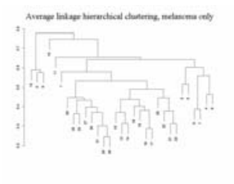
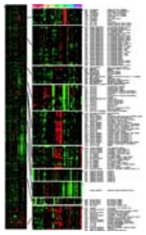
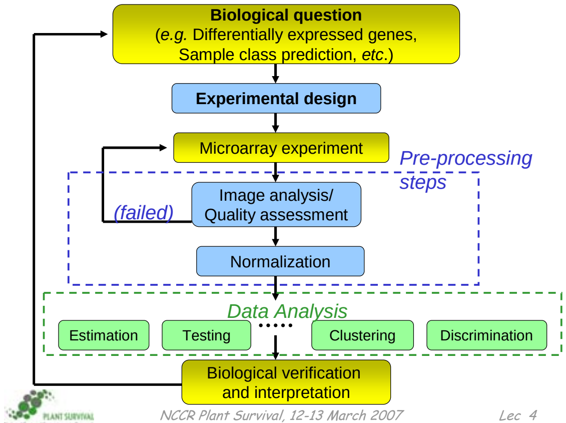
Statistics for cDNA microarrays

Experimental Design; Cluster Analysis





http://www.isrec.isb-sib.ch/~darlene/NCCR-PS/



Biological question
(*e.g.* Differentially expressed genes, Sample class prediction, *etc.*)

Experimental design

Microarray experiment

Pre-processing steps

(failed)

Image analysis/ Quality assessment

Normalization

Data Analysis

Estimation

Testing

Clustering

Discrimination

Biological verification and interpretation

Some Considerations for Microarray Experiments (I)

Scientific (Aims of the experiment)

- Specific questions and priorities
- How will the experiments answer the questions

Practical (Logistic)

- Types of mRNA samples: reference, control, treatment, mutant, etc
- Source and Amount of material (tissues, cell lines)
- *Number of slides available*

Some Considerations for Microarray Experiments (II)

Other Information

- Experimental process prior to hybridization sample isolation, mRNA extraction, amplification, labelling,…
- Controls planned: positive, negative, ratio, etc.
- Verification method: Northern, RT-PCR, in situ hybridization, etc.

Aspects of Experimental Design Applied to Microarrays (I)

Array Layout

- Which probe sequences are printed
- Spatial position

General considerations

- Replication / Sample size
- Randomization
- Blocking

Aspects of Experimental Design Applied to Microarrays (II)

Allocation of samples to slides

- A vs B: Treatment vs control
- Multiple treatments
- Factorial
- Time series

Other considerations

- Physical limitations: number of slides and amount of material
- Extensibility – linking

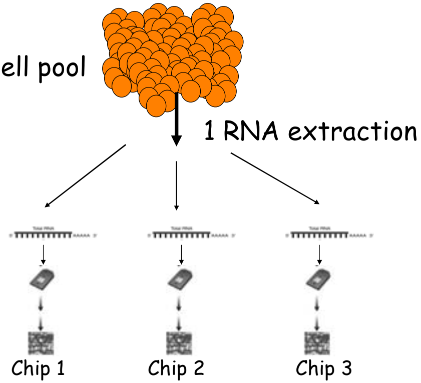
Sample Size

- More difficult than usual, as there are 1,000s of possible changes, each with its own SD
  - *Variance* of individual measurements (X)
  - *Effect size(s)* to be detected (X)
  - Acceptable *false positive rate*
  - Desired *power* (probability of detecting an effect of at least the specified size)
- *Q:* How many replicates do I need?
- *A:* As many as you can afford! (Well, almost)

Replication

- Why?
  - To reduce variability
  - To increase generalizability
- What is it?
  - Duplicate spots
  - Duplicate slides
    - *Technical replicates* – usually less desirable
    - *Biological replicates*

Triplicates preparation:

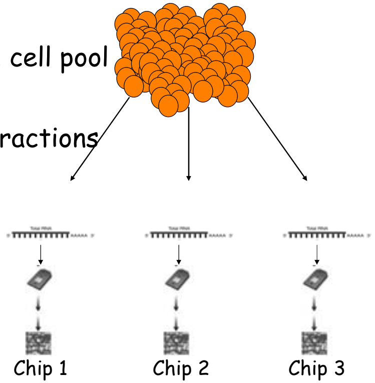


1 cell pool

1 RNA extraction

Chip 1 Chip 2 Chip 3

Triplicates preparation:

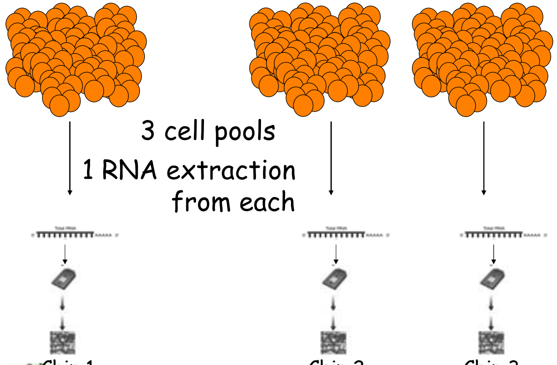


1 cell pool

3 RNA extractions

Chip 1 Chip 2 Chip 3

Triplicates preparation:



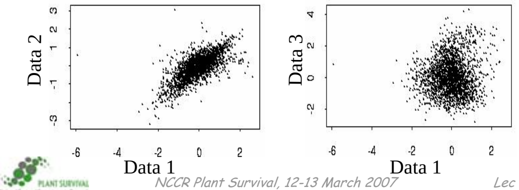
3 cell pools

1 RNA extraction from each

Chip 1 Chip 2 Chip 3

Technical Replicates: Labeling

- 3 sets of self – self hybridizations
- Data 1 and Data 2 were *labeled together* and hybridized on two slides separately
- Data 3 were labeled separately



