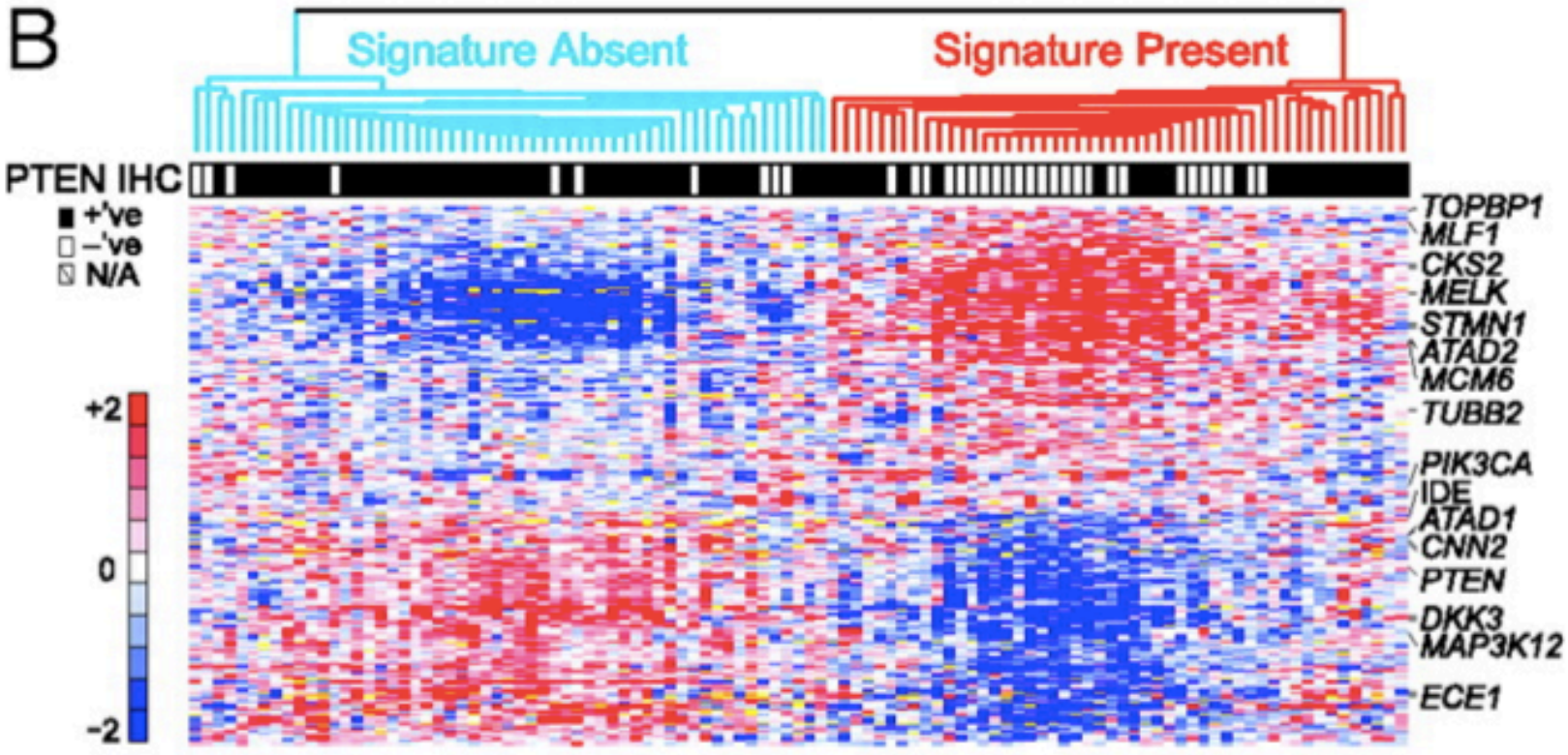
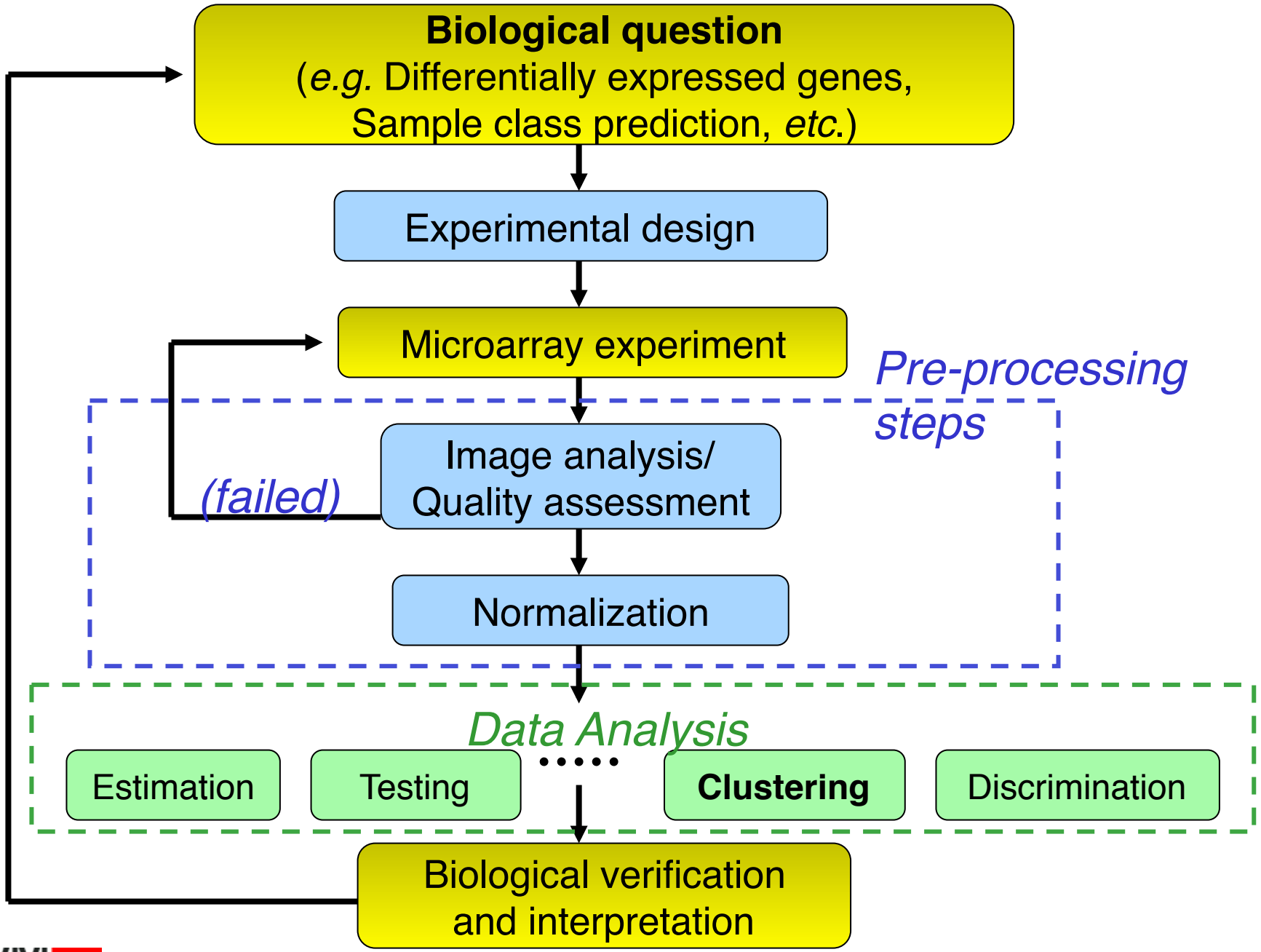
Statistics for Genomic Data Analysis

Cluster analysis



http://moodle.epfl.ch/course/view.php?id=15271


Biological question
(e.g. Differentially expressed genes,
Sample class prediction, etc.)
Experimental design
Microarray experiment
Pre-processing steps
Image analysis/
Quality assessment
(failed)
Normalization
Data Analysis
Estimation
Testing
Clustering
Discrimination
Biological verification
and interpretation

Classification

- Historically, *objects* are classified into *groups*
  - periodic table of the elements (chemistry)
  - taxonomy (zoology, botany)
- Why classify?
  - organizational convenience, convenient summary
  - prediction
  - explanation
- *Note:* these aims do not necessarily lead to the same classification; e.g. *SIZE* of object in hardware store vs. *TYPE/USE* of object

Classification, cont

- *Classification* divides objects into groups based on a set of values
- Unlike a theory, a classification is *neither true nor false*, and should be judged largely on the usefulness of results (Everitt)
- However, a classification (clustering) may be useful for suggesting a theory, which could then be tested

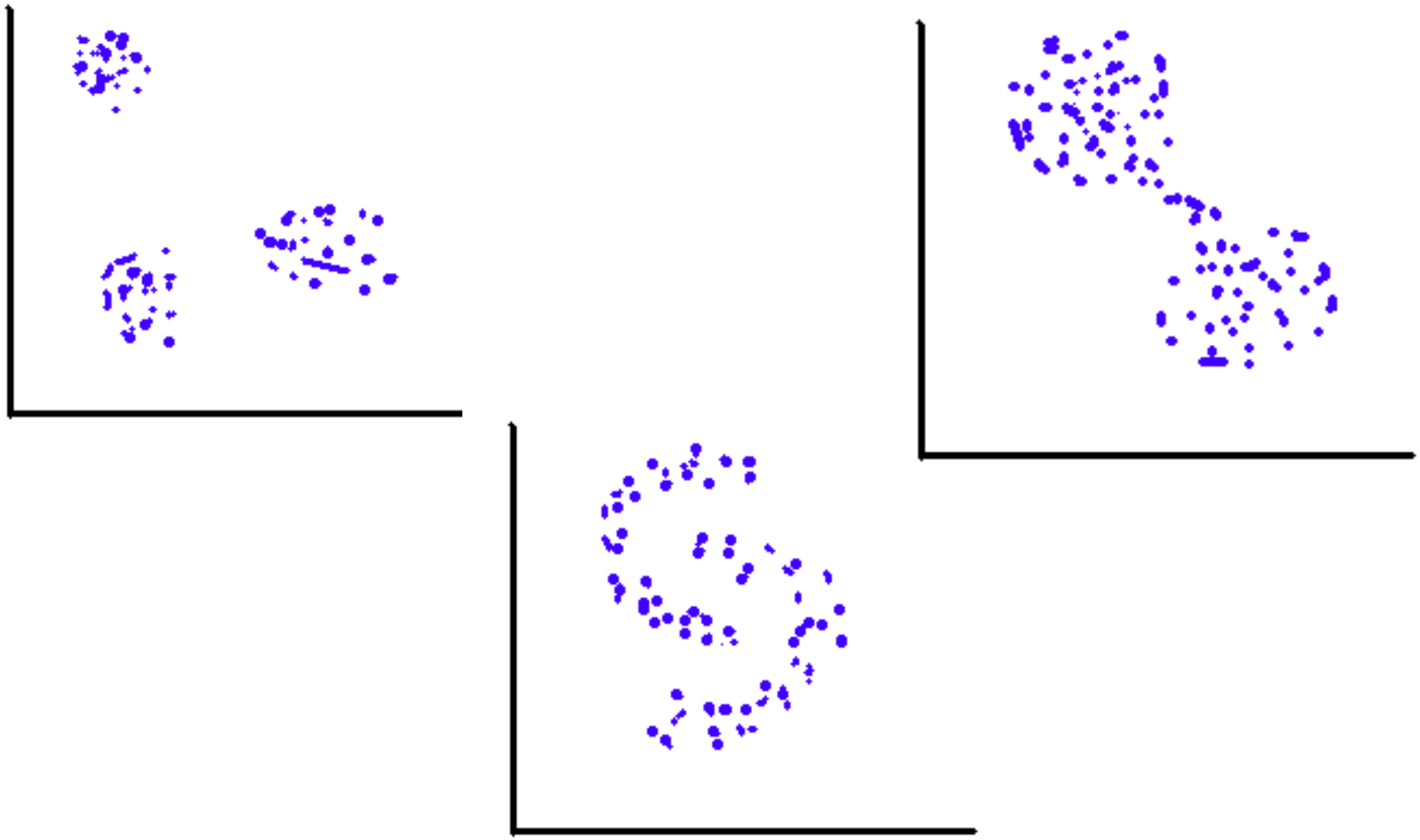
Classification

- *Task:* assign objects to classes (groups) on the basis of measurements made on the objects
- *Supervised:* classes are predefined, want to use a (training or learning) set of labeled objects to form a classifier for classification of future observations (discrimination analysis)
- *Unsupervised:* classes unknown, want to discover them from the data (cluster analysis)

Cluster analysis

- Addresses the problem: Given *n* objects, each described by *p* variables (or *features*), derive a *useful division* into a number of classes
- Often want a *partition* of objects
  - But also 'fuzzy clustering'
  - Could also take an exploratory perspective
- 'Unsupervised learning'
- Most clustering is not statistical

Difficulties in defining 'cluster'



Clustering Gene Expression Data

- Can cluster *genes* (rows), e.g. to (attempt to) identify groups of co-regulated genes
- Can cluster *samples* (columns), e.g. to identify tumors based on profiles
- Can cluster *both* rows and columns at the same time

