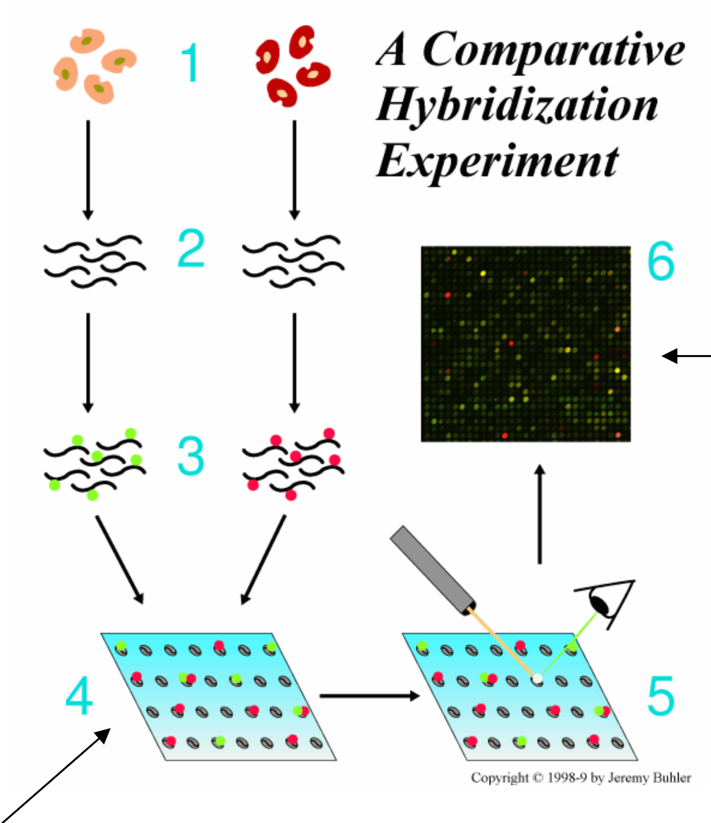
Analysis of Microarray Data

- Analysis of images
- Preprocessing of gene expression data
- Normalization of data
  - Subtraction of Background Noise
  - Global/local Normalization
  - House keeping genes (or same gene)
  - Expression in ratio (test/references) in log
- Differential Gene expression
  - Repeats and calculate significance (t-test)
  - Significance of fold used statistical method
- Clustering
  - Supervised/Unsupervised (Hierarchical, K-means, SOM)
- Prediction or Supervised Machine Learnning (SVM)

Technical

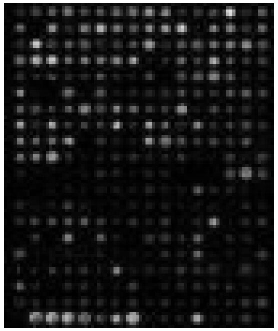

1
A Comparative Hybridization Experiment
2
3
4
5
6
Copyright © 1998-9 by Jeremy Buhler

sample
(labelled)

probe
(on chip)

pseudo-colour
image

Images from scanner



- Resolution
  - standard 10µm [currently, max 5µm]
  - 100µm spot on chip = 10 pixels in diameter

- Image format
  - TIFF (tagged image file format) 16 bit (65’536 levels of grey)
  - 1cm x 1cm image at 16 bit = 2Mb (uncompressed)
  - other formats exist e.g.. SCN (used at Stanford University)

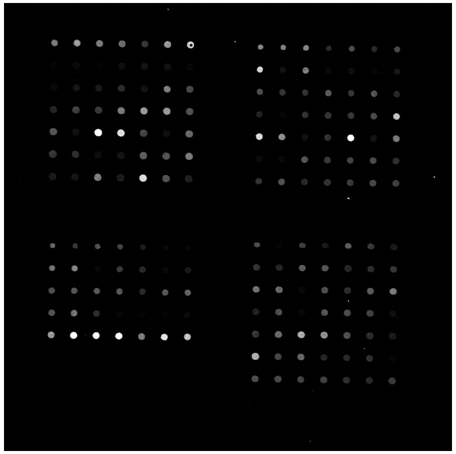
- Separate image for each fluorescent sample
  - channel 1, channel 2, etc.

