

Summary For Clustering

- Many different methods exist for finding groups and patterns in data (including some I haven't mentioned).
- Many different parameters can be used in those methods.
- Caution should be exercised in interpreting the results.

Comparing Different Clustering Methods

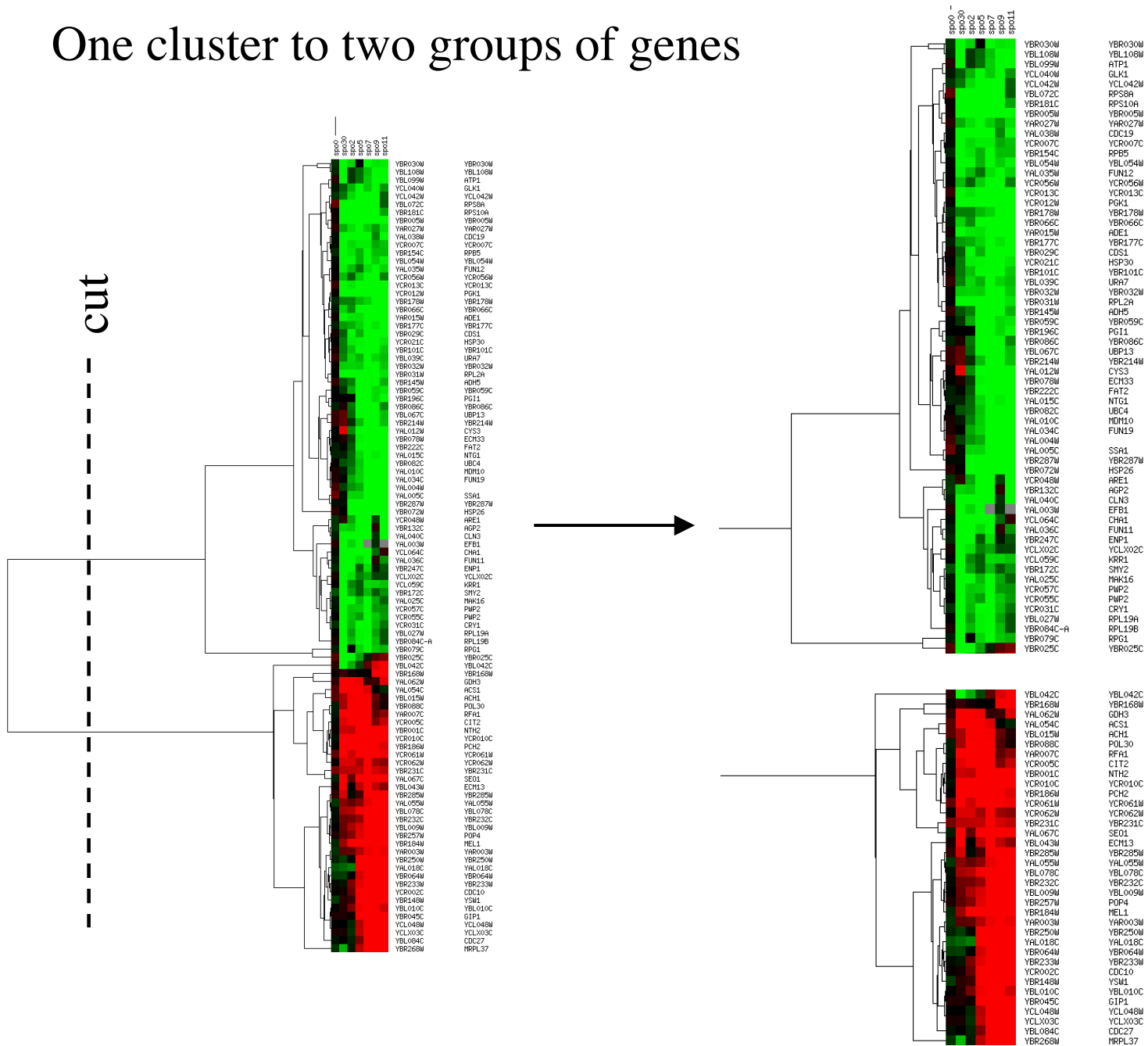
Which technique is right?