Randomization and Blocking

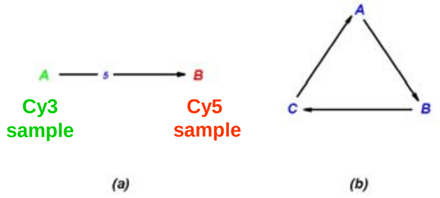
- Usually more of an issue in larger experiments (done with many samples, by different technicians, over a long period of time, ...)
- Randomization – to remove bias
- Blocking – to reduce unwanted variation
- 'Block what you can, randomize what you cannot'

Allocating samples

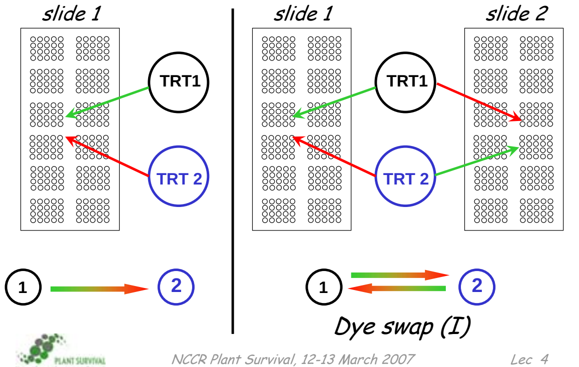
- The main issue with 2-color arrays is the use of *reference samples* (typically labeled green)
- Standard statistical design principles can lead to more efficient layouts
- Use of *dye-swaps* for some types of experiments can also help

Graphical representation

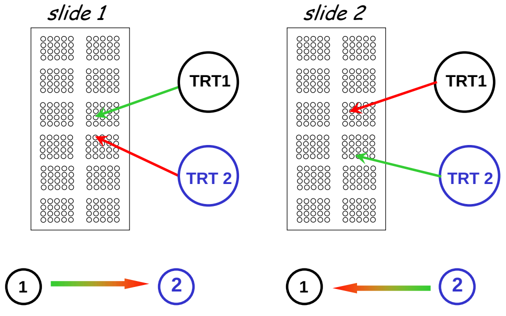
Vertices: mRNA samples
Edges: hybridization
Direction: dye assignment



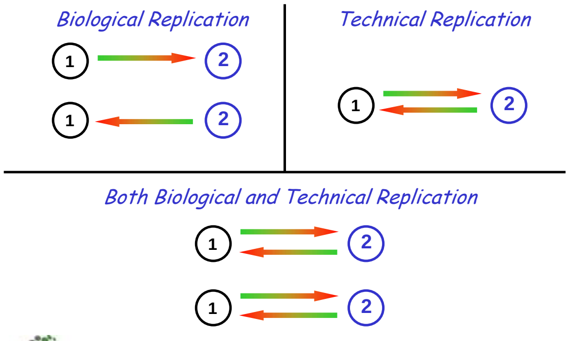
Array – graph correspondence



A different dye swap design (II)



Biological vs. technical replicates



Comparing samples

- The *structure* of the graph determines which effects can be estimated and the *precision* of the estimates
- Two mRNA samples can be compared only if there is a *path* joining the corresponding two vertices
- The precision of the estimated contrast then depends on the *number of paths* joining the two vertices and is inversely related to the *length of the paths*

Direct vs. indirect comparisons

- A comparison is *direct* when the two samples are co-hybridized to the *same* slide
- *Indirect* comparisons are those between samples on *different* slides
- The precision of the estimated effect depends on the *number of paths* joining the two vertices and is inversely related to the *length of the paths*
- Since the path between vertices is shorter for direct than indirect comparisons, direct comparisons should be *more precise*

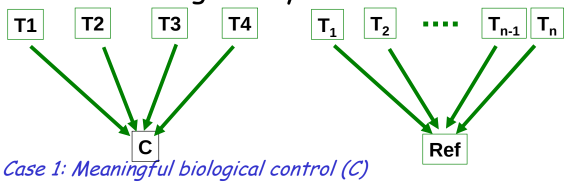
Making design decisions

- In addition to experimental constraints, design decisions should be guided by knowledge of which effects are of *most interest*
- Experimenter must decide which comparisons require the most precision
- These comparisons should be made *within slides* to the extent possible
- Direct comparisons are often *more complicated* to design and analyze, but are readily handled in a *linear modeling framework*

Some common experiments

- Comparison of *2 conditions*/types ('treatment vs. control')
  - mutant vs. wild type plants
  - liver vs. heart in mouse
- Comparison of *many treatments* to a control
- *Clinical studies* (*e.g.* cancer patients)
- *Time course* – measurements at different times
- *Factorial study* – multiple conditions varied and studied *simultaneously*

Indirect designs may be a natural choice



T1 T2 T3 T4 → C

T_1 T_2 .... T_{n-1} T_n → Ref

Case 1: Meaningful biological control (C)

Samples: Liver tissue from four mice treated by a drug.

Question 1: Which genes respond differently between T and C?

Question 2: Which genes respond similarly across two or more treatments relative to control?

Case 2: Use of universal reference (Ref)

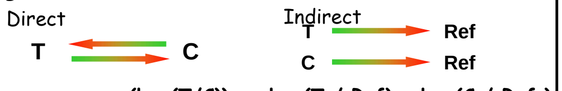
Samples: Different tumor samples.

Question: Can we discover tumor subtypes?