Images in analysis software

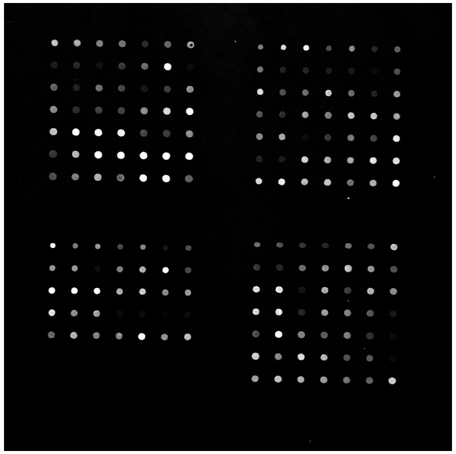
- The two 16-bit images (Cy3, Cy5) are compressed into 8-bit images
- Display fluorescence intensities for both wavelengths using a 24-bit RGB overlay image
- RGB image :
  - Blue values (B) are set to 0
  - Red values (R) are used for Cy5 intensities
  - Green values (G) are used for Cy3 intensities
- Qualitative representation of results

Images : examples

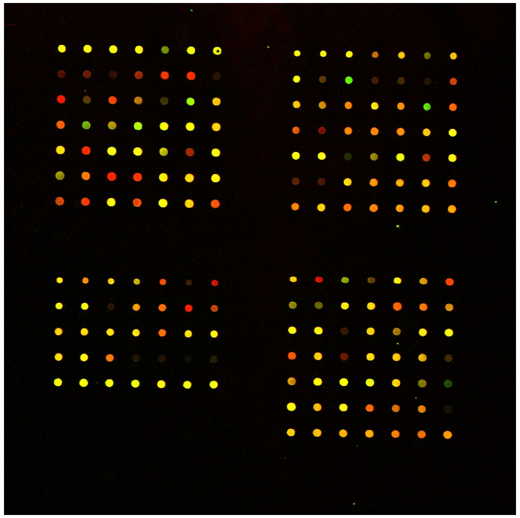
Cy3



Cy5



Pseudo-colour overlay



Spot colour	*Signal strength*	*Gene expression*
yellow	Control = perturbed	unchanged
red	Control < perturbed	induced
green	Control > perturbed	repressed

Processing of images

- Addressing or gridding
  - Assigning coordinates to each of the spots
- Segmentation
  - Classification of pixels either as foreground or as background
- Intensity determination for each spot
  - Foreground fluorescence intensity pairs (R, G)
  - Background intensities
  - Quality measures

Background intensity

- Spot’s measured intensity includes a contribution of non-specific hybridization and other chemicals on the glass

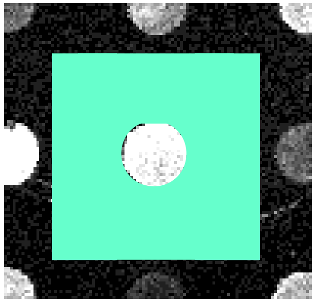
- Fluorescence from regions not occupied by DNA should by different from regions occupied by DNA
  -> one solution is to use local negative controls (spotted DNA that should not hybridize)

- Different background methods :
  - Local background
  - Morphological opening
  - Constant background
  - No adjustment

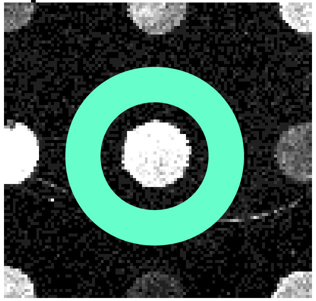
Local background

- Focusing on small regions surrounding the spot mask.
- Median of pixel values in this region