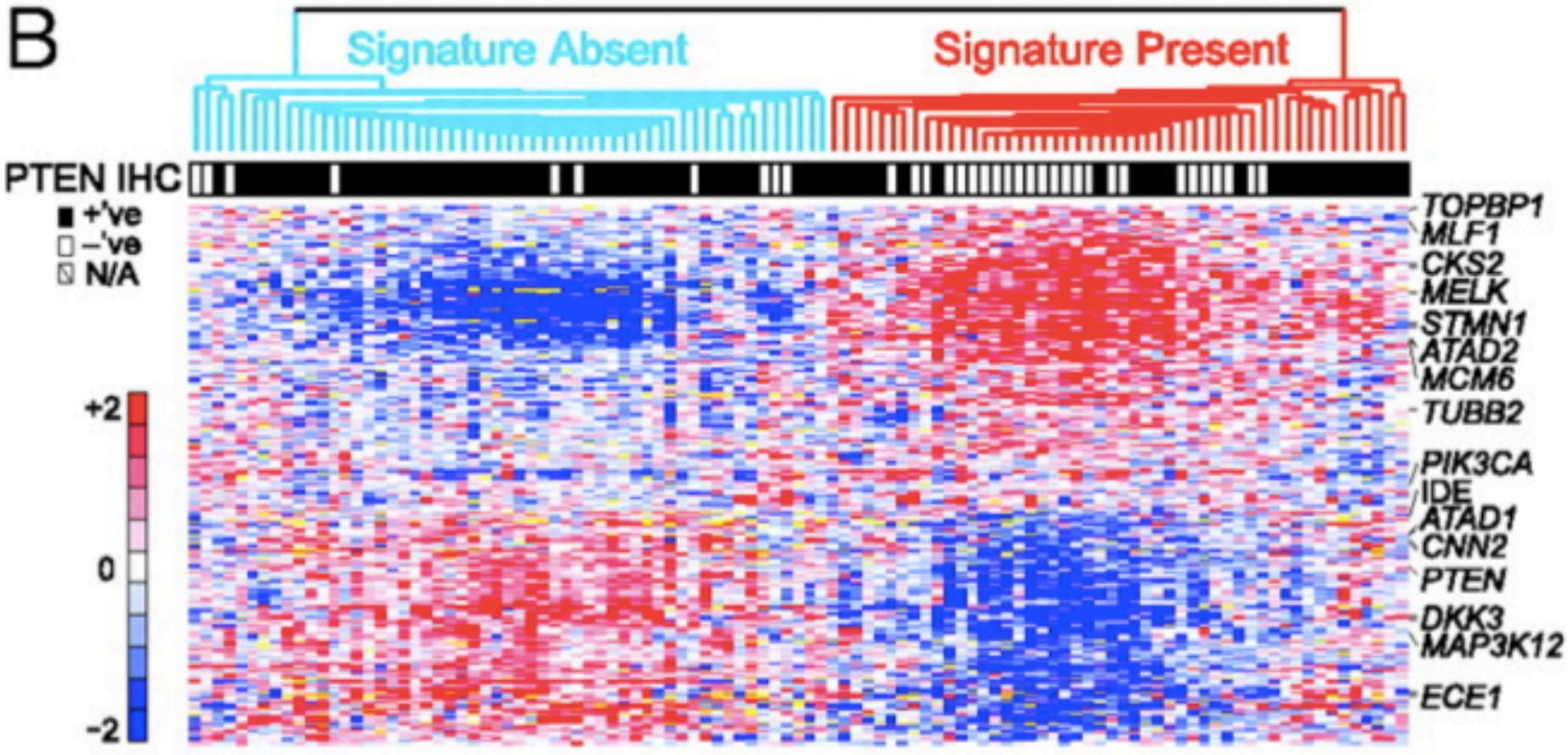
Clustering Gene Expression Data

- Leads to readily interpretable figures
- Can be helpful for identifying patterns in time or space
- Useful (essential?) when *seeking new subclasses* of samples
- Can be used for exploratory, quality assessment purposes

Visualizing Gene Expression Data

- Dendrogram (tree diagram)
- Heat Diagram (heatmap)
  - available as R function **`heatmap()`**
- Need to *reduce number of genes* first for figures to be legible/interpretable (at most a few hundred genes, not a whole array)
- A visual representation for a given clustering (e.g. dendrogram) is *not unique*
- Beware the influence of representation on *apparent structure* (e.g. color scheme)

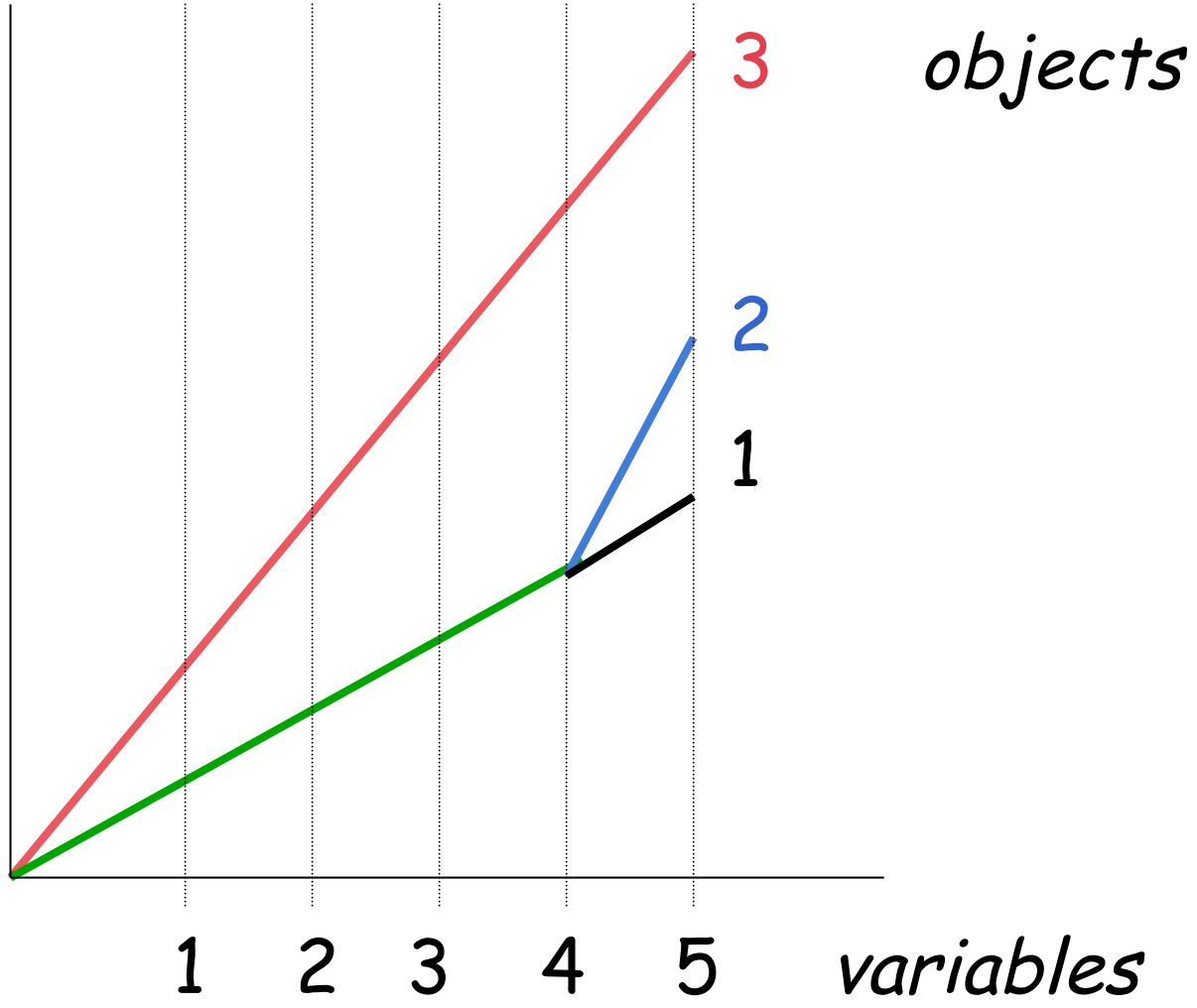
Cluster visualization



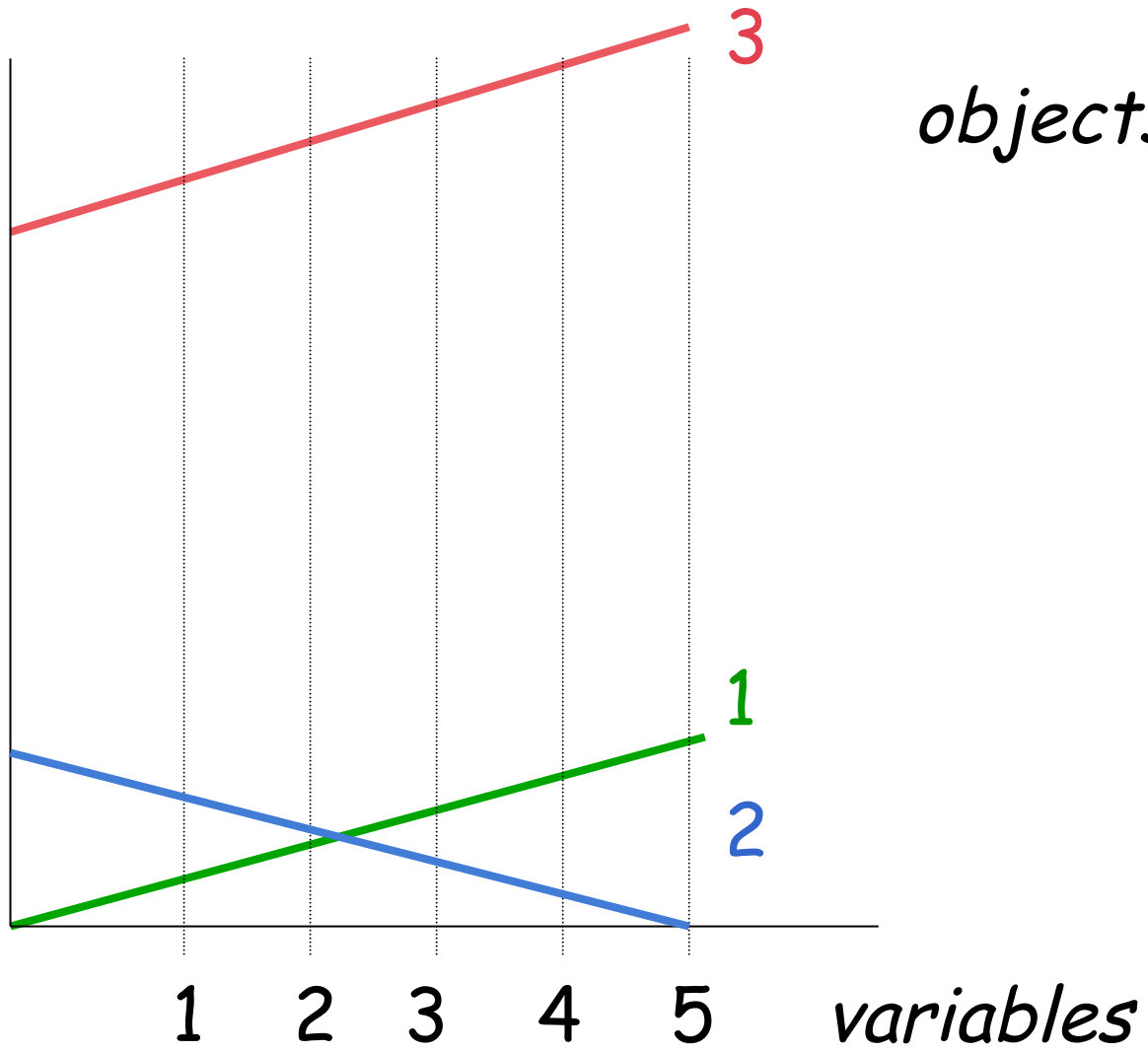
Similarity

- *Similarity* s_{ij} indicates the *strength of relationship* between two objects *i* and *j*
- Usually $0 \leq s_{ij} \leq 1$
- Correlation-based similarity ranges from –1 to 1
- Use of (1-)*correlation-based similarity* is quite common in gene expression studies but is in general contentious...

Problems using correlation



A more extreme example



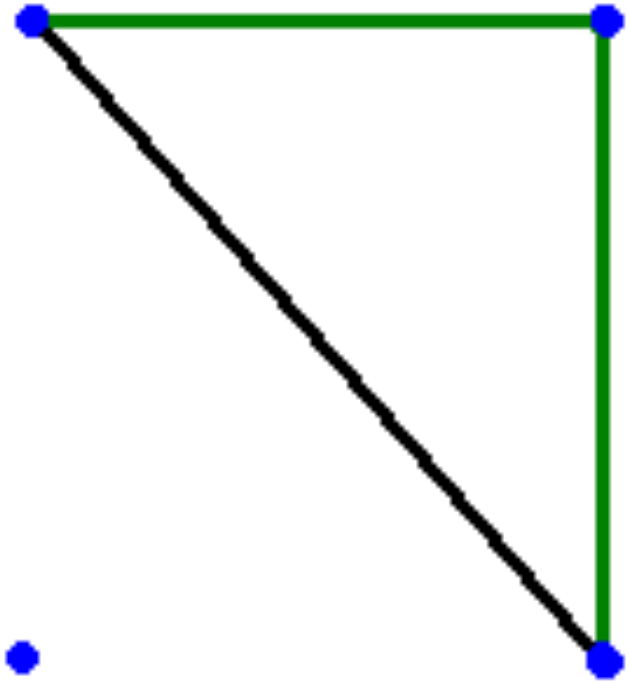
Dissimilarity and Distance

- Associated with similarity measures s_{ij} bounded by 0 and 1 is a *dissimilarity* $d_{ij} = 1 - s_{ij}$
- *Distance* measures have the metric property ($d_{ij} + d_{ik} \geq d_{jk}$)
- Many examples: Euclidean ('as the crow flies'), Manhattan ('city block'), *etc.*
- Distance measure has a large effect on performance
- Behavior of distance measure related to *scale* of measurement

Distance example

Euclidean

Manhattan



What distance should I use?

