

Clustering Gene Expression Data: *The Good, The Bad, and The Misinterpreted*

Elizabeth Garrett-Mayer

November 5, 2003

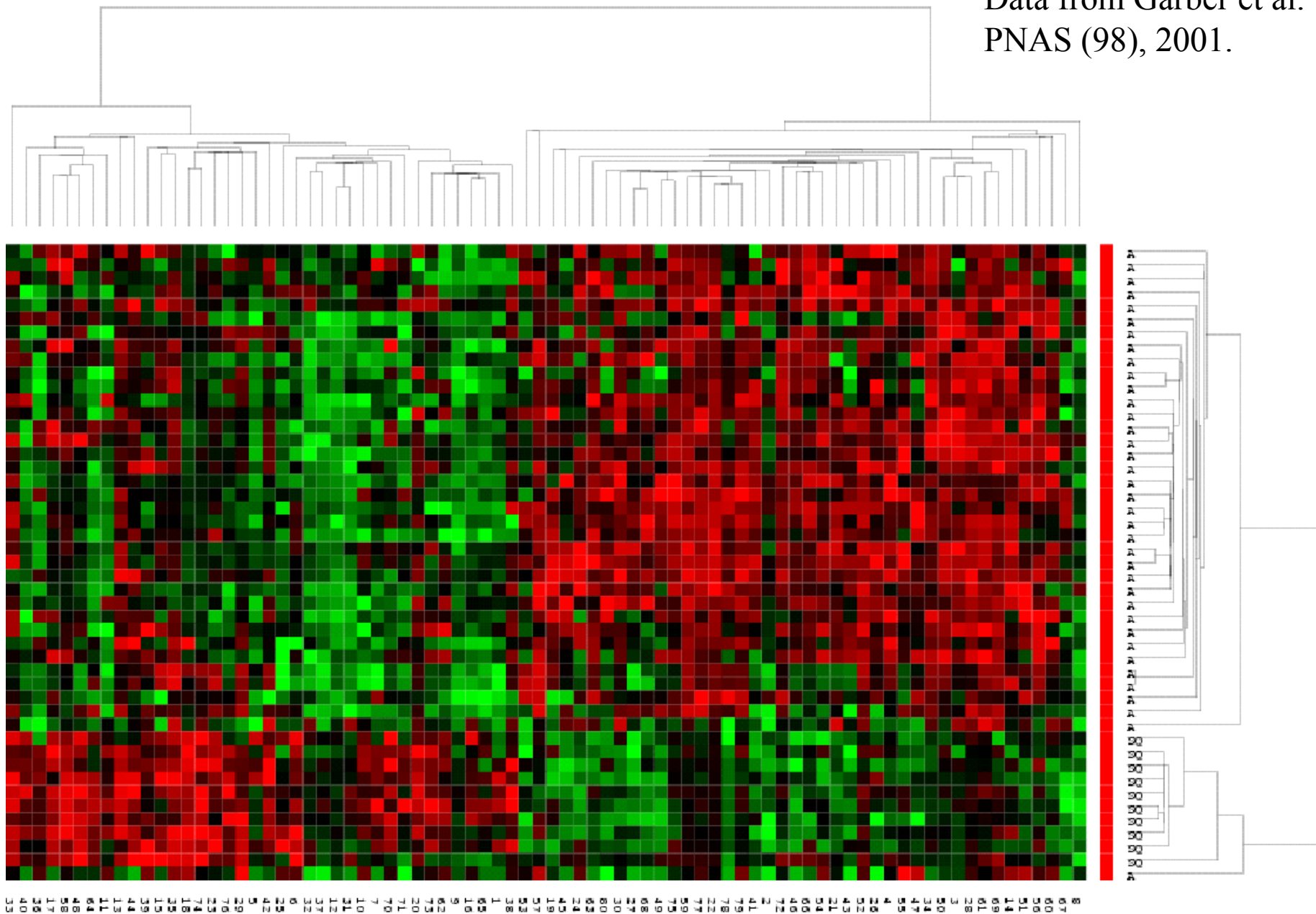
Oncology Biostatistics

Johns Hopkins University

esg@jhu.edu

Acknowledgements: Giovanni Parmigiani, David Madigan, Kevin Coombs

Data from Garber et al.
PNAS (98), 2001.



Clustering

- “Clustering” is an exploratory tool for looking at associations within gene expression data
- Hierarchical clustering dendrograms allow us to visualize gene expression data.
- These methods allow us to hypothesize about relationships between genes and classes.
- We should use these methods for visualization, hypothesis generation, selection of genes for further consideration
- We should not use these methods inferentially.
- There is no measure of “strength of evidence” or “strength of clustering structure” provided.
- Hierarchical clustering specifically: we are provided with a picture from which we can make many/any conclusions.

More specifically....

- Cluster analysis arranges samples and genes into groups based on their expression levels.
- This arrangement is determined purely by the measured distance between samples and genes.
- Arrangements are sensitive to choice of distance
 - Outliers
 - Distance mis-specification
- In hierarchical clustering, the VISUALIZATION of the arrangement (the dendrogram) is not unique!
 - Just because two samples are situated next to each other does not mean that they are similar.
 - Closer scrutiny is needed to interpret dendrogram than is usually given

A Misconception

- Clustering is not a classification method
- Clustering is ‘unsupervised:’
 - We don’t use any information about what class the samples belong to (e.g. AML vs. ALL) (Golub et al.) to determine cluster structure
 - We don’t use any information about which genes are functionally or otherwise related to determine cluster structure
 - ***Clustering finds groups in the data***
- Classification methods are ‘supervised:’
 - By definition, we use phenotypic data to help us find out which genes best classify genes
 - SAM, t-tests, CART
 - ***Classification methods finds ‘classifiers’***

Distance and Similarity

- Every clustering method is based **solely** on the measure of distance or similarity.
- E.g. correlation: measures linear association between two samples or genes.
 - What if data are not properly transformed?
 - What if there are outliers?
 - What if there are saturation effects?
- Even with large number of samples, bad measure of distance or similarity will not be helped.

Commonly Used Measures of Similarity and Distance

- Euclidean distance
- Correlation (similarity)
 - Absolute value of correlation
 - Uncentered correlation
- Spearman correlation
- Categorical measures

Limitation of Clustering

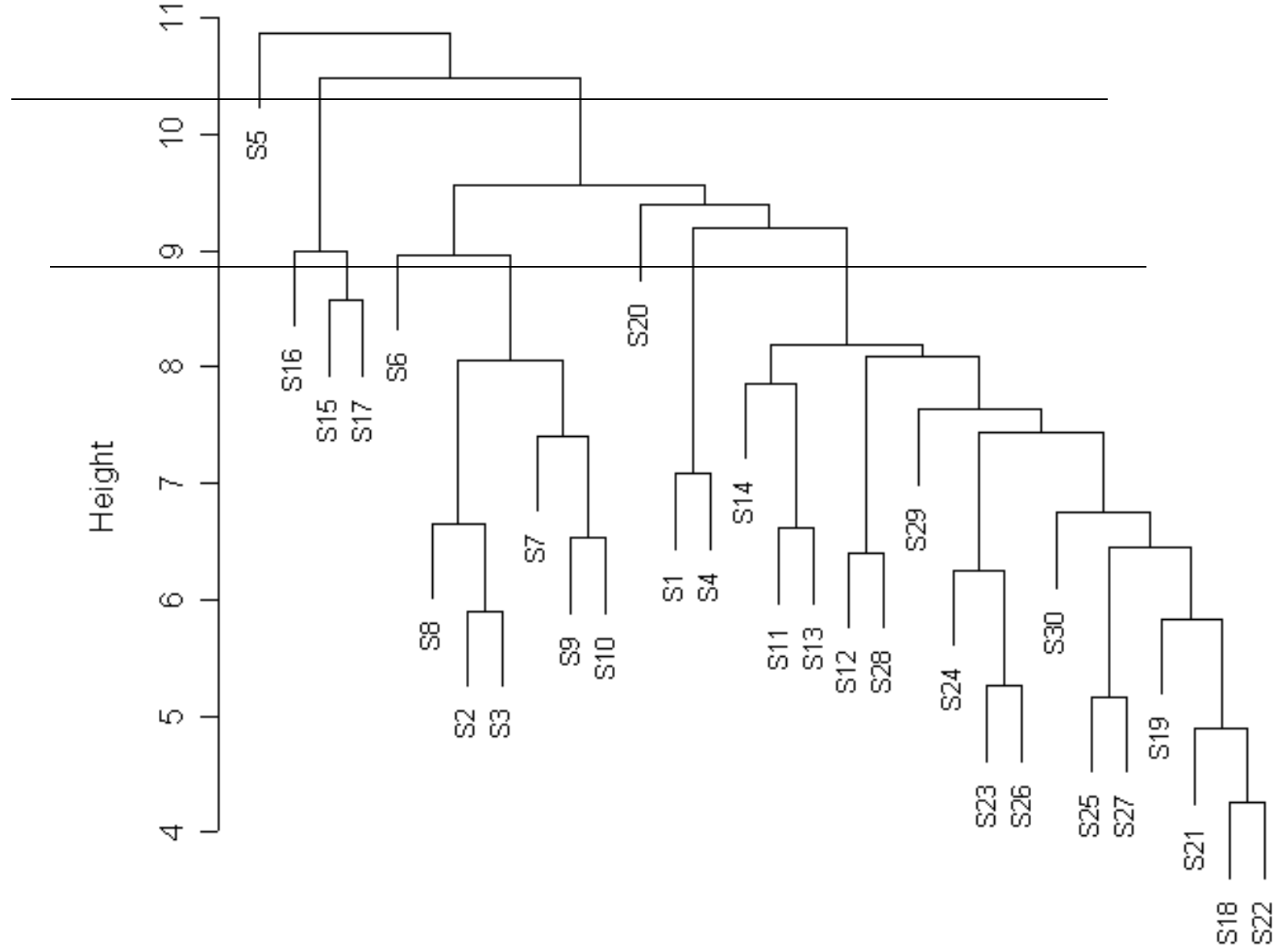
- The clustering structure can ONLY be as good as the distance/similarity matrix
- Generally, not enough thought and time is spent on choosing and estimating the distance/similarity matrix.
- “Garbage in → Garbage out”

