# Statistical Issues in the Analysis of Gene Expression Microarray Data

Lisa M. McShane

Biometric Research Branch

U.S. National Cancer Institute

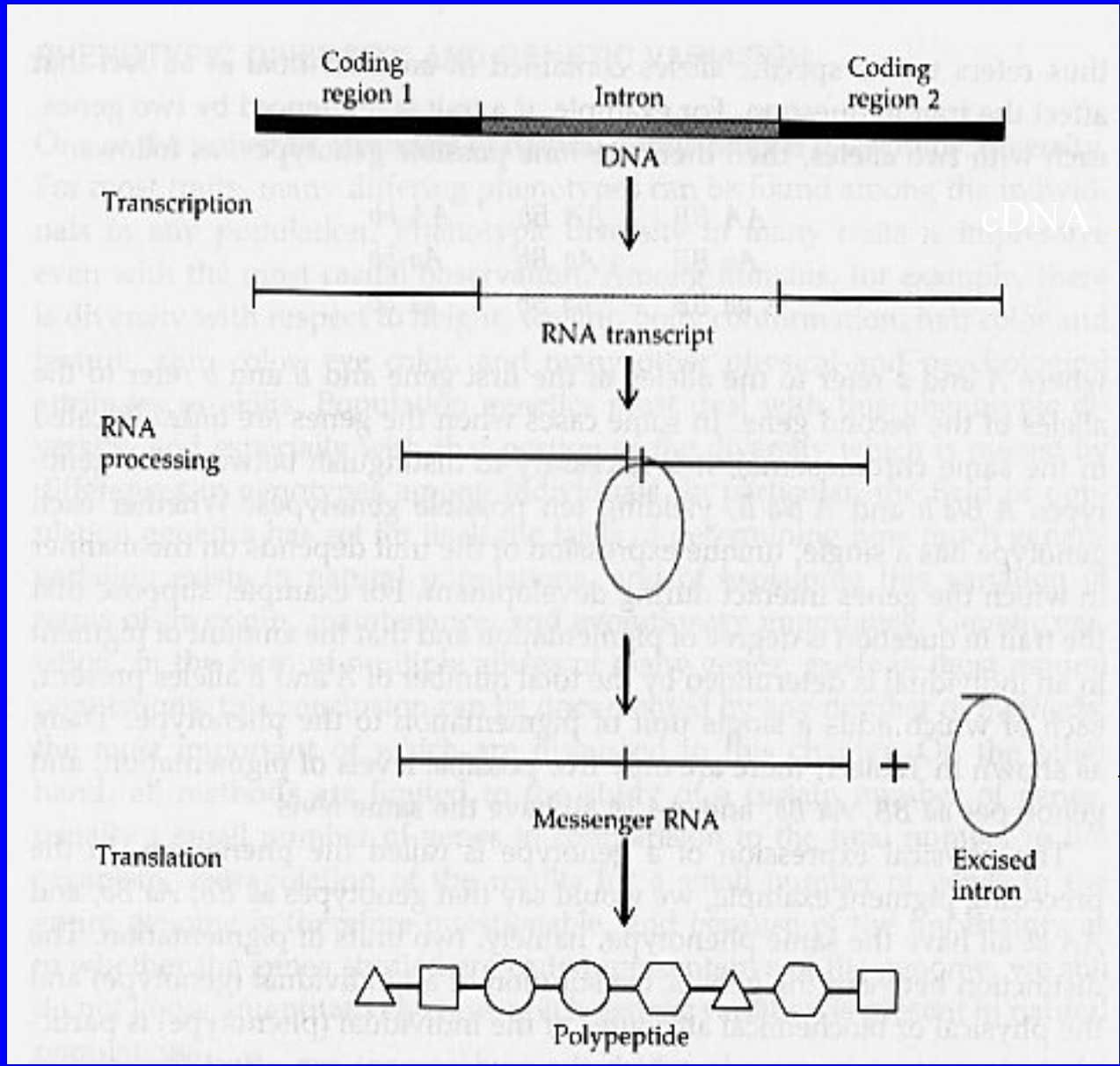
# Outline

1) Introduction: Biology & Technology
2) Data Quality, Image Processing & Normalization
3) Study Objectives & Design Considerations
4) Analysis Strategies Based on Study Objectives

# Outline

1) Introduction: Biology & Technology
2) Data Quality, Image Processing & Normalization
3) Study Objectives & Design Considerations
4) Analysis Strategies Based on Study Objectives

- All cells of a multi-cellular organism contain essentially the same DNA
- Cells differ in function based on the spectra of which genes are expressed and the level of expression
- Proteins do the work of cells and gene expression determines the intra-cellular concentration of proteins
- mRNA is an intermediate product of gene expression; a gene (DNA) is transcribed into a mRNA molecule which is then translated into a protein molecule