Treatment vs Control

Two samples

e.g. KO vs. WT or mutant vs. WT



Direct: T ⇄ C

Indirect: T → Ref; C → Ref

average (log (T/C))

log (T / Ref) – log (C / Ref)

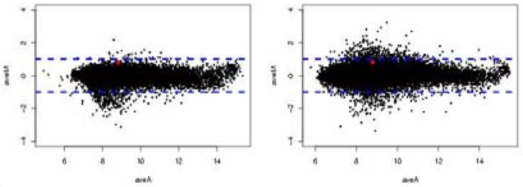
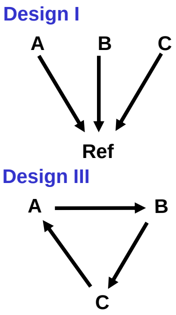
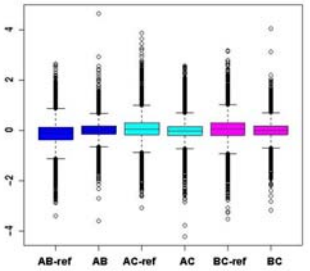


Illustration from one experiment





Box plots of log ratios: direct still ahead

Direct vs. indirect comparison

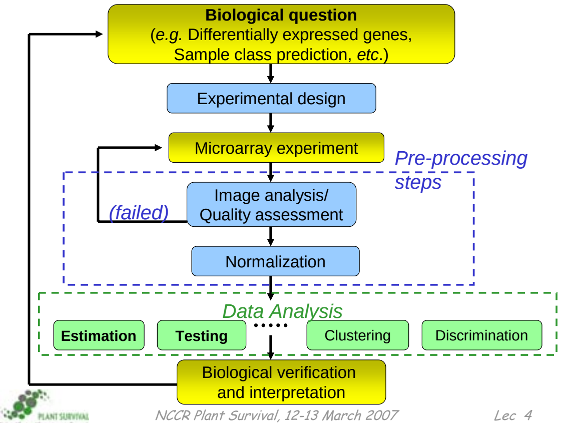
- Direct comparisons – those made *within slides* - yield more precise estimates than indirect ones between slides

Extensibility

- 'Universal' *common reference* for arbitrary undetermined number of (future) experiments
- Provides *extensibility* of the series of experiments (within and between labs)
- *Linking experiments* necessary if common reference source diminished/depleted

Summary

- Balance of *direct* and *indirect* comparisons
- Optimize precision of the estimates among comparisons of interest
- Must satisfy *scientific and physical constraints* of the experiment
- It can save you a lot of *time*, *money* and *heart-ache* to consult with an experienced analyst on design issues ***before any steps of the experiment have been carried out***



Gene expression data

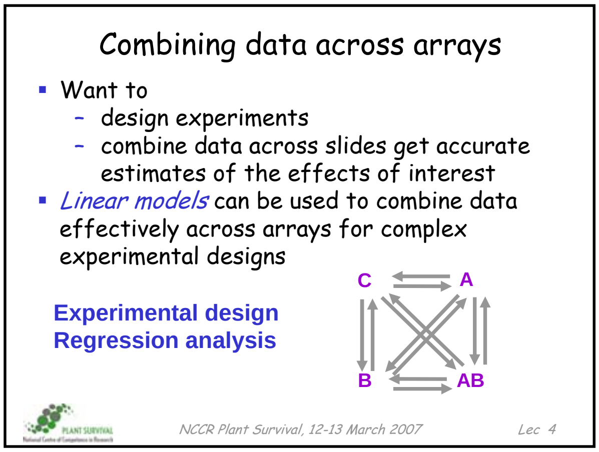
Data on *p* genes for *n* samples:

mRNA samples

Genes	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

Gene expression level of gene *i* in mRNA sample *j*

= (normalized) Log_2(Red intensity / Green intensity)

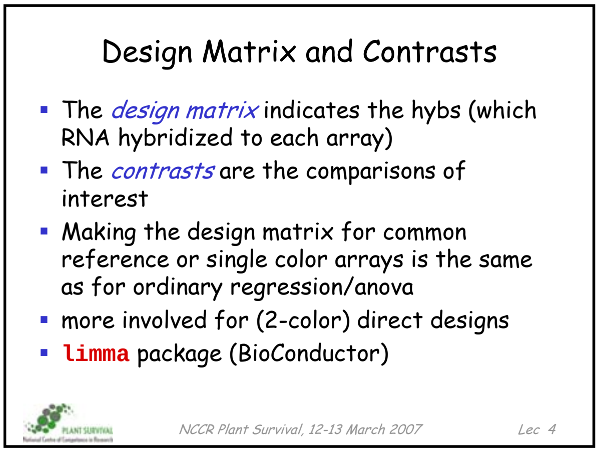


Combining data across arrays

- Want to
  - design experiments
  - combine data across slides get accurate estimates of the effects of interest
- *Linear models* can be used to combine data effectively across arrays for complex experimental designs

Experimental design
Regression analysis

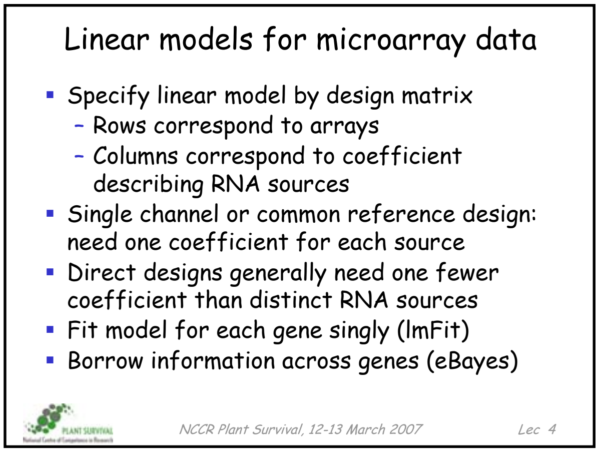
NCCR Plant Survival, 12-13 March 2007 Lec 4



Design Matrix and Contrasts

- The *design matrix* indicates the hybs (which RNA hybridized to each array)
- The *contrasts* are the comparisons of interest
- Making the design matrix for common reference or single color arrays is the same as for ordinary regression/anova
- more involved for (2-color) direct designs
- `limma` package (BioConductor)