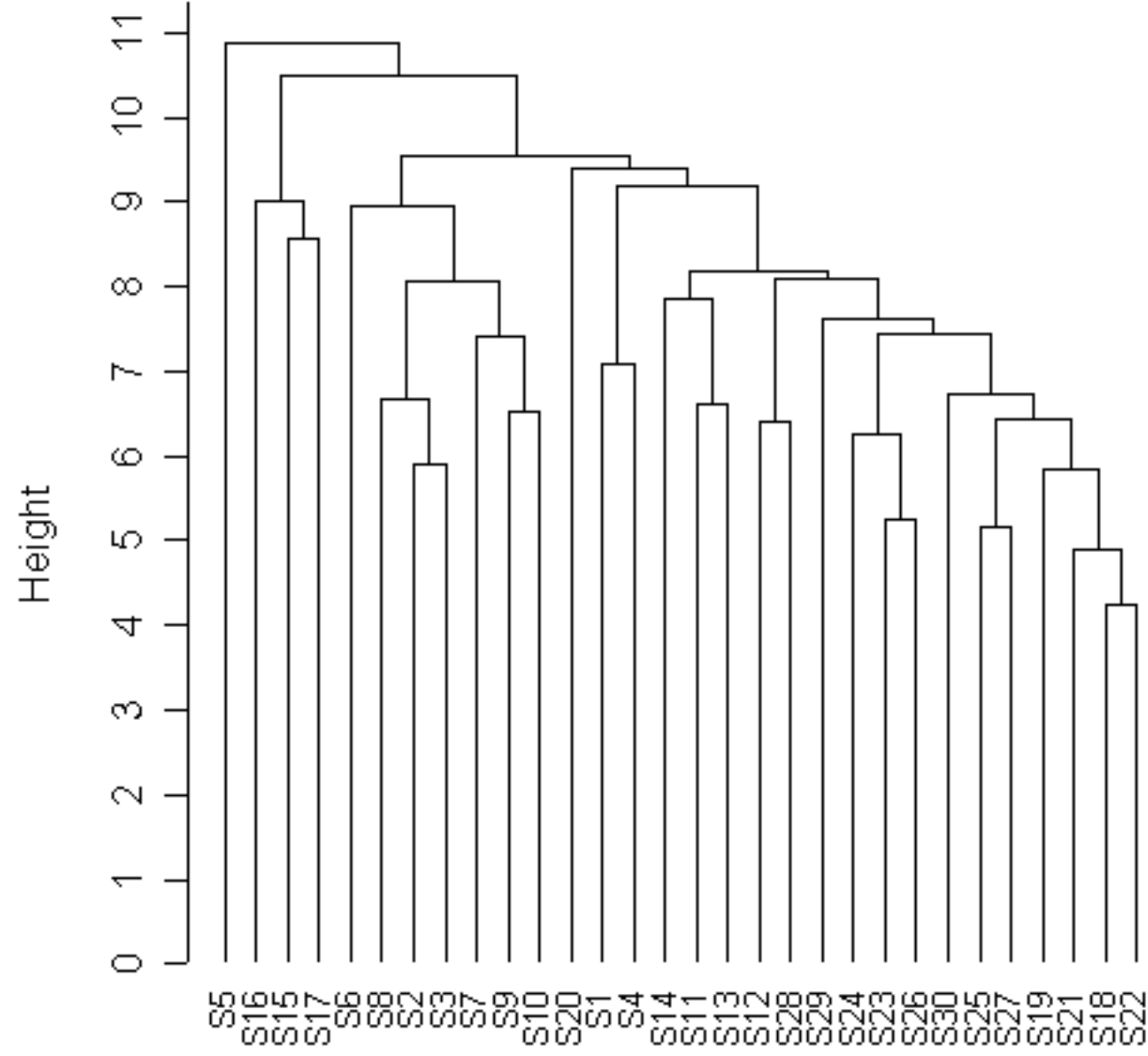
Two commonly seen clustering approaches in gene expression data analysis

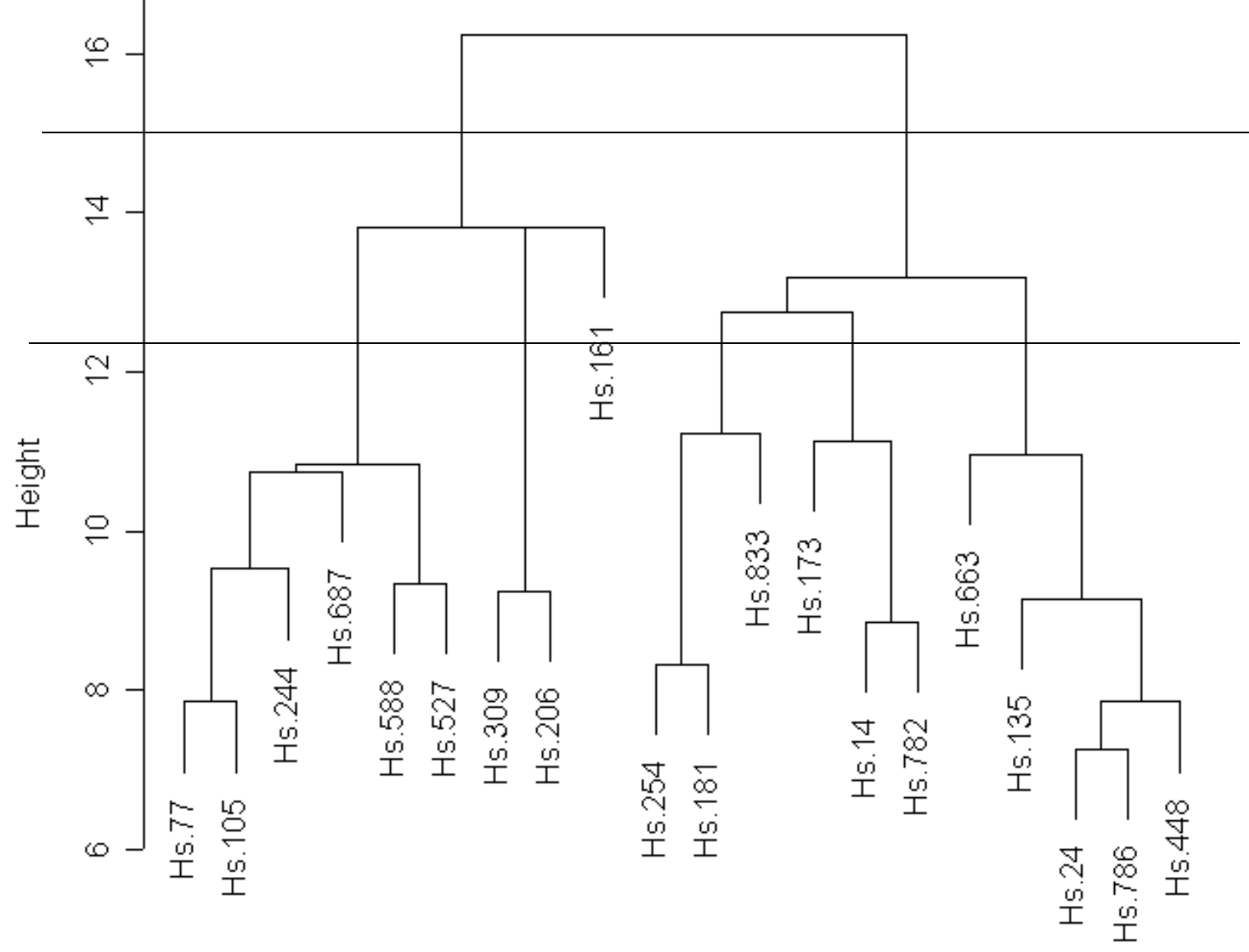
- Hierarchical clustering
 - Dendrogram (red-green picture)
 - Allows us to cluster both genes and samples in one picture and see whole dataset “organized”
- K-means/K-medoids
 - Partitioning method
 - Requires user to define K = # of clusters a priori
 - No picture to (over)interpret