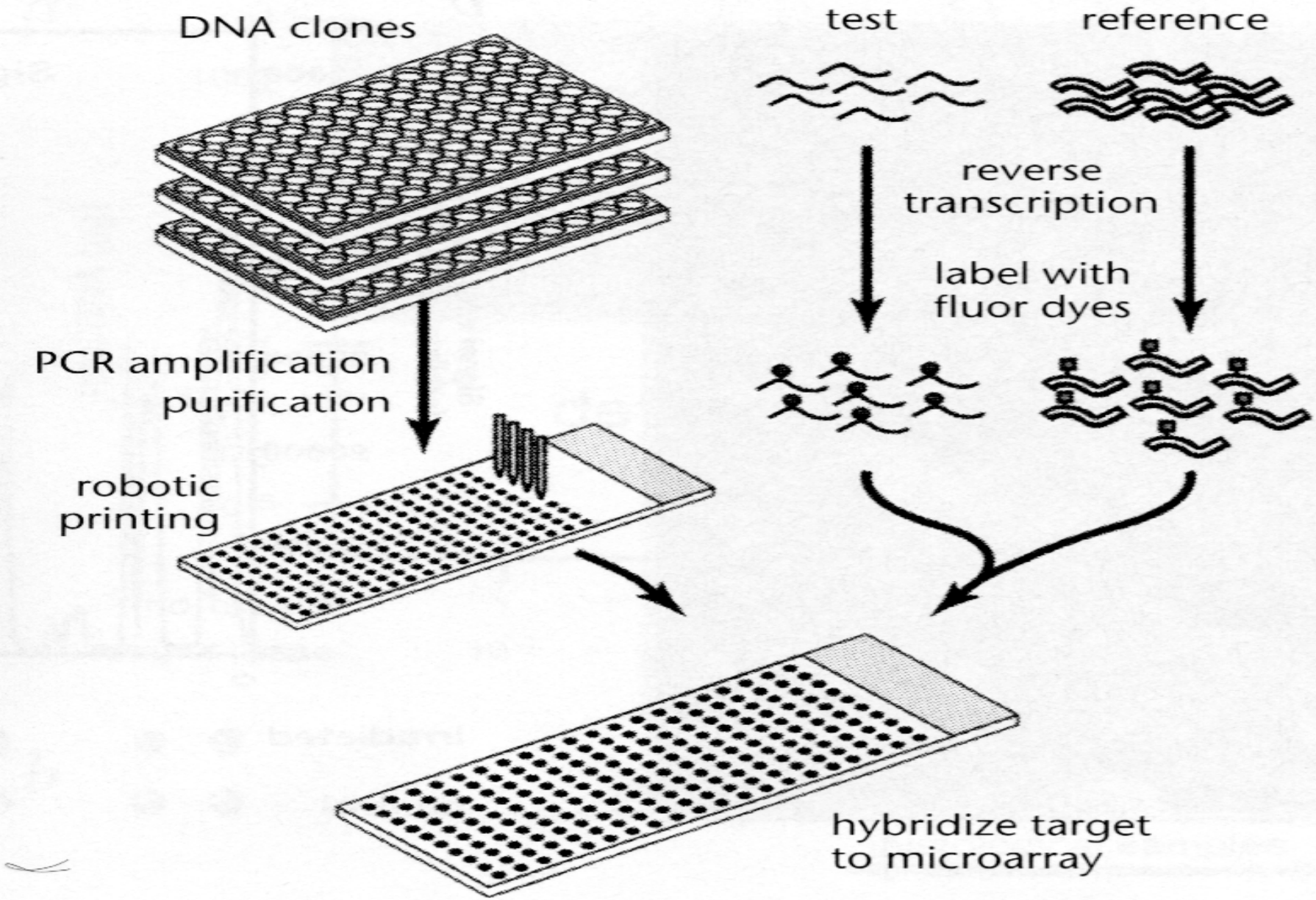


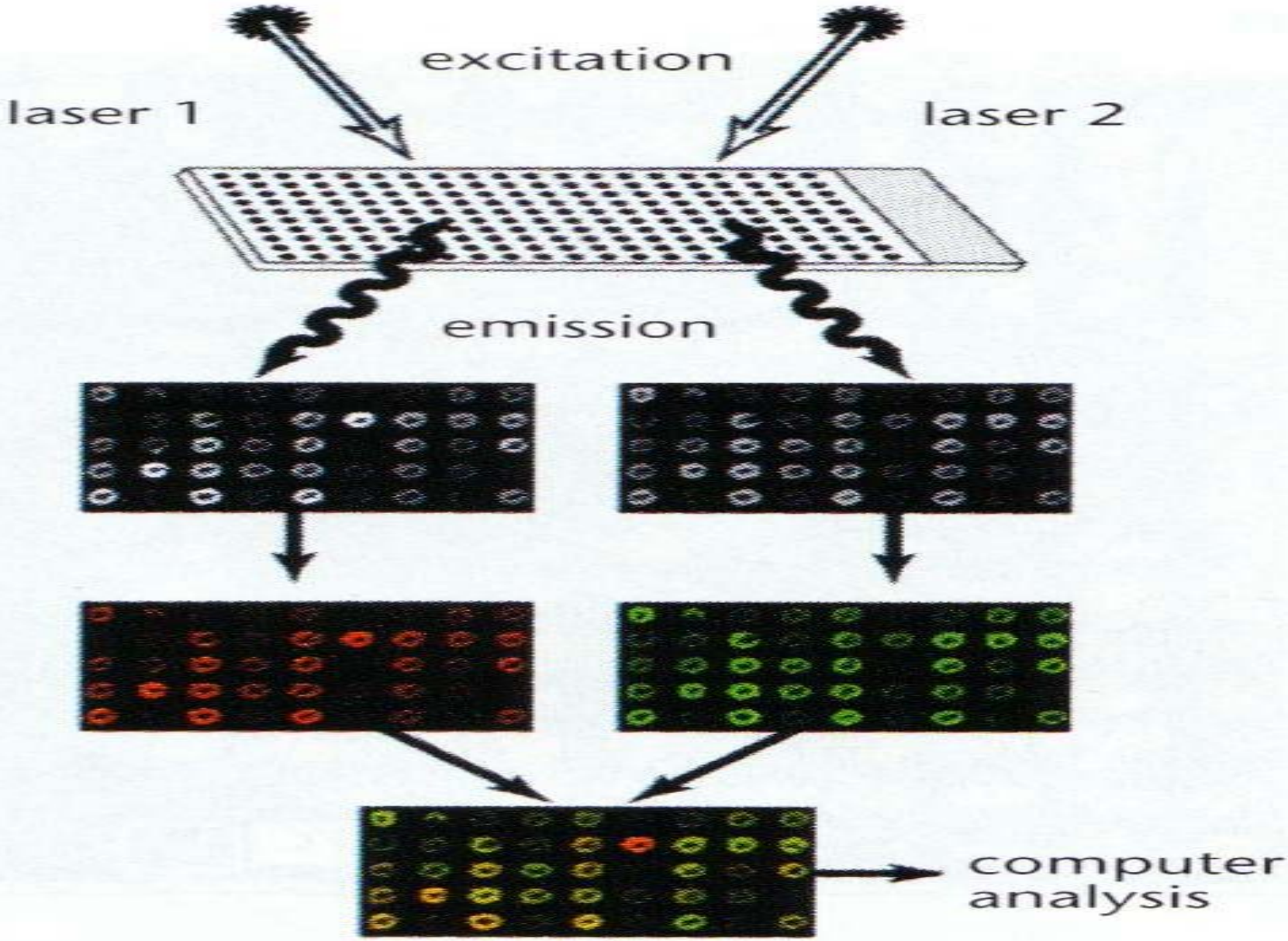

Reverse transcription

# Gene Expression Microarrays

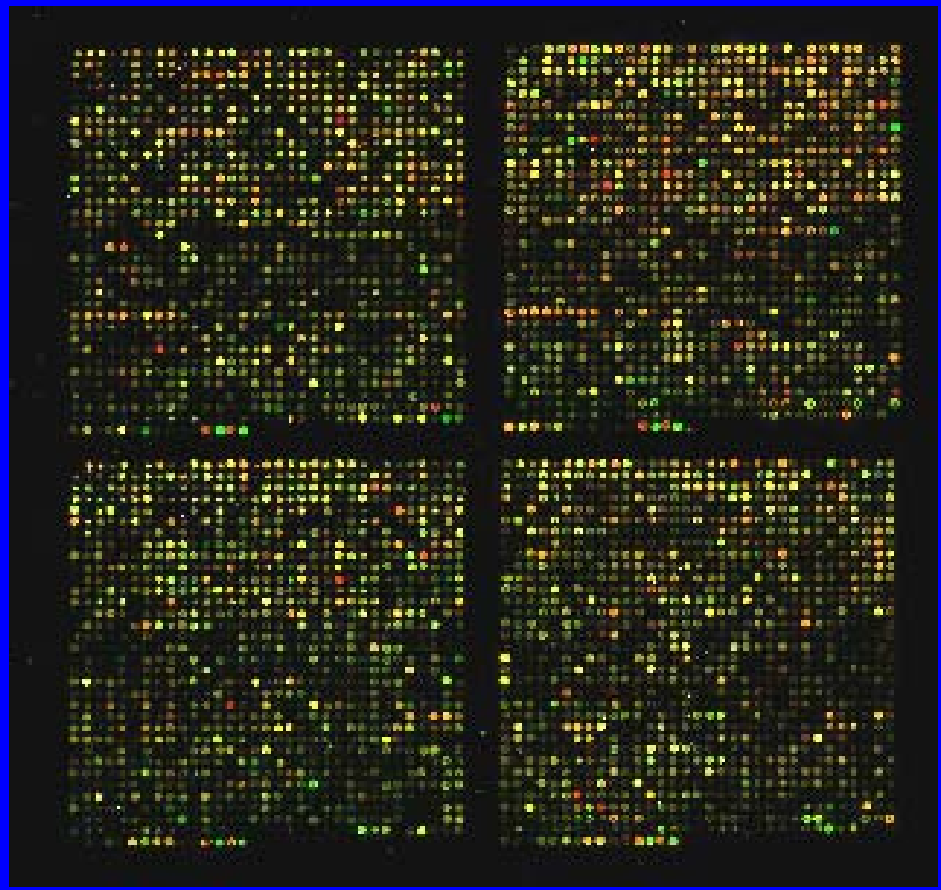
- Permit simultaneous evaluation of expression levels of thousands of genes
- Main platforms
  - cDNA arrays (glass slide, spotted)
    - Schena *et al.*, *Science,* 1995
  - Oligo arrays (glass wafer – "chip", photolithography)
    - Lockhart *et al., Nature Biotechnology,* 1996
    - Affymetrix website (http://www.affymetrix.com)
  - Spotted oligo arrays (glass slide, spotted)
  - Nylon filter arrays

# cDNA Array


DNA clones
test
reference
reverse
transcription
label with
fluor dyes
PCR amplification
purification
robotic
printing
hybridize target
to microarray