- This is like asking: *What tool should I buy?*
- It depends on what similarities you are interested in finding
- With Euclidean distance, larger values will tend to dominate; not useful if large value is simply a result of using smaller units (*e.g.*, grams vs Kilos)
- Can get around this (if desired) by scaling or standardizing variables
- Can also scale variables in *arbitrary directions* (rather than axis directions) using Mahalanobis distance

  $\sqrt{(x-y)^T S^{-1}(x-y)}$; usually S = cov. matrix

Partitioning Methods

- Partition the objects into a *prespecified* number of groups K
- Iteratively reallocate objects to clusters until some criterion is met (e.g. minimize within cluster sums of squares)
  - k-means
  - self-organizing maps (SOM)
  - partitioning around medoids (PAM; more robust and computationally efficient than k-means)
- Sometimes model-based clustering

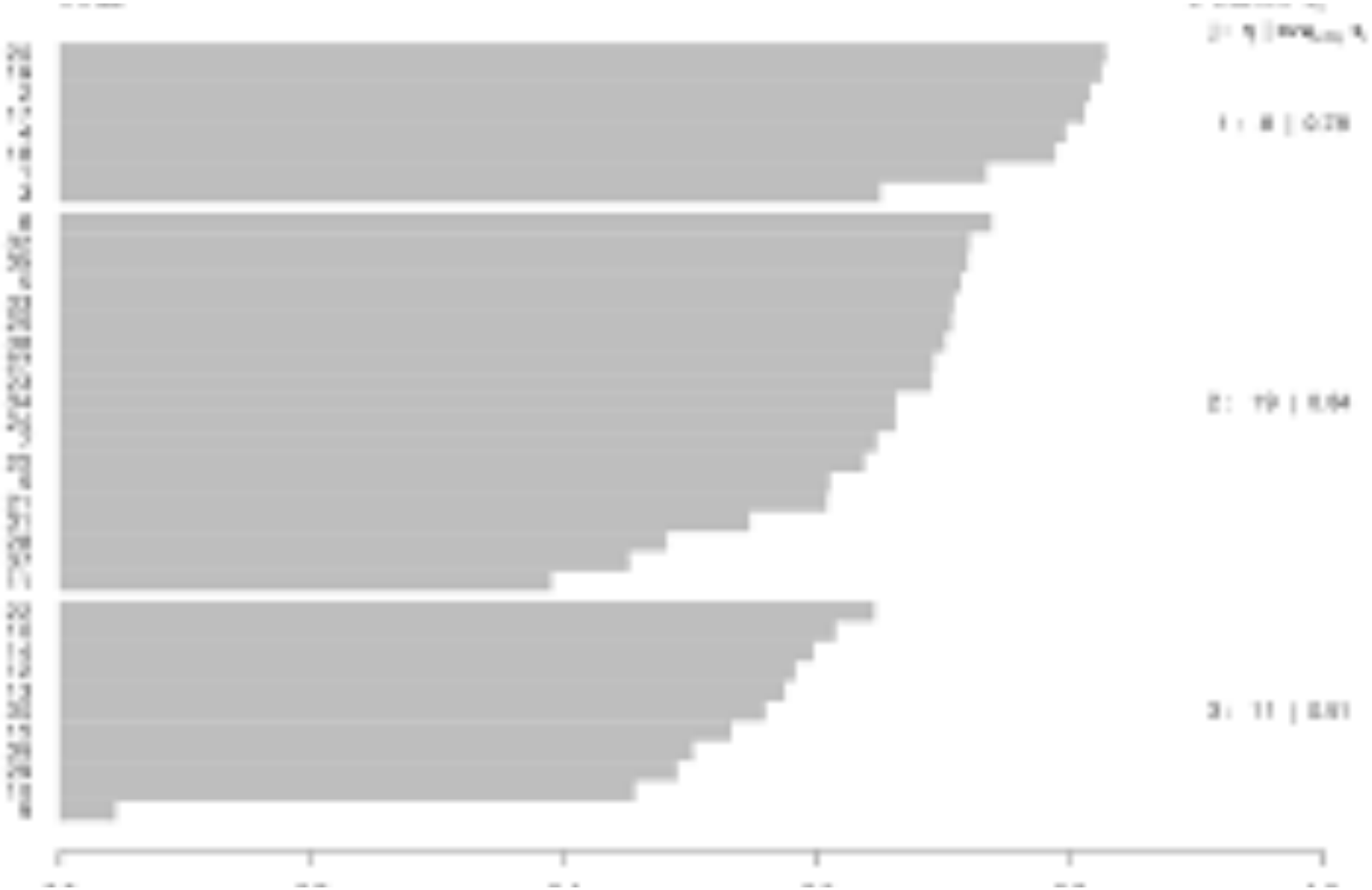
PAM - silhouette

- A measure is calculated for each observation to see how well it fits in assigned
- This is done by comparing how close the object is to other objects in *its own cluster* with how close it is to objects in *other clusters*
- Values *near 1* : observation is well placed ; *near 0* : likely the obs might really belong in another cluster
- Value displayed from smallest to largest (within cluster)

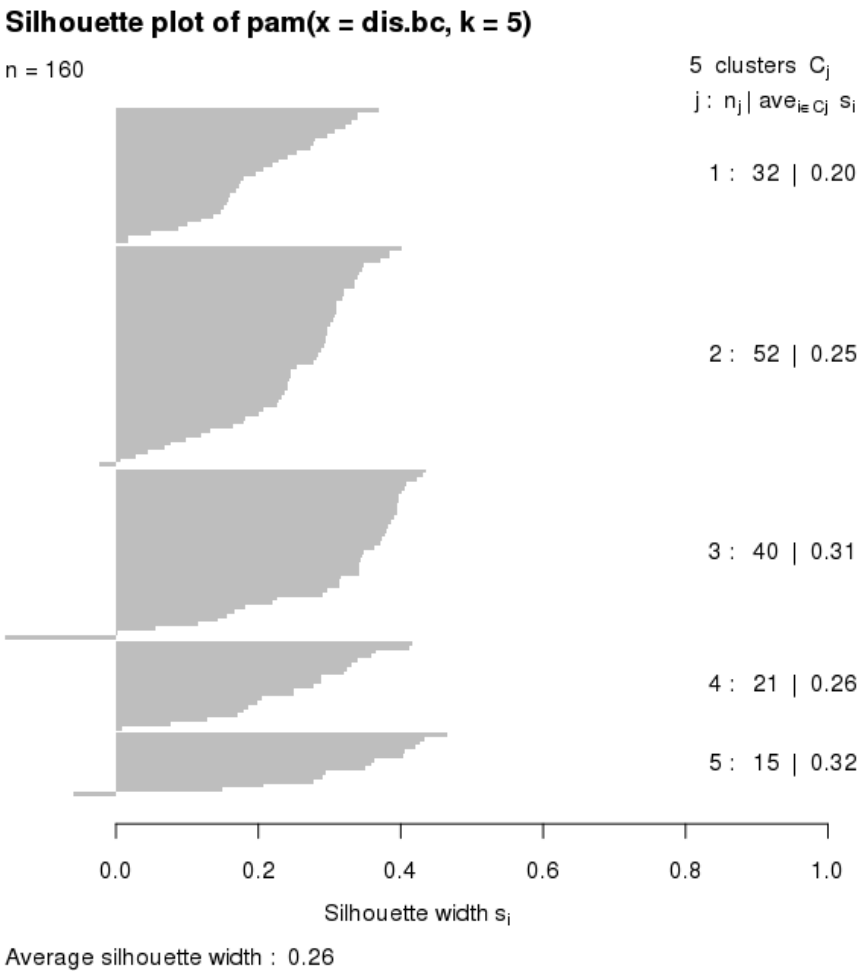
Average silhouette width

- Summary measure : Average Silhouette Width
- *Interpretation:*
  - 0.71-1.0 : strong structure
  - 0.51-0.70 : reasonably strong structure
  - 0.26-0.50 : weak structure, could be artificial
  - < 0.25 : No substantial structure found
- Number of clusters estimated by *optimum average silhouette width*

Example: 3 clusters



Example: 5 clusters



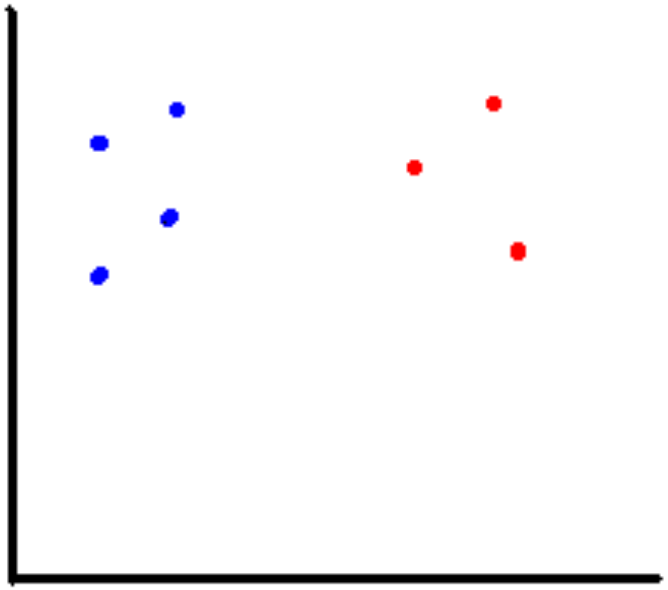
Hierarchical Clustering

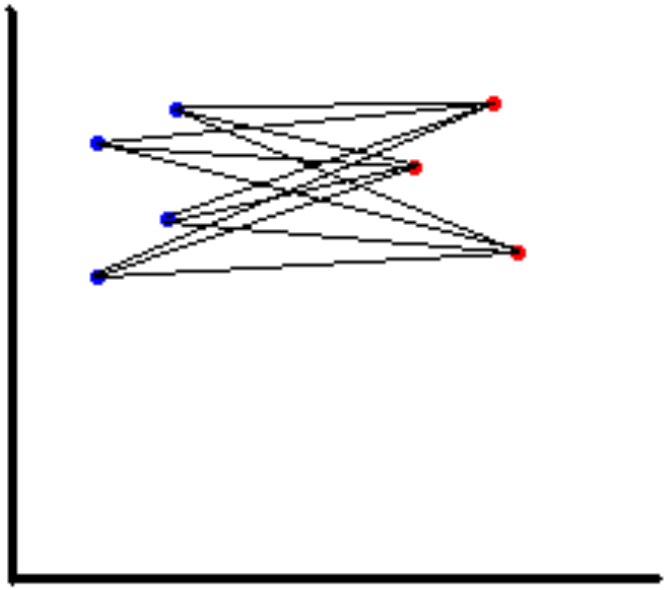
- Produce a *dendrogram* (tree diagram)
- Avoid prespecification of the number of clusters K
- The tree can be built in two distinct ways:
  - Bottom-up: *agglomerative* clustering
  - Top-down: *divisive* clustering

Agglomerative Methods

- Start with *n* mRNA sample (or G gene) clusters
- At each step, *merge* the two closest clusters using a measure of between-cluster dissimilarity
- Examples of *between-cluster* dissimilarities:
  - *Average linkage (Unweighted Pair Group Method with Arithmetic Mean (UPGMA)):* average of pairwise dissimilarities
  - *Single-link (NN):* min of pairwise dissimilarities
  - *Complete-link (FN):* max of pairwise dissimilarities
  - *Ward's method:* min information loss

Between cluster distances: avg, NN, FN





Ward's method

- Distance between two clusters is how much the sum of squares will increase when merged:

$$\begin{aligned}\Delta(A,B) &= \sum_{i\in A\cup B}\|\vec{x}_i-\vec{m}_{A\cup B}\|^2-\sum_{i\in A}\|\vec{x}_i-\vec{m}_A\|^2-\sum_{i\in B}\|\vec{x}_i-\vec{m}_B\|^2\\ &= \frac{n_A n_B}{n_A+n_B}\|\vec{m}_A-\vec{m}_B\|^2\end{aligned}$$

- where $\underline{m}_j$ is the center of cluster j, n_j is the number of points in it
- Δ = *merging cost* of combining clusters A and B
- Given two pairs of clusters whose centers are equally far apart, Ward's method prefers to merge the smaller ones

Divisive Methods

- Start with only *one* cluster
- At each step, *split* clusters into two parts
- Advantage: Obtain the main structure of the data (*i.e.* focus on upper levels of dendrogram)
- Disadvantage: Computational difficulties when considering all possible divisions into two groups

Partitioning vs. Hierarchical

- *Partitioning*
  - Advantage: Provides clusters that satisfy some optimality criterion (approximately)
  - Disadvantages: Need initial K, long computation time
- *Hierarchical*
  - Advantage: Fast computation (agglomerative)
  - Disadvantages: Rigid, cannot correct later for erroneous decisions made earlier

R: clustering

- A number of R packages contain functions to carry out clustering, including:
  - **stats: hclust**
  - **cluster (Kaufman and Rousseeuw)**
  - **fpc**
  - **mclust**
  - **E1071**
- And many more!