- Hierarchical clustering?
 - Single, Average, Complete, Centroid linkage, etc.?
- Self Organizing Maps
- K-means clustering
- Other algorithms?

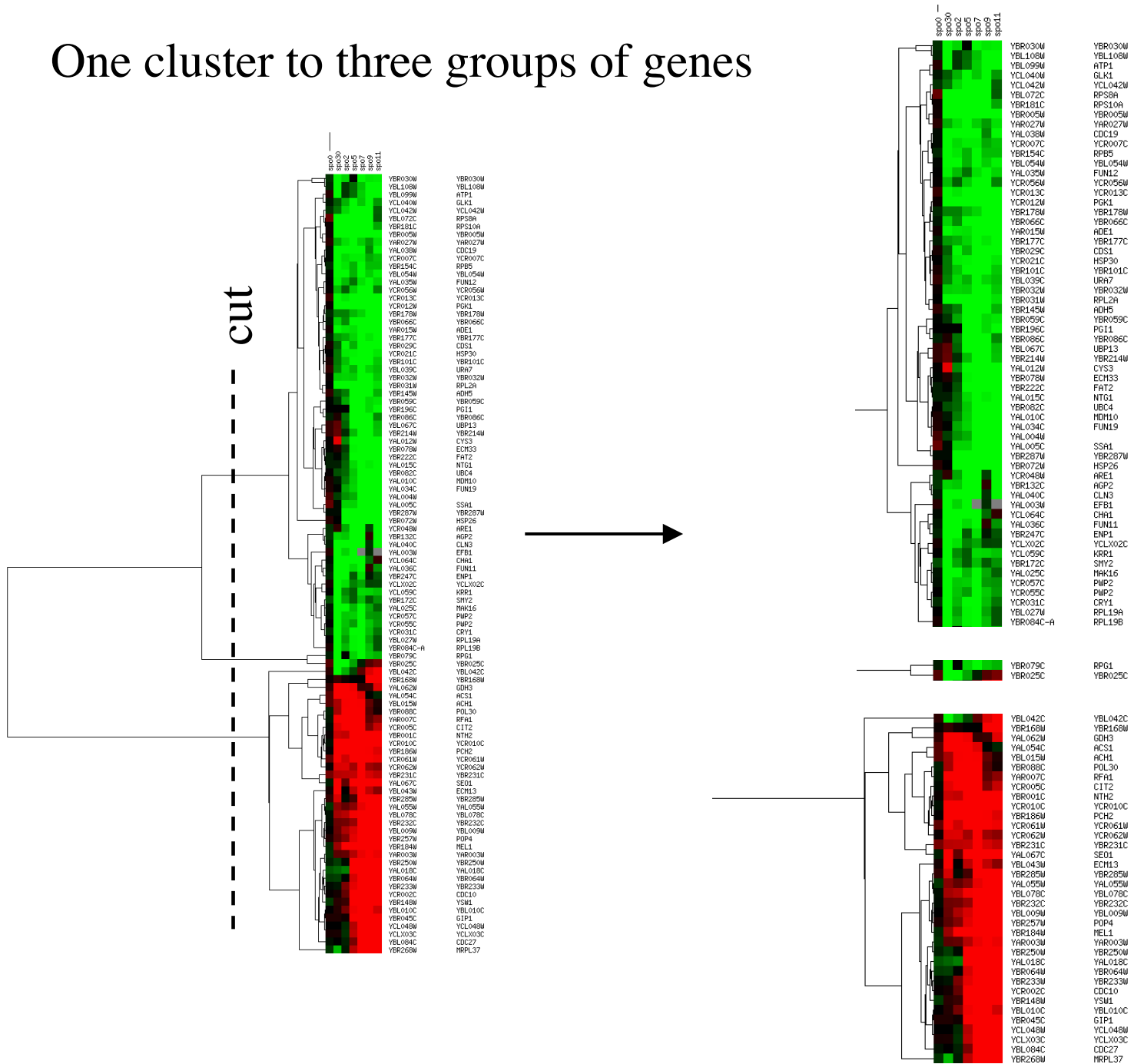
What is a 'cluster'?

