FACULTAD DE MEDICINA
UNIVERSIDAD DE CHILE

U-GENOMA
Red de Investigación y Docencia

ICBM
INSTITUTO
DE CIENCIAS
BIOMÉDICAS

Curso de Bioinformática:

Análisis funcional de expresión genética

Ricardo A. Verdugo, Ph.D.
Programa de Genética Humana, ICBM
Facultad de Medicina, U. de Chile

mayo de 2019

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->5)
ε_{ij}	residual effect

2. List of selected genes

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

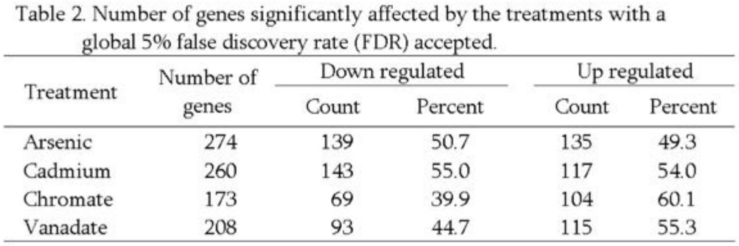
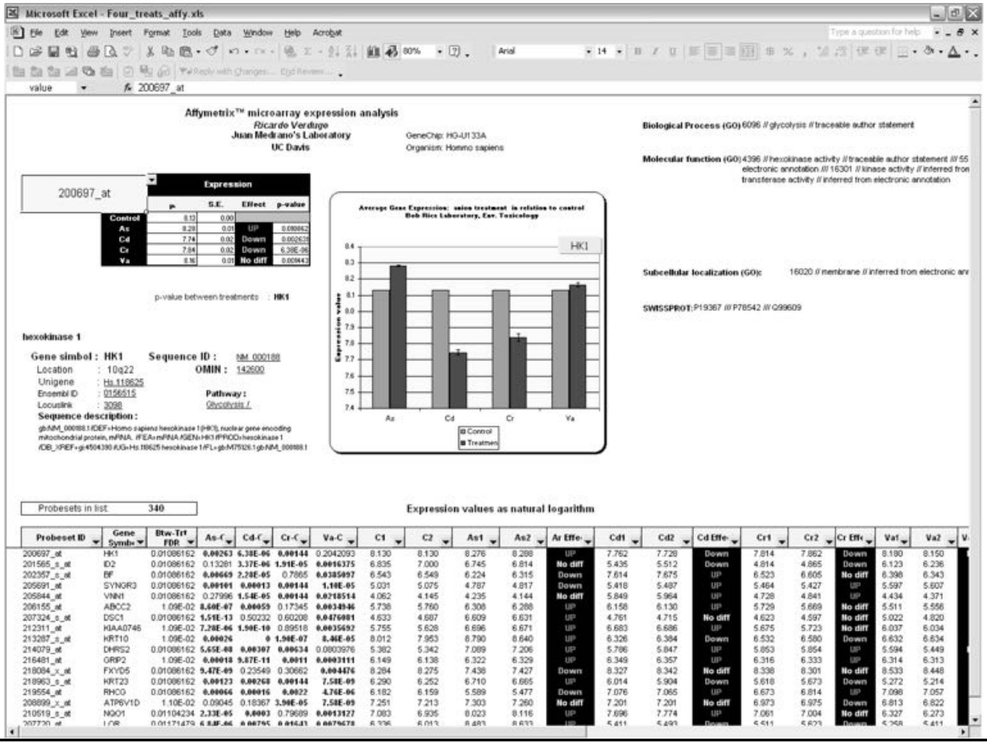


Table 2. Number of genes significantly affected by the treatments with a global 5% false discovery rate (FDR) accepted.

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


Microsoft Excel - Four_treats_affy.xls
File Edit View Insert Format Tools Data Window Help Acrobat
Type a question for help
80%
Arial
14
Reply with Changes... End Review...
value
200697_at
Average Gene Expression: anion treatment in relation to control
Bob Rice Laboratory, Env. Toxicology
HK1
Expression value
8.4
8.3
8.2
8.1
8.0
7.9
7.8
7.7
7.6
7.5
7.4
As
Cd
Cr
Va
Control
Treatmen

Affymetrix™ microarray expression analysis
Ricardo Verdugo
Juan Medrano's Laboratory
UC Davis

GeneChip: HG-U133A
Organism: Hommo sapiens

Biological Process (GO) 6096 // glycolysis // traceable author statement

Molecular function (GO) 4396 // hexokinase activity // traceable author statement /// 55
electronic annotation /// 16301 // kinase activity // inferred fron
transferase activity // inferred from electronic annotation

200697_at

Expression	p	S.E.	Effect	p-value
Control	8.13	0.00		
As	8.28	0.01	UP	0.000062
Cd	7.74	0.02	Down	0.002638
Cr	7.84	0.02	Down	6.38E-06
Va	8.16	0.01	No diff	0.008443

p-value between treatments : **HK1**

Subcellular localization (GO): 16020 // membrane // inferred fron electronic an

SWISSPROT: P19367 /// P78542 /// Q99609

hexokinase 1

Gene simbol : HK1 **Sequence ID :** NM_000188
Location : 10q22 **OMIN :** 142600
Unigene : Hs.118625
Ensembl ID : 0156515 **Pathway :**
Locuslink : 3098 Glycolysis /
Sequence description :
gb:NM_000188.1 /DEF=Homo sapiens hexokinase 1 (HK1), nuclear gene encoding mitochondrial protein, mRNA. /FEA=mRNA /GEN=HK1 /PROD=hexokinase 1 /DB_XREF=gi:4504390 /UG=Hs.118625 hexokinase 1 /FL=gb:M75126.1 gb:NM_000188.1

