Statistical Analysis of Gene Expression Microarray Data

Lisa M. McShane
Joanna H. Shih
Richard Simon

Biometric Research Branch
National Cancer Institute

Outline

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

Outline

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

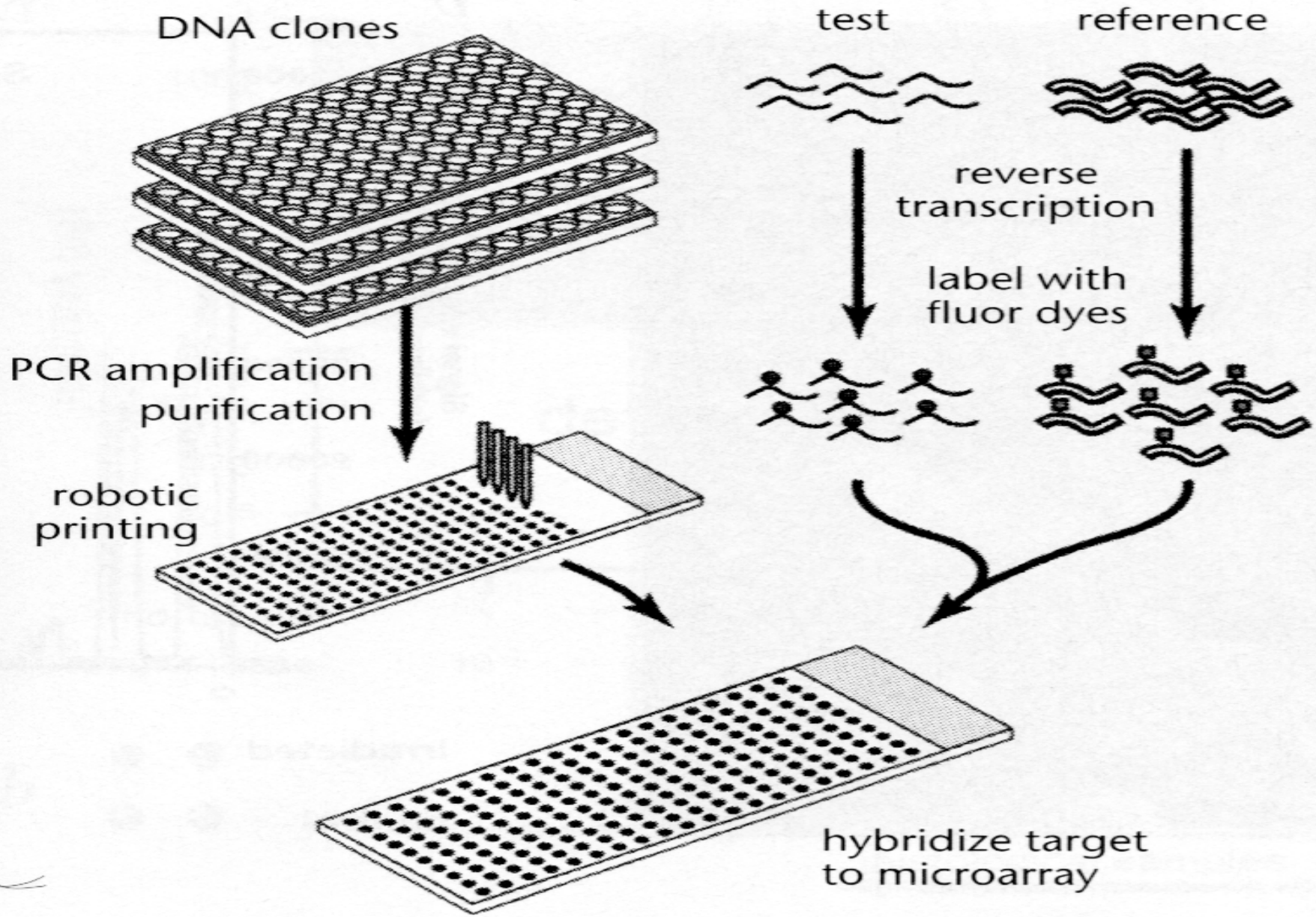