Generic Clustering Tasks

- Estimating number of clusters
- Assigning each object to a cluster
- Assessing strength/confidence of cluster assignments for individual objects
- Assessing cluster homogeneity
- *(Interpretation of the resulting clusters)*

Estimating how many clusters

- Many suggestions for how to decide this!
- Indices based on homogeneity and/or separation (within and between cluster sums of squares)
- Milligan and Cooper (Psychometrika 50:159-179, 1985) studied performance of 30 such methods in a large simulation
- R package `fpc` (Christian Hennig) has function `cluster.stats` which computes many of these

Additional methods

- Model-based criteria (AIC, BIC, MDL) when using model-based clustering
- GAP, GAP-PC (Tibshirani et al.)
- Average silhouette width (Kaufman and Rousseuw)
- mean silhouette split (Pollard and van der Laan)
- clest (Dudoit and Fridlyand)

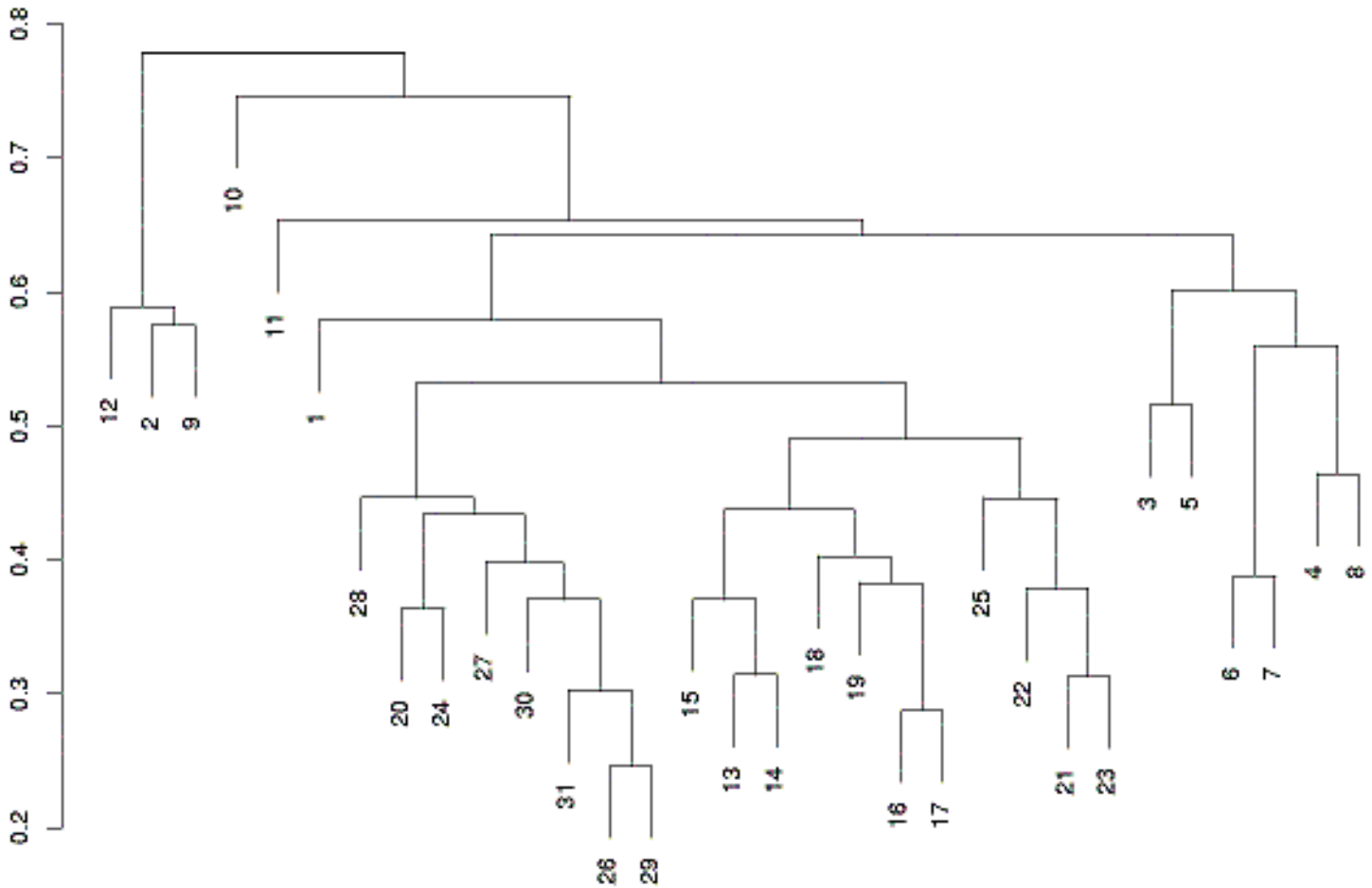
(BREAK)

Example: Bittner et al.

It has been proposed (by many) that a *cancer taxonomy* can be identified from *gene expression experiments*.

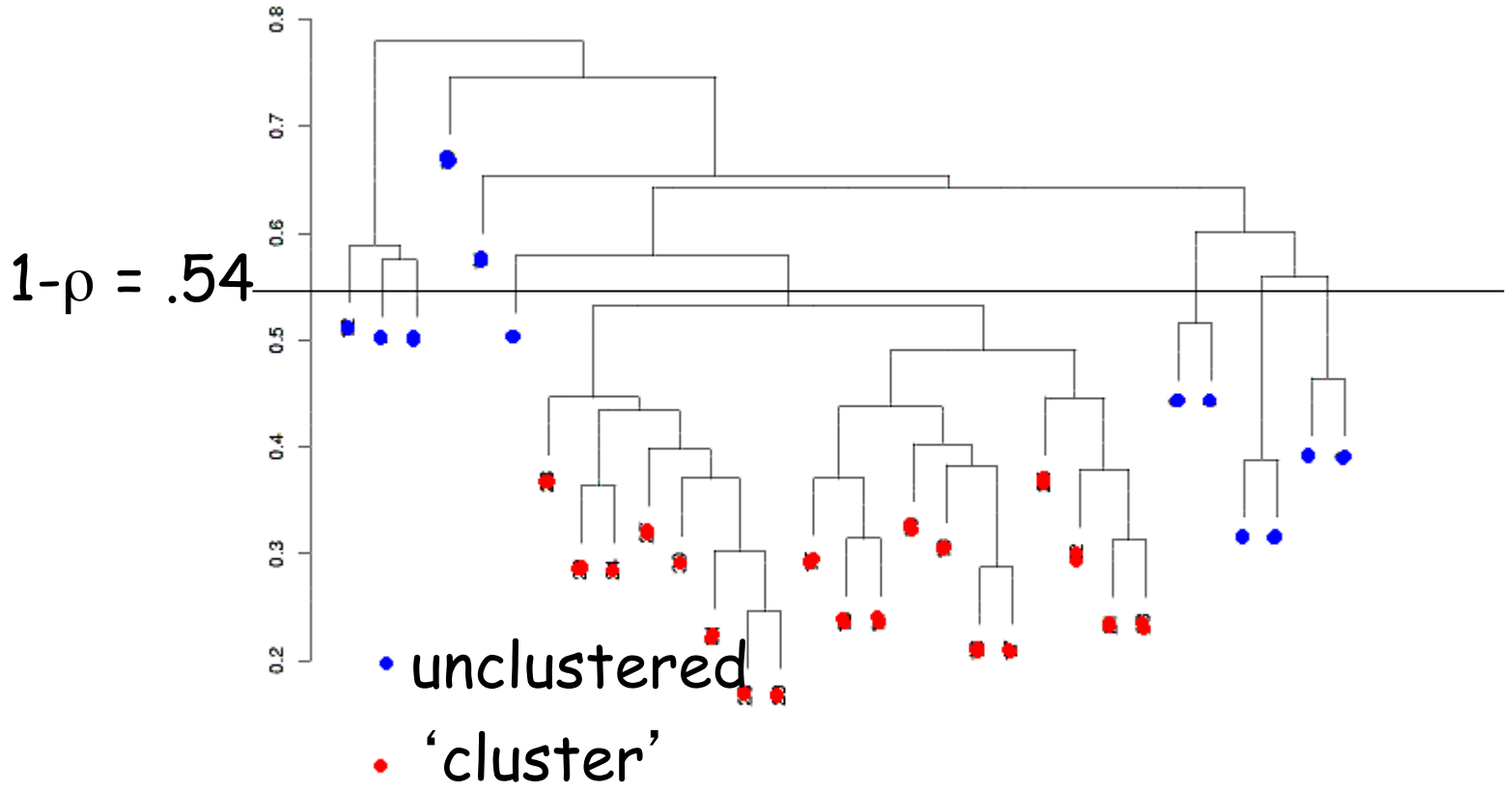
- 31 melanomas (from a variety of tissues/cell lines)
- 7 controls
- 8150 cDNAs
- 6971 unique genes
- 3613 genes 'strongly detected'

Average linkage hierarchical clustering, melanoma only



How many clusters are present?

Average linkage, melanoma only



Issues in Clustering

- Pre-processing (Image analysis and Normalization)
- Which *variables* are used
- Which *samples* are used
- Which *distance measure* is used
- Which *algorithm* is applied
- How to decide the *number of clusters K*

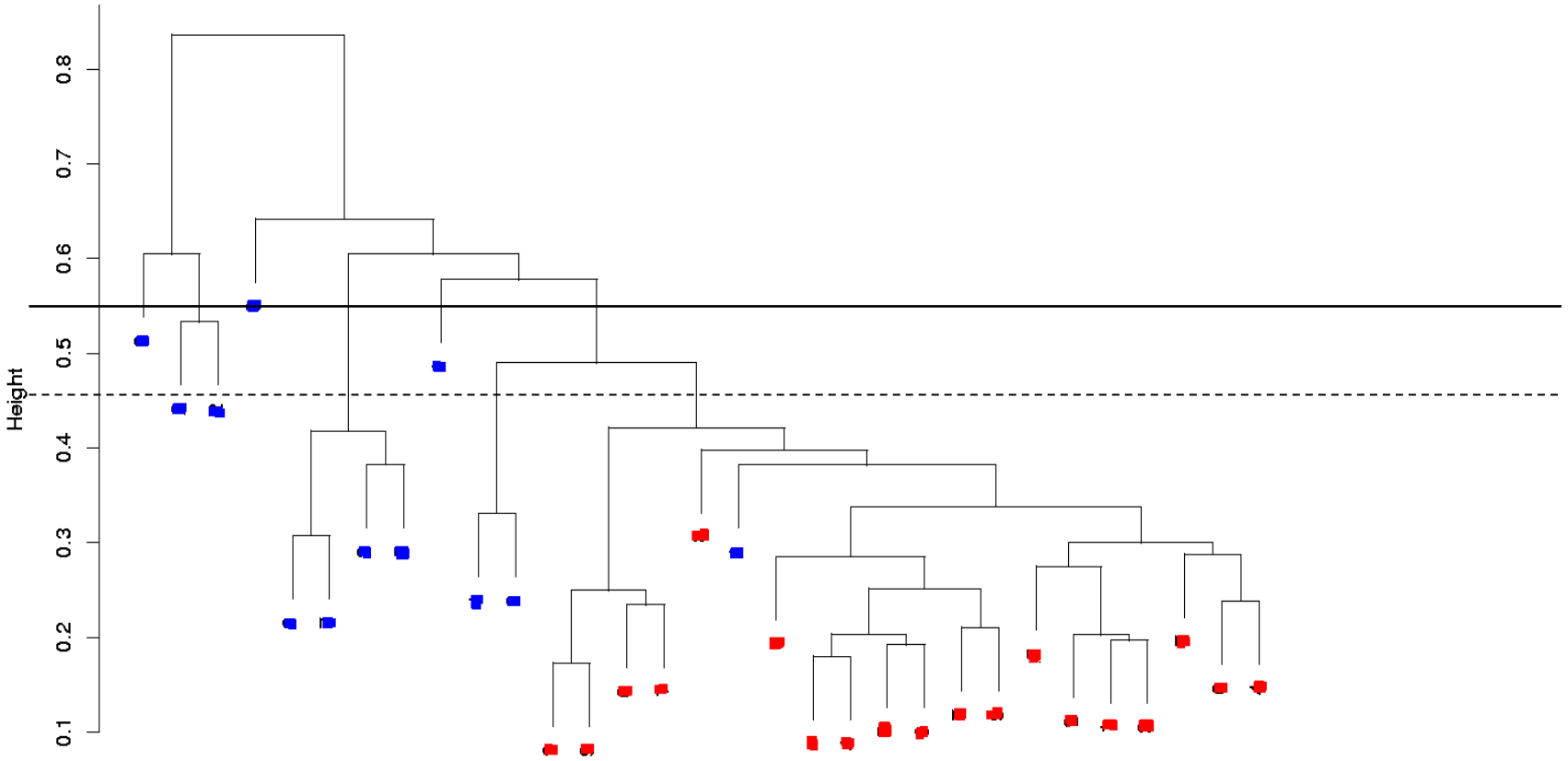
Issues in Clustering

- Pre-processing (Image analysis and Normalization)
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

Filtering Genes

- All genes (i.e. don't filter any)
- At least k (or a proportion p) of the samples must have expression values larger than some specified amount, A
- Genes showing 'sufficient' variation
  - a gap of size A in the central portion of the data
  - a interquartile range of at least B
  - 'large' SD, CV, ...