Probesets in list 340

Expression values as natural logarithm

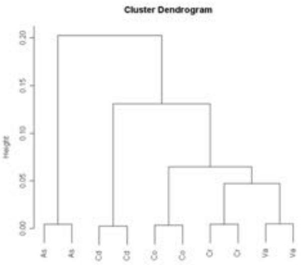
Probeset ID	Gene Symbo	Btw-Trt FDR	As-C	Cd-C	Cr-C	Va-C	C1	C2	As1	As2	Ar Effe	Cd1	Cd2	Cd Effe	Cr1	Cr2	Cr Effe	Va1	Va2	V
200697_at	HK1	0.01086162	**0.00263**	**6.38E-06**	**0.00144**	0.2042093	8.130	8.130	8.276	8.288	UP	7.762	7.728	Down	7.814	7.862	Down	8.180	8.150	
201565_s_at	ID2	0.01086162	0.13281	**3.37E-06**	**1.91E-05**	**0.0016375**	6.835	7.000	6.745	6.814	No diff	5.435	5.512	Down	4.814	4.865	Down	6.123	6.236	
202357_s_at	BF	0.01086162	**0.00669**	**2.28E-05**	0.7865	**0.0385097**	6.543	6.549	6.224	6.315	Down	7.614	7.675	UP	6.523	6.605	No diff	6.398	6.343	
205691_at	SYNGR3	0.01086162	**0.00101**	**0.00013**	**0.00144**	**1.10E-05**	5.031	5.075	4.787	4.817	Down	5.418	5.487	UP	5.464	5.427	UP	5.597	5.607	
205844_at	VNN1	0.01086162	0.27996	**1.54E-05**	**0.00144**	**0.0218514**	4.062	4.145	4.235	4.144	No diff	5.849	5.964	UP	4.728	4.841	UP	4.434	4.371	
206155_at	ABCC2	1.09E-02	**8.60E-07**	**0.00059**	0.17345	**0.0034946**	5.738	5.760	6.308	6.288	UP	6.158	6.130	UP	5.729	5.669	No diff	5.511	5.556	
207324_s_at	DSC1	0.01086162	**1.51E-13**	0.50232	0.60208	**0.0476081**	4.633	4.687	6.609	6.631	UP	4.761	4.715	No diff	4.623	4.597	No diff	5.022	4.820	
212311_at	KIAA0746	1.09E-02	**7.28E-06**	**1.90E-10**	0.89518	**0.0035692**	5.755	5.628	6.696	6.671	UP	6.683	6.686	UP	5.675	5.723	No diff	6.037	6.034	
213287_s_at	KRT10	1.09E-02	**0.00026**	**0**	**1.90E-07**	**8.46E-05**	8.012	7.953	8.790	8.640	UP	6.326	6.384	Down	6.532	6.580	Down	6.632	6.634	
214079_at	DHRS2	0.01086162	**5.65E-08**	**0.00307**	**0.00634**	0.0803976	5.382	5.342	7.089	7.206	UP	5.786	5.847	UP	5.853	5.854	UP	5.594	5.449	
216481_at	GRIP2	1.09E-02	**0.00018**	**9.87E-11**	**0.0011**	**0.0003111**	6.149	6.138	6.322	6.329	UP	6.349	6.357	UP	6.316	6.333	UP	6.314	6.313	
218084_x_at	FXYD5	0.01086162	**9.47E-09**	0.23549	0.30662	**0.004476**	8.284	8.275	7.438	7.427	Down	8.327	8.342	No diff	8.338	8.301	No diff	8.533	8.448	
218963_s_at	KRT23	0.01086162	**0.00123**	**0.00268**	**0.00144**	**7.58E-09**	6.290	6.252	6.710	6.665	UP	6.014	5.904	Down	5.618	5.673	Down	5.272	5.214	
219554_at	RHCG	0.01086162	**0.00066**	**0.00016**	**0.0022**	**4.76E-06**	6.182	6.159	5.589	5.477	Down	7.076	7.065	UP	6.673	6.814	UP	7.098	7.057	
208899_x_at	ATP6V1D	1.10E-02	0.09045	0.18367	**3.90E-05**	**7.58E-09**	7.251	7.213	7.303	7.260	No diff	7.201	7.201	No diff	6.973	6.975	Down	6.813	6.822	
210519_s_at	NQO1	0.01104234	**2.33E-05**	**0.0003**	0.79689	**0.0013127**	7.083	6.935	8.023	8.116	UP	7.696	7.774	UP	7.061	7.004	No diff	6.327	6.273	
207720_at	LOR	0.01171479	**6.84E-06**	**0.00795**	**0.01641**	**0.0079678**	6.336	6.013	8.483	8.633	UP	5.411	5.493	Down	5.511	5.623	Down	5.258	5.411	

Clustering Analysis

Finds groups of similar genes

Measures of distance:

- •Manhattan
- •Euclidean
- •Minkowski distance
- •Chebychev
- •Correlation complement



Measures of distance

- Manhattan

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

(from Kaufam and Rousseeuw, 1990. pp 11)

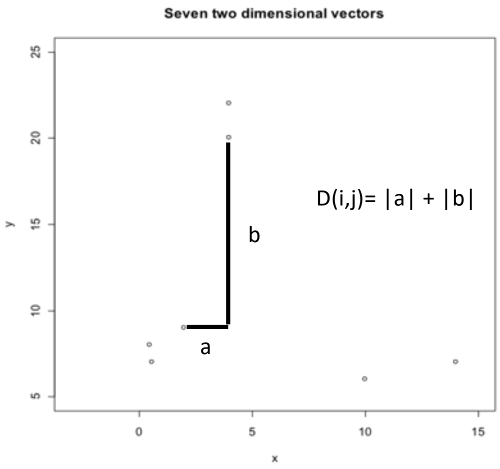
Measures of distance

- Euclidean

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

(from Kaufam and Rousseeuw, 1990. pp 11)

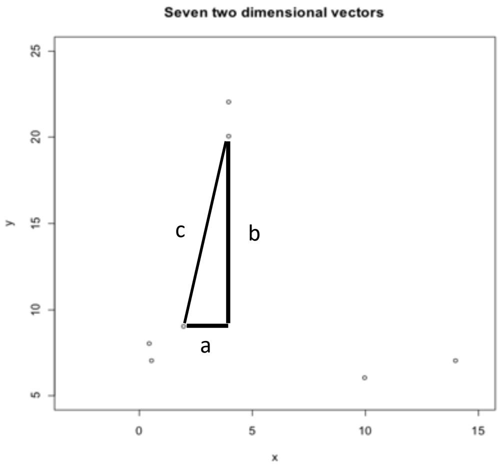
Manhattan or city block



D(i,j)= |a| + |b|

Euclidean

D(i,j)= √a2 + b2 = c



Measures of distance

- Minkowsky

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

(from Drăghici, 2003. pp 265-276)

Measures of distance

- Chebyshev

$$d(\mathbf{i,j}) = \max_{k} | x_{i_k} - x_{jk} |$$

(from Drăghici, 2003. pp 265-276)

Measures of distance

- Correlation complement

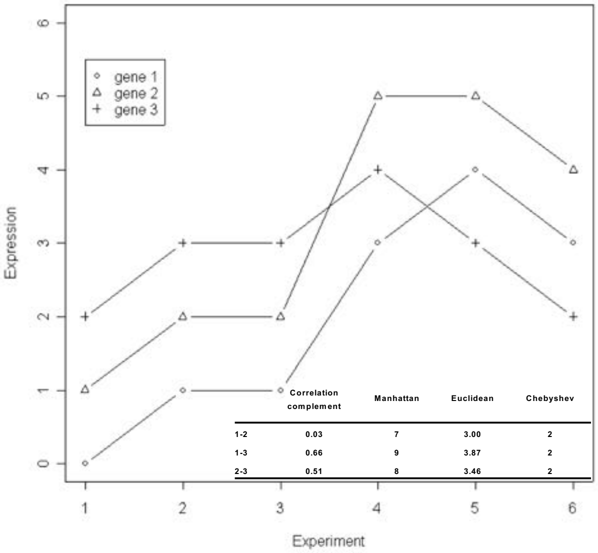
d(**i,j**) = 1 - r

Where

r Pearson Correlation

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

Distance measures

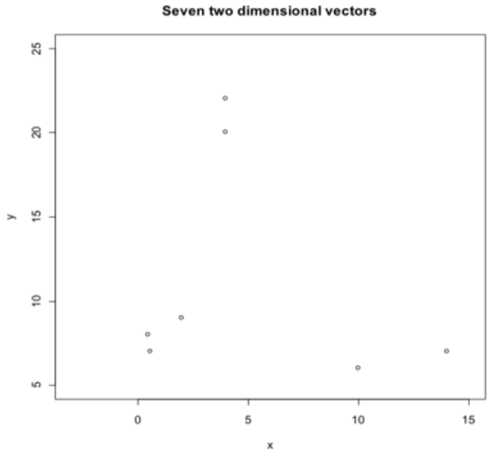


	Correlation complement	Manhattan	Euclidean	Chebyshev
1-2	0.03	7	3.00	2
1-3	0.66	9	3.87	2
2-3	0.51	8	3.46	2