Gene Expression Microarrays

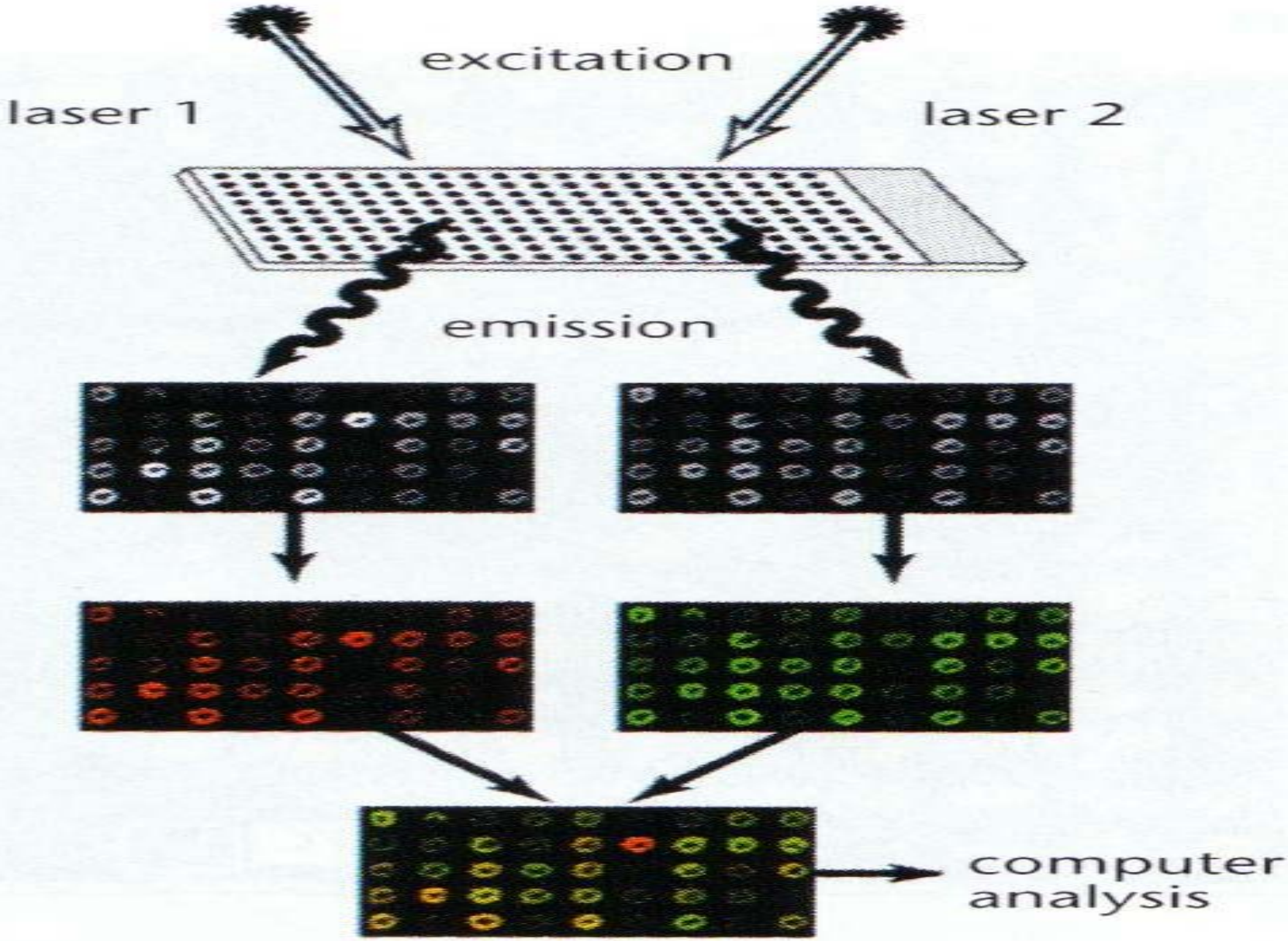
- Permit simultaneous evaluation of expression levels of thousands of genes
- Main platforms
  - Spotted cDNA arrays (2-color)
  - Affymetrix GeneChip (1-color)
  - Spotted oligo arrays (2-color or 1-color)
  - Nylon filter arrays