Hierarchical Clustering

- The most overused statistical method in gene expression analysis
- Gives us pretty red-green picture with patterns
- But, pretty picture tends to be pretty unstable.
- Many different ways to perform hierarchical clustering
- Tend to be sensitive to small changes in the data
- Provided with clusters of every size: where to “cut” the dendrogram is user-determined







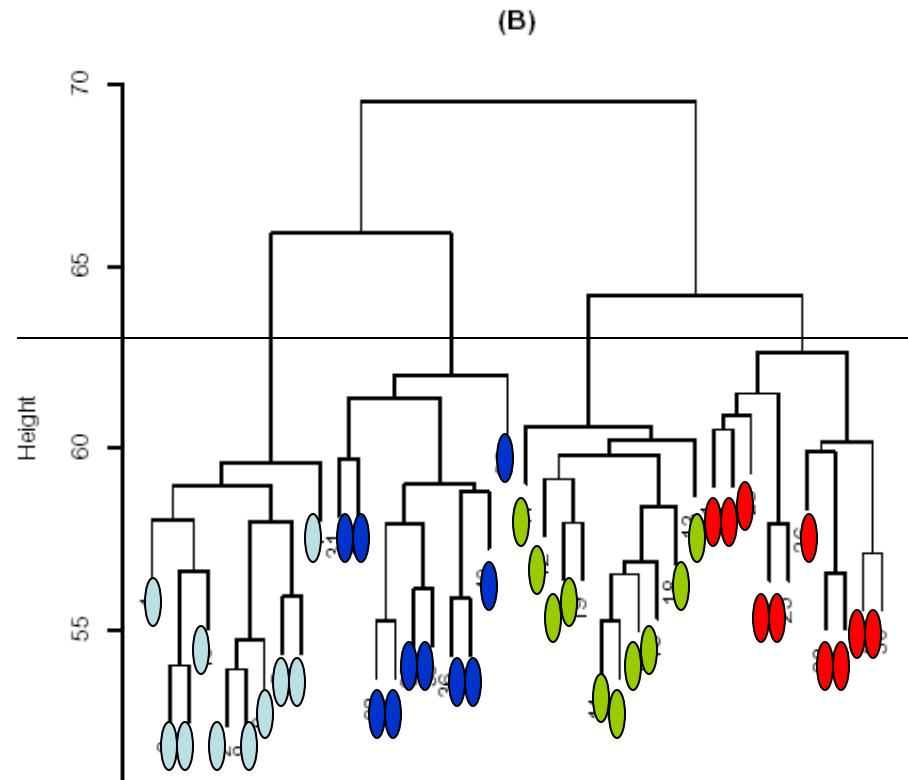
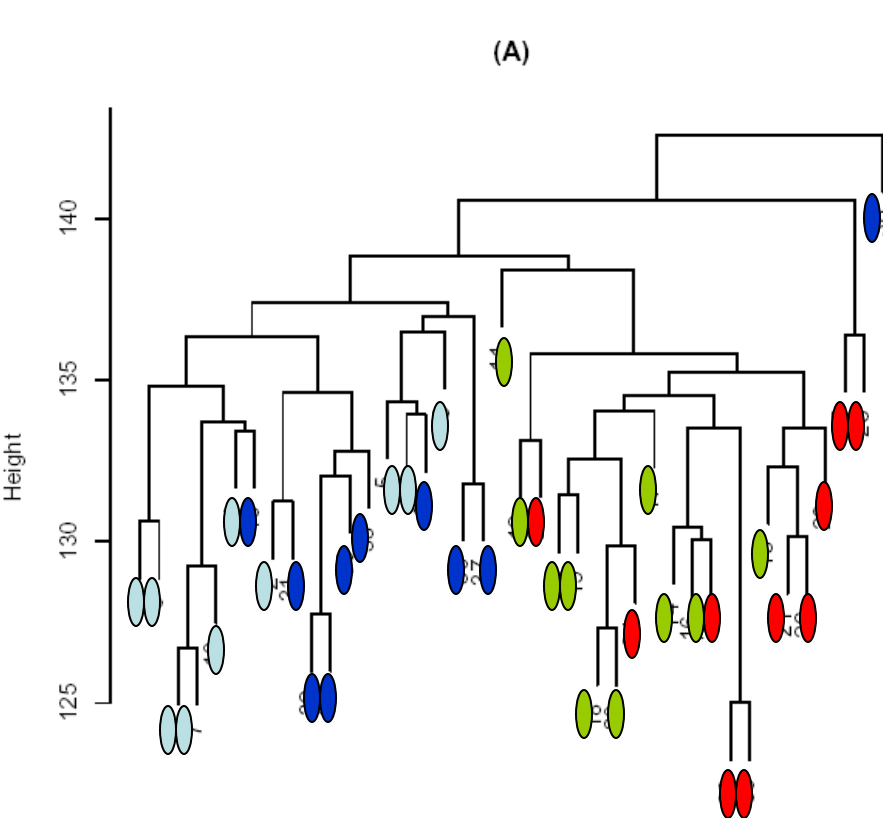
How to make a hierarchical clustering

1. Choose samples and genes to include in cluster analysis
2. Choose similarity/distance metric
3. Choose clustering direction (top-down or bottom-up)
4. Choose linkage method (if bottom-up)
5. Calculate dendrogram
6. Choose height/number of clusters for interpretation
7. Assess cluster fit and stability
8. Interpret resulting cluster structure

1. Choose samples and genes to include

- Important step!
- Do you want housekeeping genes included?
- What to do about replicates from the same individual/tumor?
- Genes that contribute noise will affect your results.
- Including all genes: dendrogram can't all be seen at the same time.
- Perhaps screen the genes?

Simulated Data with 4 clusters: 1-10, 11-20, 21-30, 31-40



A: 450 relevant genes plus
450 “noise” genes.

B: 450 relevant genes.

2. Choose similarity/distance matrix

- Think hard about this step!
- Remember: garbage in → garbage out
- The metric that you pick should be a valid measure of the distance/similarity of genes.
- Examples:
 - Applying correlation to highly skewed data will provide misleading results.
 - Applying Euclidean distance to data measured on categorical scale will be invalid.
- Not just “wrong”, but which makes most sense

Some correlations to choose from

- Pearson Correlation:

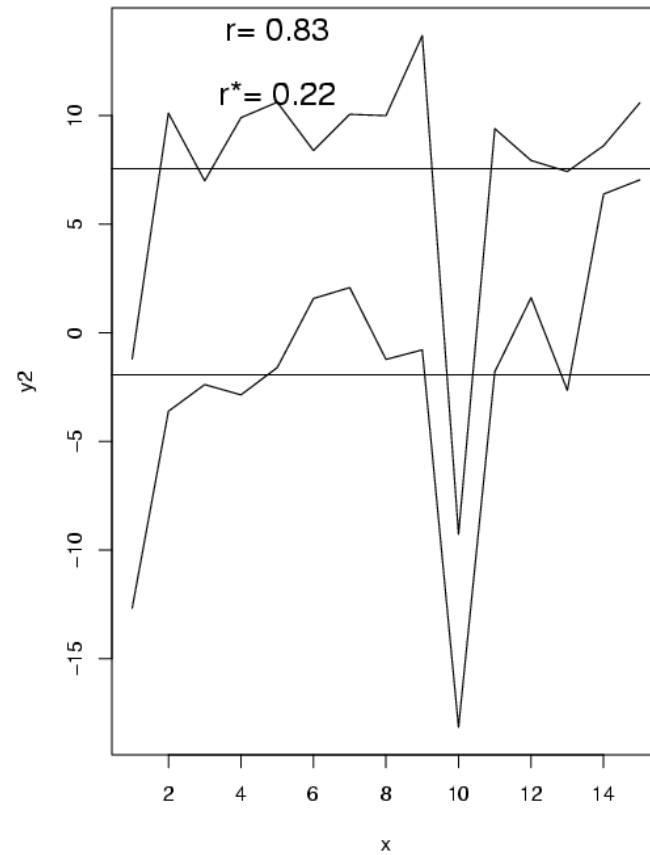
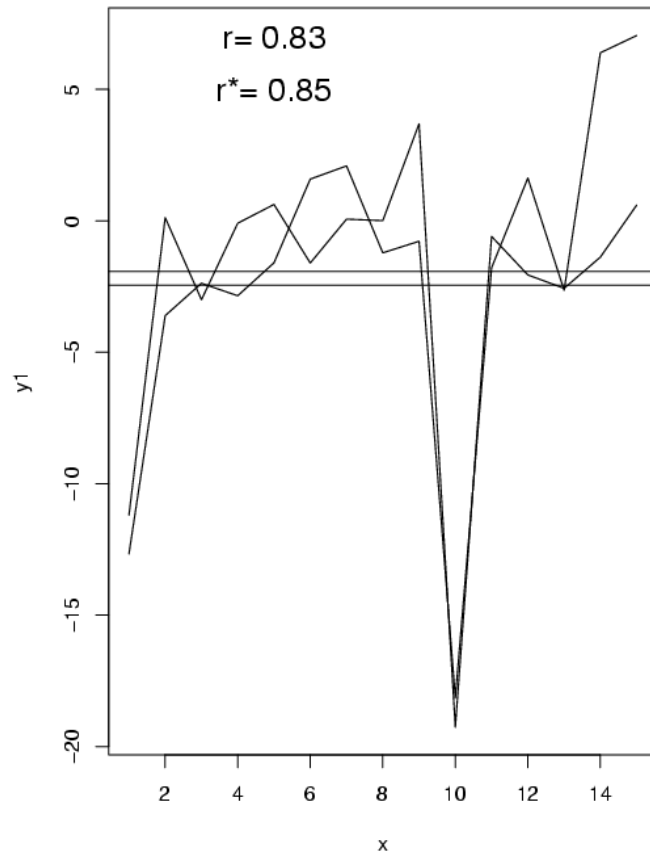
$$s(x_1, x_2) = \frac{\sum_{k=1}^K (x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2)}{\sqrt{\sum_{k=1}^K (x_{1k} - \bar{x}_1)^2 \sum_{k=1}^K (x_{2k} - \bar{x}_2)^2}}$$

- Uncentered Correlation:

$$s(x_1, x_2) = \frac{\sum_{k=1}^K x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^K x_{1k}^2 \sum_{k=1}^K x_{2k}^2}}$$

- Absolute Value of Correlation:

$$s(x_1, x_2) = \left| \frac{\sum_{k=1}^K (x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2)}{\sqrt{\sum_{k=1}^K (x_{1k} - \bar{x}_1)^2 \sum_{k=1}^K (x_{2k} - \bar{x}_2)^2}} \right|$$



The difference is that, if you have two vectors X and Y with identical shape, but which are offset relative to each other by a fixed value, they will have a standard Pearson correlation (centered correlation) of 1 but will not have an uncentered correlation of 1.

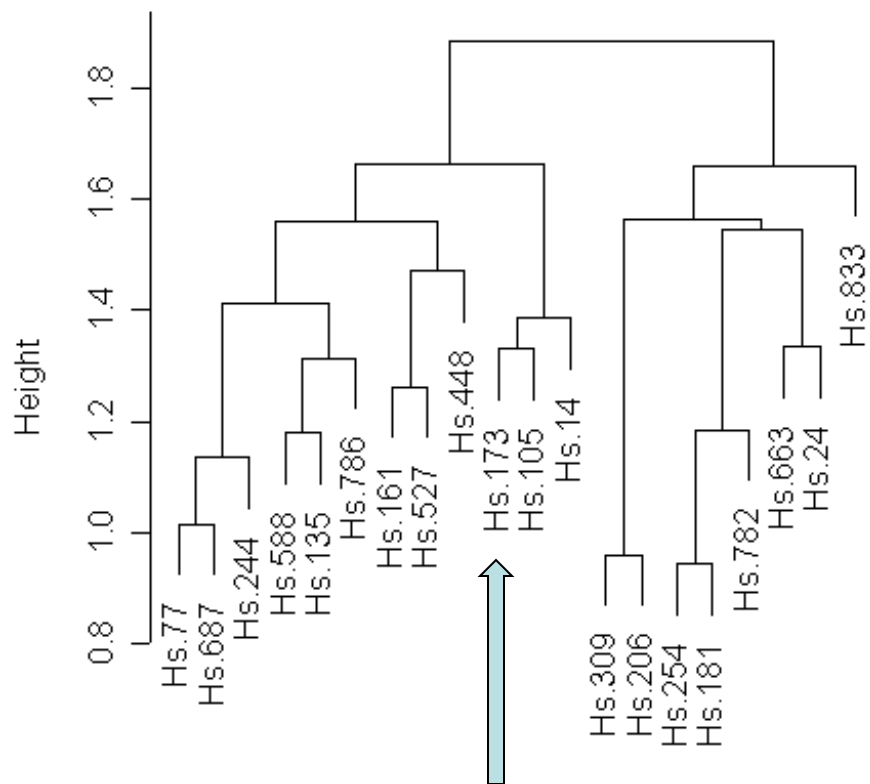
3. Choose clustering direction (top-down or bottom-up)

- Agglomerative clustering (bottom-up)
 - Starts with as each gene in its own cluster
 - Joins the two most similar clusters
 - Then, joins next two most similar clusters
 - Continues until all genes are in one cluster
- Divisive clustering (top-down)
 - Starts with all genes in one cluster
 - Choose split so that genes in the two clusters are most similar (maximize “distance” between clusters)
 - Find next split in same manner
 - Continue until all genes are in single gene clusters

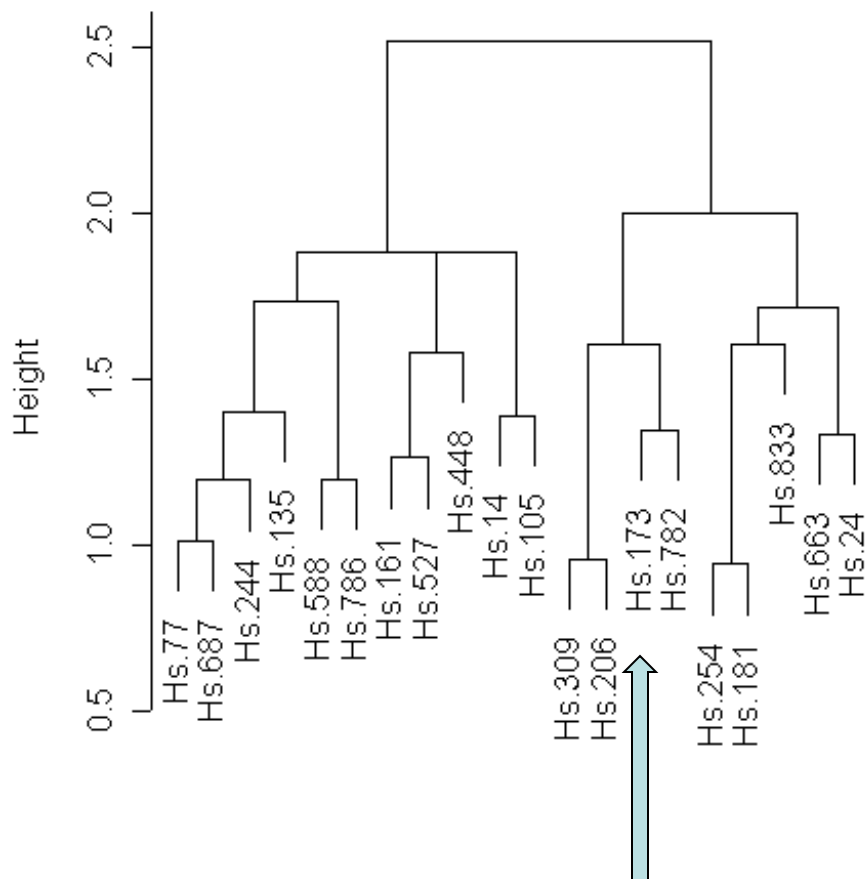
Which to use?