Clustering analysis

- Hierarchical clustering

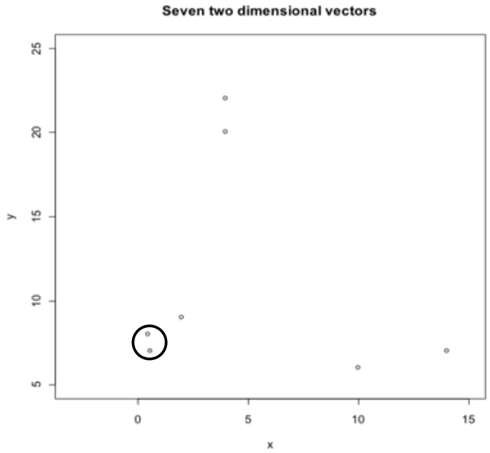
Single linkage



Clustering analysis

- Hierarchical clustering

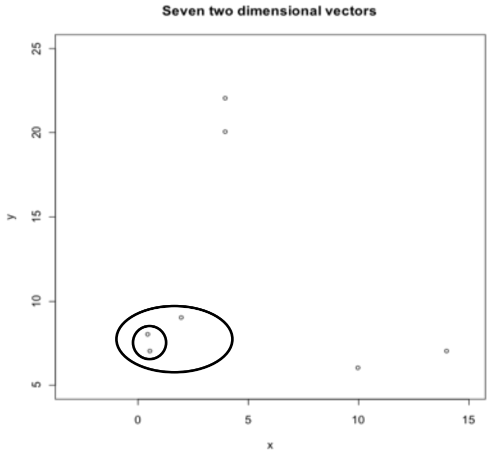
Single linkage



Clustering analysis

- Hierarchical clustering

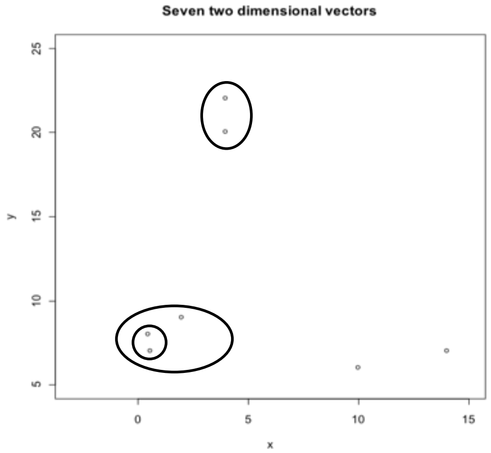
Single linkage



Clustering analysis

- Hierarchical clustering

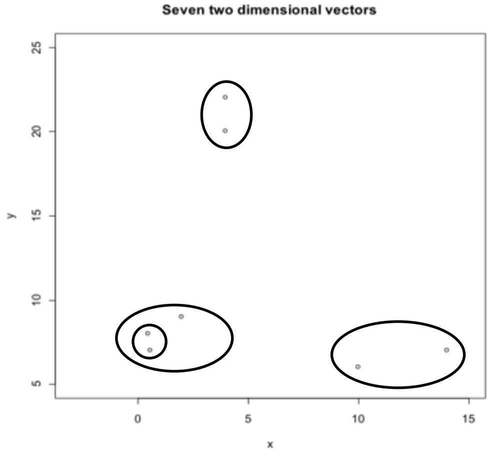
Single linkage



Clustering analysis

- Hierarchical clustering

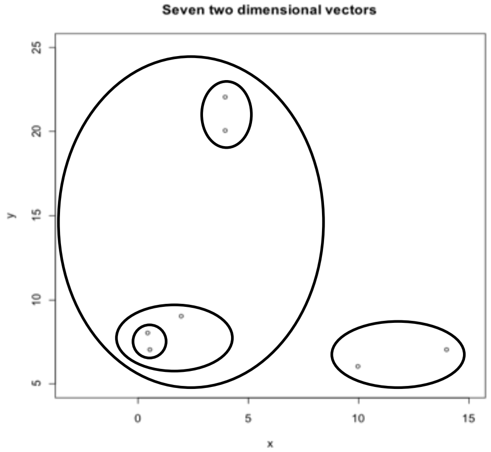
Single linkage



Clustering analysis

- Hierarchical clustering

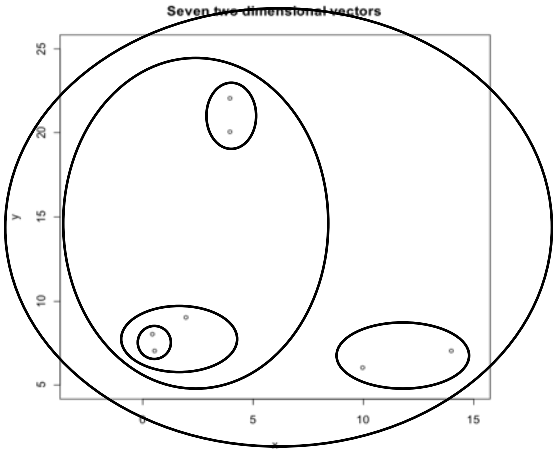
Single linkage



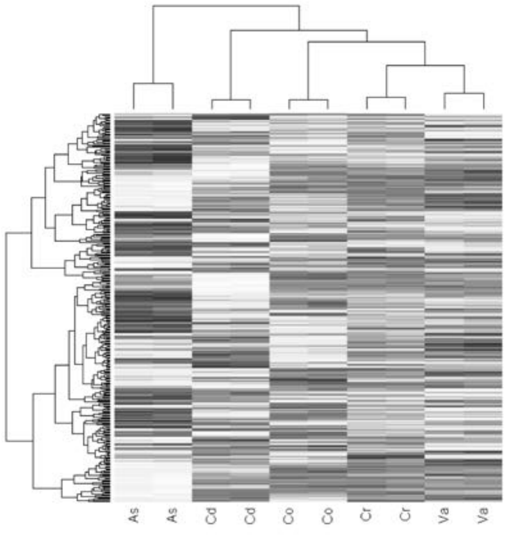
Clustering analysis

- Hierarchical clustering

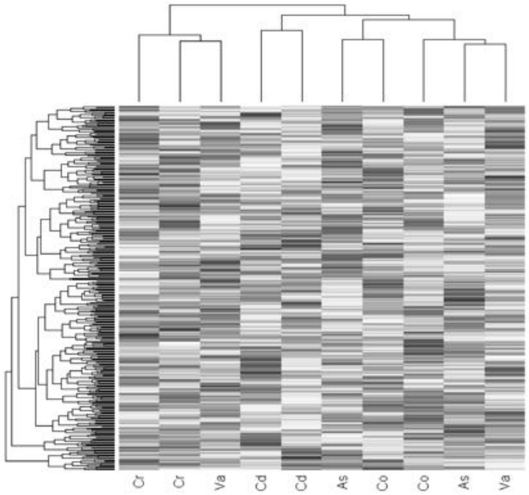
Single linkage



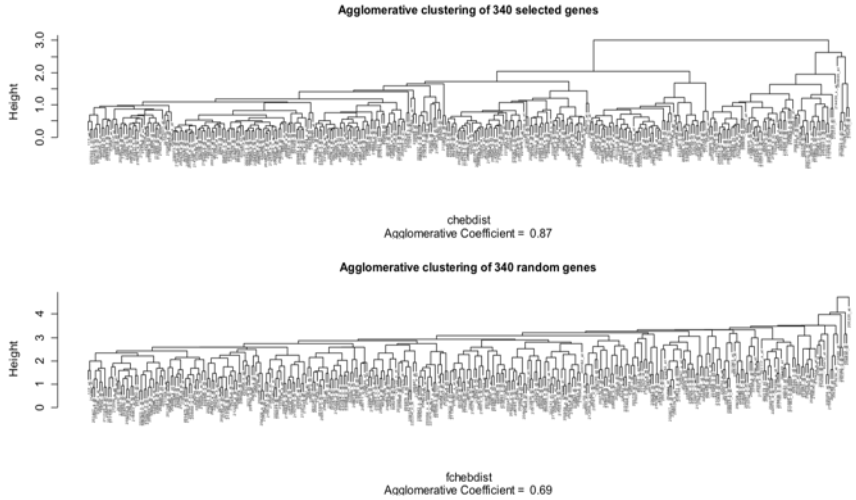
Hierarchical clustering of 340 selected genes