Spotted cDNA Arrays (and other 2-color spotted arrays)

- cDNA arrays: Schena *et al.*, *Science,* 1995
- Each gene represented usually by one spot (occasionally multiple)
- Two-color (two-channel) system
  - Two colors represent the two samples competitively hybridized
  - Each spot has “red” and “green” measurements associated with it

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

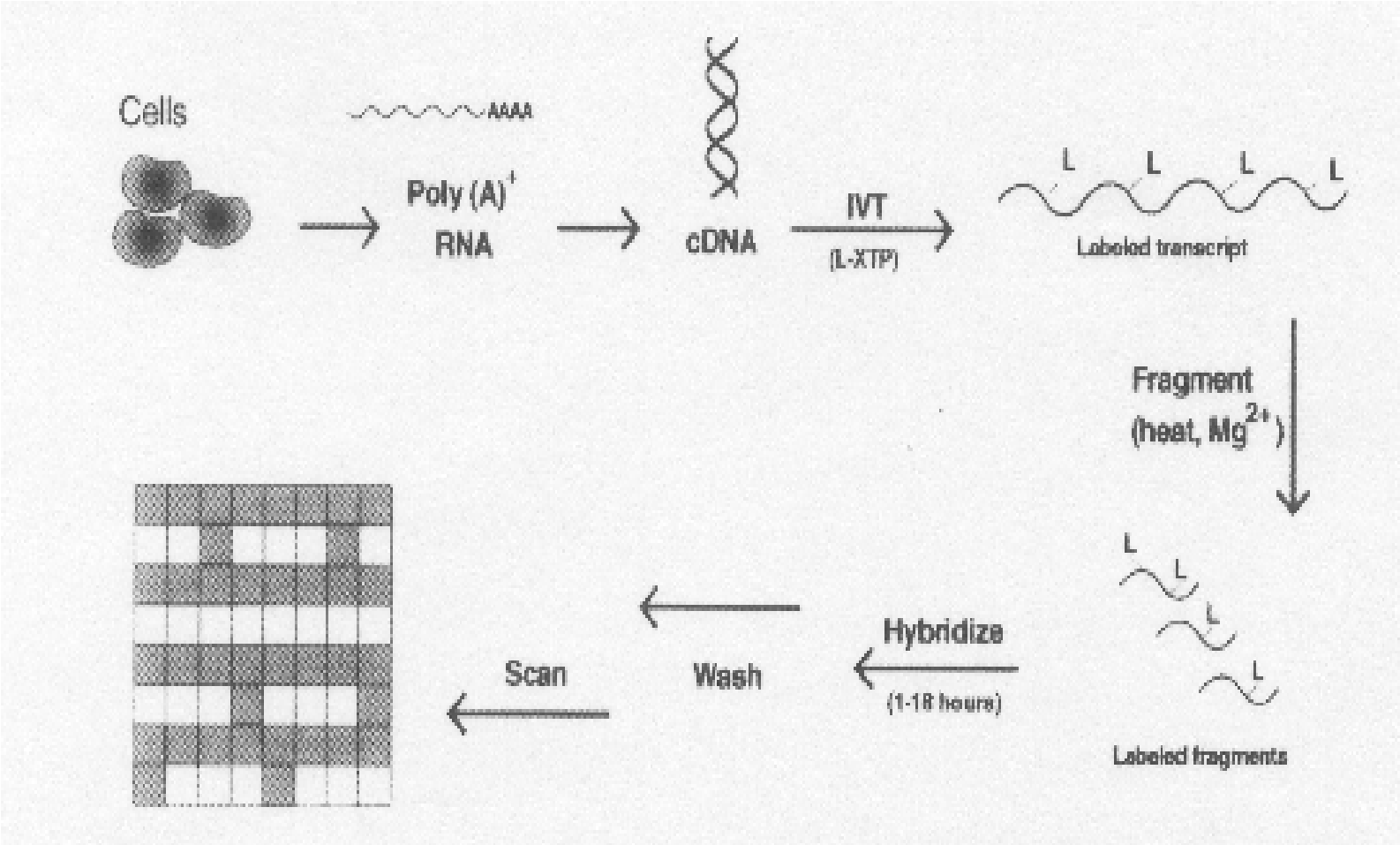
Affymetrix GeneChip Arrays

- Lockhart *et al., Nature Biotechnology,* 1996
- Affymetrix: http://www.affymetrix.com
- Glass wafer ("chip") – photolithography, oligonucleotides synthesized on chip
- 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)