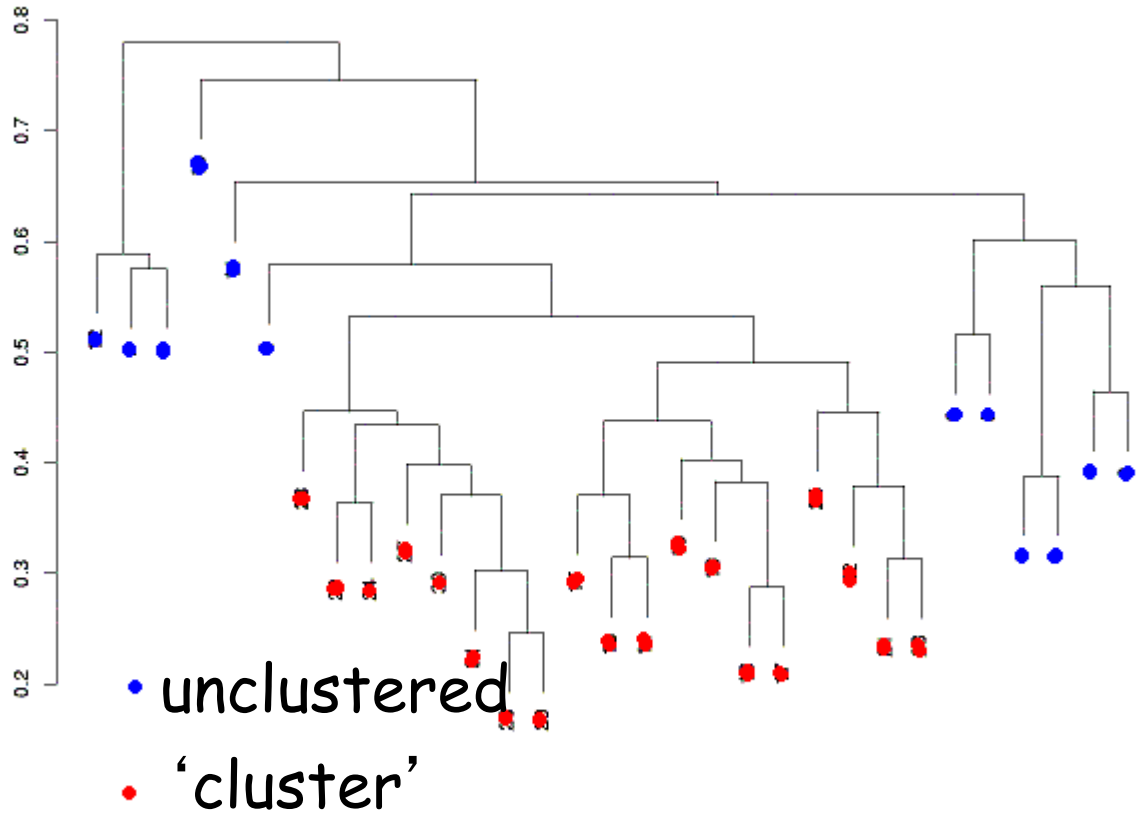
Average linkage, top 300 genes in SD



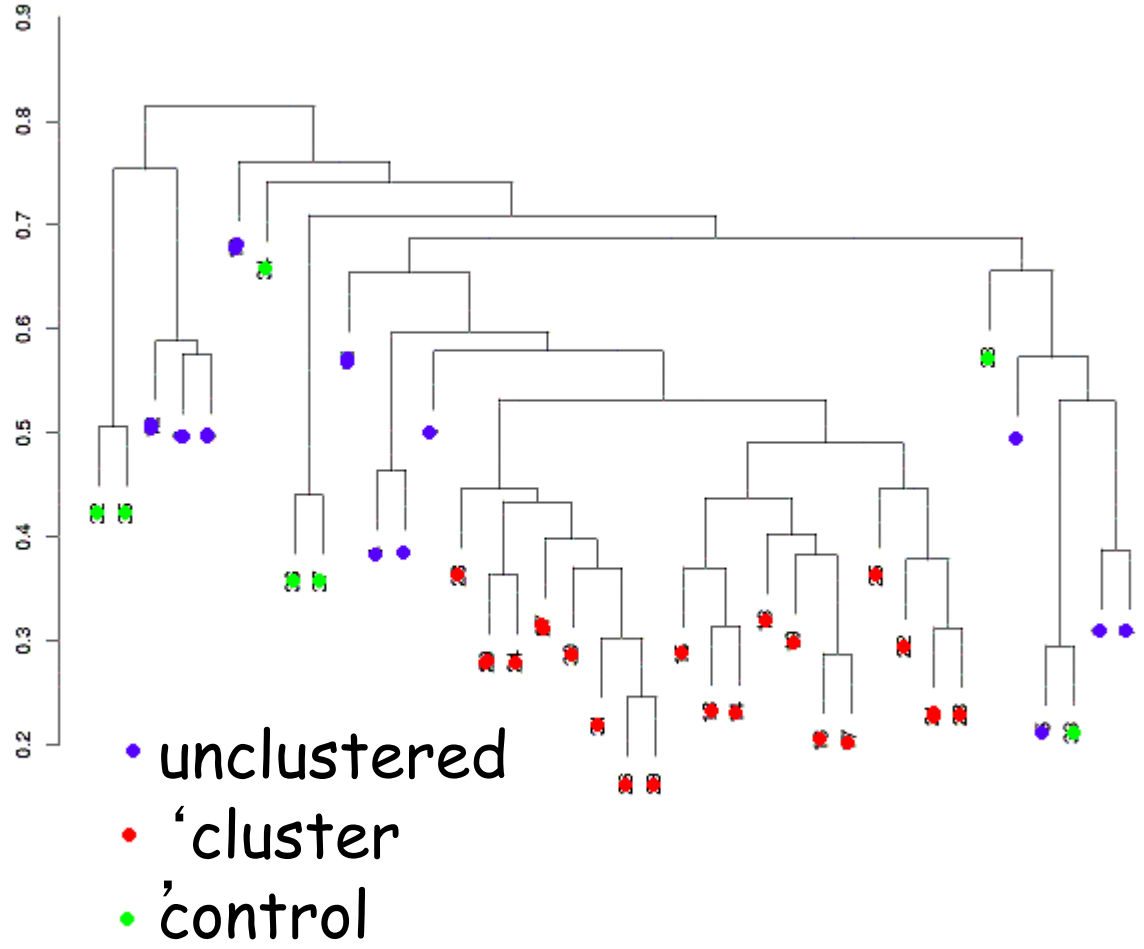
Issues in Clustering

- Pre-processing (Image analysis and Normalization)
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