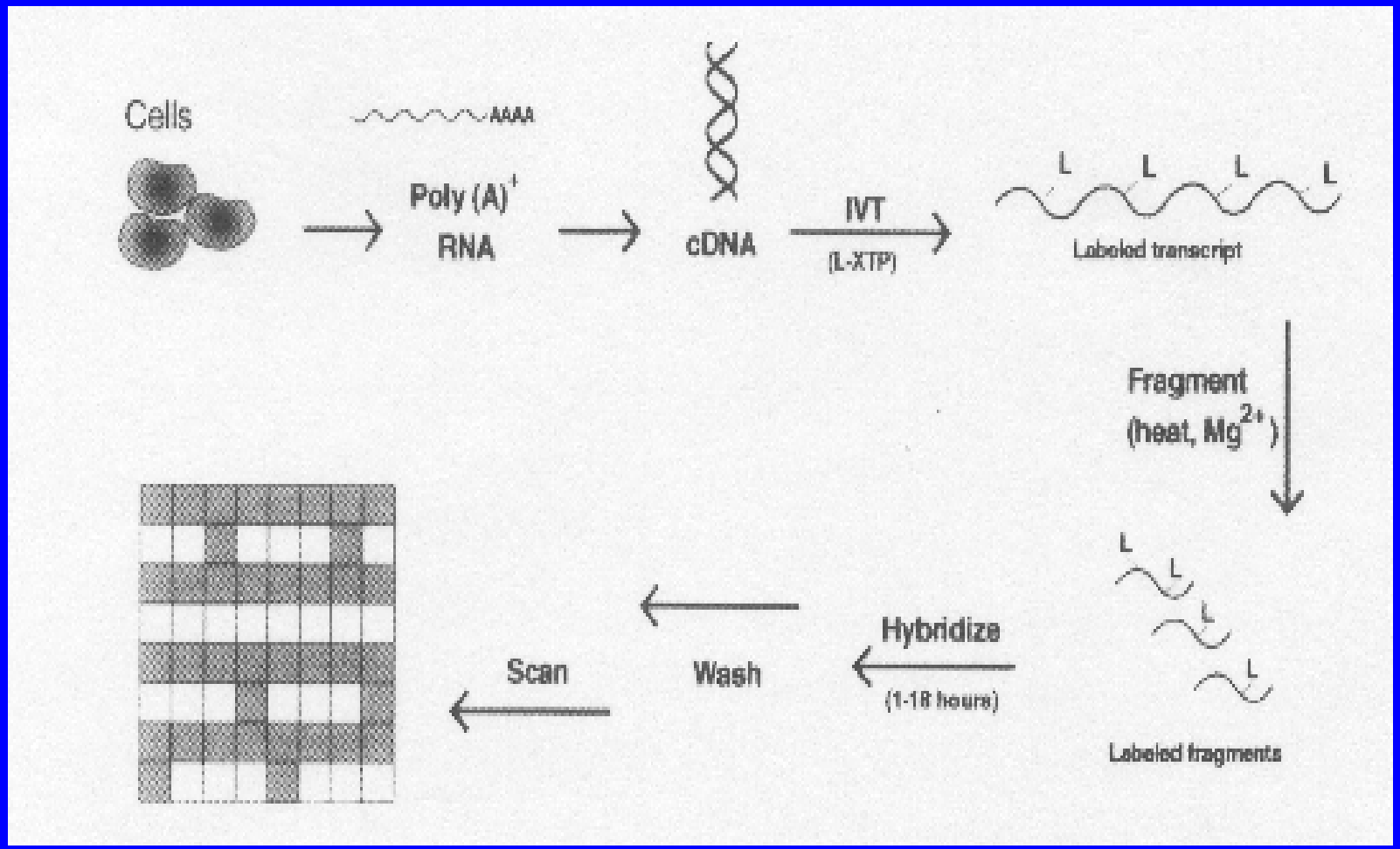

excitation
laser 1
laser 2
emission
computer
analysis

# cDNA Microarray Image
## (overlaid "red" and "green" images)



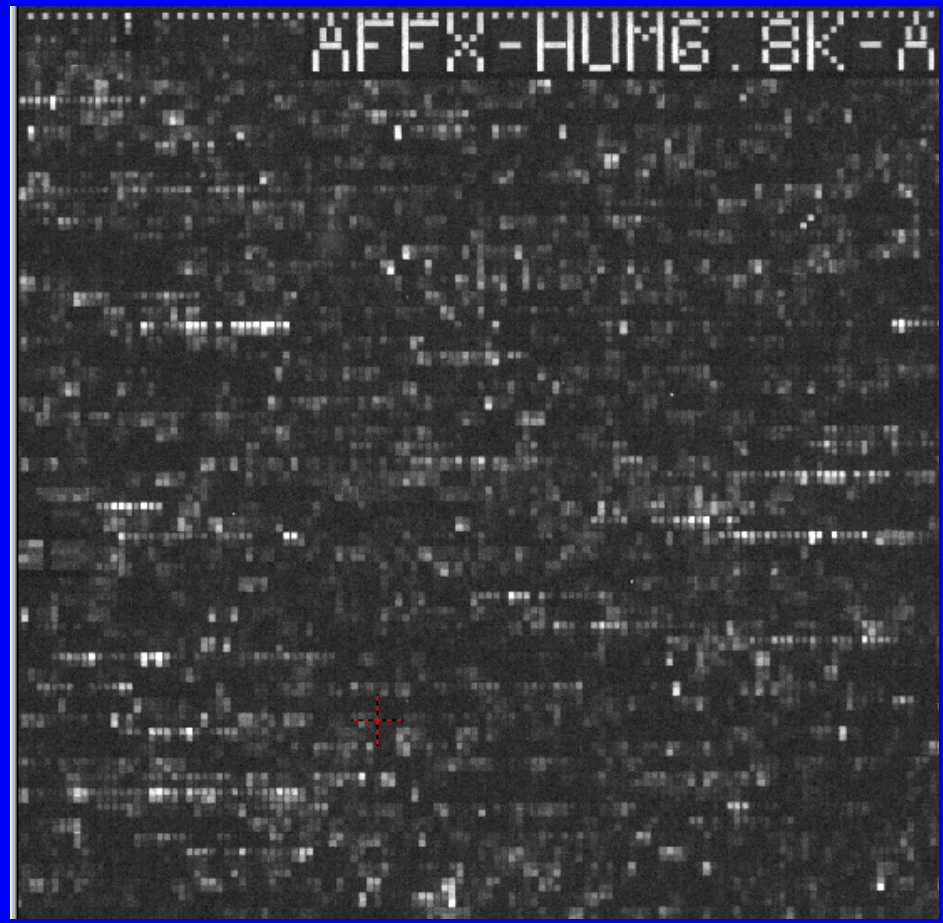
# [Affymetrix] Hybridization Oligo Array

# Oligo Array: Assay procedure

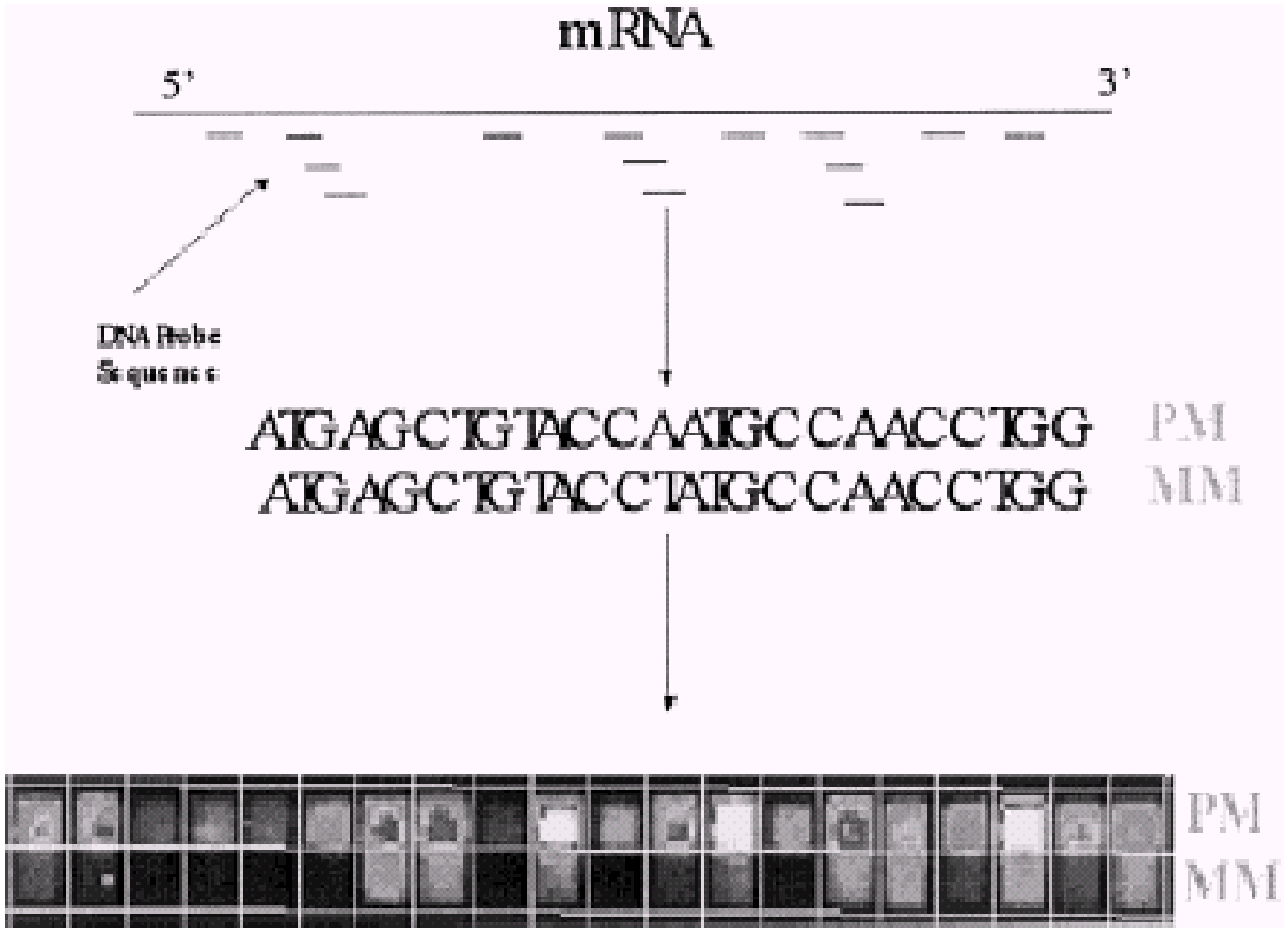



(Figure 1 from Lockhart *et al., Nature Biotechnology,* 1996)

# Image of a Scanned Affymetrix Gene Chip



# Oligo Arrays: Perfect Match - Mismatch Probe Pairs




(Figure 2 from Schadt *et al., Journal of Cellular Biochemistry,* 2000)

# Oligo Arrays

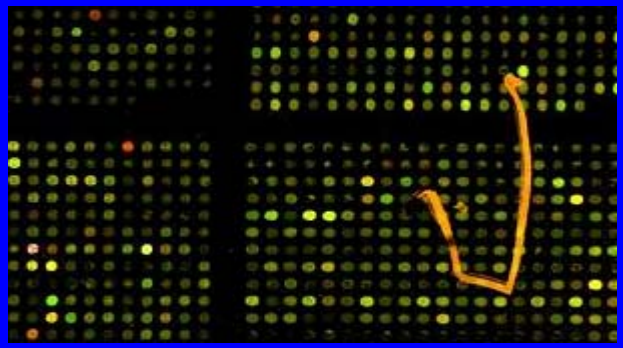
- Single sample hybridized to each array
- Each gene represented by a "probe set"
  - One probe type per array "cell"
  - Typical probe is a 25-mer oligo
  - 11-20 PM:MM pairs per probe set

    (PM = perfect match, MM = mismatch)

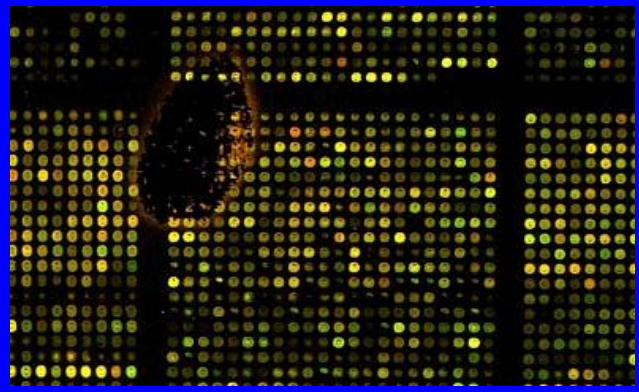
# Outline

1) Introduction: Biology & Technology
2) Data Quality, Image Processing & Normalization
3) Study Objectives & Design Considerations
4) Analysis Strategies Based on Study Objectives