Hierarchical clustering of 340 simulated genes



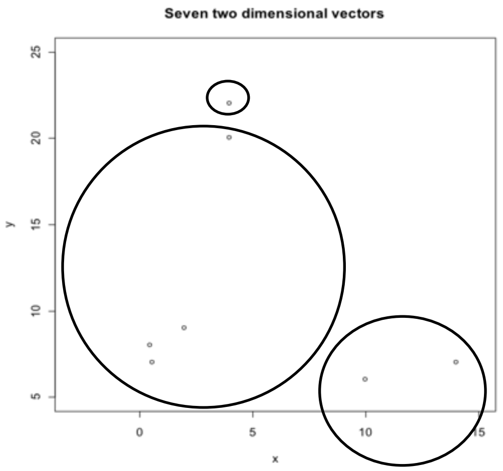
Hierarchical clustering comparison



Partitioning around medoids

Iteration 1

Step 1

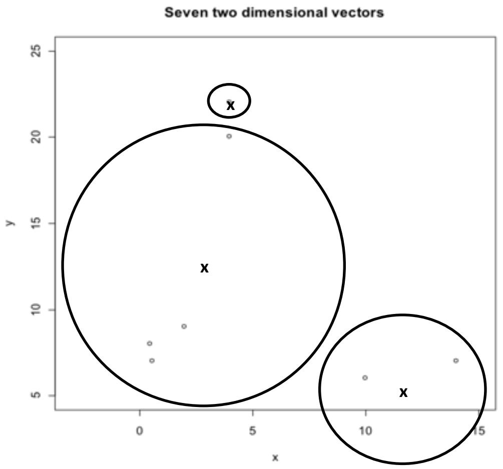


K=3

Partitioning around medoids

Iteration 1

Step 2

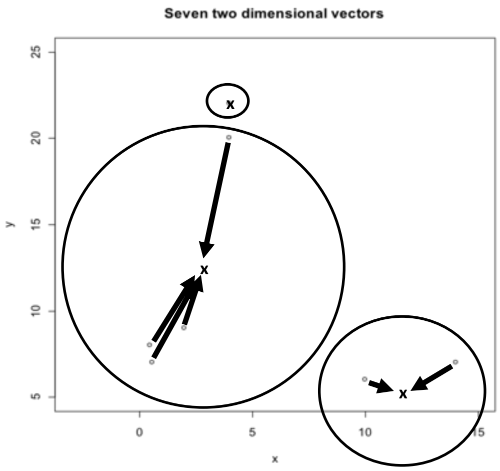


K=3

Partitioning around medoids

Iteration 1

Step 3



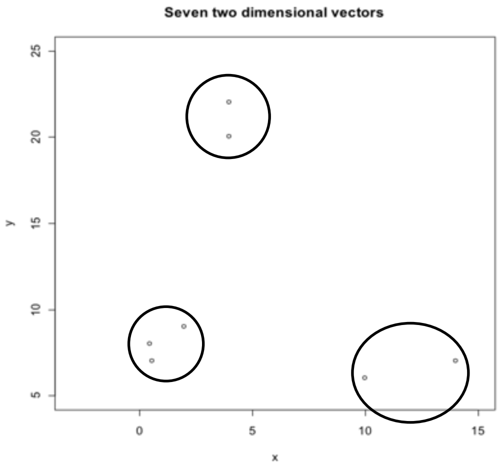
K=3

Mean distance =10

Partitioning around medoids

Iteration 2

Step 1

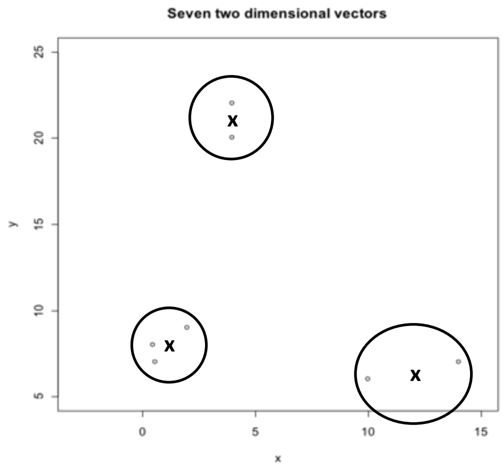


K=3