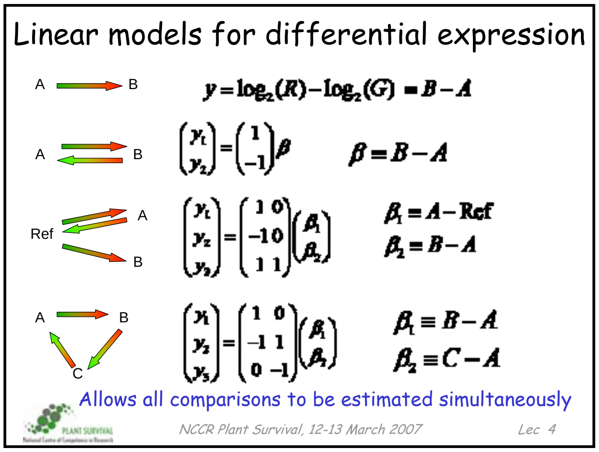
NCCR Plant Survival, 12-13 March 2007 Lec 4



Linear models for microarray data

- Specify linear model by design matrix
  - Rows correspond to arrays
  - Columns correspond to coefficient describing RNA sources
- Single channel or common reference design: need one coefficient for each source
- Direct designs generally need one fewer coefficient than distinct RNA sources
- Fit model for each gene singly (lmFit)
- Borrow information across genes (eBayes)

NCCR Plant Survival, 12-13 March 2007 Lec 4



Linear models for differential expression

$$y = \log_2(R) - \log_2(G) = B - A$$

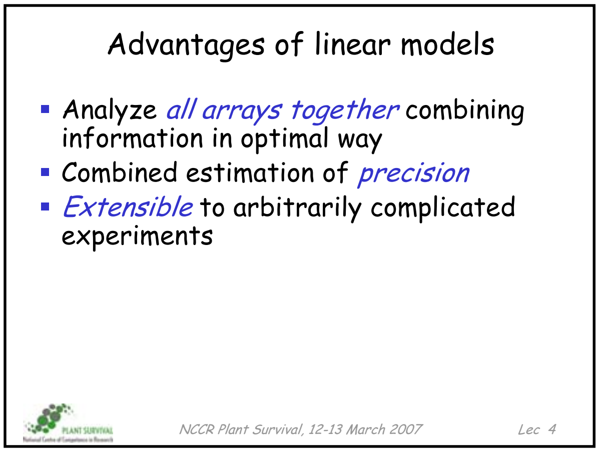
$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \beta \qquad \beta = B - A$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \qquad \begin{aligned} \beta_1 &= A - \text{Ref} \\ \beta_2 &= B - A \end{aligned}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \qquad \begin{aligned} \beta_1 &\equiv B - A \\ \beta_2 &\equiv C - A \end{aligned}$$

Allows all comparisons to be estimated simultaneously

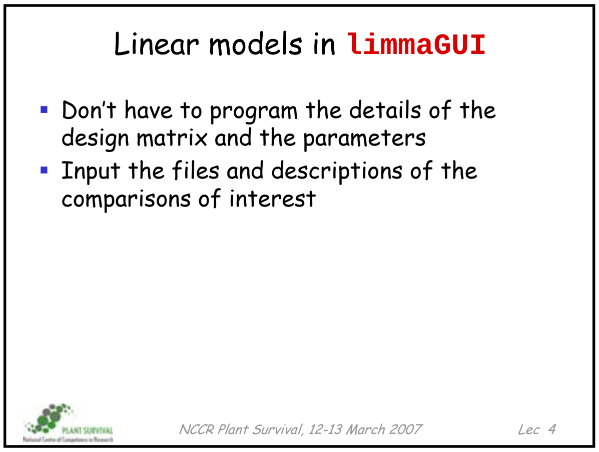
NCCR Plant Survival, 12-13 March 2007 Lec 4



Advantages of linear models

- Analyze *all arrays together* combining information in optimal way
- Combined estimation of *precision*
- *Extensible* to arbitrarily complicated experiments

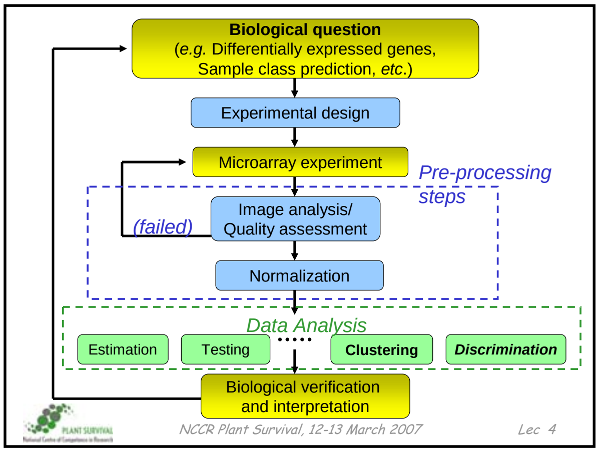
NCCR Plant Survival, 12-13 March 2007 Lec 4



Linear models in `limmaGUI`

- Don't have to program the details of the design matrix and the parameters
- Input the files and descriptions of the comparisons of interest

NCCR Plant Survival, 12-13 March 2007 Lec 4



Biological question
(*e.g.* Differentially expressed genes, Sample class prediction, *etc.*)

Experimental design

Microarray experiment

Pre-processing steps

(failed)