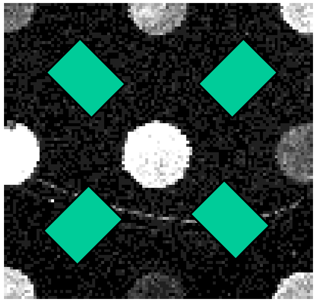
- Most software package implement such an approach



ScanAlyze



ImaGene

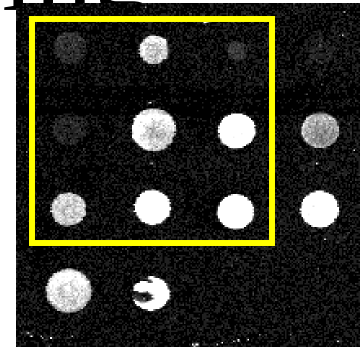


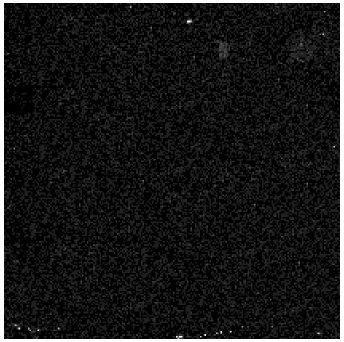
Spot, GenePix

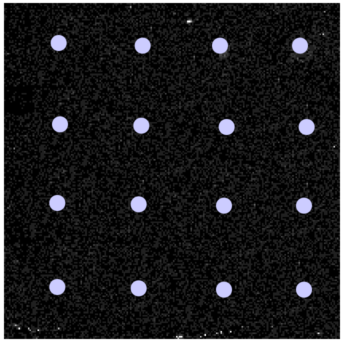
  - By not considering the pixels immediately surrounding the spots, the background estimate is less sensitive to the performance of the segmentation procedure

Morphological opening

- Non-linear filtering, used in Spot
- Use a square structuring element with side length at least twice as large as the spot separation distance
- Compute local minimum filter, then compute local maximum filter
  - This removes all the spots and generates an image that is an estimate of the background for the entire slide
- For individual spots, the background is estimated by sampling this background image at the nominal center of the spot
- Lower background estimate and less variable







Constant background

- Global method which subtracts a constant background for all spots

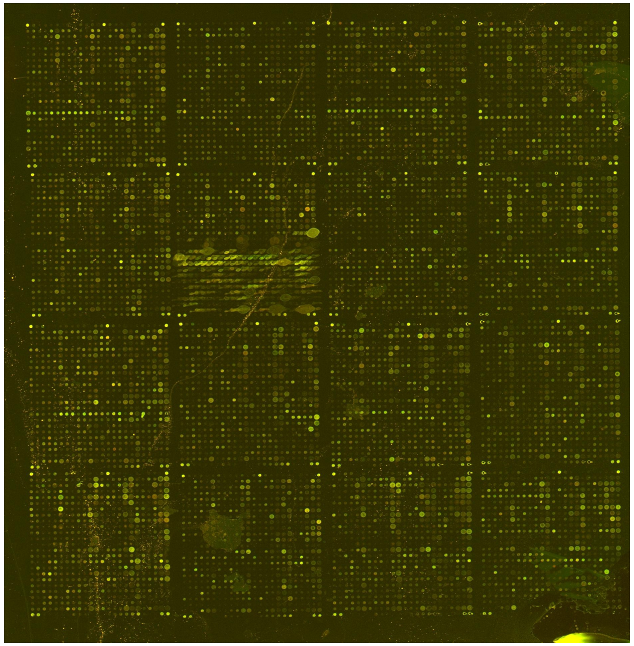
- Some evidence that the binding of fluorescent dyes to ‘negative control spots’ is lower than the binding to the glass slide