- Both are **only** ‘step-wise’ optimal: at each step the optimal split or merge is performed
- This does not imply that the final cluster structure is optimal!
- Agglomerative/Bottom-Up
 - Computationally simpler, and more available.
 - More “precision” at bottom of tree
 - When looking for small clusters and/or many clusters, use agglomerative
- Divisive/Top-Down
 - More “precision” at top of tree.
 - When looking for large and/or few clusters, use divisive
- **In gene expression applications, divisive makes more sense.**
- Results ARE sensitive to choice!

C: Agglom,Cor,Average

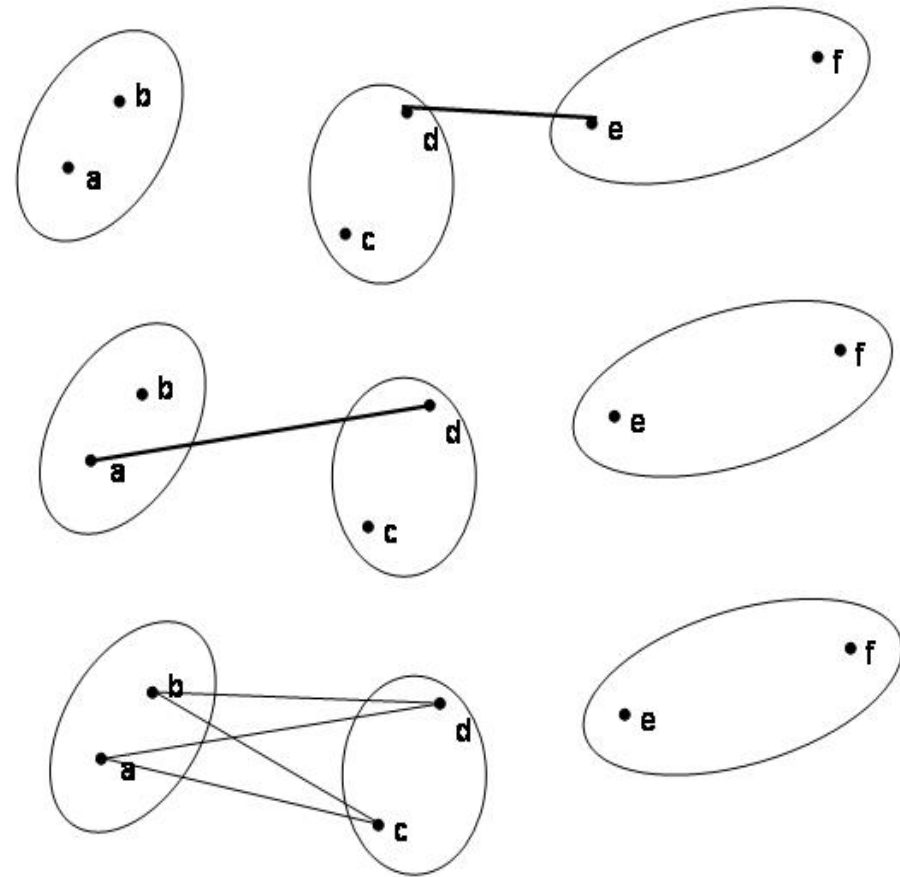


G: Div,Cor



4. Choose linkage method (if bottom-up)

- **Single Linkage:** join clusters whose distance between closest genes is smallest (elliptical)
- **Complete Linkage:** join clusters whose distance between furthest genes is smallest (spherical)
- **Average Linkage:** join clusters whose average distance is the smallest.



5. Calculate dendrogram

6. Choose height/number of clusters for interpretation

- In gene expression, we don't see "rule-based" approach to choosing cutoff very often.
- Tend to look for what makes a good story.
- There are more rigorous methods. (more later)
- "Homogeneity" and "Separation" of clusters can be considered. (Chen et al. Statistica Sinica, 2002)
- Other methods for assessing cluster fit can help determine a reasonable way to "cut" your tree.

7. Assess cluster fit and stability

- One approach is to try different approaches and see how tree differs.
 - Use average instead of complete linkage
 - Use divisive instead of agglomerative
 - Use Euclidean distance instead of correlation
- More later on assessing cluster structure

K-means and K-medoids

- Partitioning Method
- Don't get pretty picture
- MUST choose number of clusters K a priori
- More of a “black box” because output is most commonly looked at purely as assignments
- Each object (gene or sample) gets assigned to a cluster
- Begin with initial partition
- Iterate so that objects within clusters are most similar

K-means (continued)

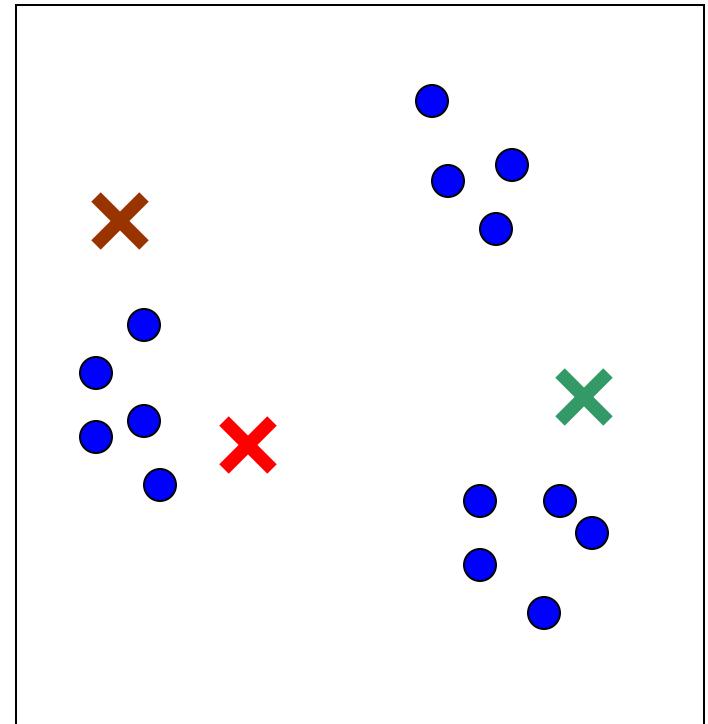
- Euclidean distance most often used
- Spherical clusters.
- Can be hard to choose or figure out K.
- Not unique solution: clustering can depend on initial partition
- No pretty figure to (over)interpret

How to make a K-means clustering

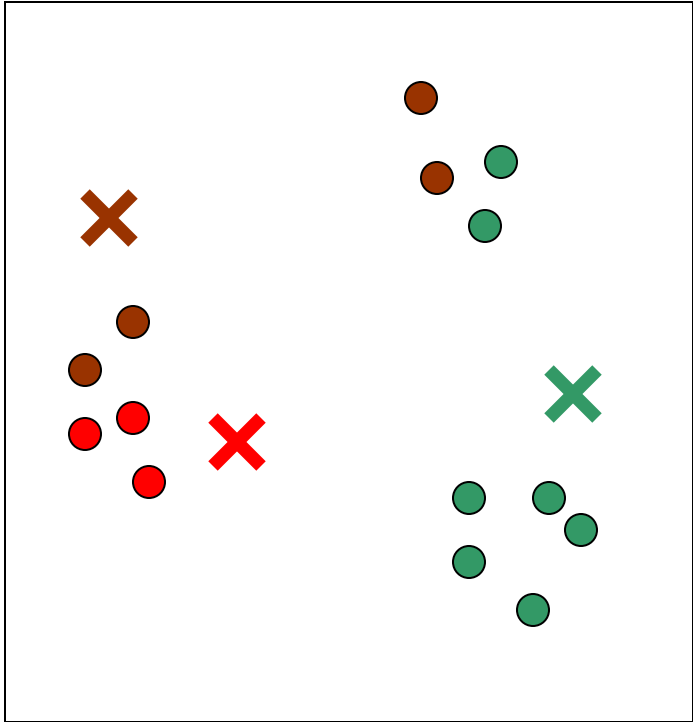
1. Choose samples and genes to include in cluster analysis
2. Choose similarity/distance metric (generally Euclidean)
3. Choose K.
4. Perform cluster analysis.
5. Assess cluster fit and stability
6. Interpret resulting cluster structure