Image analysis/ Quality assessment

Normalization

Data Analysis

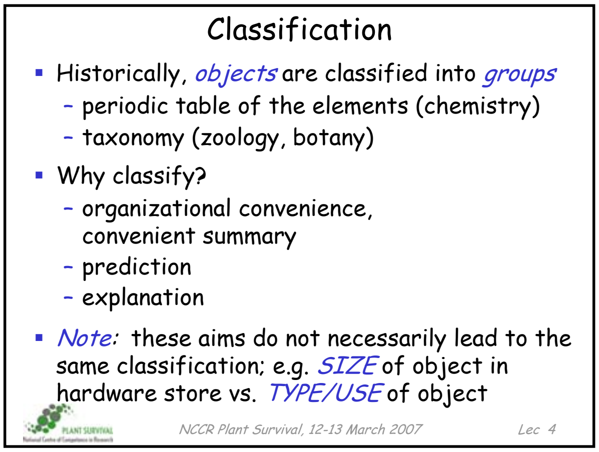
Estimation

Testing

Clustering

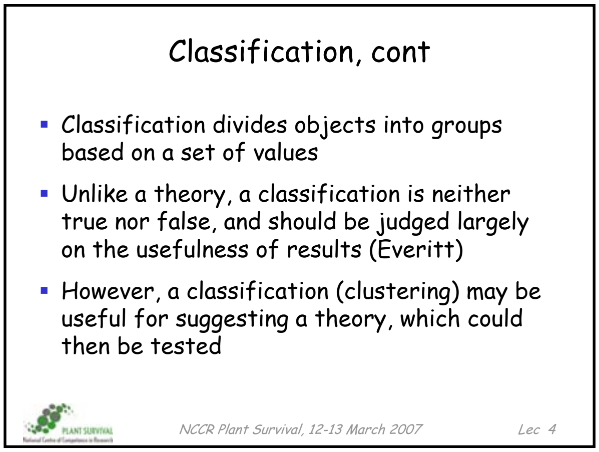
Discrimination

Biological verification and interpretation



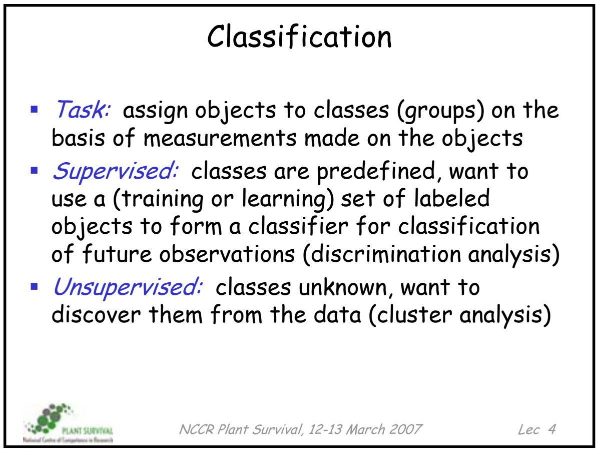
Classification

- 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



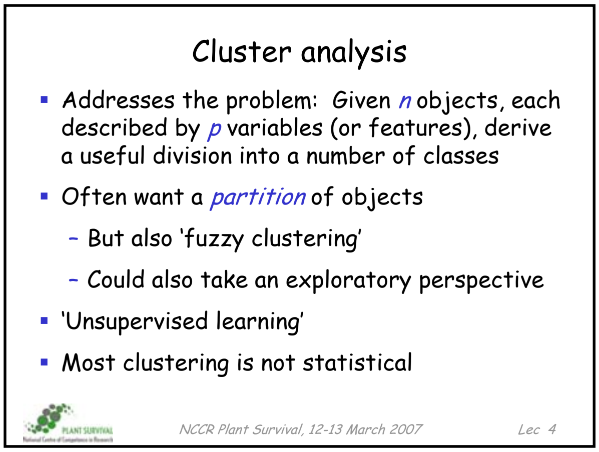
Classification, cont

- 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



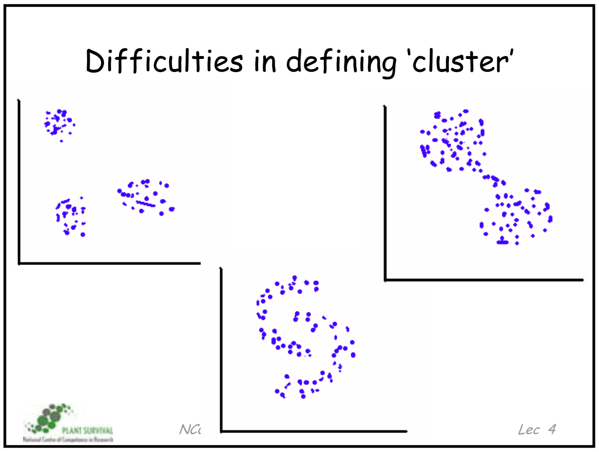
Classification

- *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

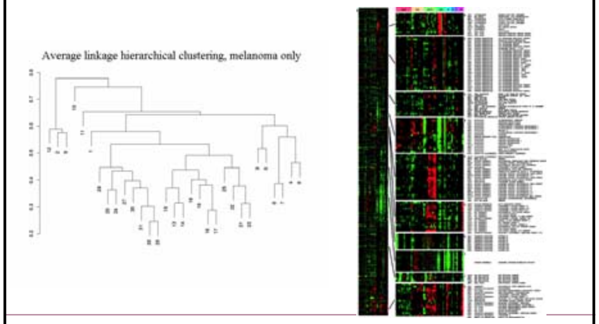
Clustering Gene Expression Data

- 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
  - available as R function `heatmap()`
  - http://rana.lbl.gov/EisenSoftware.htm
- 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

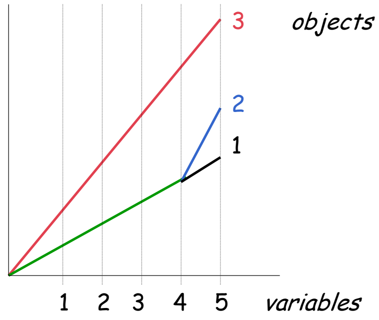


Eisen, Michael B. et al. (1998)
Proc. Natl. Acad. Sci. USA 95, 14863-14868

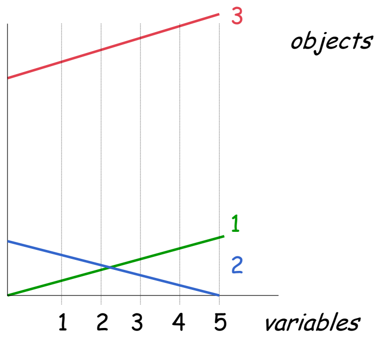
Similarity

- *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 correlation-based similarity is quite common in gene expression studies but is in general contentious...