Partitioning around medoids

Iteration 2

Step 2

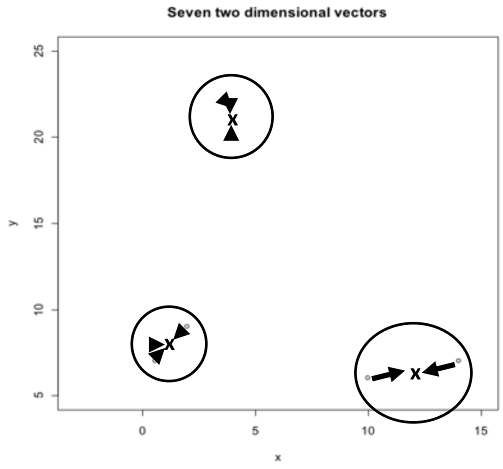


K=3

Partitioning around medoids

Iteration 2

Step 3



K=3

Mean distance =2

Cluster quality

- Homogeneity

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

Chen et al., 2002

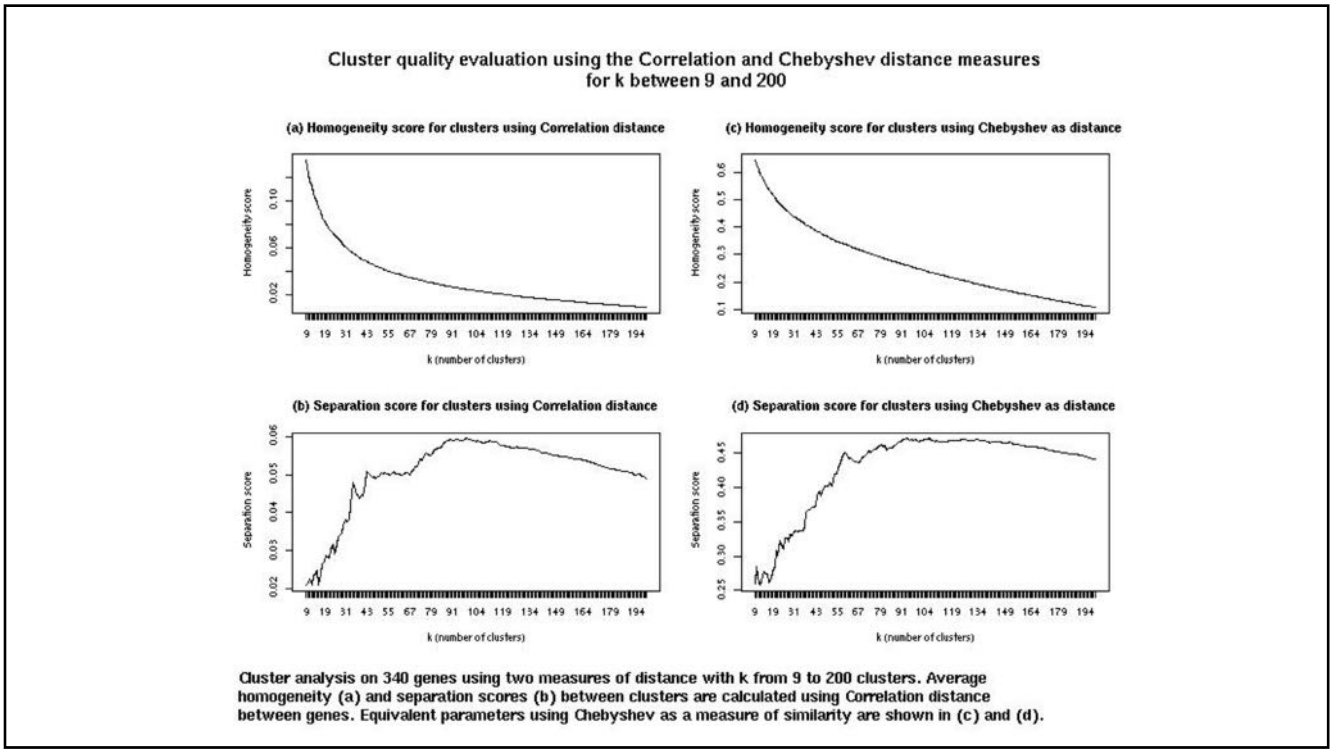
g_j is ith gene

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

Cluster quality

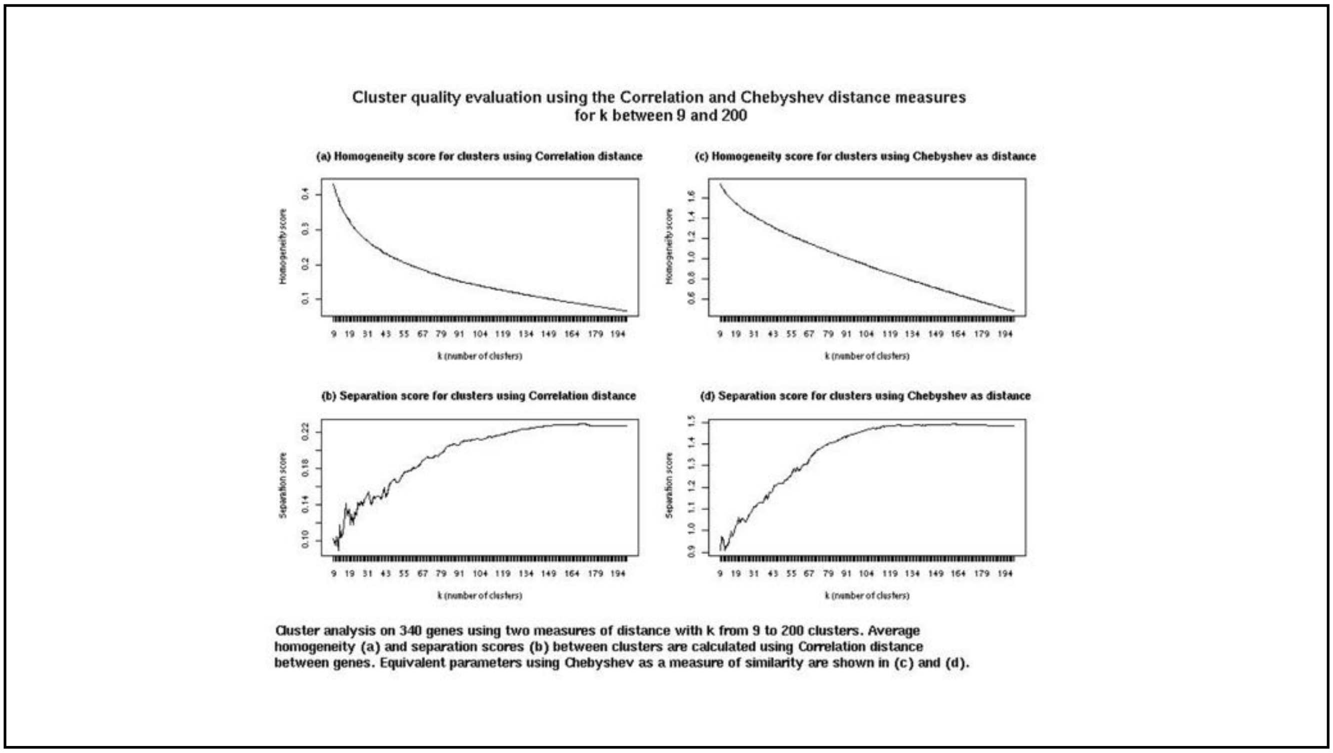
- Separation

$$S_{min} = \frac{1}{N_{cluster}} \sum \min_{i \neq k,} (D(g_{ij} , \underset{j}{g}_{kl}))$$