K-means Algorithm

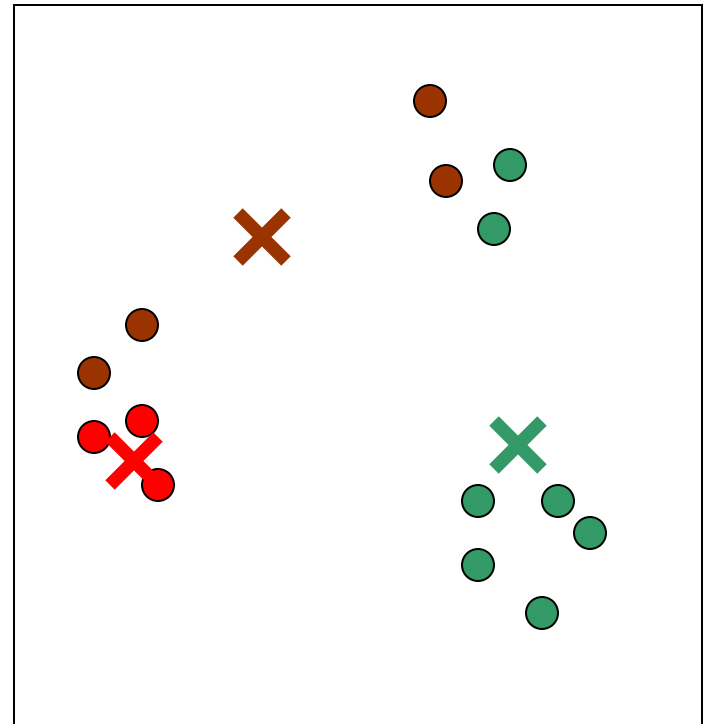
1. Choose K centroids at random
2. Make initial partition of objects into k clusters by assigning objects to closest centroid
3. Calculate the centroid (mean) of each of the k clusters.
4.
 - a. For object i , calculate its distance to each of the centroids.
 - b. Allocate object i to cluster with closest centroid.
 - c. If object was reallocated, recalculate centroids based on new clusters.
4. Repeat 3 for object $i = 1, \dots, N$.
5. Repeat 3 and 4 until no reallocations occur.
6. Assess cluster structure for fit and stability



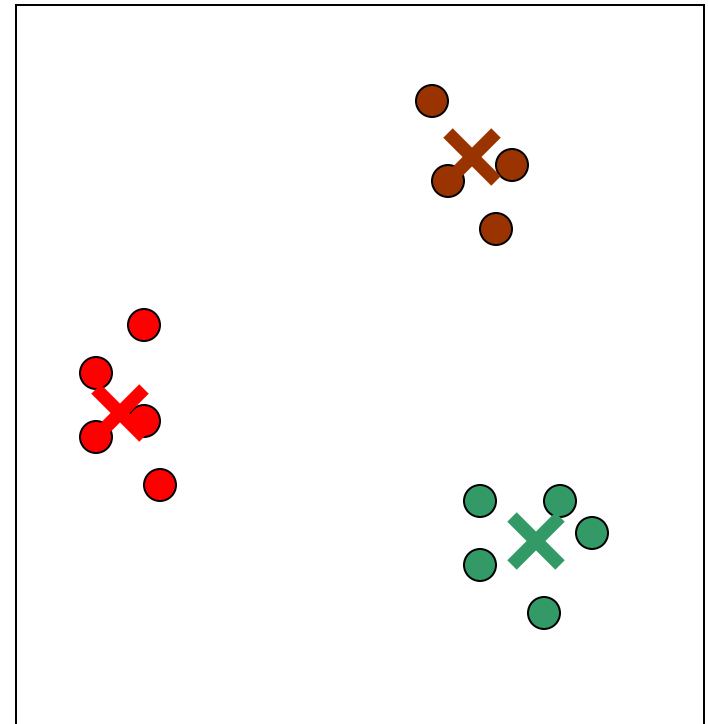
Iteration = 0



Iteration = 1



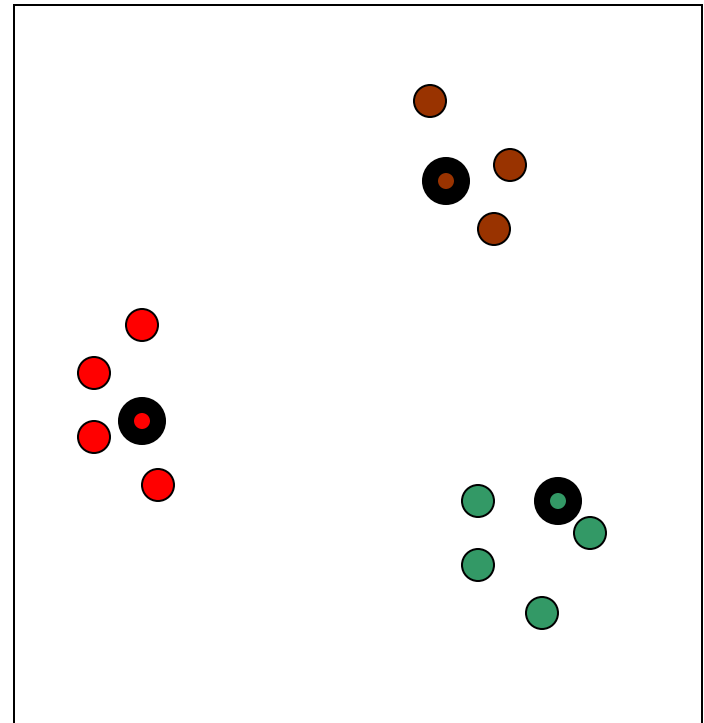
Iteration = 2



Iteration = 3

K-medoids

- A little different
- Centroid: The average of the samples within a cluster
- Medoid: The “representative object” within a cluster.
- Initializing requires choosing medoids at random.



7. Assess cluster fit and stability

- PART OF THE MISUNDERSTOOD!
- Most often ignored.
- Cluster structure is treated as reliable and precise
- BUT! Usually the structure is rather unstable, at least at the bottom.
- Can be VERY sensitive to noise and to outliers
- Homogeneity and Separation
- Cluster Silhouettes and Silhouette coefficient: how similar genes within a cluster are to genes in other clusters (composite separation and homogeneity) (more later with K-medoids) (Rousseeuw Journal of Computation and Applied Mathematics, 1987)

Assess cluster fit and stability (continued)

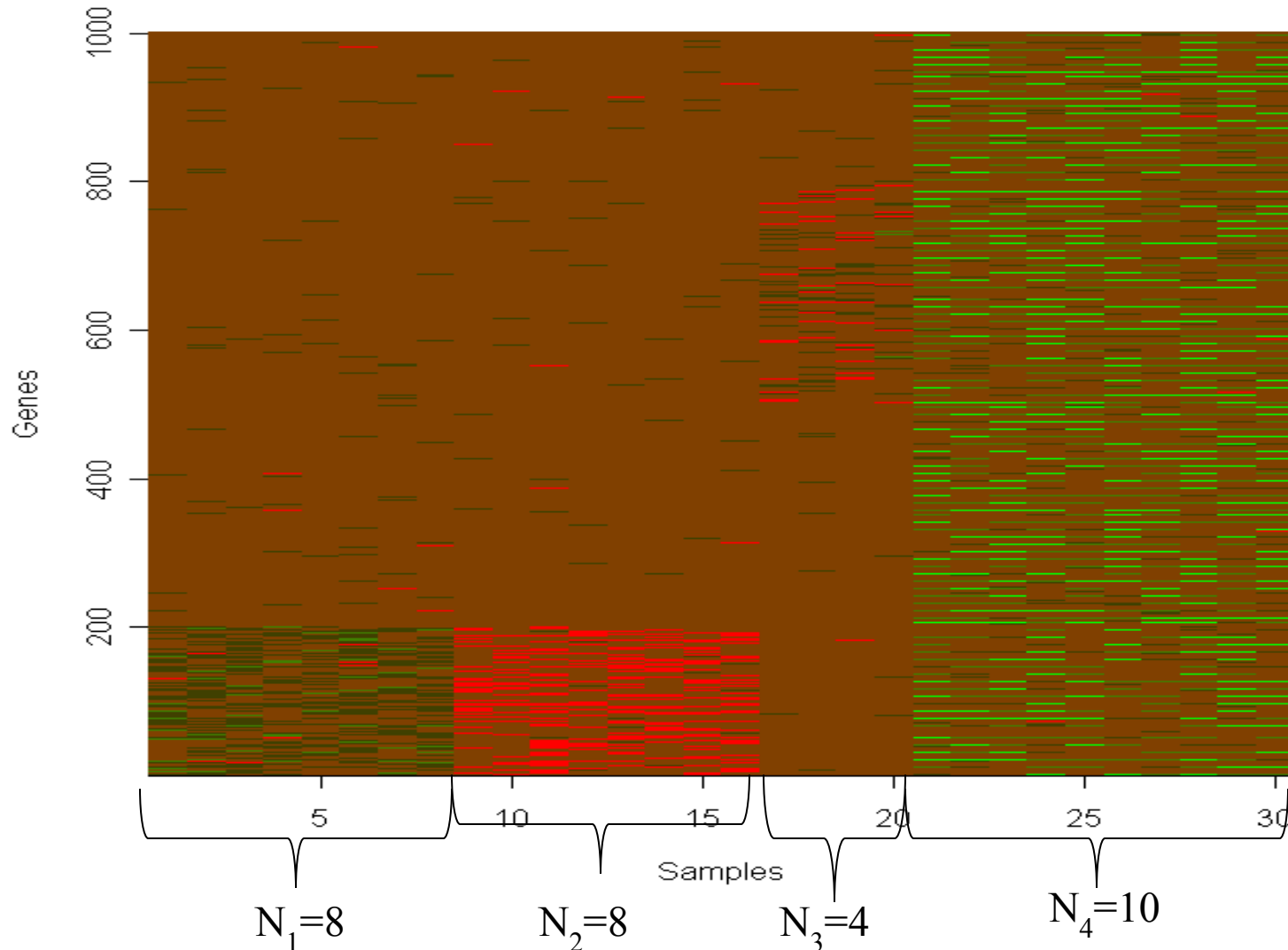
- WADP: Weighted Average Discrepant Pairs
 - Bittner et al. Nature, 2000
 - Fit cluster analysis using a dataset
 - Add random noise to the original dataset
 - Fit cluster analysis to the noise-added dataset
 - Repeat many times.
 - Compare the clusters across the noise-added datasets.
- Consensus Trees
 - Zhang and Zhao Functional and Integrative Genomics, 2000.
 - Use parametric bootstrap approach to sample new data using original dataset
 - Proceed similarly to WADP.
 - Look for nodes that are in a “majority” of the bootstrapped trees.
- More not mentioned.....