Problems using correlation



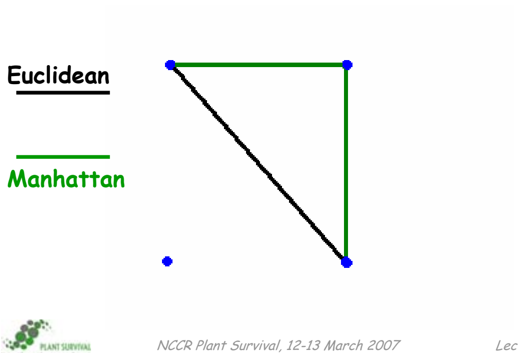
A more extreme example



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



What distance should I use?

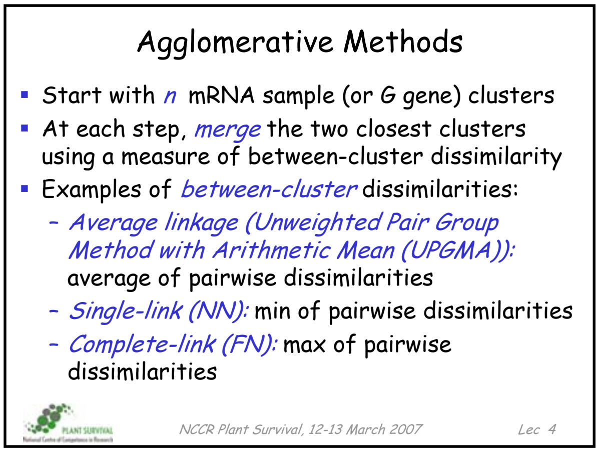
- This is like asking: *What tool should I buy?*
- It depends on what similarities you are interested in finding
- With Euclidean distance, larger values will tend to dominate; not useful if large value is simply a result of using smaller units (*e.g.*, grams vs Kilos)
- Can get around this (if desired)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

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)
- Examples: k-means, self-organizing maps (SOM), partitioning around medoids (PAM; more robust and computationally efficient than k-means), model-based clustering