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

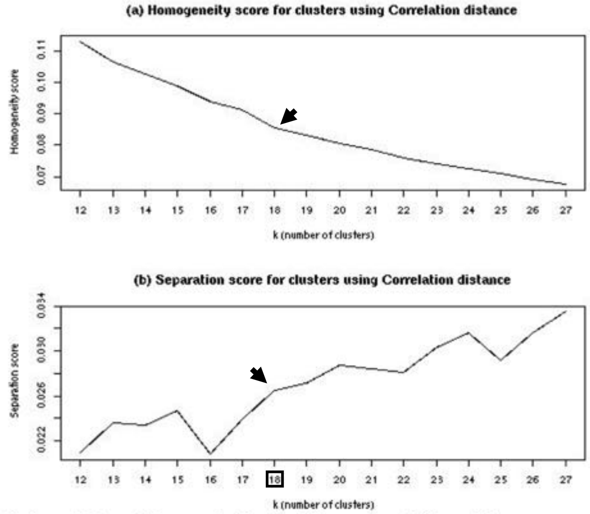
Cluster analysis on 340 genes using two measures of distance with k from 9 to 200 clusters. Average homogeneity (a) and separation scores (b) between clusters are calculated using Correlation distance between genes. Equivalent parameters using Chebyshev as a measure of similarity are shown in (c) and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

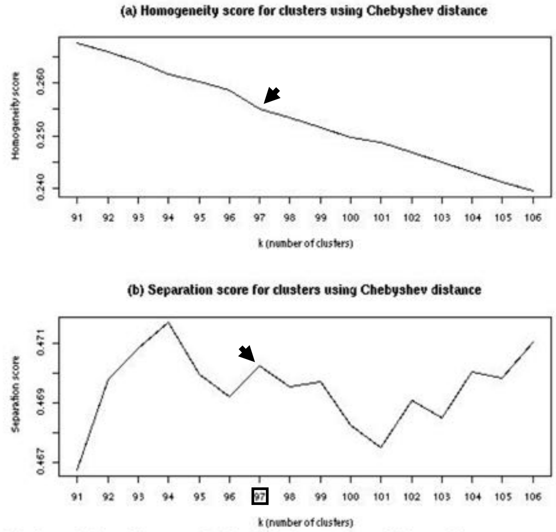
Cluster quality evaluation using the Correlation distance measure
for k between 12 and 27


(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 this genes have in common?

Functional Analysis

Are these clusters meaningful?

1. Chi-square test for association between clusters and functional annotations

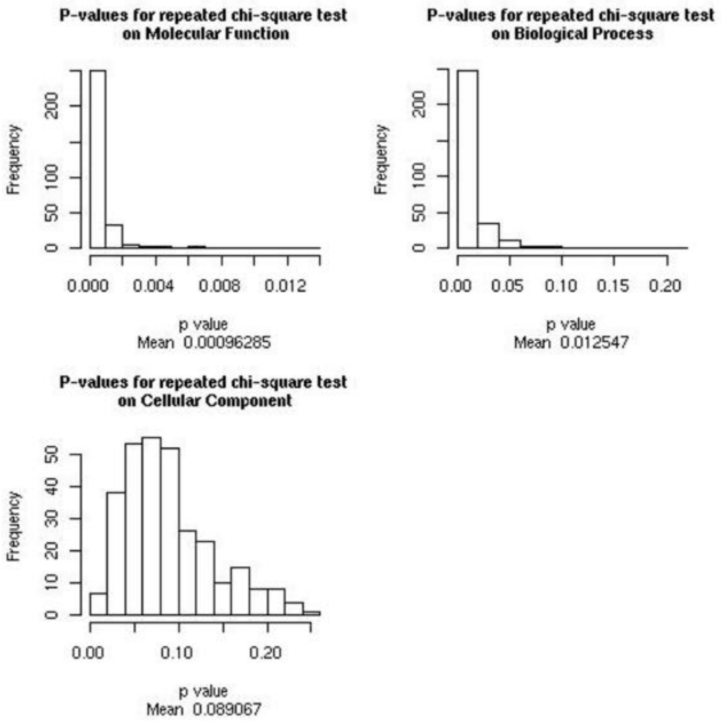
2. Gene Ontology: Molecular function, Biological Process, Subcellular Localization

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-vale
    - Use the mean p-value as an estimate



P-values for repeated chi-square test on Molecular Function

P-values for repeated chi-square test on Biological Process

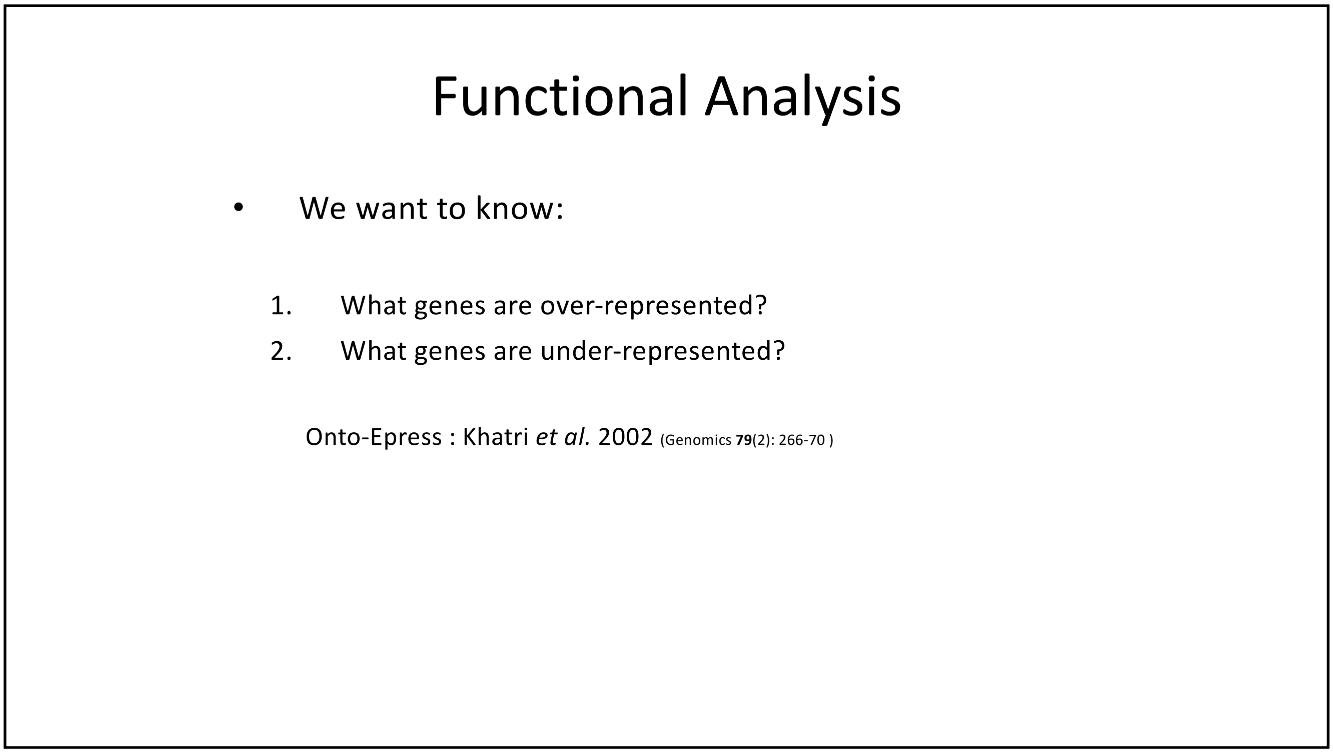
P-values for repeated chi-square test on Cellular Component