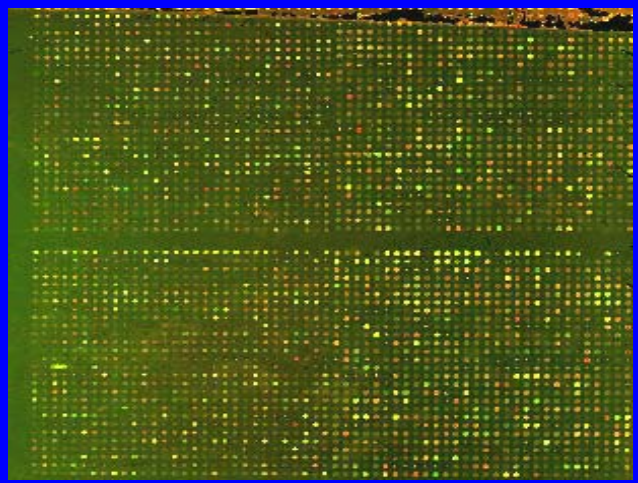
# cDNA Arrays: Slide Quality



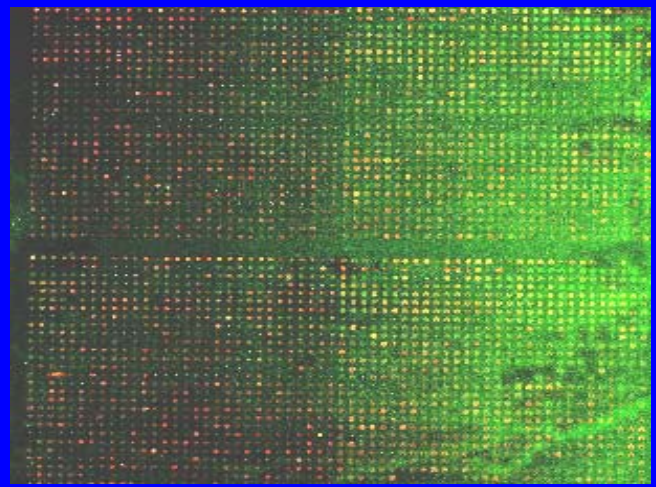
Fiber or scratch?



Bubble

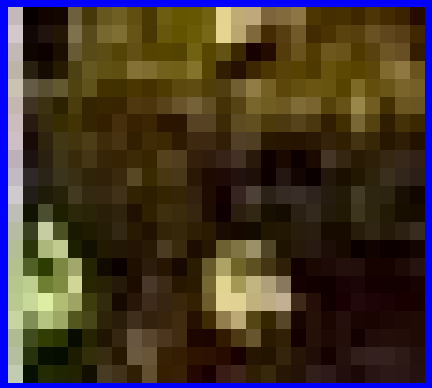


Edge effect

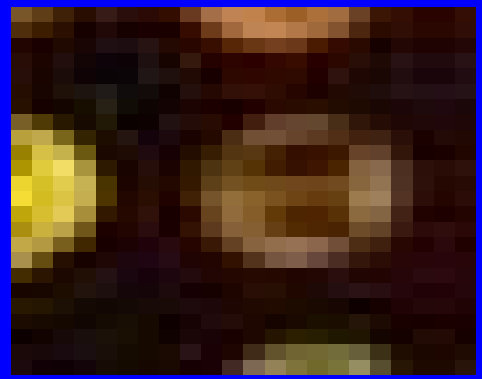


Background haze

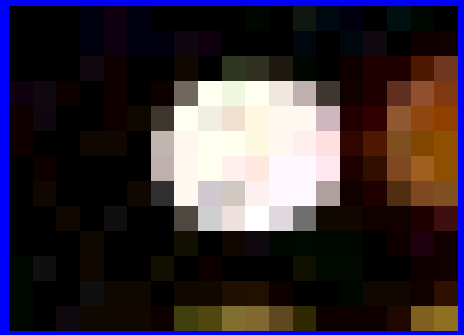
# cDNA Arrays: Spot Quality



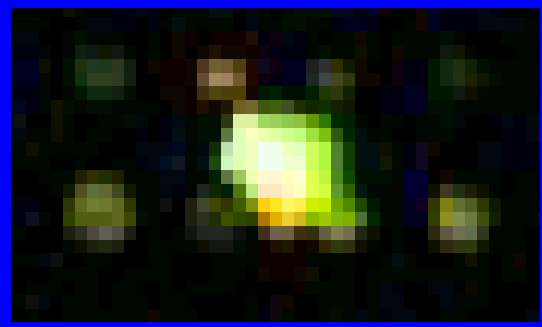
Poorly defined borders



Large holes



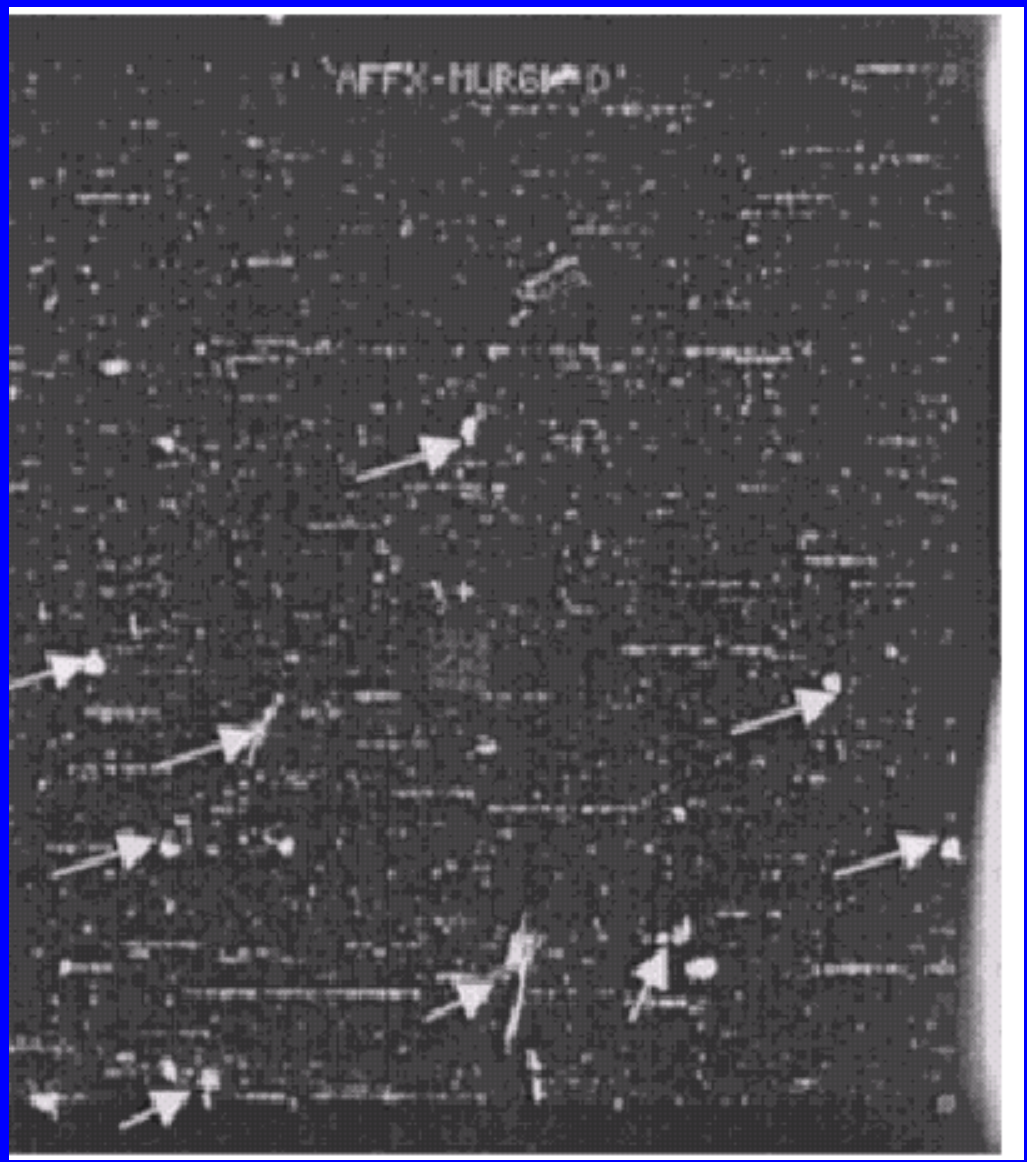
Saturated spot



Dust specs

# Oligo Arrays: Quality problems due to debris

(Figure 1 from Schadt *et al., Journal of Cellular Biochemistry,* 2000)



# Image Processing

- Start with image consisting of millions of pixels
- Spot-background segmentation or grid alignment
- Signal calculation (summarize pixels within a spot or cell, background adjust, …)
- Spot flagging criteria
- Spot or cell-level summaries
  - cDNA: red and green signals for each spot
    - Red/green ratio (common for reference design)
  - Oligo: single signal per cell

# Oligo Arrays: Probe Set (Gene) Summaries

- $AvDiff_i = \Sigma(PM_{ij}-MM_{ij})/n_i$ for each probe set i
  (original Affymetrix algorithm)
- $MBEI_i = \theta_i$ estimated from
  $PM_{ij}- MM_{ij} = \theta_i \phi_j + \varepsilon_{ij} \Rightarrow$ weighted average difference
  (Model-Based Expression Index, Li and Wong, *PNAS,* 2001)
- Other algorithms – e.g. address issues of negative or outlier differences
  - Corrected or global backgrounds, robust measures, etc.
    - "New" Affymetrix algorithm
    - Irizarry *et al.,* 2002
      (http://biosun01.biostat.jhsph.edu/~ririzarr/papers)
  - PM only (Naef *et al.* referenced in Irizarry *et al.,* 2002)