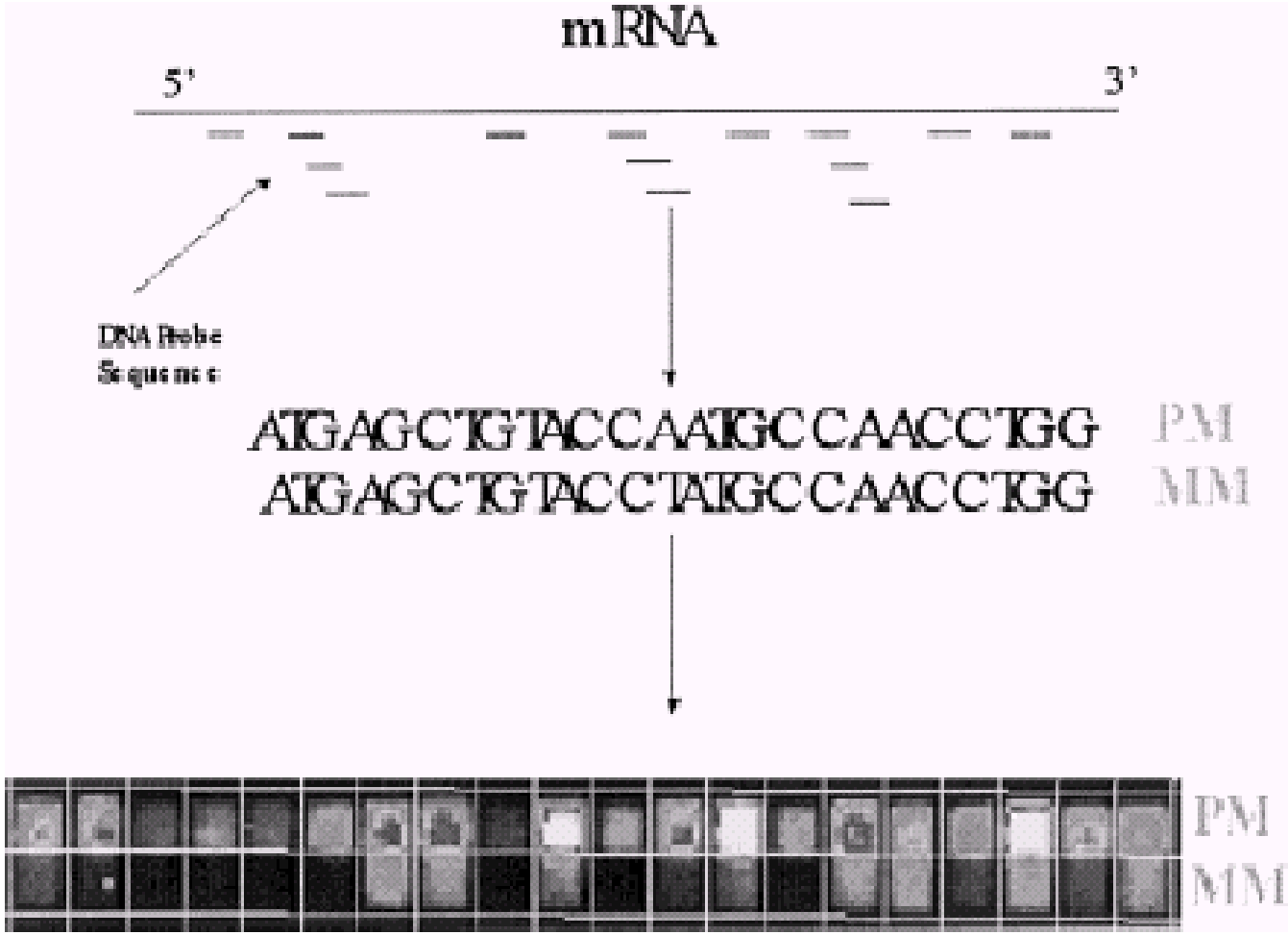
[Affymetrix] Hybridization Oligo “GeneChip” Array