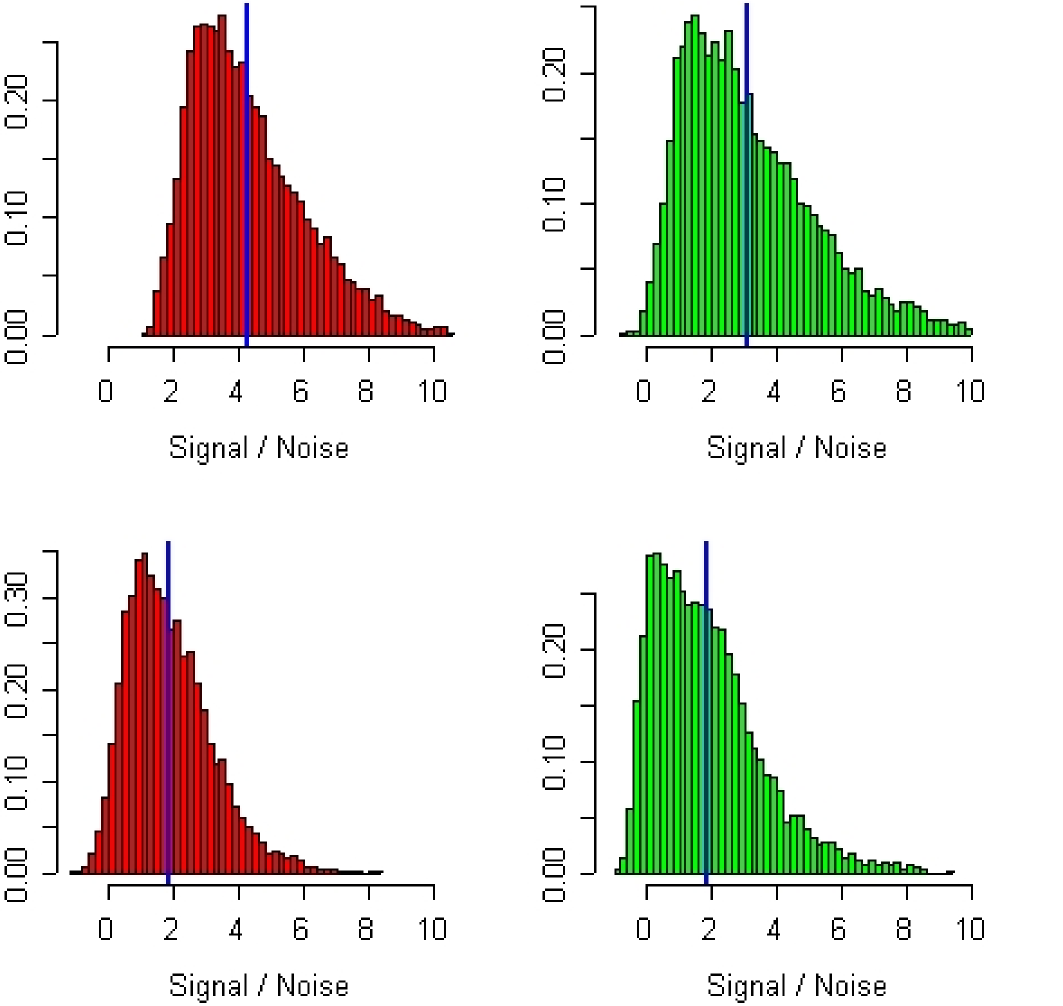
- -> More meaningful to estimate background based on a set of negative control spots
  - If no negative control spots :
    approximation of the average background =
    third percentile of all the spot foreground values