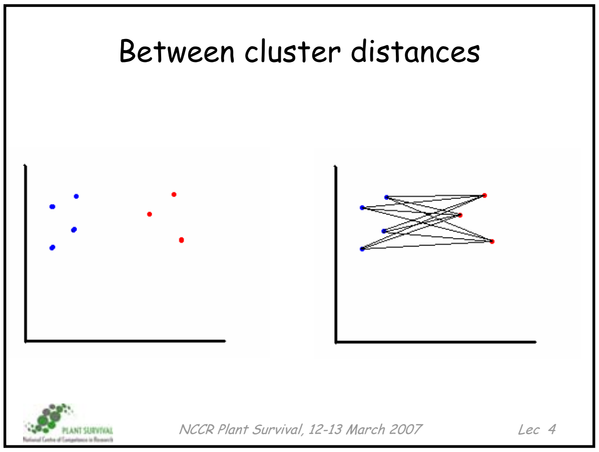
Hierarchical Clustering

- 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

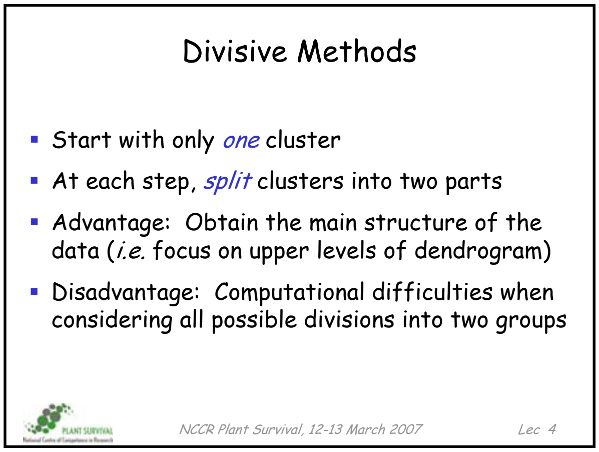


Agglomerative Methods

- 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

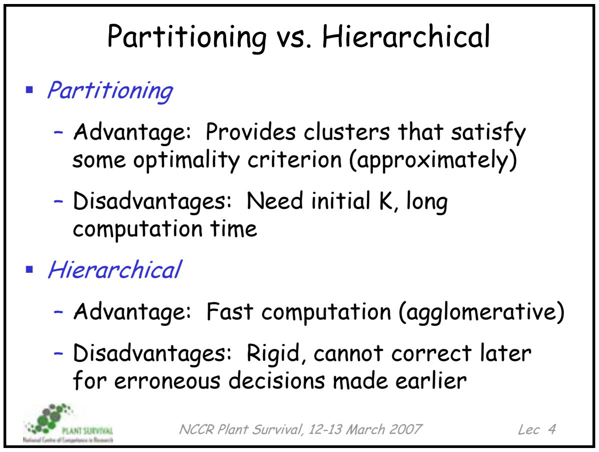


Between cluster distances



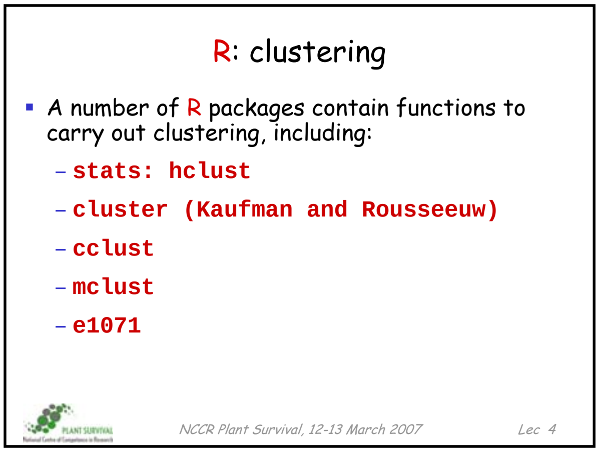
Divisive Methods

- 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



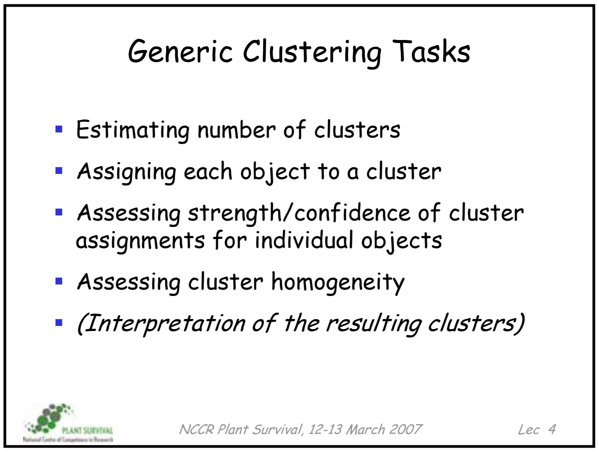
Partitioning vs. Hierarchical

- *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)`
  - `cclust`
  - `mclust`
  - `e1071`



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

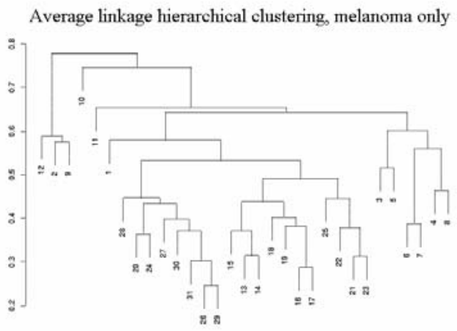
Additional methods

- 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)

Example: Bittner et al.

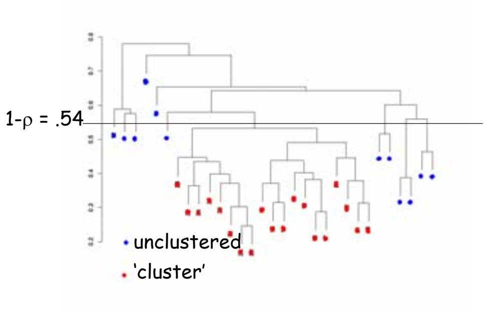
It has been proposed (by many) that a *cancer taxonomy* can be identified from *gene expression experiments*.

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



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*

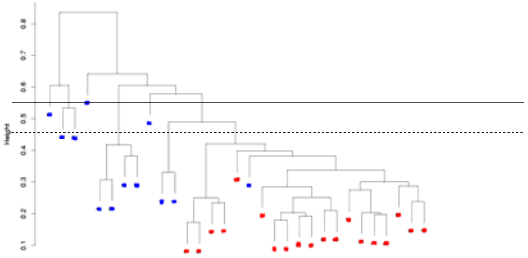
Issues in Clustering

- 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, ...

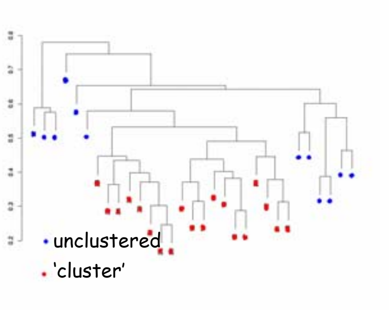
Average linkage, top 300 genes in SD



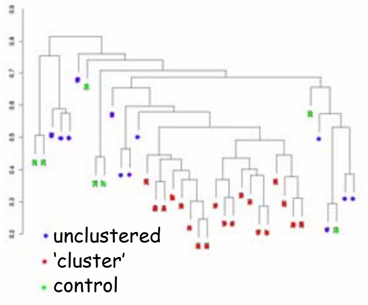
Issues in Clustering

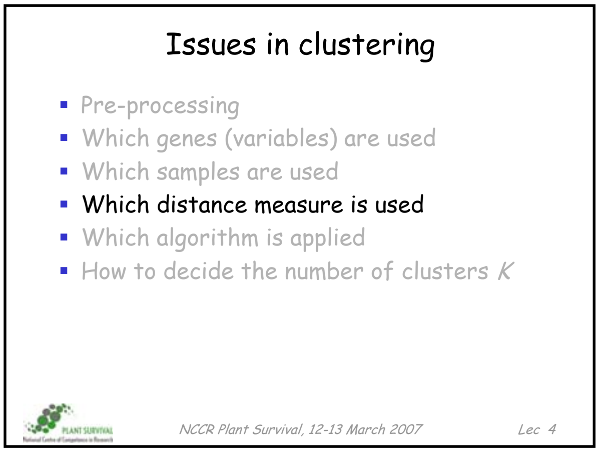
- Pre-processing (Image analysis and Normalization)
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