- And how do we know if it's any good, or if one technique for producing clusters is better than another?
- Rather than think simply of clustering, think of all these methods as capable of producing groups of genes:

One cluster to two groups of genes



One cluster to three groups of genes



Now what?

- Try many methods, and demand they each produce the same number of groups of genes.
- Is there a metric that says which did best for a given number of groups?
- Can we come up with a metric for the best number of groups?

What do we think that co-expression means?

- Our general assumption is guilt by association:
i.e. genes with similar expression patterns are more likely to participate in the same biological process.
- Therefore, we can exploit the Gene Ontology to assess our clusters:

How do we measure how 'good' the annotation is?

- Use a score that measures how coherent the level of annotation is compared to what would be expected from random clusters.
 - see Gibbons and Roth (2002). *Genome Research* **12**, 1574-1581.
 - Developed system, such that the higher the score, the better the annotation fit the clustering.

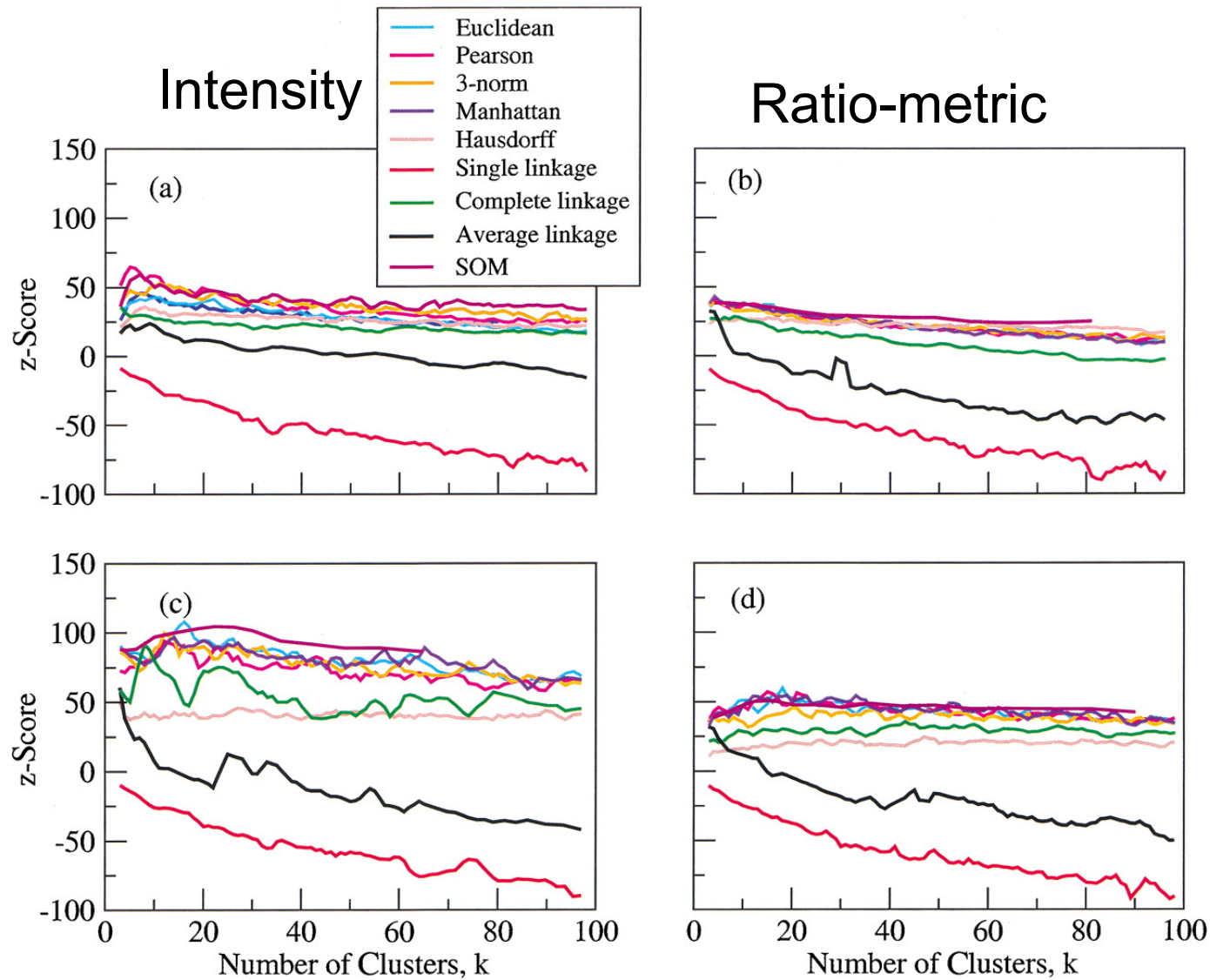


Figure 2. Four data sets clustered using k -means, hierarchical, and self-organized map algorithms. The horizontal axis shows the number of clusters desired, and the vertical axis shows z-scores. Data sets are (a) Cho, (b) CJRR, (c) Gasch, and (d) Spellman.