# Data Normalization

- Needed due to variations in
  - Chip or slide properties, batch
  - Hybridization and environmental conditions
  - Amount of sample
  - Scanner setting
  - Unequal dye incorporation (cDNA arrays)
- Methods
  - Simple global mean or median adjustments
  - Complex (dependent on intensity, region, pin, ...)

    cDNA - Yang *et al.* (http://oz.berkeley.edu/users/terry/zarray)

    Oligo – Affymetrix doumentation, Li and Wong papers

# Outline

1) Introduction: Biology & Technology
2) Data Quality, Image Processing & Normalization
3) Study Objectives & Design Considerations
4) Analysis Strategies Based on Study Objectives

# Design and Analysis Methods Should Be Tailored to Study Objectives

- Class Comparison (supervised)
  - For predetermined classes, establish whether gene expression profiles differ, and identify genes responsible for differences
- Class Discovery (unsupervised)
  - Discover clusters among specimens or among genes
- Class Prediction (supervised)
  - Prediction of phenotype using information from gene expression profile

# Class Comparison Examples

- Establish that expression profiles differ between two histologic types of cancer
- Identify genes whose expression level is altered by exposure of cells to an experimental drug

# Class Discovery Examples

- Discover previously unrecognized subtypes of lymphoma
- Identify co-regulated genes

# Class Prediction Examples

- Predict from expression profiles which patients are likely to experience severe toxicity from a new drug versus who will tolerate it well
- Predict which breast cancer patients will relapse within two years of diagnosis versus who will remain disease free

# Design Considerations

- Controls and replicates within array
- Level of replication (multiple array)
  - Aliquots from single RNA batch
  - RNA extractions
  - Subjects
- Additional considerations for cDNA arrays
  - Allocation of samples to (cDNA) array experiments
    - Common reference design
    - Kerr and Churchill, *Biostatistics,* 2001 ("loop design")
    - Dobbin and Simon, *Bioinformatics, in press* (comparisons)
  - Reverse fluor experiments – gene specific dye biases

# Outline

1) Introduction: Biology & Technology
2) Data Quality, Image Processing & Normalization
3) Study Objectives & Design Considerations
4) Analysis Strategies Based on Study Objectives

# Analysis Strategies for Class Comparisons

- Model-based methods (ANOVA/mixed models)
- Global tests
- Multiple testing procedures to identify differentially expressed genes

# Mixed Model and ANOVA Methods for cDNA Arrays

(based on red and green signals rather than ratios)

- Kerr *et al.*, *Journal of Computational Biology*, 2000
- Lee *et al.*, *PNAS*, 2000
- Kerr and Churchill, *Biostatistics*, 2001
- Wolfinger *et al.*, *Journal of Computational Biology*, 2001

# Global Tests for Differences Between Classes

- Choice of summary measure of difference

  Examples:

  - Sum of squared univariate t-statistics
  - Number of genes univariately significant at 0.001 level

- Statistical testing by permutation test or bootstrap

# Multiple testing procedures:
## Identifying differentially expressed genes while controlling for false discoveries*

- *Expected Number* of False Discoveries – E(FD)
- *Expected Proportion* of False Discoveries – E(FDP) = False Discovery Rate (FDR)
- *Actual Number* of False Discoveries - FD
- *Actual Proportion* of False Discoveries - FDP

*False discovery = declare gene as differentially expressed (reject test) when in truth it is not differentially expressed

# Simple Procedures

- Control E(FD) $\leq u$
  - Conduct each of k tests at level $u$/k
- Control E(FDP) $\leq \gamma$
  - FDR procedure (Benjamini and Hochberg)
- Bonferroni control of familywise error (FWE) rate at level $\alpha$
  - Conduct each of k tests at level $\alpha$/k
  - At least (1-$\alpha$)100% confident that FD = 0

# Problems With Simple Procedures

- Bonferroni control of FWE is very conservative and allowing no false discoveries may be too restrictive when testing so many genes
- Controlling *expected* number or proportion of false discoveries may not provide adequate control on *actual* number or proportion

# Additional Procedures

- "SAM" - Significance Analysis of Microarrays
  - Tusher *et al.*, *PNAS,* 2001
  - Estimate FDR
- Empirical Bayes
  - Efron *et al.*, *JASA,* 2001
  - Related to FDR
- Step-down permutation procedures
  - Korn *et al.,* 2001 (http://linus.nci.nih.gov/~brb)
  - Control number or proportion of false discoveries with stated confidence

# Class Discovery

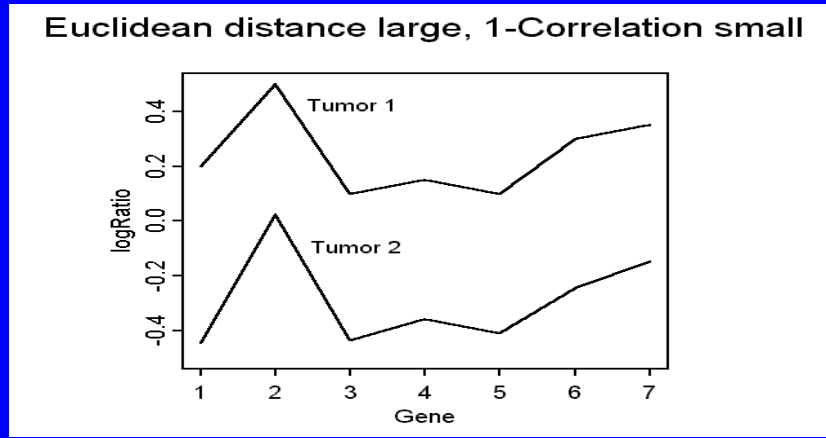
- Cluster analysis algorithms (Gordon, 1999)
  - Hierarchical
  - K-means
  - Self-Organizing Maps
  - Maximum likelihood/mixture models
  - Multitude of others
- Graphical displays
  - Hierarchical clustering
    - Dendrogram
    - "Ordered" color image plot
  - Multidimensional scaling plot

# Hierarchical Agglomerative Clustering Algorithm

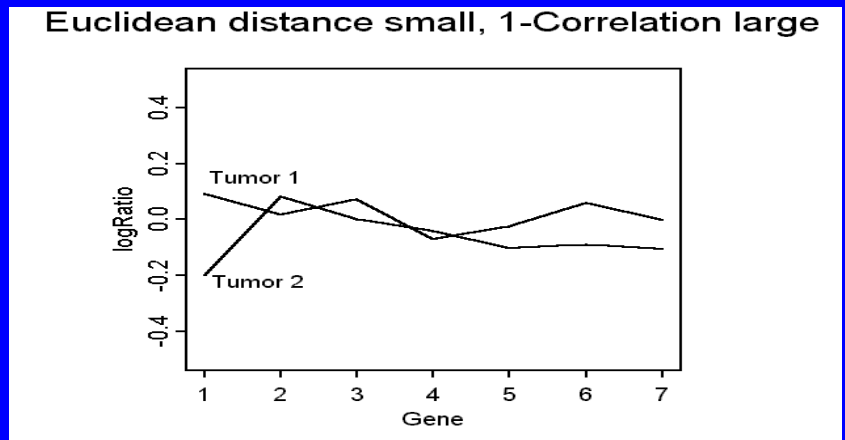
- Merge two closest observations into a cluster.
  - How is distance between individual observations measured?
- Continue merging closest clusters/observations.
  - How is distance between clusters measured?
    - Average linkage
    - Complete linkage
    - Single linkage
    - Many others

# Common Distance Metrics for Hierarchical Clustering

- Euclidean distance
  - Measures absolute distance (square root of sum of squared differences)



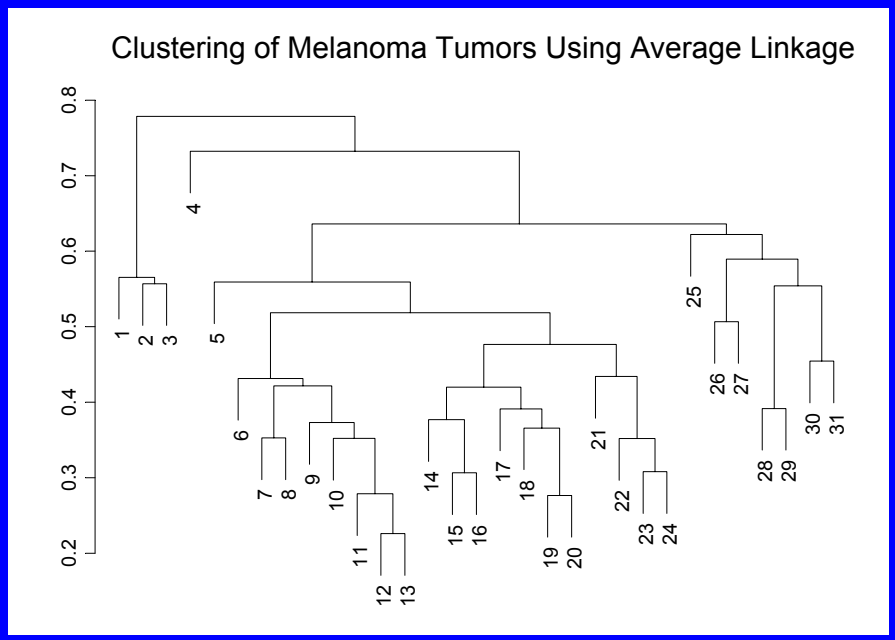

- 1-Correlation
  - Large values reflect lack of linear association (pattern dissimilarity)