Affymetrix: Assay procedure



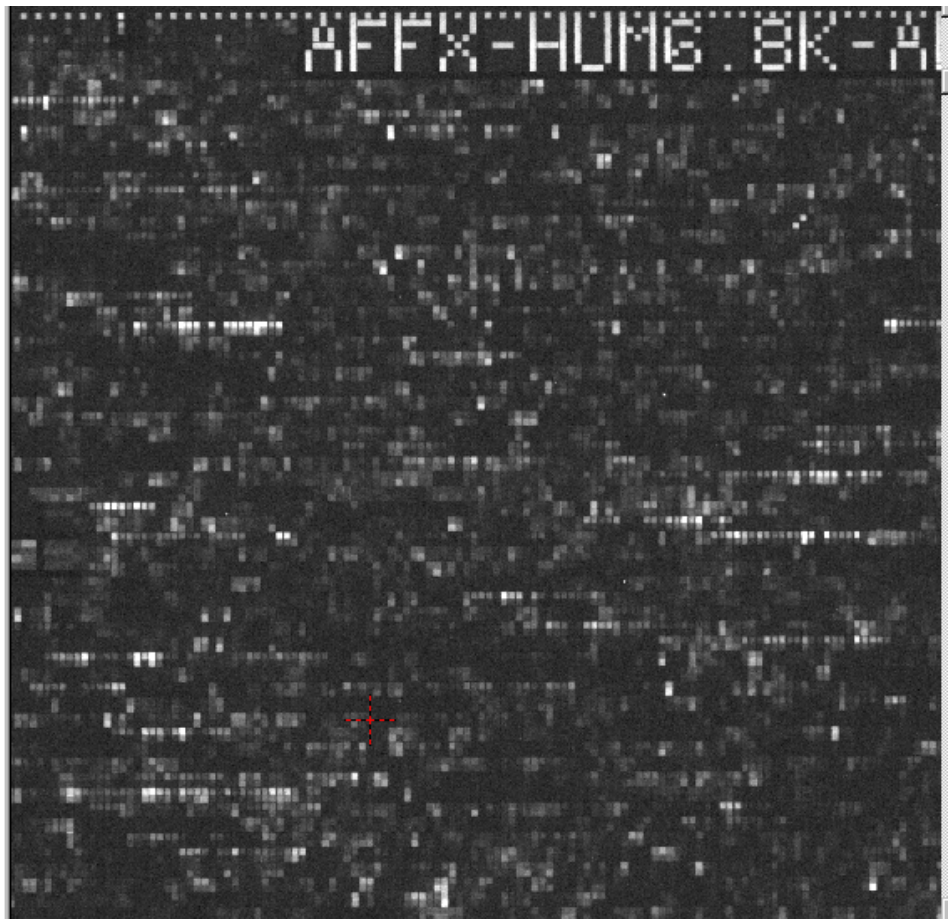
(Figure 1 from Lockhart *et al., Nature Biotechnology,* 1996)

