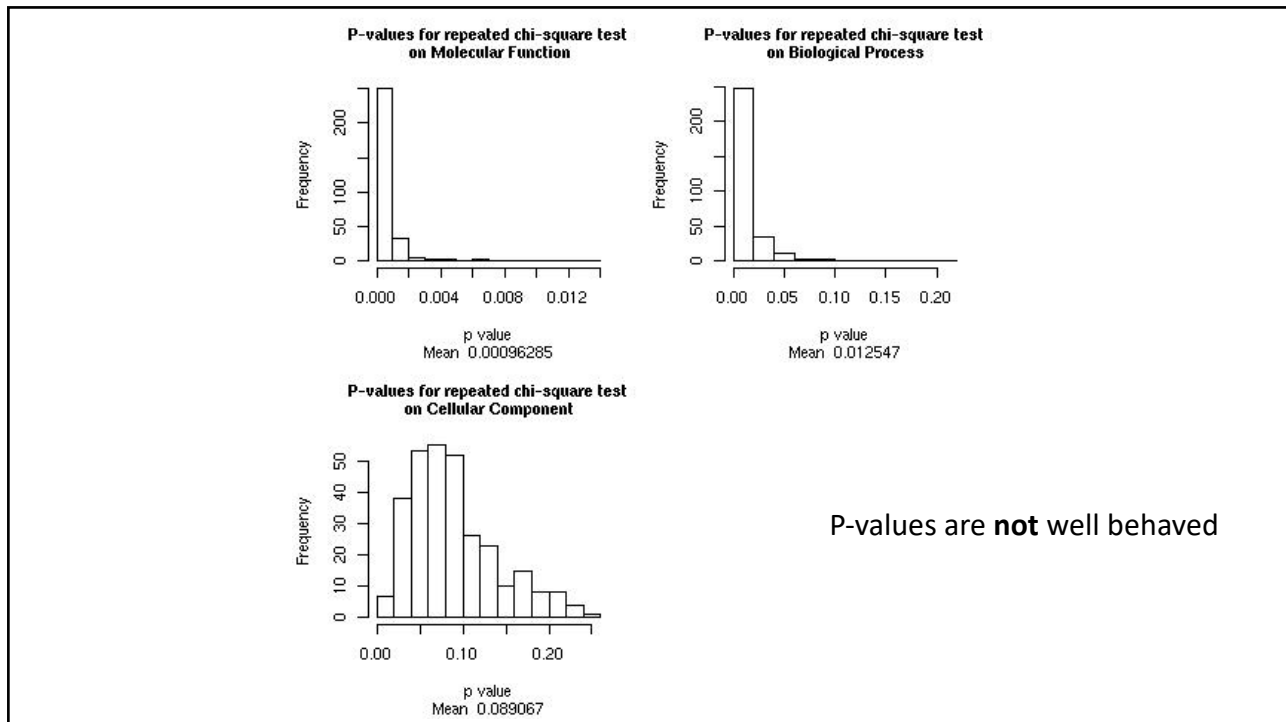
1. Chi-square test for association between clusters and functional annotations
2. Gene Ontology: Molecular function, Biological Process, Subcellular Localization

Functional Analysis

- Problems
 - Chi-square test does not perform well with small samples ($n=5$ or more)
 - Simulate expected values from the data

Functional Analysis

- Problems
 - One gene can have more than one function (domain) and it can be located in more than one compartment
 - Randomly choose one
 - Repeat hundreds of times
 - Keep the p-value
 - Use the mean p-value as an estimate



Functional Analysis

- We want to know:
 1. What genes are over-represented?
 2. What genes are under-represented?

Onto-Epress : Khatri *et al.* 2002 (Genomics 79(2): 266-70)

Functional testing

- First (naïve) approach: Fisher Exact test for over representation:

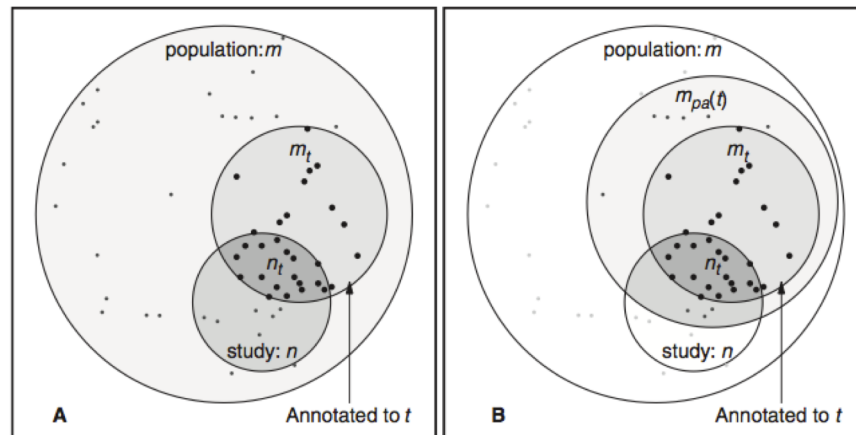
$$\text{— Ho: } \frac{\text{selected in group}}{\text{selected}} = \frac{\text{gene in group}}{\text{genes universe}}$$

- No consideration for overlap among functional groups
- Many redundant signals
- Easy to use, DAVID (david.abcc.ncifcrf.gov)

Functional Testing 2.0

- Parent-Child:
 - Accounts for groups hierarchy (GO)
 - Fisher test on conditional on parent group (pa)
 - H0: $\frac{\text{selected in group} \cap \text{pa}}{\text{selected} \cap \text{pa}} = \frac{\text{gene in group} \cap \text{pa}}{\text{genes universe} \cap \text{pa}}$
 - R-package: topGO
 - Grossman 2008, Bioinformatics 23:3024

Parent-Child enrichment test



Functional Testing 2.0

- Elim algorithm:
 - Accounts for groups hierarchy (GO)
 - Fisher test on conditional on enriched children groups (chi)
 - H0: $\frac{\text{selected in group} / \text{chi}}{\text{selected} / \text{chi}} = \frac{\text{gene in group} / \text{chi}}{\text{genes universe} / \text{chi}}$
 - R-package: topGO
 - Grossman 2008, Bioinformatics 23:3024

There are more methods available

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

Table 1: *Algorithms currently supported by topGO.*

Functional tests: conclusions

- There are several algorithms that can remove redundancy from list of enrichments in GO
- Tests other than Fishers' exact can better use quantitative scores per gene (p-values instead of DE classification)
- topGO packages provides a consistent framework for comparing methods
- topGO is well documented, give give it a try!

References

- Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**: 289-300.
- Durbin, B. and D. M. Rocke (2003). Estimation of transformation parameters for microarray data. *Bioinformatics* **19**(11): 1360-7.
- Everitt, B. (1980). *Cluster Analysis*. NY, USA.
- Kaufman, L. and P. Rosseeuw (1990). *Finding Groups in Data: An introduction to cluster analysis*. USA, John Wiley & Sons, Inc.: 68-119.
- Khatri, P., S. Draghici, et al. (2002). Profiling gene expression using onto-express. *Genomics* **79**(2): 266-70.