Average linkage, *melanoma only*



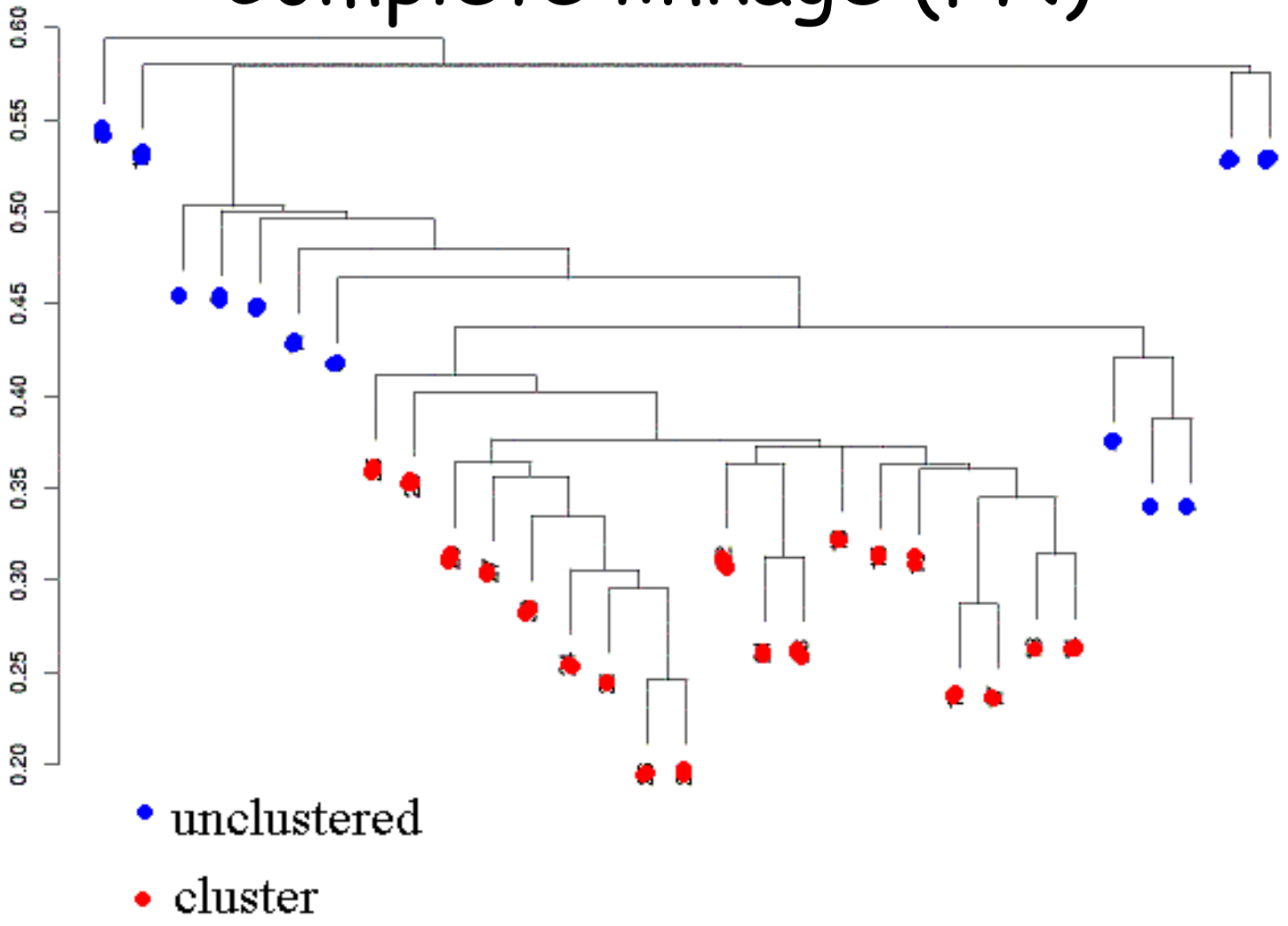
Average linkage, *melanoma & controls*



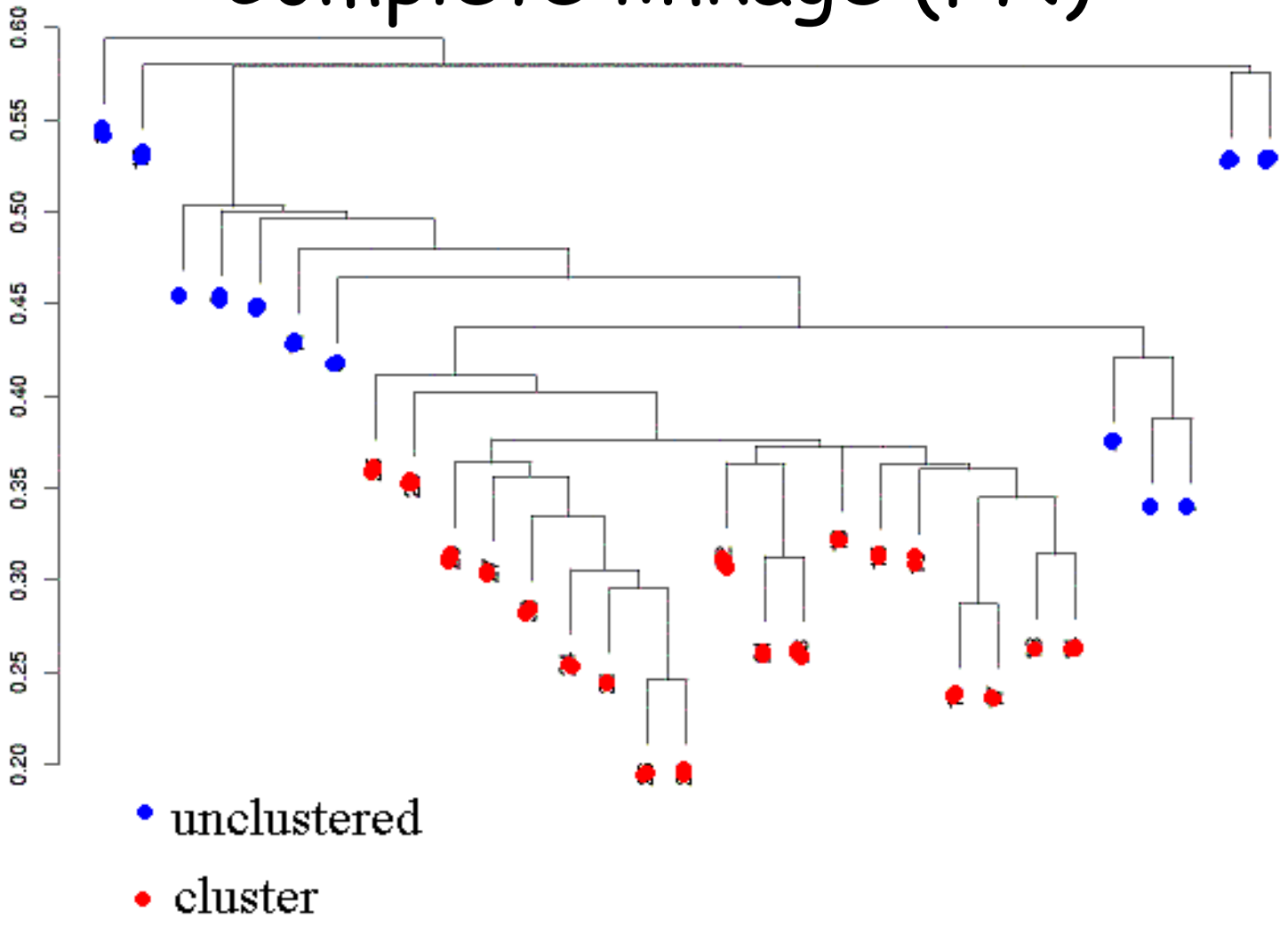
Issues in clustering

- Pre-processing
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

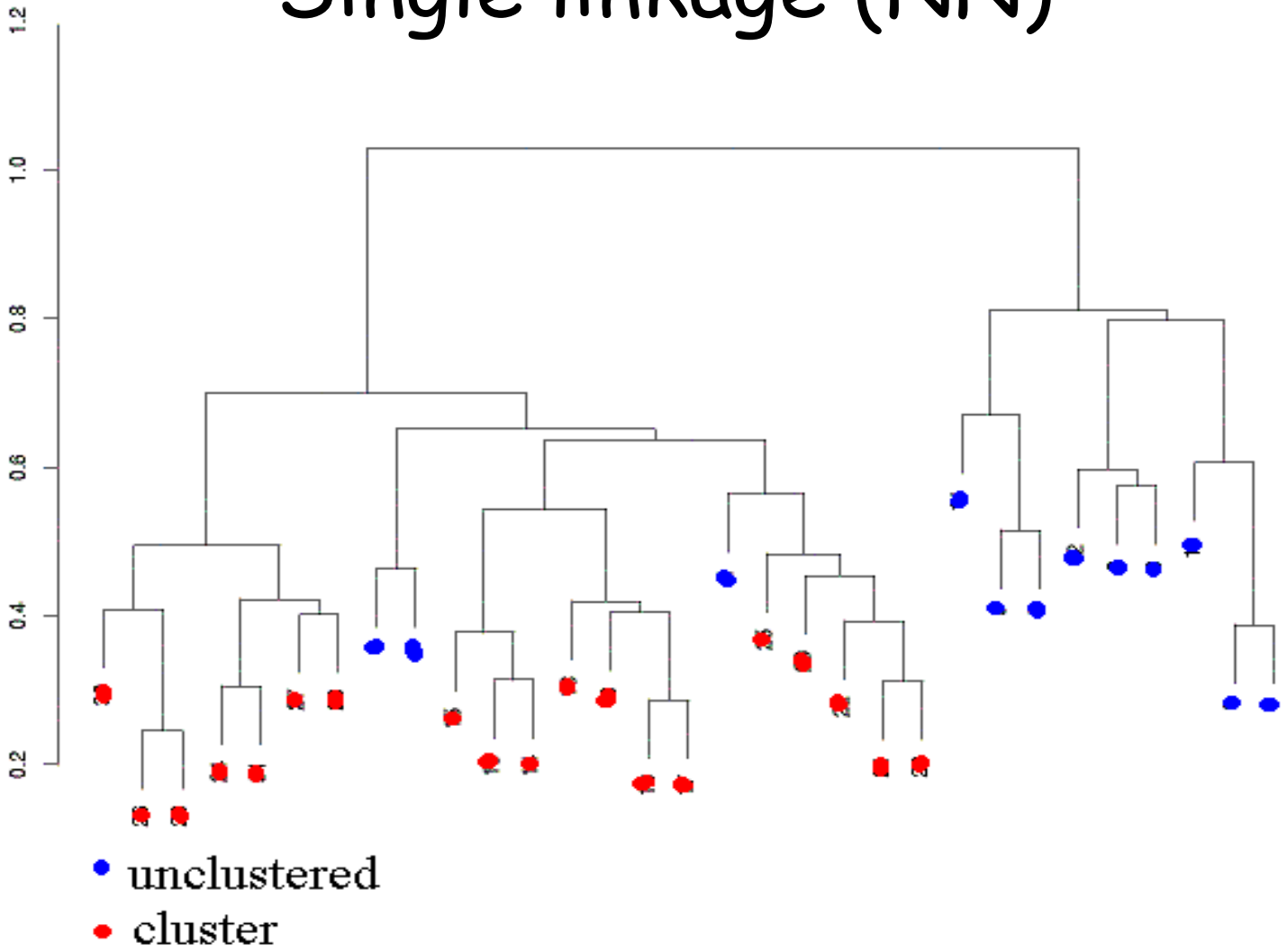
Complete linkage (FN)



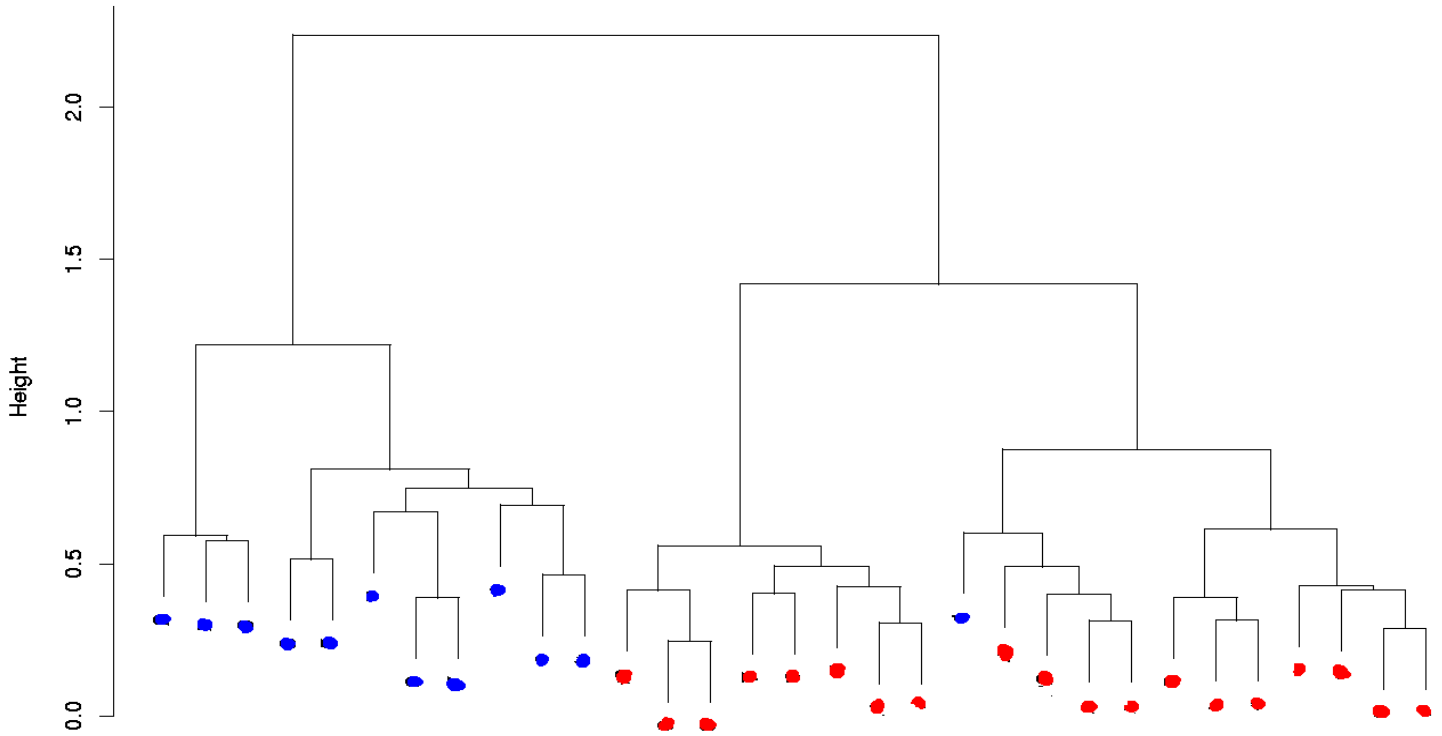
Complete linkage (FN)



Single linkage (NN)



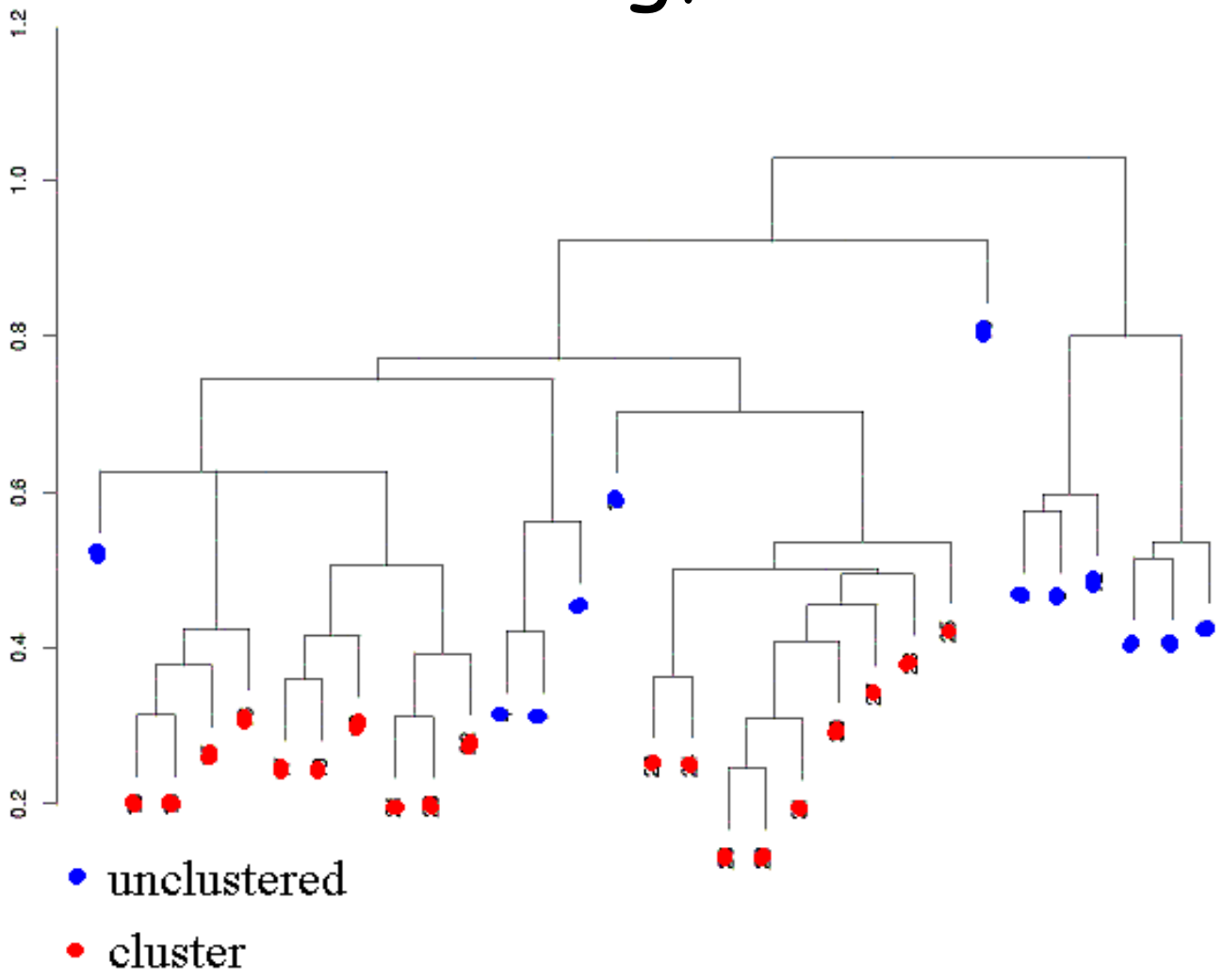
Ward's method (information loss)