Careful though....

- Some validation approaches are more suited to some clustering approaches than others.
- Most of the methods require us to define number of clusters, even for hierarchical clustering.
 - Requires choosing a cut-point
 - If true structure is hierarchical, a cut tree won't appear as good as it might truly be.

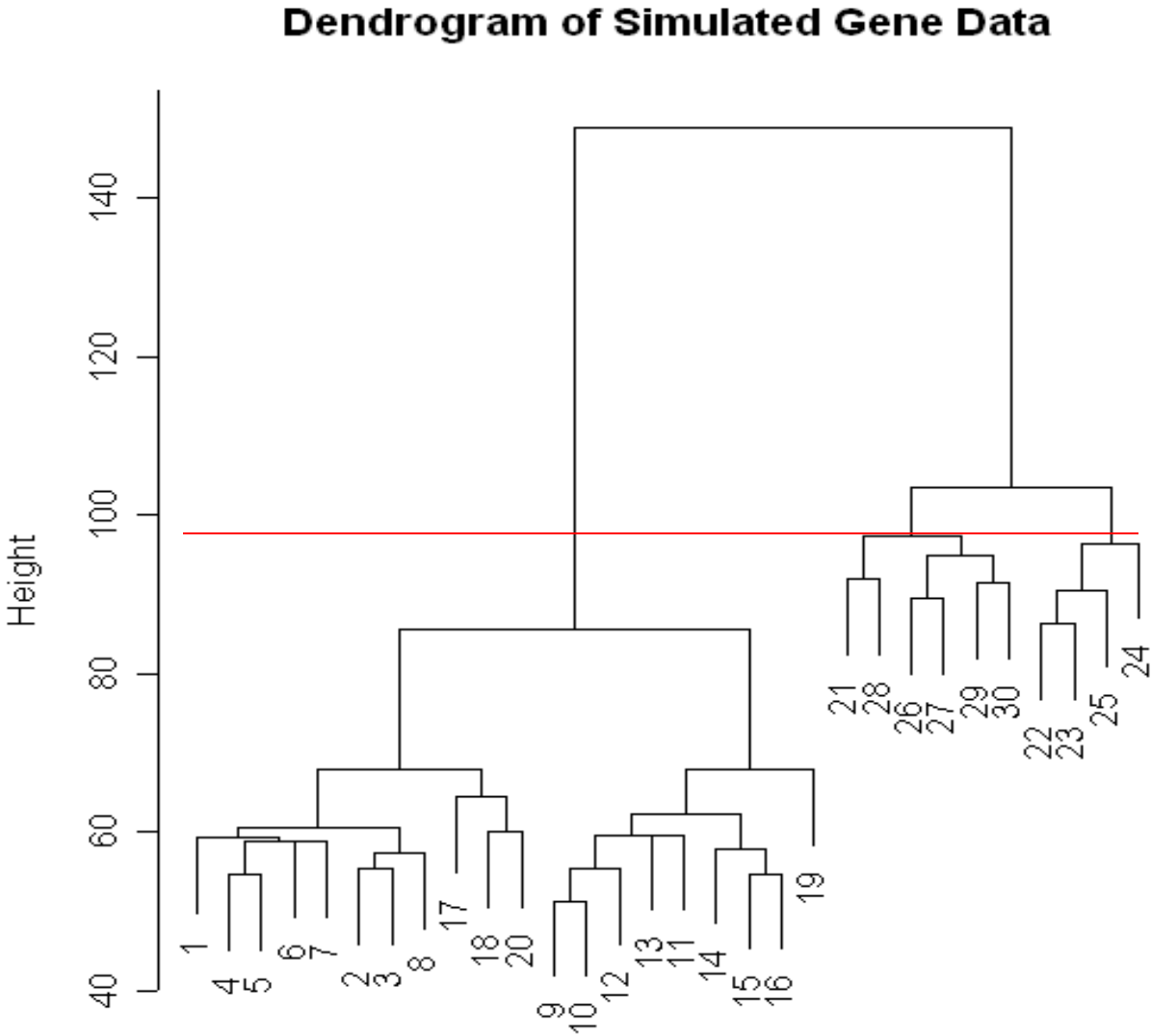
Example with Simulated Gene Expression Data

Four groups of samples: determined by k-means type assumptions.



K-Means

True Class	cluster			
	1	2	3	4
1	8	0	0	0
2	0	8	0	0
3	3	1	0	0
4	0	0	7	3



Silhouettes

- Silhouette of gene i is defined as:

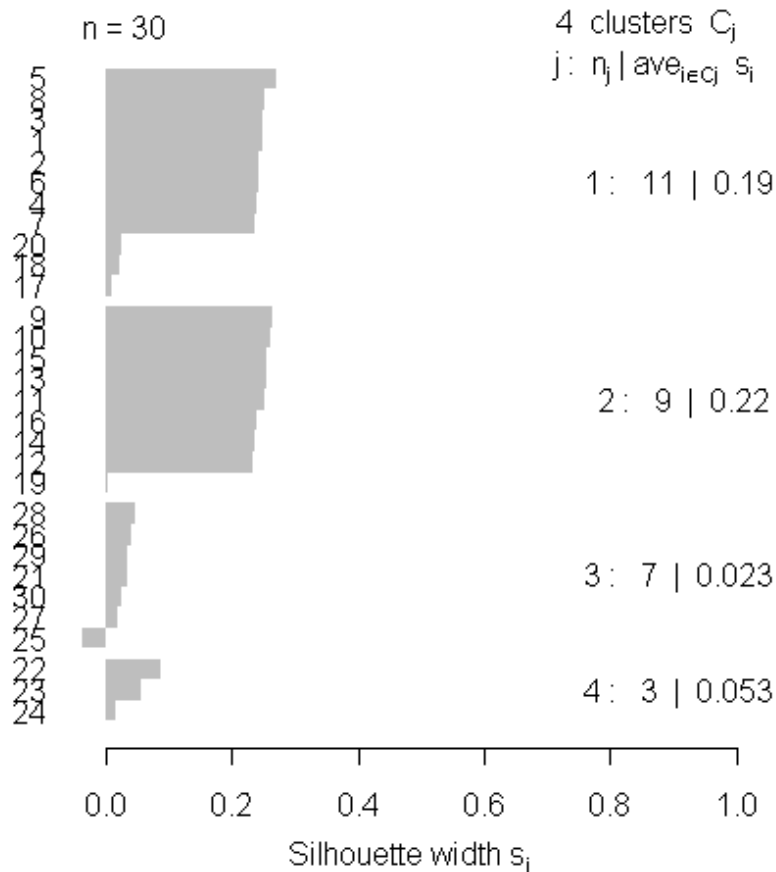
$$s(i) = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- a_i = average distance of sample i to other samples in same cluster
- b_i = average distance of sample i to genes in its nearest neighbor cluster

Silhouette Plots (Kaufman and Rousseeuw)