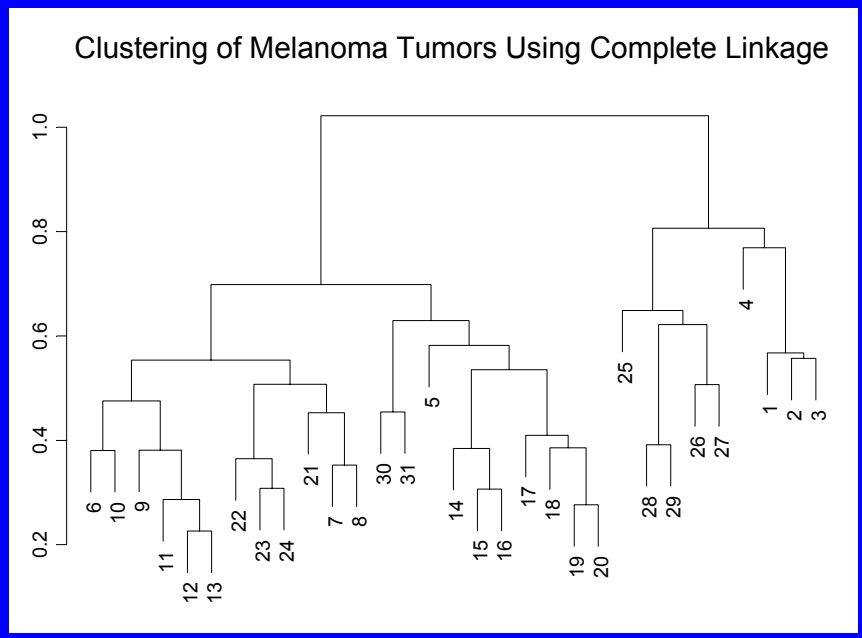


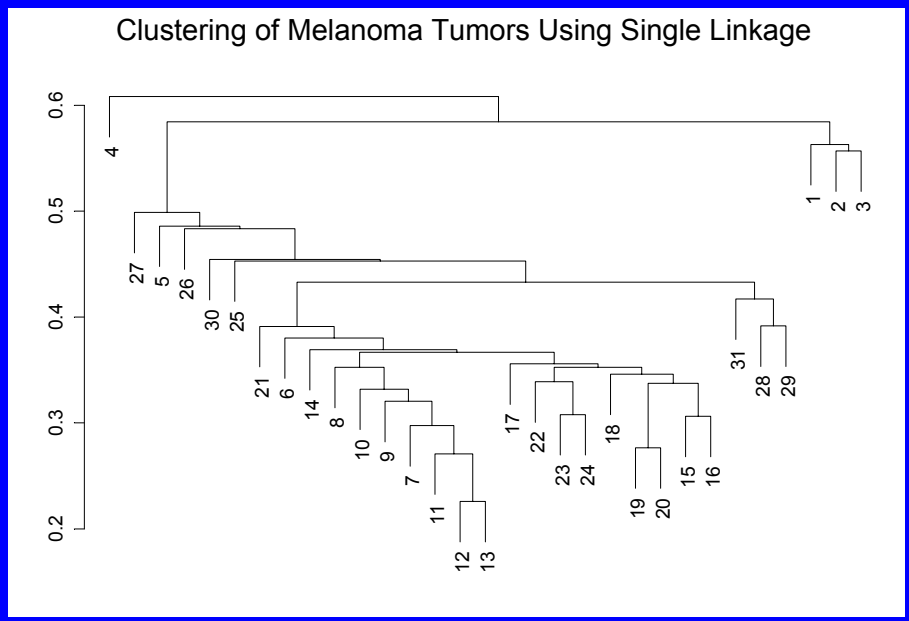

Others: Mahalanobis distance, angular distance, etc.

# Linkage Methods

- Average Linkage
  - Merge clusters whose average distance between all pairs of items (one item from each cluster) is minimized
  - Particularly sensitive to distance metric
- Complete Linkage
  - Merge clusters to minimize the maximum distance within any resulting cluster
  - Tends to produce compact clusters
- Single Linkage
  - Merge clusters at minimum distance from one another
  - Sensitive to noise and prone to “chaining”



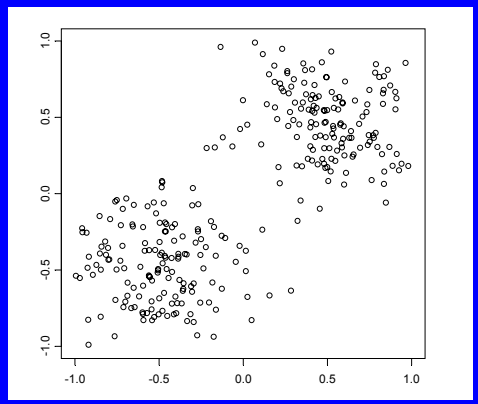

Dendrograms using 3 different linkage methods,
distance = 1-correlation

(Data from Bittner *et al.*, *Nature*, 2000)

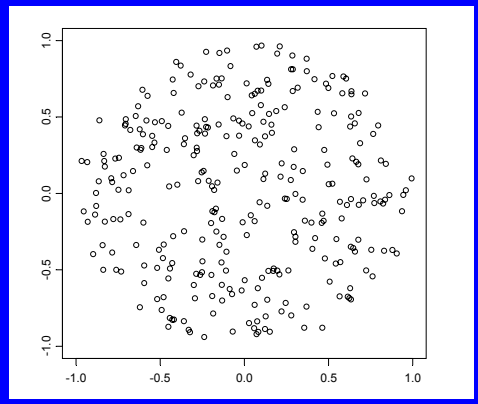
# Interpretation of Cluster Analysis Results

- Cluster analyses always produce cluster structure
  - Are there any "real" clusters?
  - Where to "cut" the dendrogram?
  - Assessing individual clusters
- Different clustering algorithms may find different structure using the same data.
- Which clusters do we believe?
  - Reproducible between methods
  - Reproducible within a method

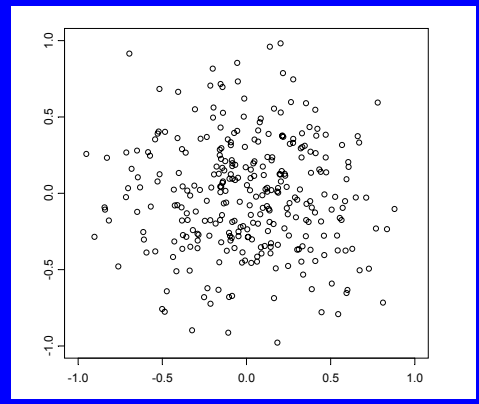
# Global Test for Clustering




Observed data



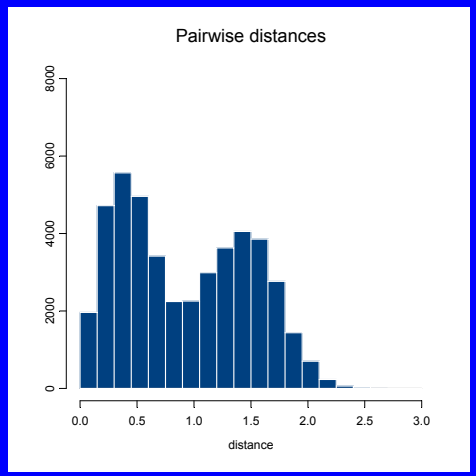

Null data 1

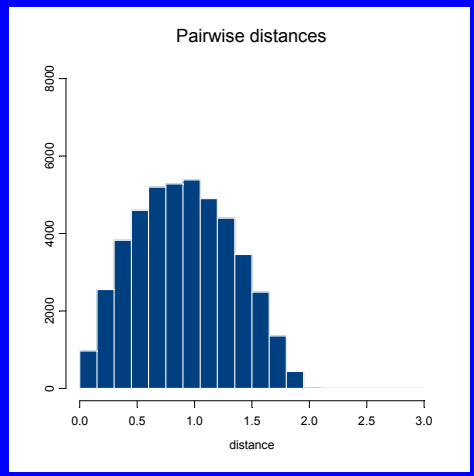


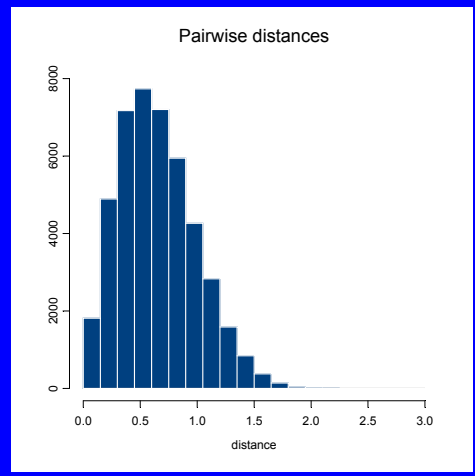

Null data 2

Quantify and compare degree of clustering

(e.g. compare shape of pairwise distance distribution)