Average linkage, *melanoma only*



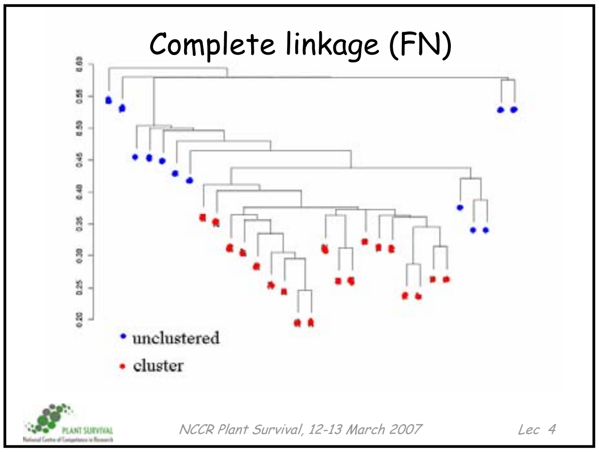
Average linkage, *melanoma & controls*



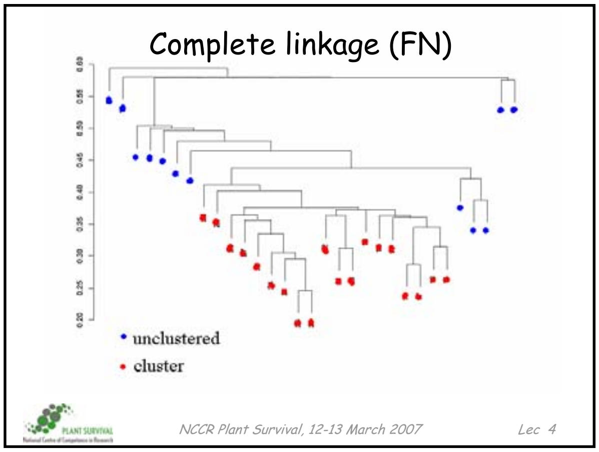


Issues in clustering

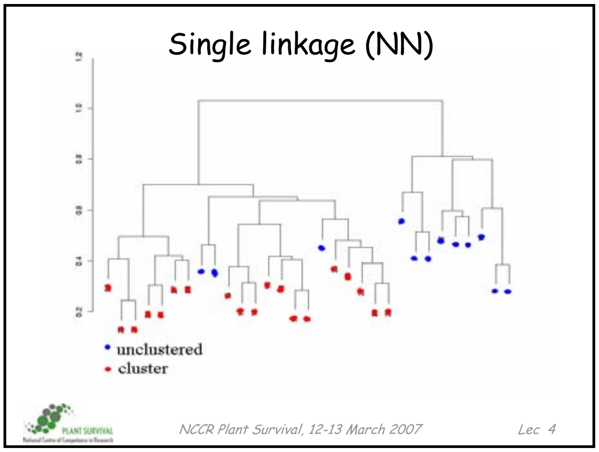
- Pre-processing
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*



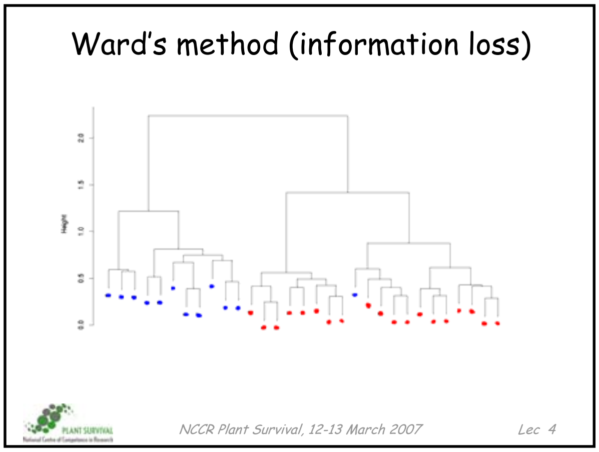
Complete linkage (FN)



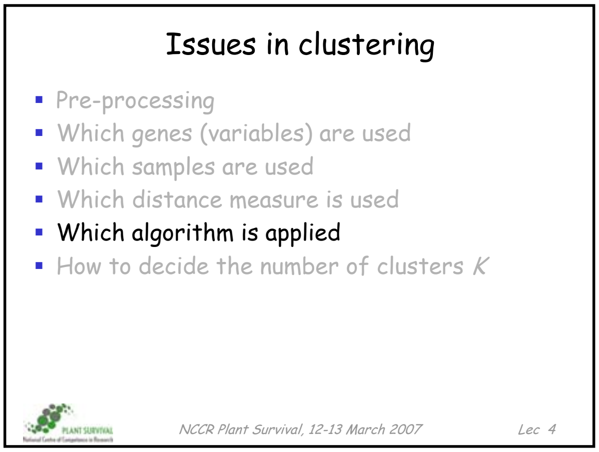
Complete linkage (FN)



Single linkage (NN)



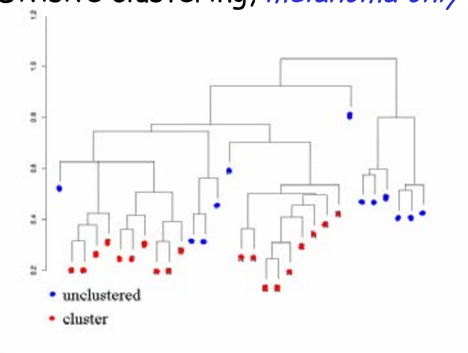
Ward's method (information loss)



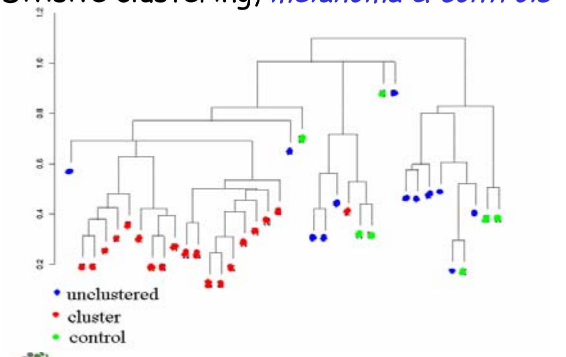
Issues in clustering

- 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*



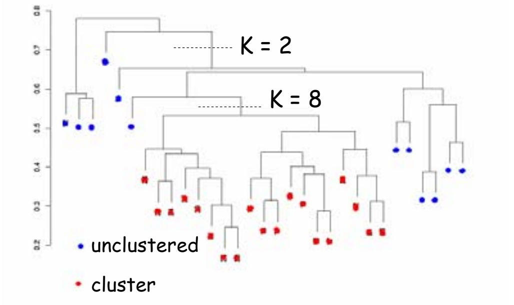
Issues in clustering

- 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



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