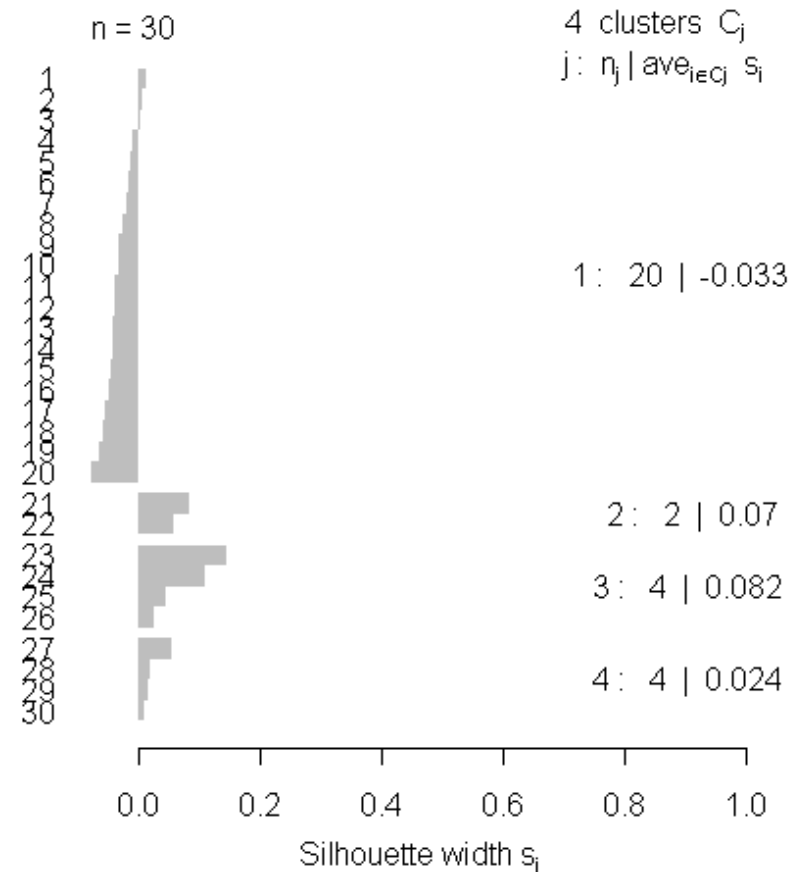
Assumes 4 classes

Silhouette Plot: K-medoids



Average silhouette width : 0.15

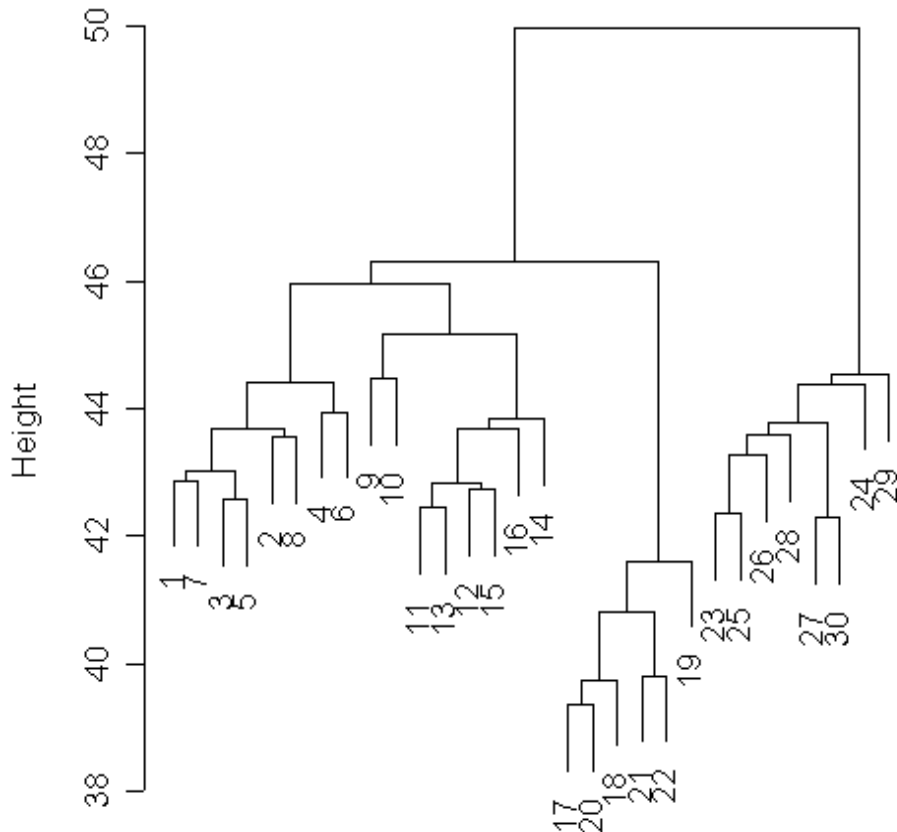
Silhouette Plot: Divisive Clustering



Average silhouette width : 0

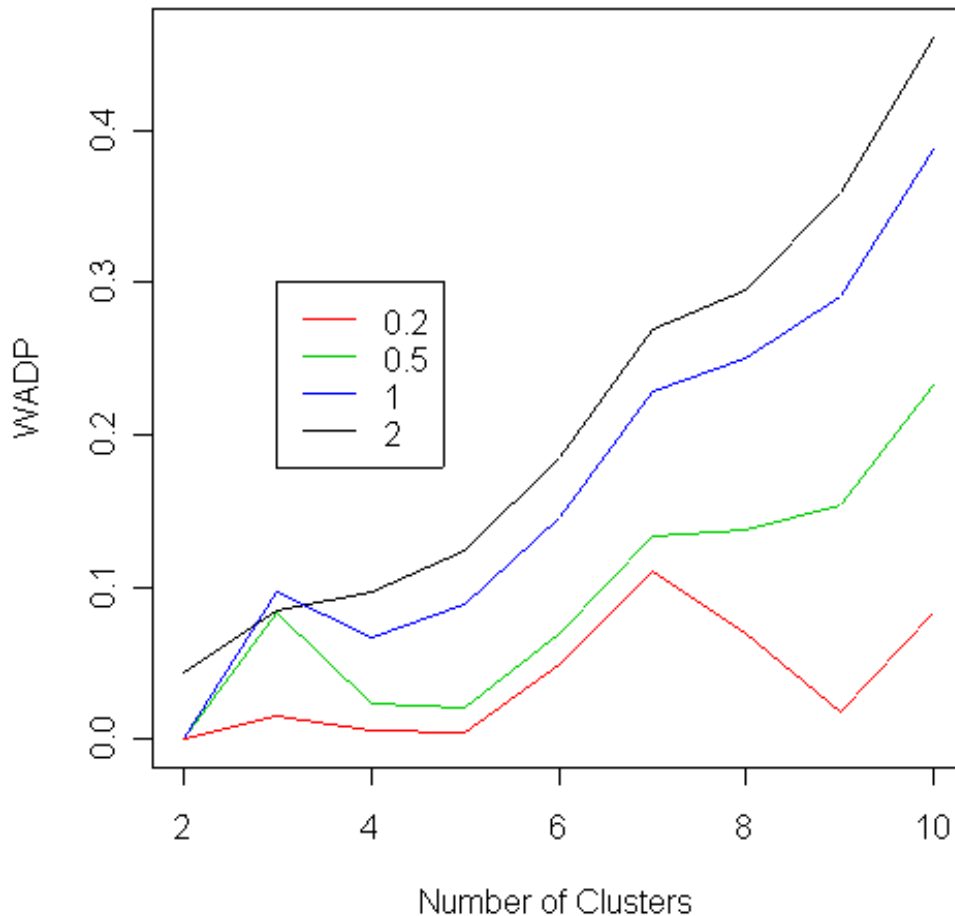
WADP: Weighted Average Discrepancy Pairs

Agglomerative Clustering



- Add perturbations to original data
- Calculate the number of paired samples that cluster together in the original cluster that didn't in the perturbed
- Repeat for every cutoff (i. e. for each k)
- Do iteratively
- Estimate for each k the proportion of discrepant pairs.

WADP



- Different levels of noise have been added
- By Bittner's recommendation, 1.0 is appropriate for our dataset
- But, not well-justified.
- External information would help determine level of noise for perturbation
- We look for largest k before WADP gets big.

Some Take-Home Points

- Clustering can be a useful exploratory tool
- Cluster results are very sensitive to noise in the data
- It is crucial to assess cluster structure to see how stable your result is
- Different clustering approaches can give quite different results
- For hierarchical clustering, interpretation is almost always subjective