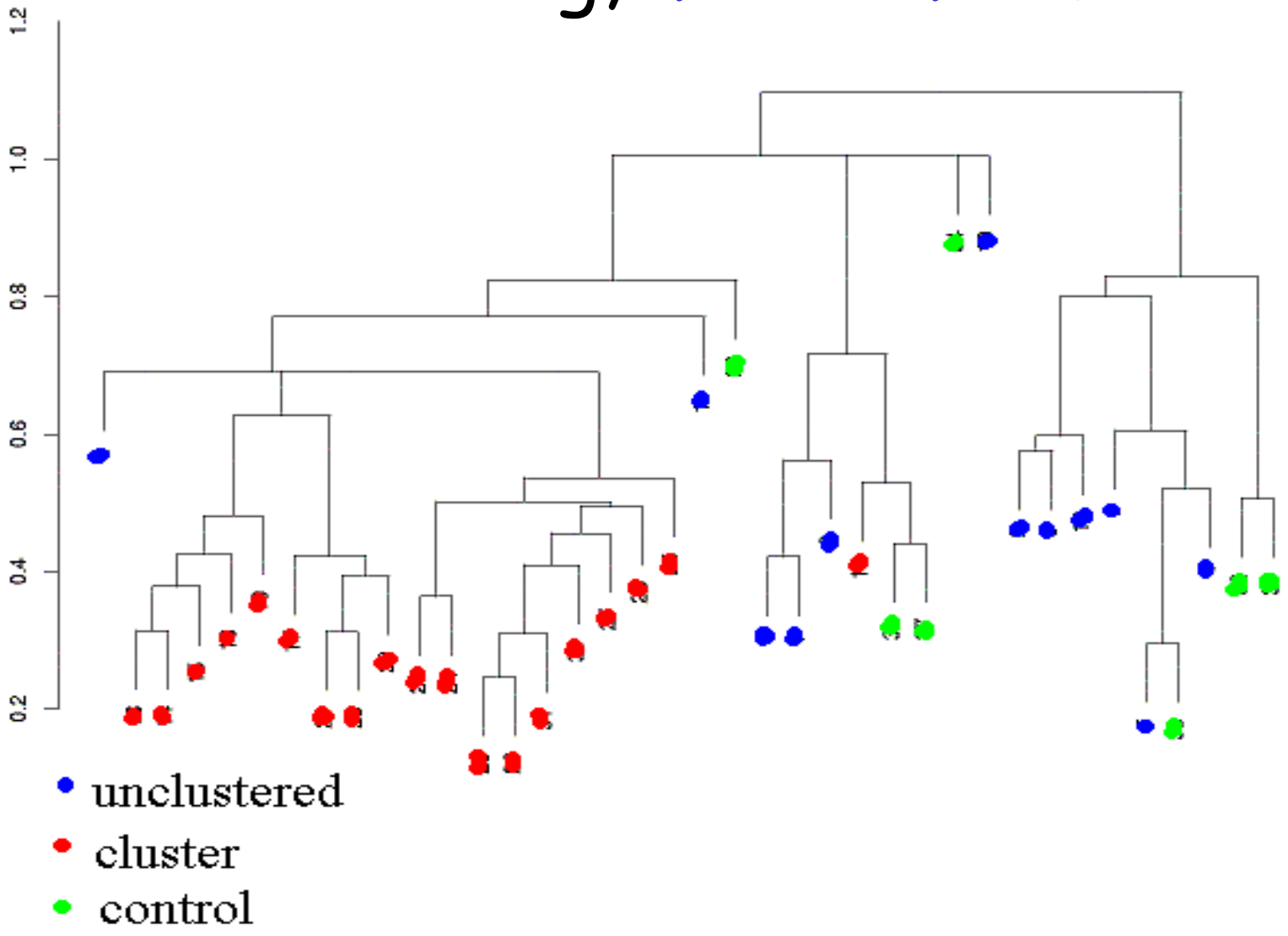
Issues in clustering

- Pre-processing
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

Divisive clustering, *melanoma only*



Divisive clustering, *melanoma & controls*



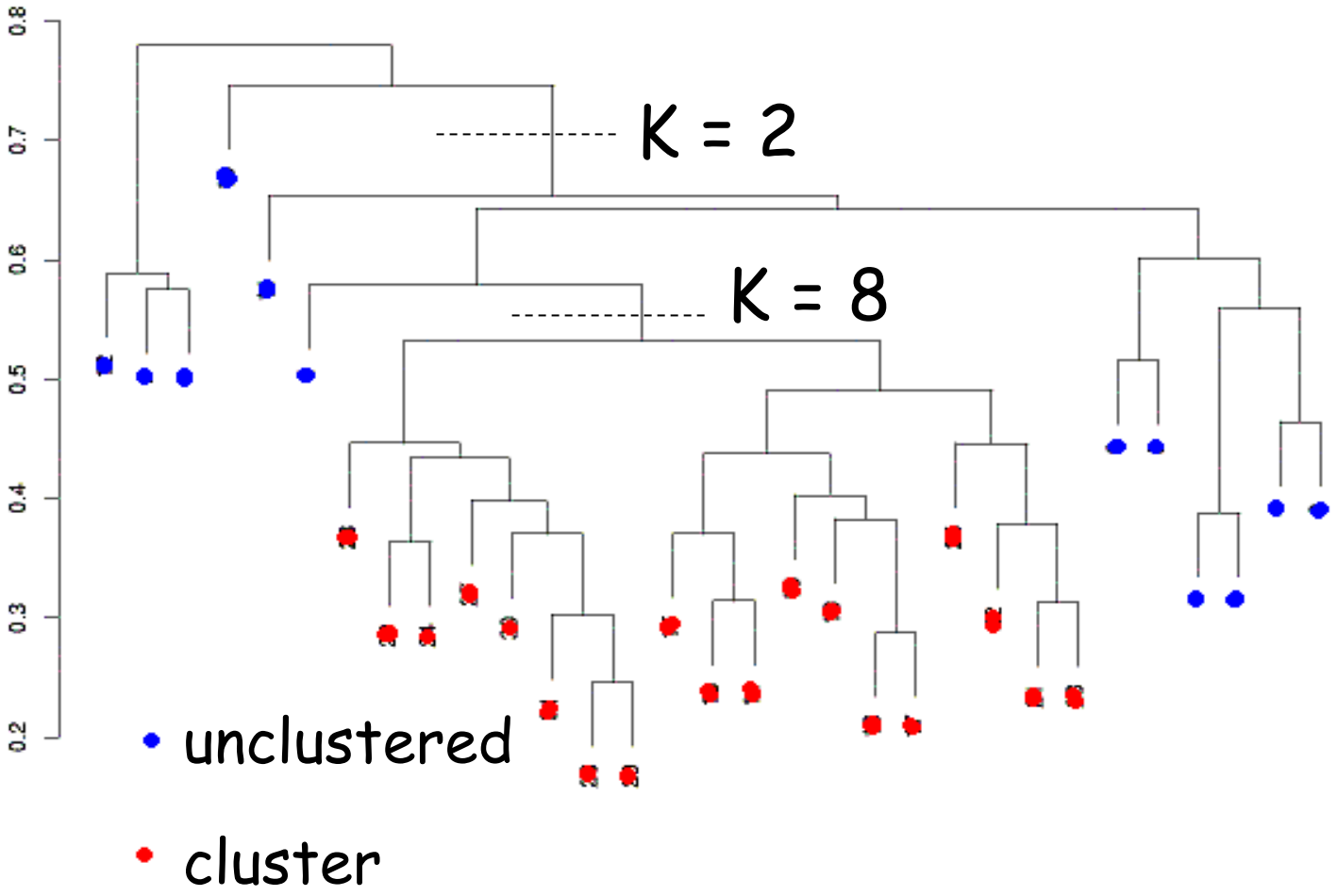
Issues in clustering

- Pre-processing
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

How many clusters *K*?

- Applying *several methods* yielded estimates of
  - *K* = 2 (largest cluster has 27 members)
  - to *K* = 8 (largest cluster has 19 members)

Average linkage, melanoma only



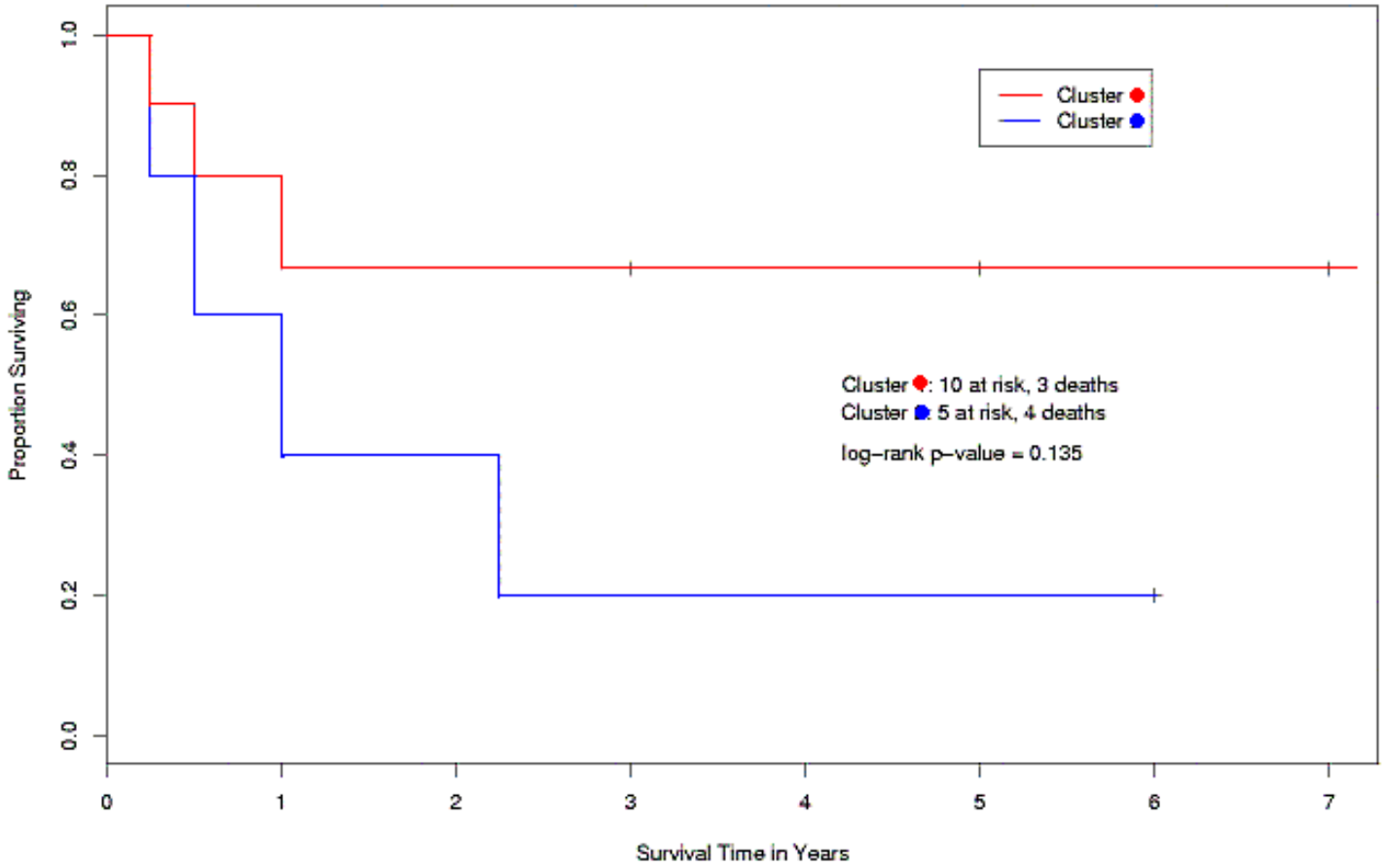
Association of Variables

- *Variables* tested for association with *cluster* :
  - Sex *(p = .68, n = 16 + 11 = 27)*
  - ❖Age *(p = .14, n = 15 + 10 = 25)*
  - ❖Mutation status *(p = .17, n = 12 + 7 = 19)*
  - Biopsy site *(p = .88, n = 14 + 10 = 24)*
  - Pigment *(p = .26, n = 13 + 9 = 22)*
  - Breslow thickness *(p = .26, n = 6 + 3 = 9)*
  - Clark level *(p = .44, n = 6 + 5 = 11)*
  - ❖Specimen type *(p = .11, n = 11 + 12 = 23)*

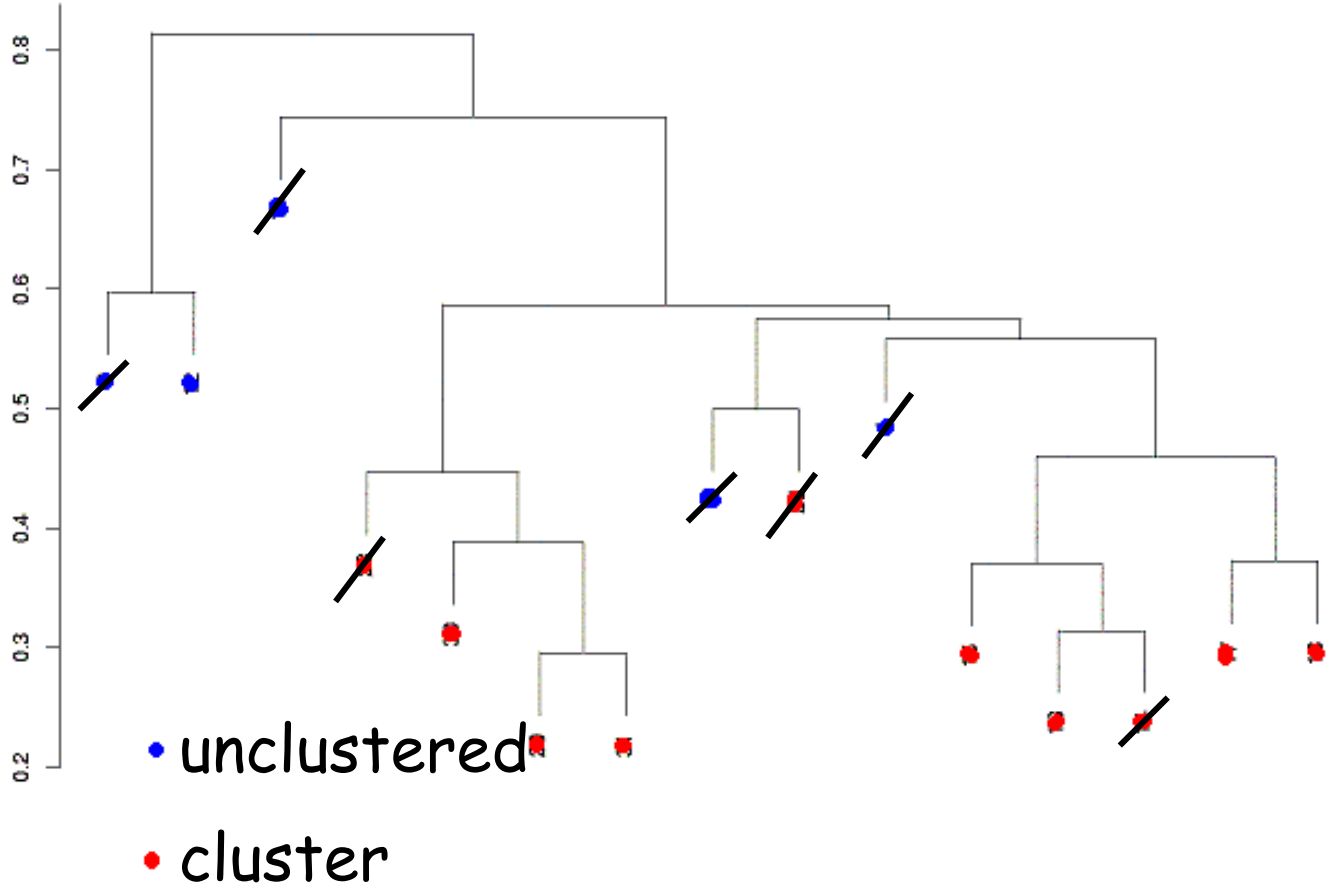
Survival analysis: Bittner et al.