Perfect Match - Mismatch Probe Pairs



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

Image of a Scanned Affymetrix GeneChip

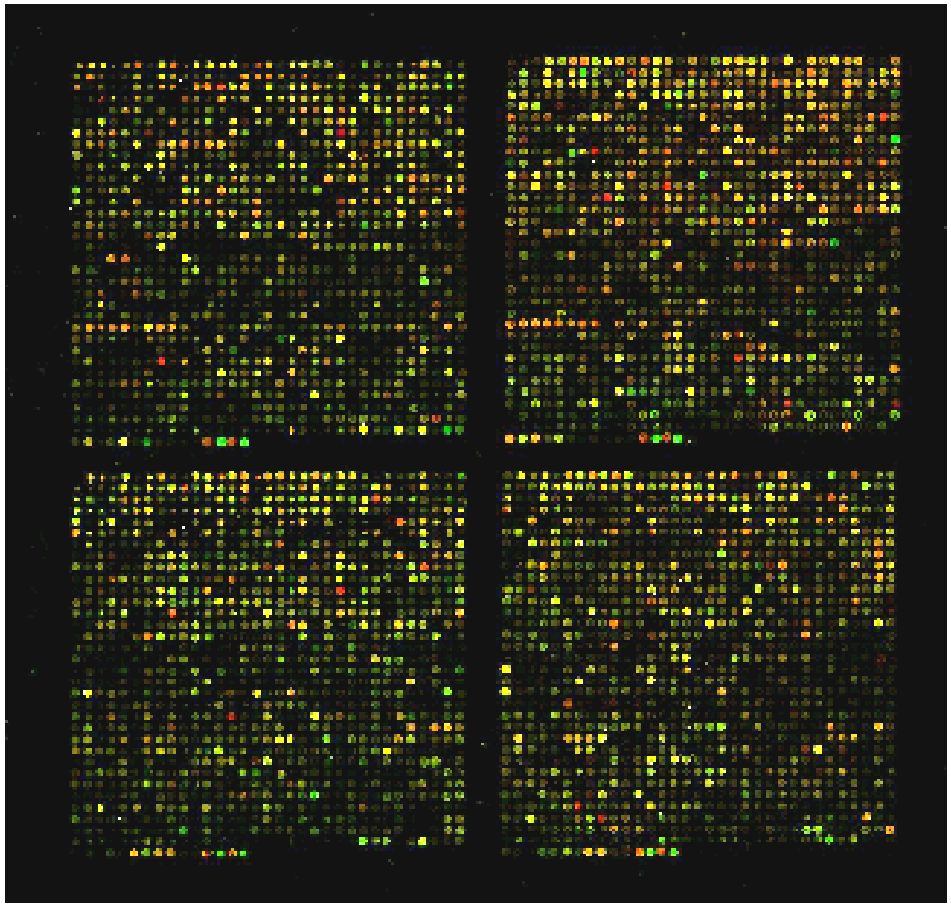


Outline

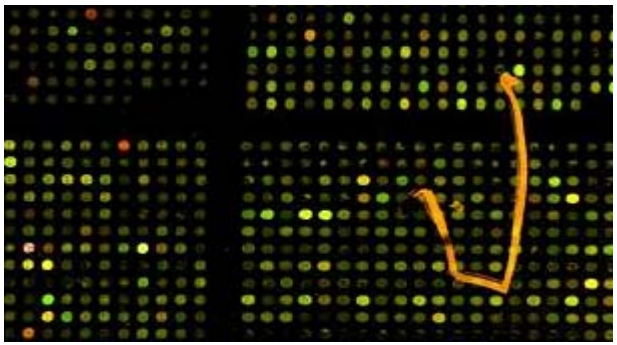
1) Introduction: Technology
2) Data Quality & Image Processing
3) Normalization & Filtering
4) Study Objectives & Design Considerations
5) Analysis Strategies Based on Study Objectives