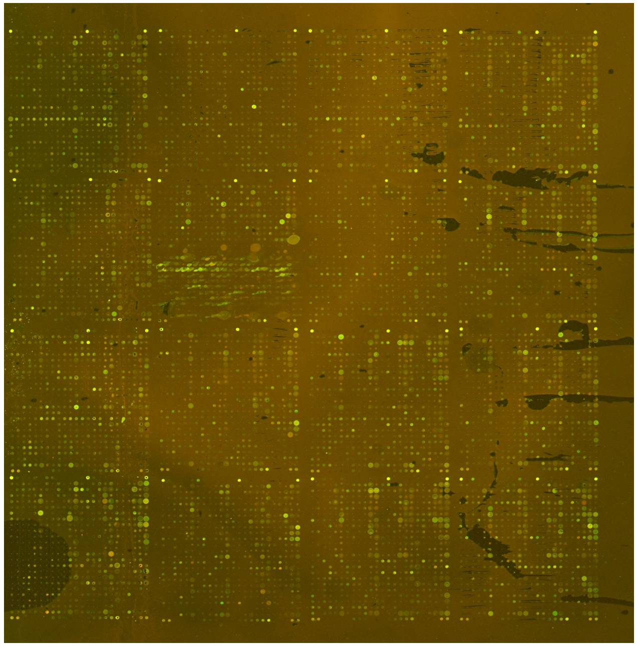
No background adjustment

- Do not consider the background
  - Probably not accurate, but may be better than some forms of local background determination!

Histograms







Signal/Noise = $\log_2$(spot intensity/background intensity)

Preprocessing of Gene expression Data

- Scale transformation
  - CY3/CY5
  - LOG(CY3/CY5)
- Replicates handling
  - Inconsistent replicate removal
  - Replicate merging
- Missing value handling
  - Removal of patterns having excess of missing values
  - Value of missing points
- Flat pattern filtering
- Unknown Gene Removing

Preprocessing: Normalization

- Why?

  To correct for systematic differences between samples on the same slide, or between slides, which do not represent true biological variation between samples.

- How do we know it is necessary?

  By examining self-self hybridizations, where no true differential expression is occurring.

  We find dye biases which vary with overall spot intensity, location on the array, plate origin, pins, scanning parameters,….

Normalization Techniques

- Global normalization
  - Divide channel value by means
- Control spots
  - Common spots in both channels
  - House keeping genes
  - Ratio of intensity of same gene in two channel is used for correction
- Iterative linear regression
- Parametric nonlinear nomalization
  - log(CY3/CY5) vs log(CY5))
  - Fitted log ratio – observed log ratio
- General Non Linear Normalization
  - LOESS
  - curve between log(R/G) vs log(sqrt(R.G))

Pre-processed cDNA Gene Expression Data

On p genes for n slides: p is O(10,000), n is O(10-100), but growing,

Slides

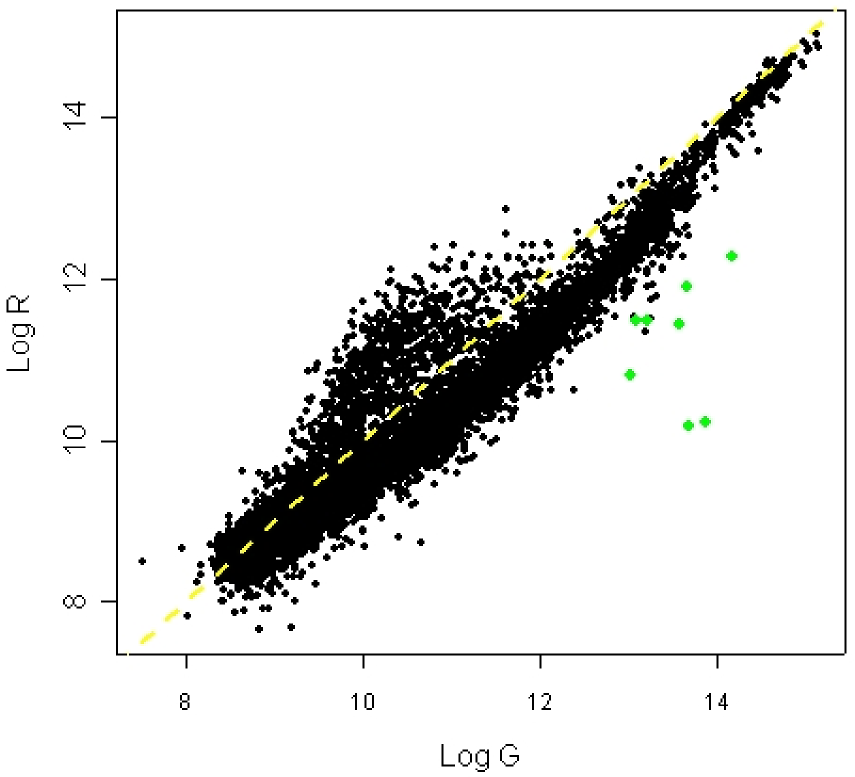
		slide 1	**slide 2**	**slide 3**	**slide 4**	**slide 5**	**…**
	1	0.46	0.30	0.80	1.51	0.90	...
	2	-0.10	0.49	0.24	0.06	0.46	...
Genes	3	0.15	0.74	0.04	0.10	0.20	...
	4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	5	-0.06	1.06	1.35	1.09	-1.09	...