Bioconductor example

```
# Find EG for Illumina probe "6840019"
> get("6840019", illuminaMousev1ProbeIDENTREZID)
[1] "78928”
```

What if expression values are associated to Target ID?

```
E.g. “18S_rRNA_X00686_523-S”
Illumina genelist Mouse-6 V1b: NM_133779
EntreGene: 78928
                 OR
> probes2target=read.csv(“Mouse-6_V1b.csv")
> probes2target=unique(probes2target[,c("Target", "ProbeId")])
> probes2target[match("18S_rRNA_X00686_523-S", probes2target$Target),]
                   Target ProbeId
374 18S_rRNA_X00686_523-S 6840019
```

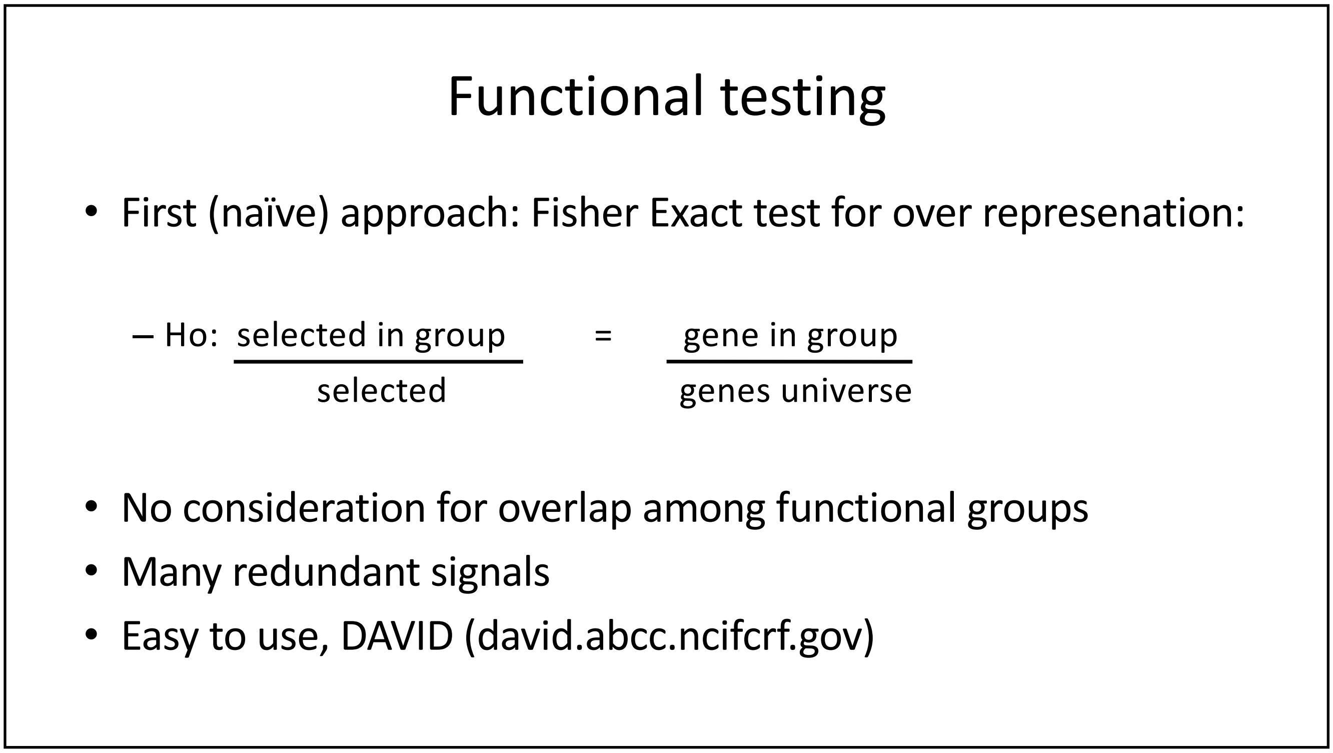


Functional Analysis

- 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)