Slide Quality

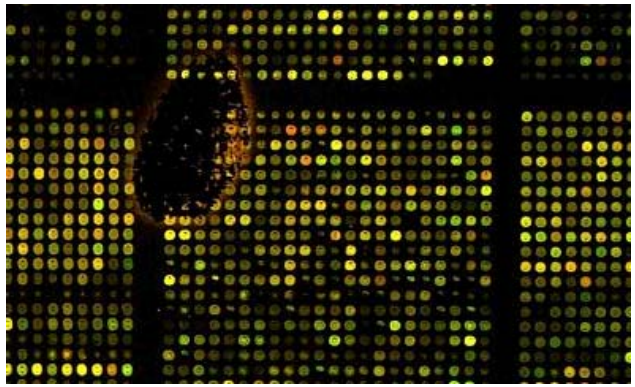
A “good” quality cDNA array



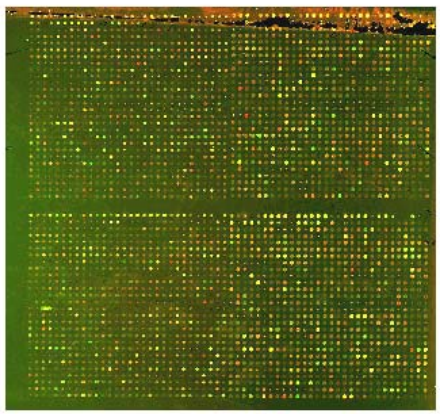
cDNA/Spotted Arrays: Slide Quality



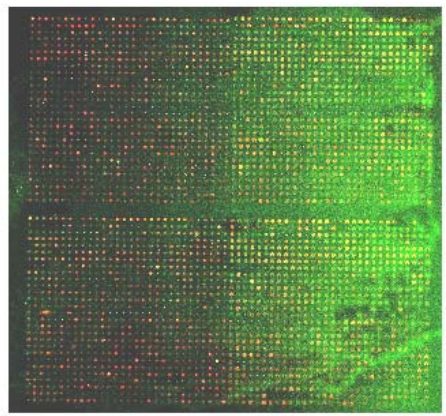
Fiber or scratch?