# Determining the "Optimal" Number of Clusters

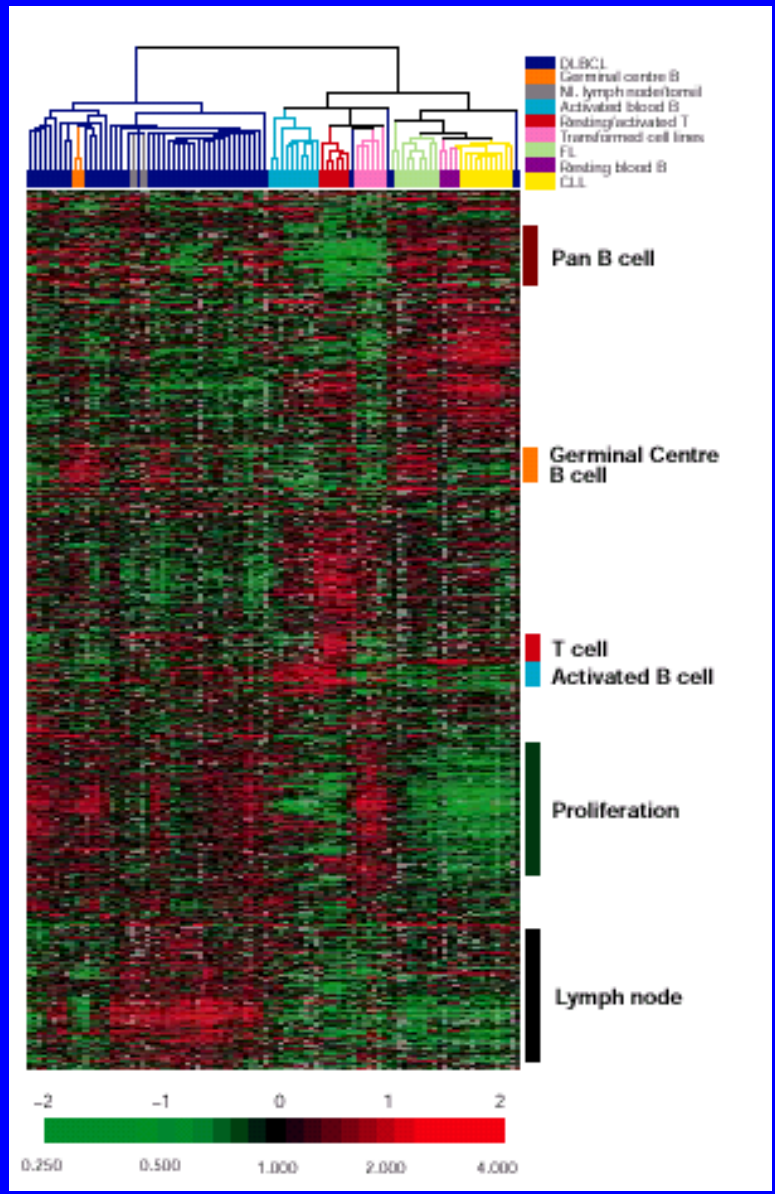
- Comparison of methods for estimating number of clusters in small dimension cases (Milligan and Cooper, *Psychometrika,* 1985)
- Gap Statistic (Tibshirani *et al*., *JRSS B*, 2002)
- Generally do NOT work well as global tests (sometimes not even defined for 1 cluster)
- Performance on very high dimensional data may not have been tested

# Assessing Individual Clusters: Data Perturbation Methods

- Most believable clusters are those that persist given small perturbations of the data.
  - Perturbations represent an anticipated level of noise in gene expression measurements.
  - Re-cluster perturbed data and compare to original clustering
  - References specific to microarrays
    - McShane *et al*. (*Bioinformatics, in press*) – Gaussian errors
    - Kerr and Churchill (*PNAS*, 2001) – Bootstrap residual errors

# Graphical Displays: Color Image Plot

Hierarchical
Clustering
of Lymphoma Data
(Alizadeh *et al.*,
*Nature*, 2000)

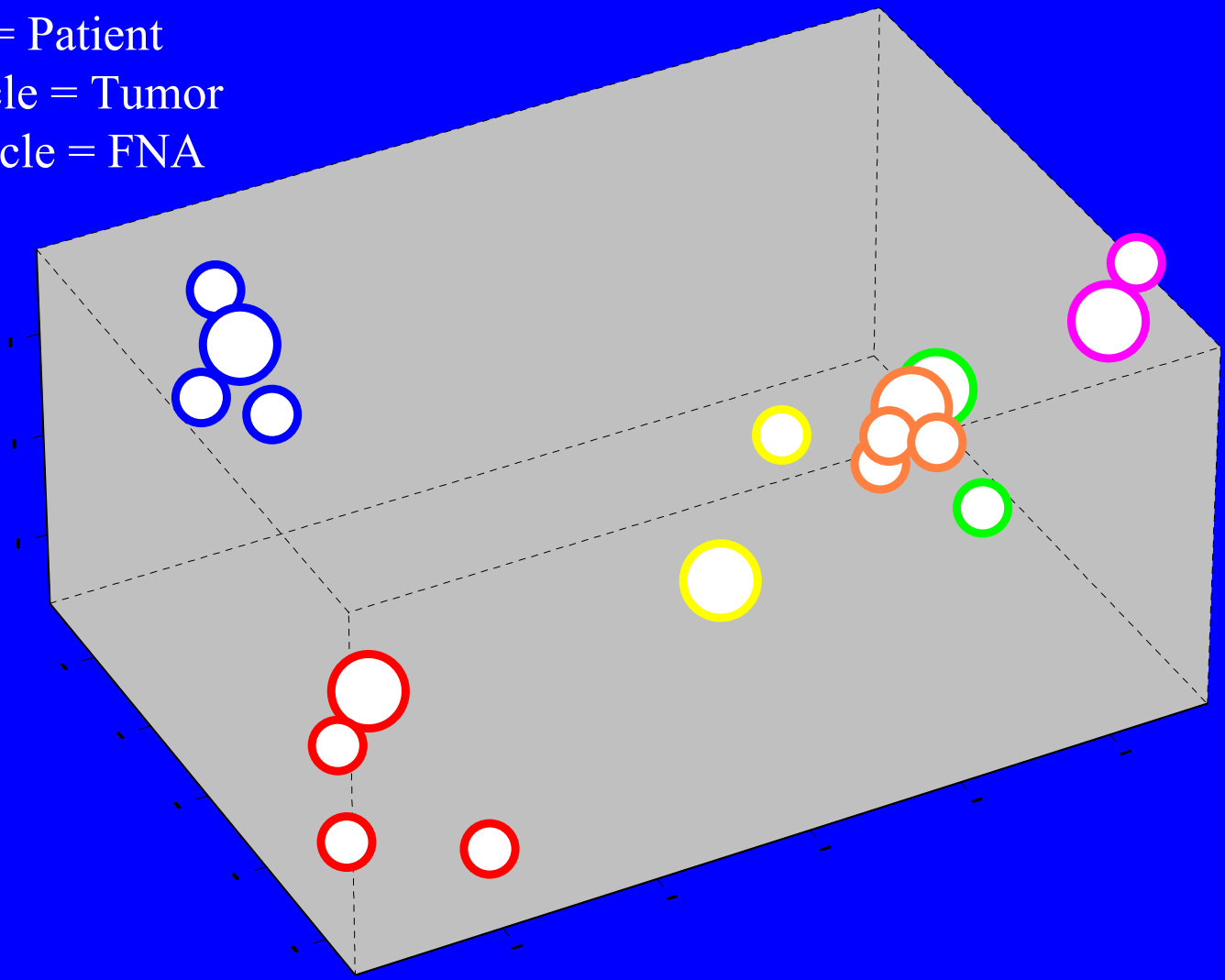


# Graphical Displays: Multidimensional Scaling (MDS)

- High-dimensional (e.g. 5000-D) data points are represented in a lower-dimensional space (e.g. 3-D)
  - Principal components (classical) or optimization methods
  - Depends only on pairwise distances (Euclidean, 1-correlation, . . .) between points
  - "Relationships" need not be well-separated clusters

# MDS: Breast Tumor and FNA Samples

Color = Patient
Large circle = Tumor
Small circle = FNA



(Assersohn *et al., Clinical Cancer Research*, 2002)

# Class Prediction Methods

**Comparison of linear discriminant analysis, NN classifiers, classification trees, bagging, and boosting:** tumor classification based on gene expression data (Dudoit, *et al.*, *JASA*, 2002)

**Weighted voting method:** distinguished between subtypes of human acute leukemia (Golub *et al.*, *Science*, 1999)

**Compound covariate prediction:** distinguished between mutation positive and negative breast cancers (Hedenfalk *et al.*, *NEJM*, 2001)

**Support vector machines:** classified ovarian tissue as normal or cancerous (Furey *et al.*, *Bioinformatics*, 2000)

**Neural Networks:** distinguished among diagnostic subcategories of small, round, blue cell tumors in children (Khan *et al.*, *Nature Medicine*, 2001)

# Pitfalls in Class Prediction for Microarray Data

"Note that when classifying samples, we are confronted with a problem that there are many more attributes (genes) than objects (samples) that we are trying to classify. This makes it always possible to find a perfect discriminator if we are not careful in restricting the complexity of the permitted classifiers. To avoid this problem we must look for very simple classifiers, compromising between simplicity and classification accuracy." (Brazma & Vilo, *FEBS Letters*, 2000)

Validation! Validation! Validation!

# The Compound Covariate Predictor (CCP)

(Tukey, *Controlled Clinical Trials*, 1993)

- Select "differentially expressed" genes by two-sample *t*-test with small $\alpha$.

$$\text{CCP}_i = \sum_j t_j x_{ij}$$

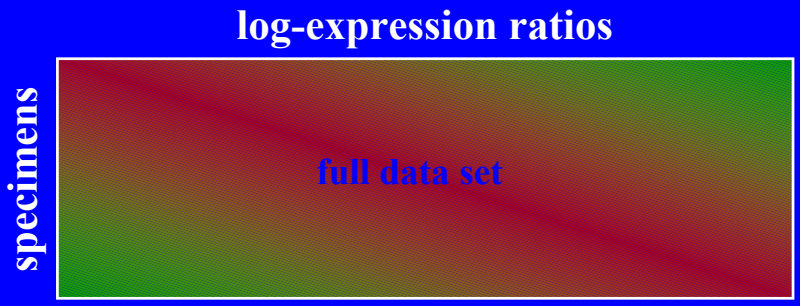
$t_j$ is the two-sample *t*-statistic for gene *j*.