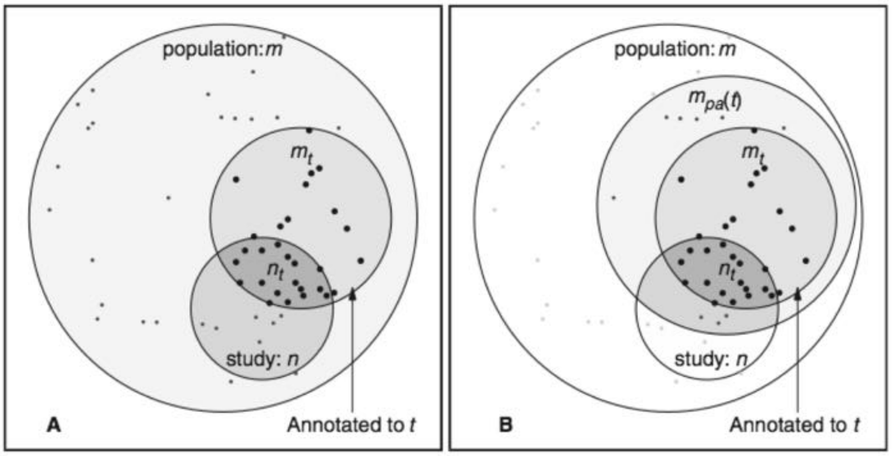
Functional testing

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

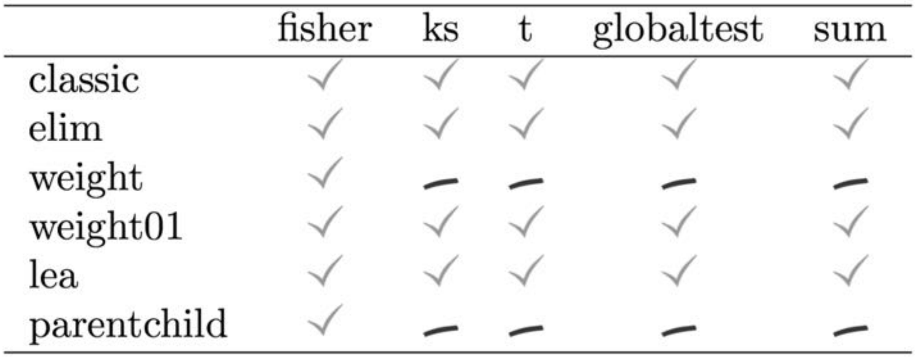
Parent-Child enrichment test



Functional Testing 2.0

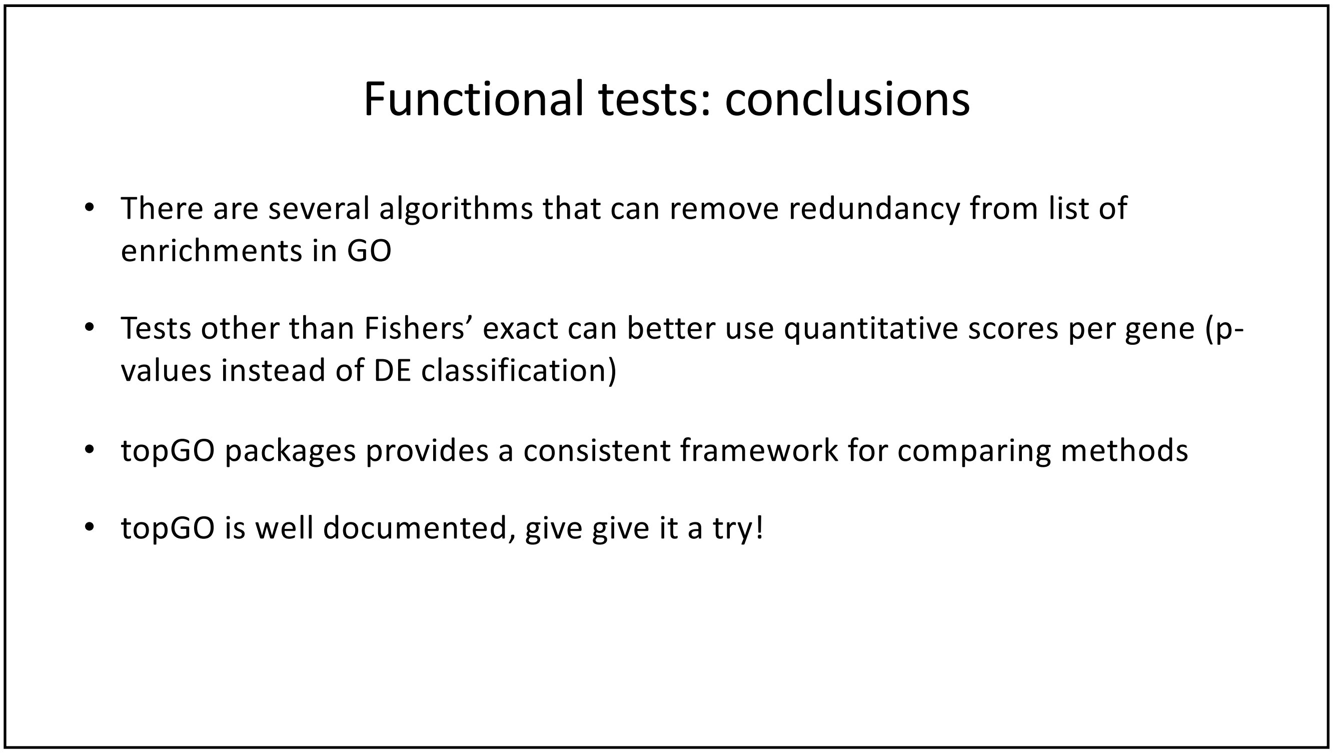
- Elim algorithm:
  - Accounts for groups hierarchy (GO)
  - Fisher test on conditional on enriched children groups (chi)
  - H0: $\frac{\text{selected in group / chi}}{\text{selected / chi}} = \frac{\text{gene in group / chi}}{\text{genes universe / chi}}$
  - R-package: topGO
  - Grossman 2008, Bioinformatics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

There are more methods available



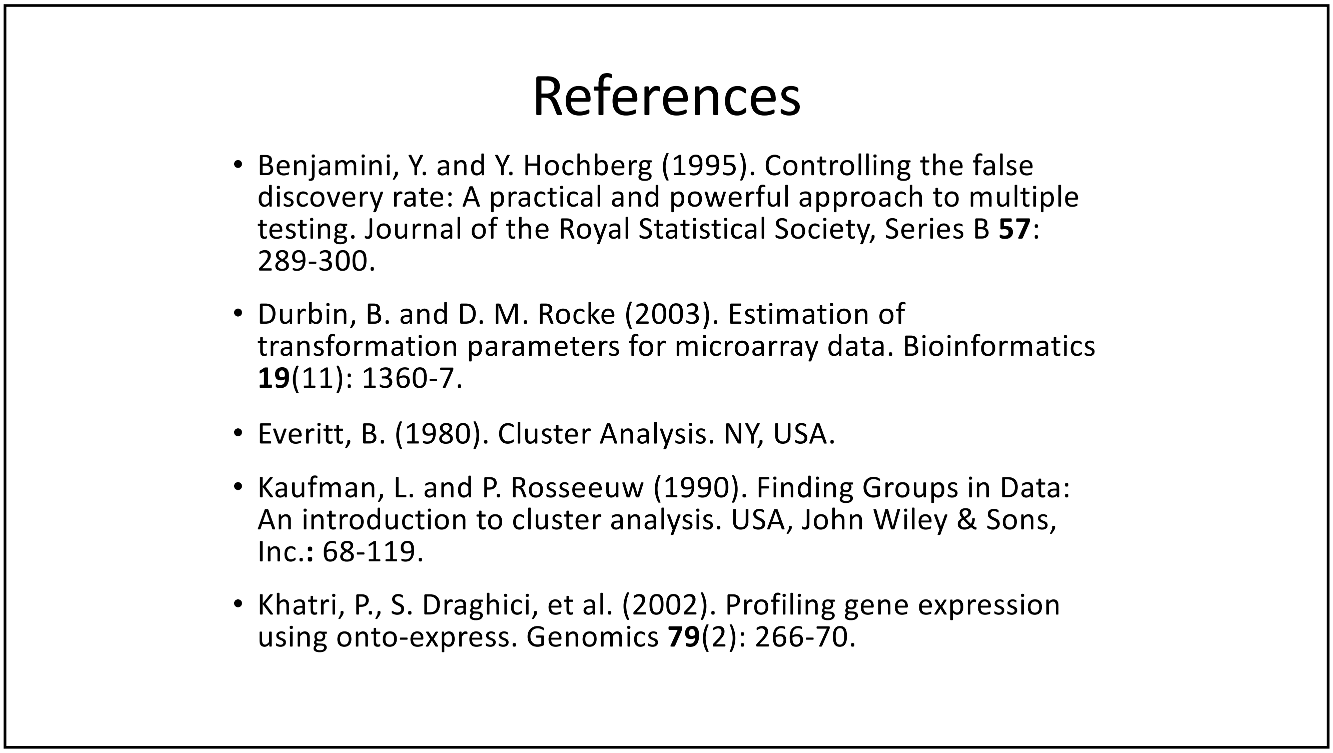
	fisher	ks	t	globaltest	sum
classic	✓	✓	✓	✓	✓
elim	✓	✓	✓	✓	✓
weight	✓	–	–	–	–
weight01	✓	✓	✓	✓	✓
lea	✓	✓	✓	✓	✓
parentchild	✓	–	–	–	–

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.