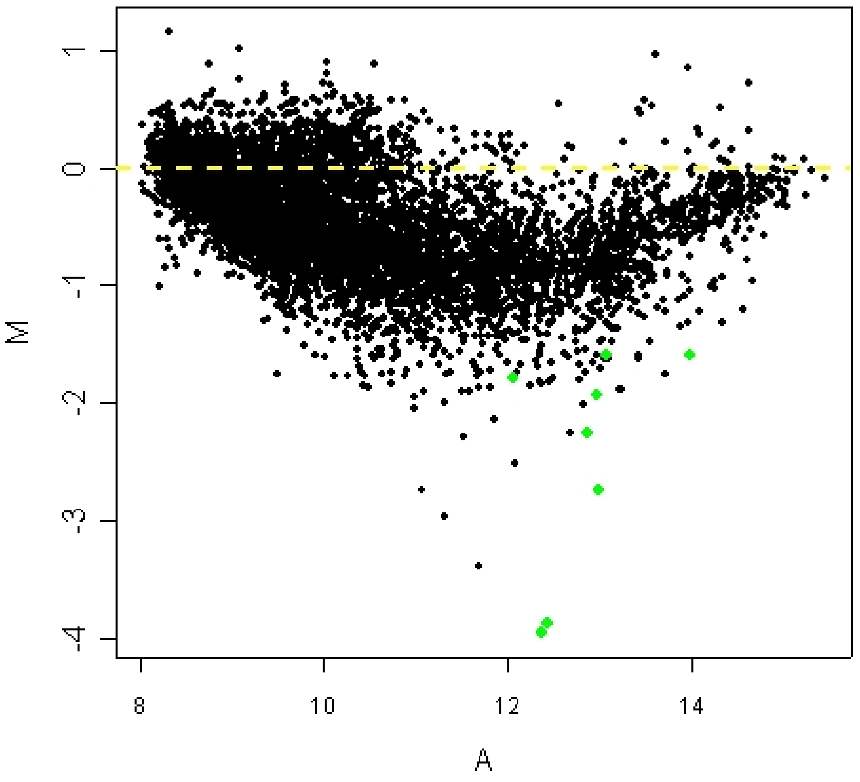
Gene expression level of gene *5* in slide 4

= **Log$_2$(Red intensity / Green intensity)**

These values are conventionally displayed on a red (>0) yellow (0) green (<0) scale.