$x_{ij}$ is the log-ratio measure of sample *i* for gene *j*.

Sum is over all differentially expressed genes.

- Threshold of classification: midpoint of the CCP means for the two classes.

# Non-Cross-Validated Prediction




1. Prediction rule is built using full data set.
2. Rule is applied to each specimen for class prediction.

# Cross-Validated Prediction (Leave-One-Out Method)


1. Full data set is divided into training and test sets (test set contains 1 specimen).
2. Prediction rule is built using the training set.
3. Rule is applied to the specimen in the test set for class prediction.
4. Process is repeated until each specimen has appeared once in the test set.

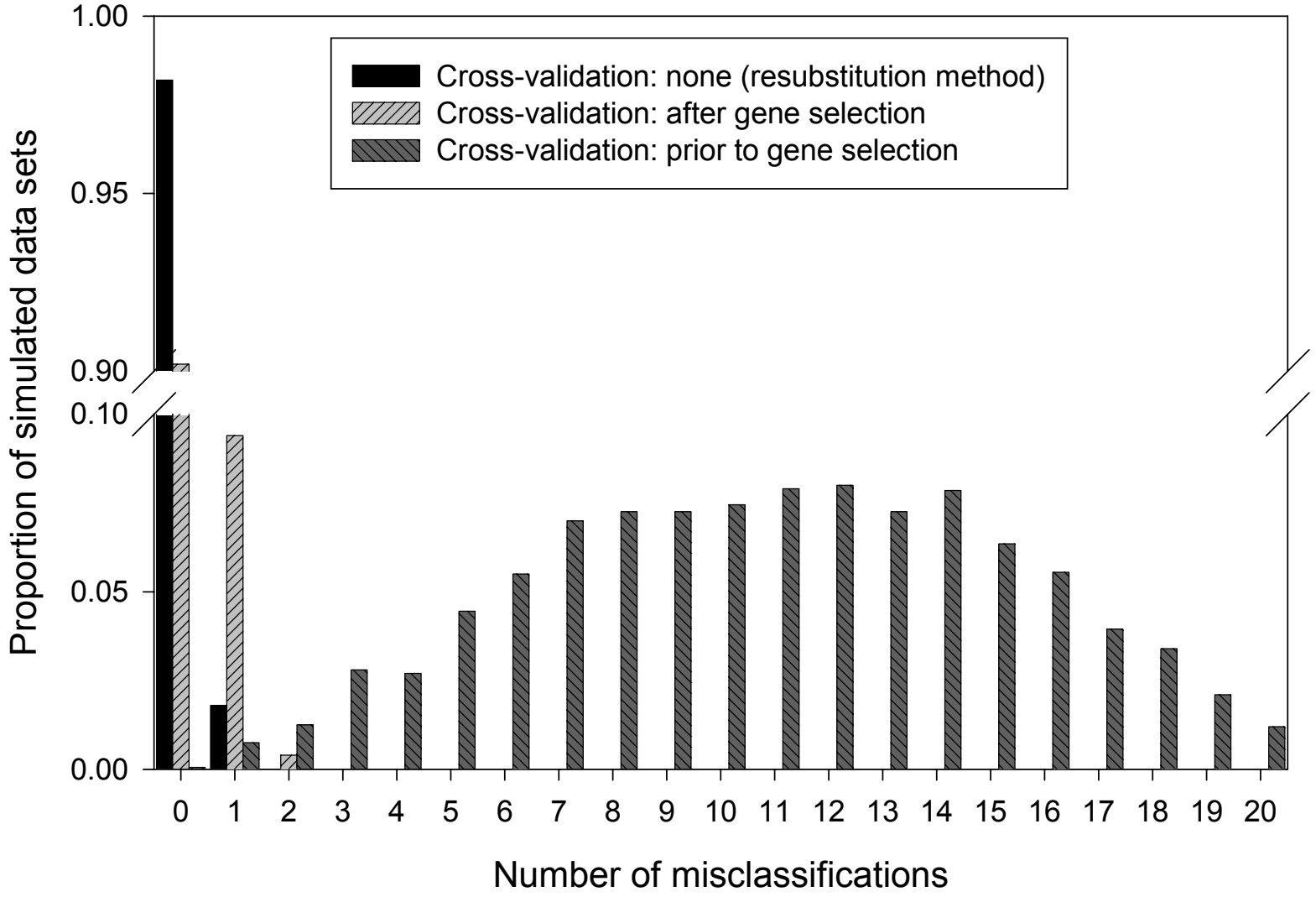
# Prediction on Simulated Null Data

**Generation of Gene Expression Profiles**

- 20 specimens ($P_i$ is the expression profile for specimen $i$)
- Log-ratio measurements on 6000 genes
- $P_i \sim \text{MVN}(\mathbf{0}, \mathbf{I}_{6000})$
- Can we distinguish between the first 10 specimens (Class 1) and the last 10 (Class 2)? (class distinction is totally artificial since all 20 profiles were generated from the same distribution)

**Prediction Method**

- Compound covariate prediction
- Compound covariate built from the log-ratios of the 10 most differentially expressed genes.

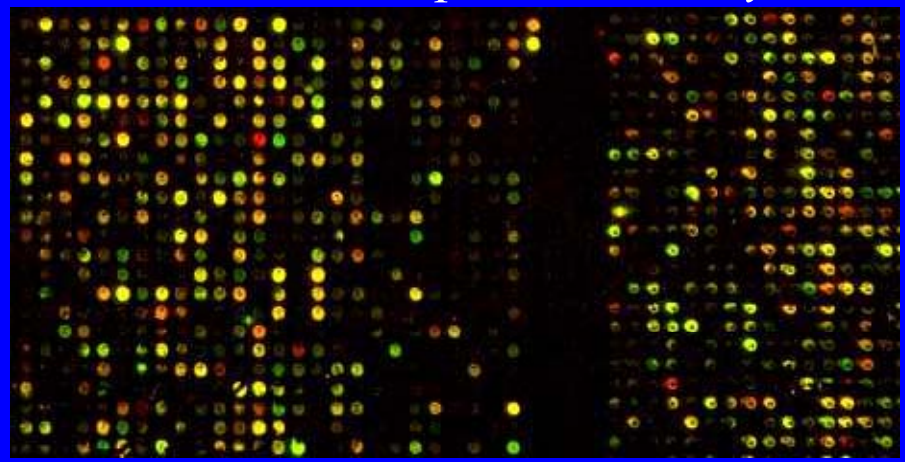

Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
Proportion of simulated data sets
1.00
0.95
0.90
0.10
0.05
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of misclassifications

# Gene-Expression Profiles in Hereditary Breast Cancer

( Hedenfalk *et al.*, *NEJM*, 2001)

**cDNA Microarrays**

*Parallel Gene Expression Analysis*



- Breast tumors studied:
  - 7 *BRCA1*+ tumors
  - 8 *BRCA2*+ tumors
  - 7 sporadic tumors
- Log-ratios measurements of 3226 genes for each tumor after initial data filtering

**RESEARCH QUESTION**

Can we distinguish *BRCA1*+ from *BRCA1*– cancers and *BRCA2*+ from *BRCA2*– cancers based solely on their gene expression profiles?

# Results of leave-one-out cross-validation

Classification of hereditary breast cancers with the compound covariate predictor

| Class labels | Number of differentially expressed genes | $m$ = number of misclassifications | Proportion of random permutations with $m$ or fewer misclassifications |
|---|---|---|---|
| *BRCA1*$^+$ vs. *BRCA1*$^-$ | 9 | 1 (0 *BRCA1*$^+$, 1 *BRCA1*$^-$) | 0.004 |
| *BRCA2*$^+$ vs. *BRCA2*$^-$ | 11 | 4 (3 *BRCA2*$^+$, 1 *BRCA2*$^-$) | 0.043 |

# Validation of Predictor on Independent Data

- Potential pitfalls of estimated prediction accuracy from leave-one-out cross-validation on a single data set
  - High variance of LOO CV error rate for small samples
  - Peculiarities of the training set may influence the prediction rule
- Independent data set for validation
  - Should be fairly large (e.g., as big as training set)
  - Similar proportions of specimens for the classes as exist in the population

# Summary Remarks

- Data quality assessment and pre-processing are important.
- Different study objectives will require different statistical analysis approaches.
- Different analysis methods may produce different results. Thoughtful application of *multiple* analysis methods may be required.
- Chances for spurious findings are enormous, and validation of any findings on larger independent collections of specimens will be essential.
- Analysis tools are not an adequate substitute for collaboration with professional data analysts.

# Helpful Websites

- NCI: http://linus.nci.nih.gov/~brb
  - Tech reports, talk slides
  - BRB-ArrayTools software
- Berkeley: http://www.stat.berkeley.edu/users/terry/zarray/Html/index.html
- Harvard: http://www.dchip.org
- Hopkins: http://biosun01.biostat.jhsph.edu/~ririzarr/Raffy/
- Jackson Labs: http://www.jax.org/research/churchill/
- Stanford:
  - http://genome-www5.stanford.edu/MicroArray/SMD/restech.html
  - http://www-stat.stanford.edu/~tibs/ (R. Tibshirani)