- 15 of the 31 melanomas had associated *survival times*
- Bittner et al. also looked at differences in survival between the two groups (the 'cluster' and the 'unclustered' samples)
- 'Cluster' seemed associated with longer survival

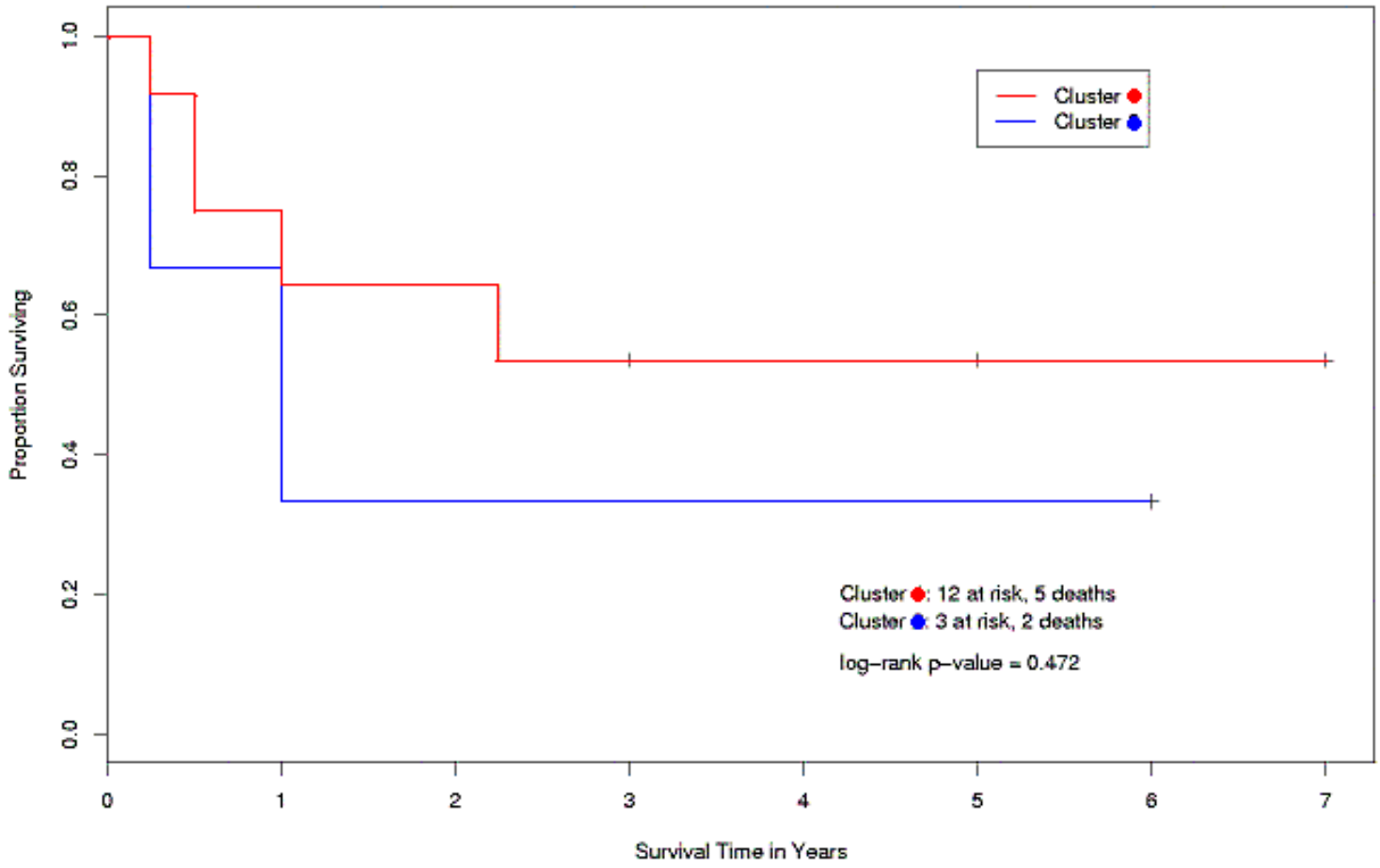
Kaplan-Meier Survival Curves



Average Linkage Hierarchical Clustering, survival samples only



Kaplan-Meier Survival Curves, new grouping



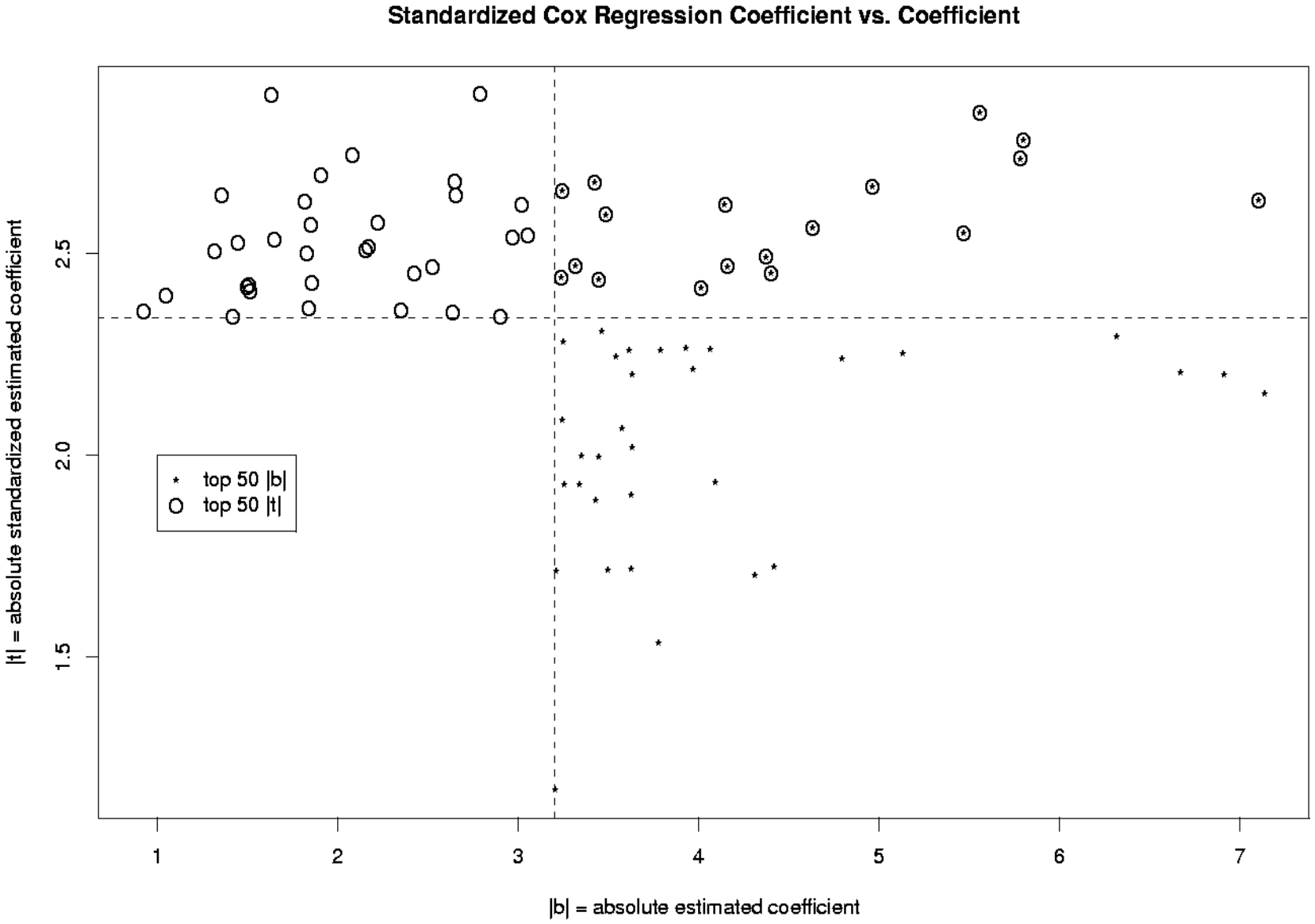
Identification of Genes Associated with Survival

For each gene j, $j = 1, \ldots, 3613$, model the *instantaneous failure rate*, or hazard function, $h(t)$ with the Cox proportional hazards model:

$$h(t) = h_0(t) \exp(\beta_j x_{ij})$$

and look for genes with *both* :

- large effect size $\hat{\beta}_j$
- large *standardized* effect size $\hat{\beta}_j / SE(\hat{\beta}_j)$


Standardized Cox Regression Coefficient vs. Coefficient
|t| = absolute standardized estimated coefficient
|b| = absolute estimated coefficient
top 50 |b|
top 50 |t|
1.5
2.0
2.5
1
2
3
4
5
6
7

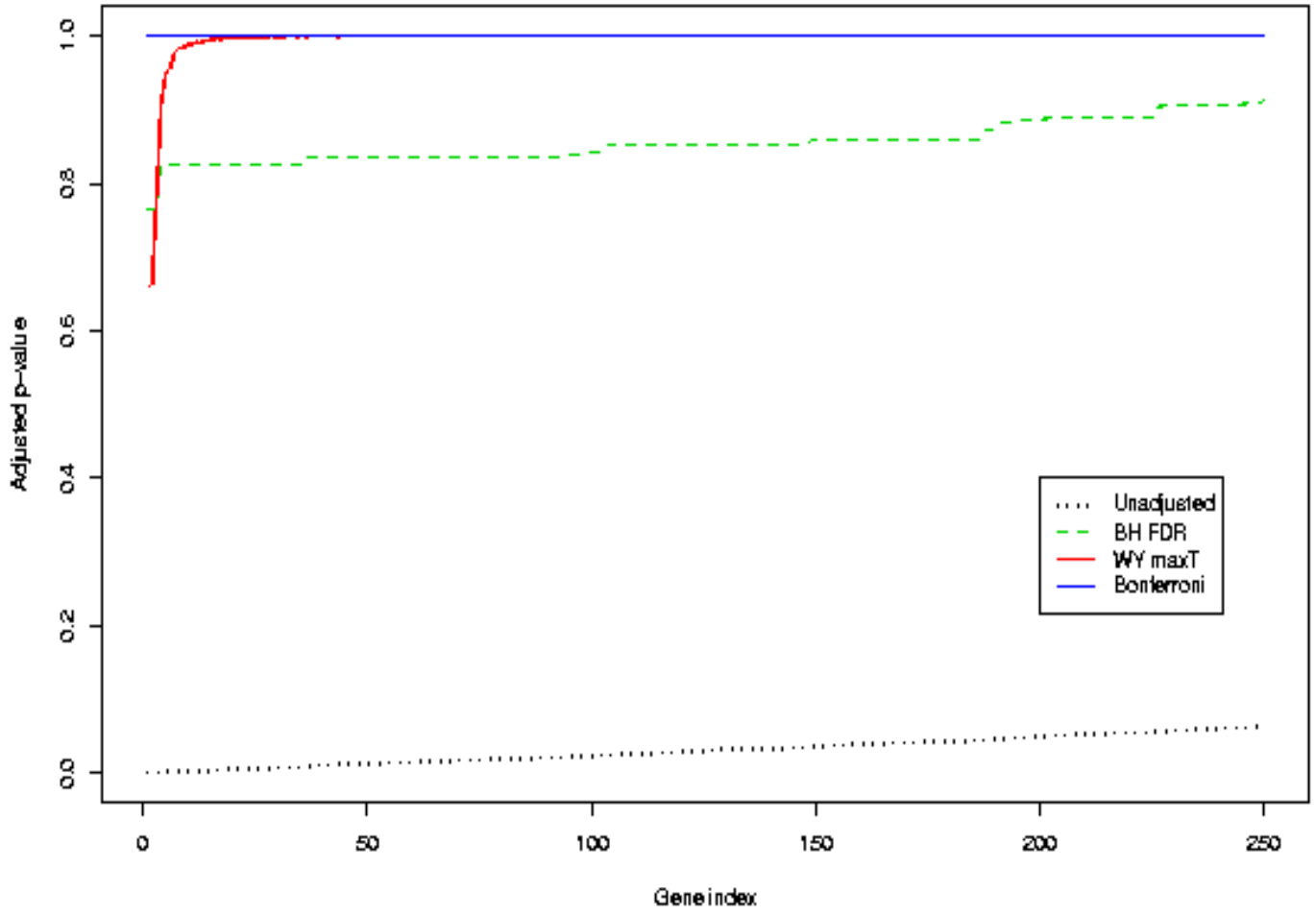
Sites Potentially Influencing Survival

Image Clone ID	UniGene Cluster	UniGene Cluster Title
137209	Hs.126076	Glutamate receptor interacting protein
240367	Hs.57419	Transcriptional repressor
838568	Hs.74649	Cytochrome c oxidase subunit Vlc
825470	Hs.247165	ESTs, Highly similar to topoisomerase
841501	Hs.77665	KIAA0102 gene product

Findings

- Top 5 genes by this method not in Bittner et al. 'weighted gene list' - Why?
- weighted gene list based on entire sample; our method only used half
- weighting relies on Bittner et al. cluster assignment
- other possibilities?

Statistical Significance of Cox Model Coefficients



Advantages of Modeling

- Can address questions of interest *directly*
  - Contrast with what has become the 'usual' (and indirect) approach with microarrays: clustering, followed by tests of association between cluster group and variables of interest
- Great deal of *existing machinery*
- *Quantitatively* assess strength of evidence

Limitations of Single Gene Tests

- May be too noisy in general to show much
- Do not reveal coordinated effects of positively correlated genes
- Hard to relate to pathways

Not Covered…

- Careful followup
  - Assessment of *proportionality*
  - Inclusion of *combinations* of genes, interactions
  - Consideration of alternative models
- Power assessment
  - Not worth it here, there can't be much!

Summary

- Buyer beware – results of cluster analysis should be treated with GREAT CAUTION and ATTENTION TO SPECIFICS, because…
- Many things can vary in a cluster analysis
- If covariates/group labels are known, then clustering is usually inefficient