Scatterplots: always log, always rotate





log$_2$R vs log$_2$G

$M=\log_2 R/G$ vs $A=\log_2\sqrt{RG}$

Classification

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

Cluster analysis

- Used to find groups of objects when not already known
- “Unsupervised learning”
- Associated with each object is a set of measurements (the feature vector)
- Aim is to identify groups of similar objects on the basis of the observed measurements

Example: Tumor Classification

- Reliable and precise classification essential for successful cancer treatment

- Current methods for classifying human malignancies rely on a variety of morphological, clinical and molecular variables

- Uncertainties in diagnosis remain; likely that existing classes are heterogeneous

- Characterize molecular variations among tumors by monitoring gene expression (microarray)

- Hope: that microarrays will lead to more reliable tumor classification (and therefore more appropriate treatments and better outcomes)

Nearest Neighbor Classification

- Based on a measure of distance between observations (e.g. Euclidean distance or one minus correlation)
- k-nearest neighbor rule (Fix and Hodges (1951)) classifies an observation **X** as follows:
  - find the k observations in the learning set closest to **X**
  - predict the class of **X** by majority vote, i.e., choose the class that is most common among those k observations.
- The number of neighbors k can be chosen by cross-validation

Hierarchical Clustering

- Produce a dendrogram
- 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

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

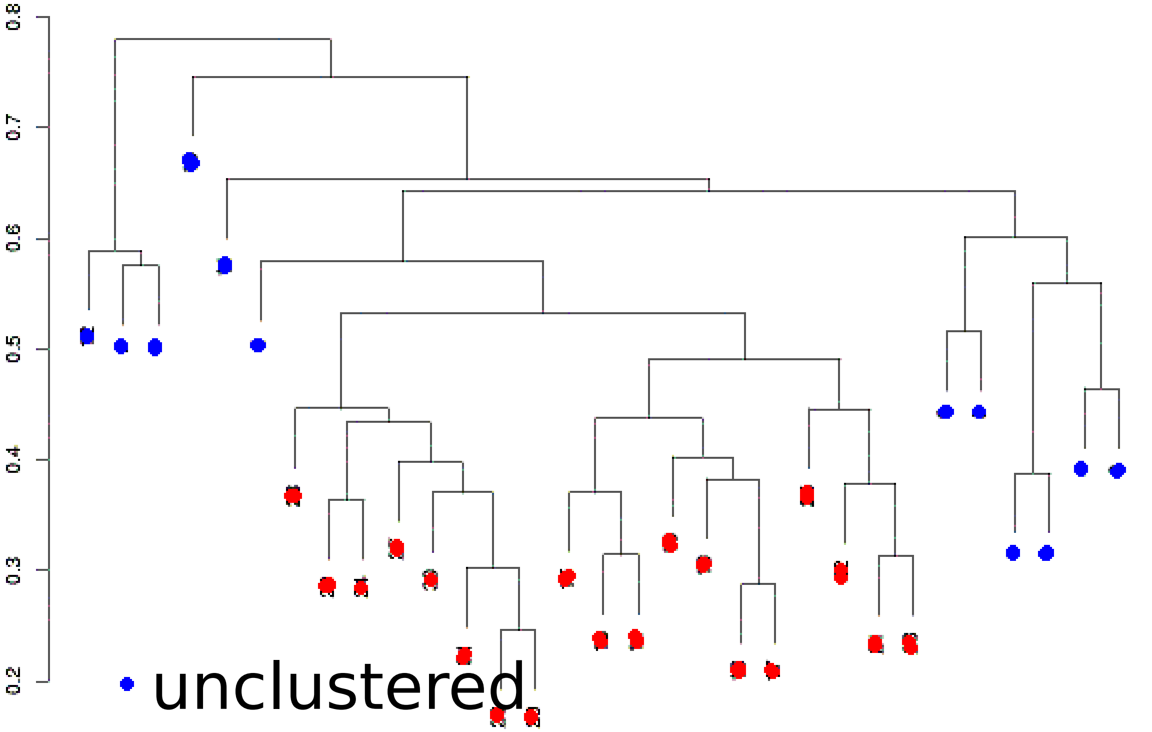
Issues in Clustering

- 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
- Filter based on statistical comparison
  - t-test
  - ANOVA
  - Cox model, etc.

Average linkage hierarchical clustering, melanoma only



● unclustered

● ‘cluster’