Characterization of clusters

- Now we have groups of genes that best fit their annotation, find the best annotation(s) that fits those groups.
- Calculate P-values for each GO term's association to a cluster, and choose those that are most significant.

Using the Gene Ontology to assess clusters

- Many microarray analyses result in a list of interesting genes
- Typically biologists can make up a story about any random list
- So, look at all GO annotations for the genes in a list, and see if the number of annotations for any GO node is significant

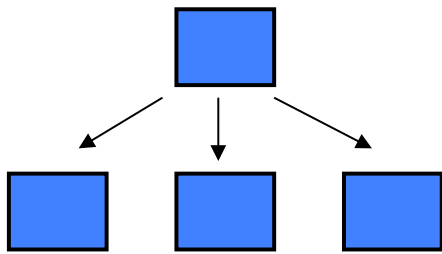
The Categories of GO

(The Gene Ontology)

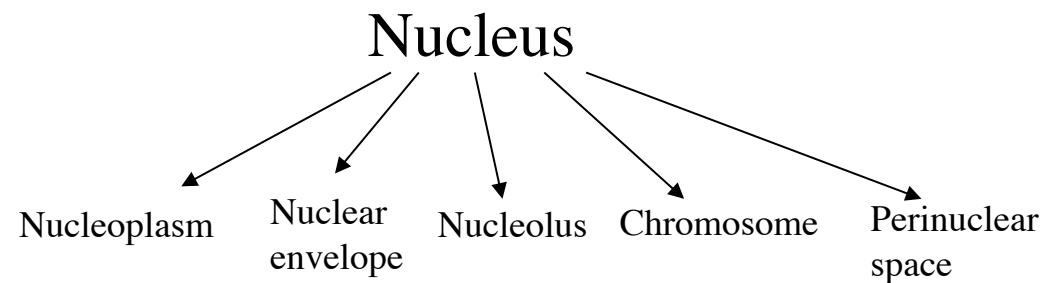
- *Biological Process* = goal or objective (Why)
(e.g. DNA replication, Cell Cycle Control, Cell adhesion)
- *Molecular Function* = elemental activity/task (What)
(e.g. Transcription factor, polymerase, protein kinase)
- *Cellular Component* = location or complex (Where)
(e.g. pre-replication complex, kinetochore, membrane)

Each Category is a structured, controlled vocabulary

Parent-Child Relationships



A child is a subset of a parent's elements



The cell component term *Nucleus* has 5 children

Determining P-values for GO annotation for a list of genes

We can calculate the probability of having x of n genes having an annotation to a GO node, given that in the genome, M of N genes have that annotation, using the *hypergeometric distribution*, as:

$$p = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$$

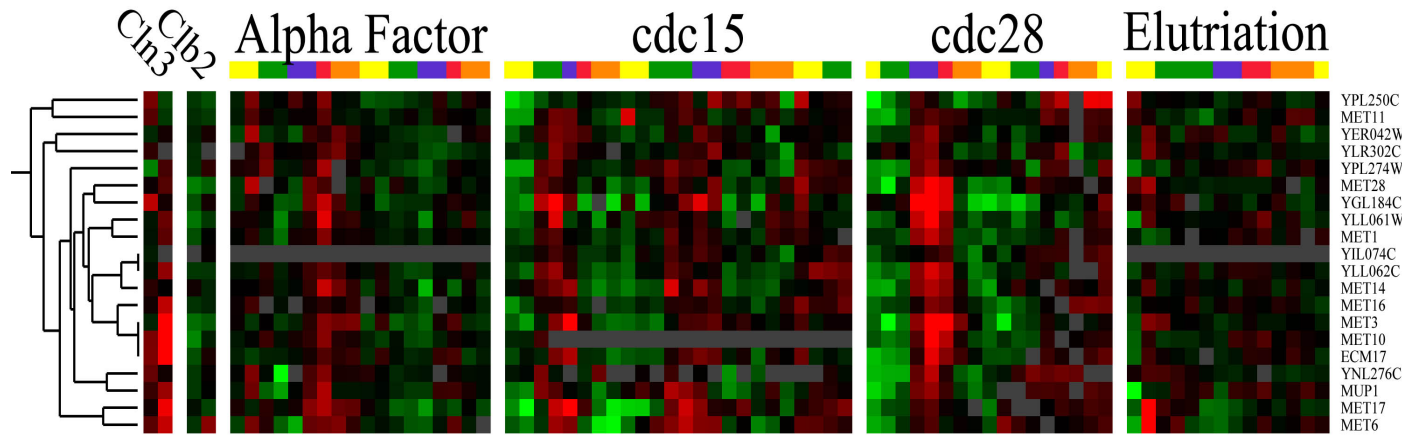
Determining GO significance

To calculate a P-value, we calculate the probability of having *at least* x of n annotations:

$$\text{P-value} = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{i}}$$

Then do multiple hypothesis correction on the p-values

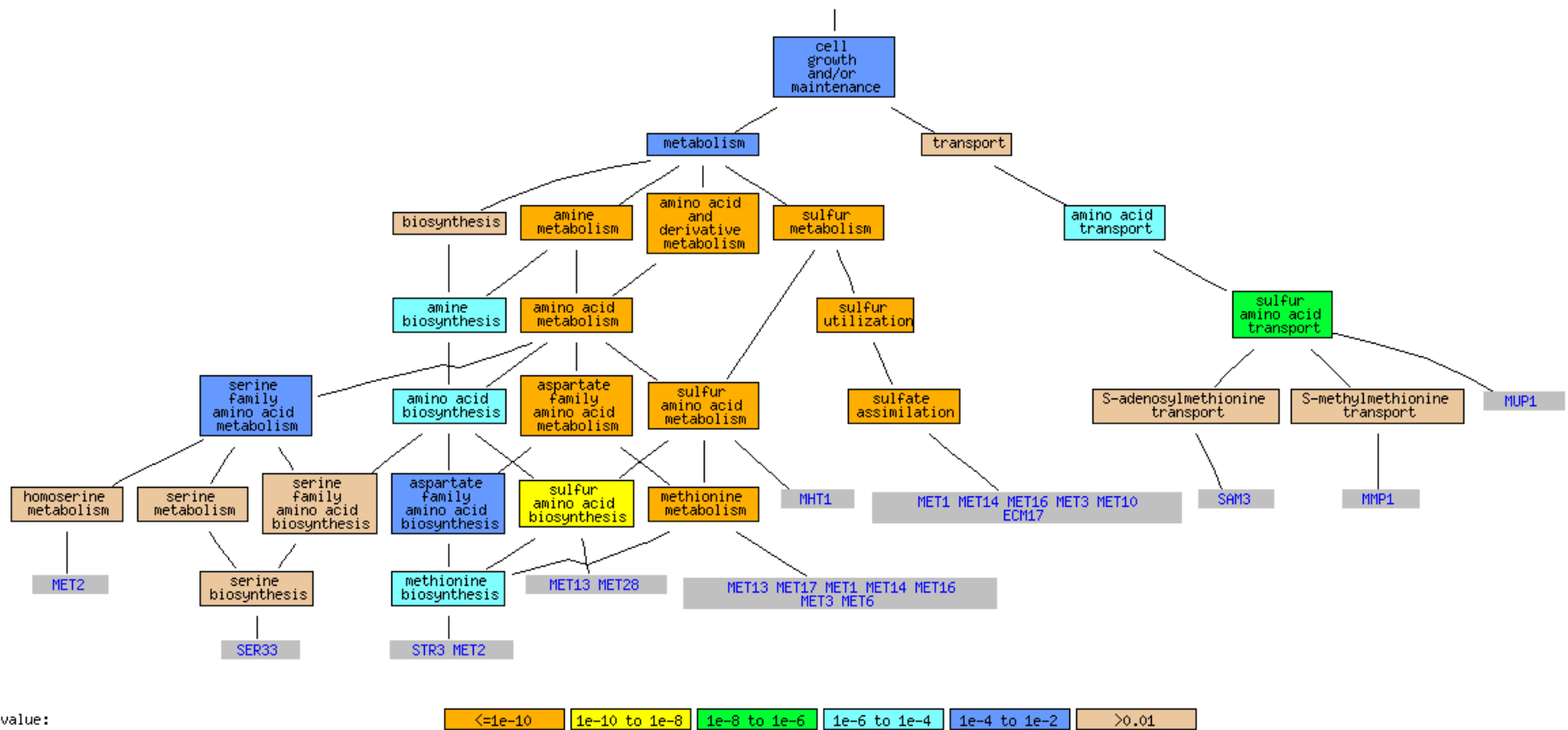
Methionine Cluster



YPL250C
 MET11
 YER042W
 YLR302C
 YPL274W
 MET28
 YGL184C
 YLL061W
 MET1
 YIL074C
 YLL062C
 MET14
 MET16
 MET3
 MET10
 ECM17
 YNL276C
 MUP1
 MET17
 MET6

GO Annotations

- sulfur metabolic process : 2.43e-19 (12/18 vs 66/6608)
- methionine metabolic process : 1.40e-14 (10/18 vs 24/6608)



Recommended reading : Clustering

- **Eisen MB, Spellman PT, Brown PO, Botstein D.** (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**(25):14863-8.
- **Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR** (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* **96** (6):2907.
- **Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM** (1999). Systematic determination of genetic network architecture. *Nat Genet.* **22**(3):281-5.
- **Tusher VG, Tibshirani R, Chu G** (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* **98**(9):5116-21
- **Slonim DK.** (2002). From patterns to pathways: gene expression data analysis comes of age. *Nat Genet.* **32** Suppl:502-8.
- **McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R.** (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* **18**(11):1462-9.
- **Bryan J** (2004). Problems in gene clustering based on gene expression data. *Journal of Multivariate Analysis* **90**, 44–66.
- **D'haeseleer P** (2005). How does gene expression clustering work? *Nat Biotechnol.* **23** (12):1499-501.
- **Chipman H and Tibshirani R** (2006). Hybrid Hierarchical Clustering with Applications to Microarray Data. *Biostatistics*, **7**(2):286-301.

Recommended reading for Cluster Validation / Analysis

- **Yeung KY, Haynor DR, Ruzzo WL.** (2001). Validating clustering for gene expression data. *Bioinformatics* **17**, 309-318.
- **Gibbons FD, Roth FP.** (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* **12**(10):1574-81.
- **Slonim DK.** (2002). From patterns to pathways: gene expression data analysis comes of age. *Nat Genet.* **32** Suppl:502-8.
- **McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R.** (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* **18**(11):1462-9.
- **Zhou X, Kao MC, Wong WH.** (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U S A.* **99**(20):12783-8.
- **Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC.** (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet.* **34**, 267-73.
- **Breitling R, Amtmann A, Herzyk P** (2004). Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics* **5**(1):34.
- **Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G** (2004). GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics.* **20**(18):3710-5.
- **Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP** (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* **102**, 15545-50.
- **Handl J, Knowles J, Kell DB** (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics.* **21**(15):3201-12.