Bubble



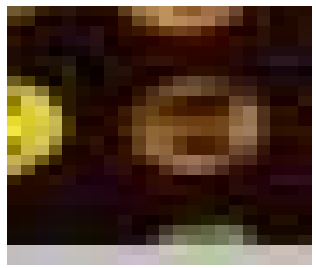
Edge effect



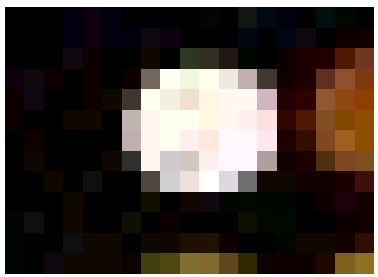
Background haze

cDNA/Spotted Arrays: Spot Quality

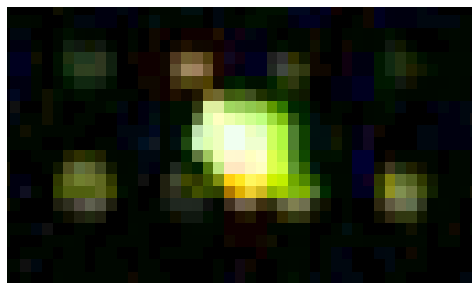
Poorly defined borders



Large holes



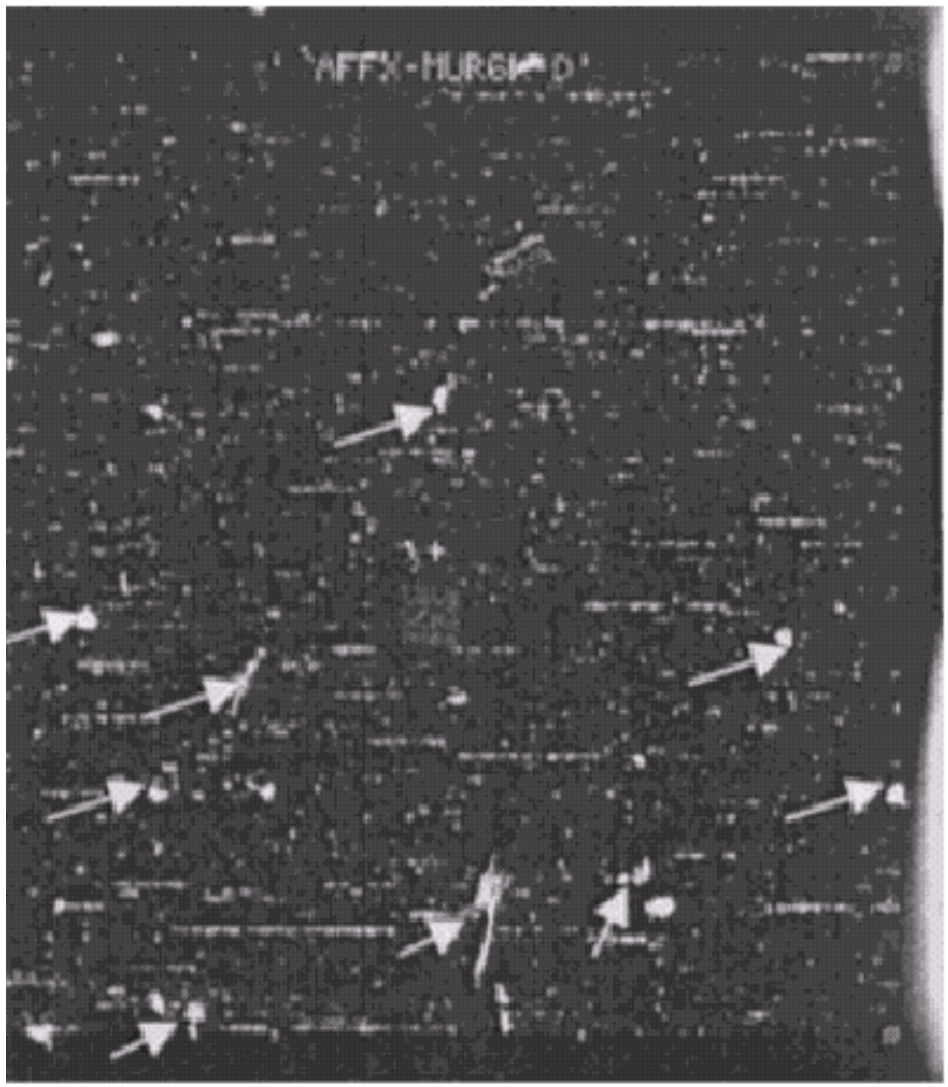
Saturated spot



Dust specs

Affymetrix Arrays: Quality Problems

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



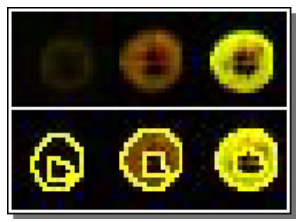
cDNA/2-color spotted arrays: Image Processing

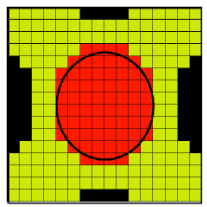
(Yang, *et al., J. Computational and Graphical Statistics*, 2002)

- Segmentation
- Background correction & signal calculation
- Spot flagging criteria
- Gene-level summaries

cDNA/2-color spotted arrays: Segmentation

- Segmentation - separation of feature (F) from background (B) for each spot.





(See software documentation)

- Summary measures computed for F
  - Intensity: mean or median over pixels
  - Additional measures: SD, # pixels (size), etc.

cDNA/2-color spotted arrays: Background Correction & Signal Calculation

- No background correction

  Signal = F intensity

- Local background correction

  Signal = F intensity - B_{local}

- Regional background correction

  Signal = F intensity - $B_{regional}$

cDNA/2-color spotted arrays: Flagging Spots for Exclusion

A spot is excluded from analysis if "signal" or "signal-to-noise" measure(s) at that spot fail to exceed a threshold. Several criteria can be used:

- F
- F-B
- F/B
- (F-B)/SD(B)
- Spot Size

Excluding Entire Arrays or Regions

- Too many spots flagged
- Narrow range of intensities
- Uniformly low signals

cDNA/2-color spotted arrays: Gene-level Summaries

- Model-based methods
  - Work directly on signals from two channels (two colors)
- Ratio methods
  - Red signal/Green signal

Affymetrix Arrays: Image Processing

- Grid alignment to probe cells
- Background correction
- Summarize over probe sets to get gene expression indices
  - Detection calls - present/absent

See Affymetrix documentation:

- Affymetrix website (http://www.affymetrix.com)
- *Affymetrix Microarray Suite User Guide*

Affymetrix Arrays: Probe Set (Gene) Summaries

- $AvDiff_i = \Sigma(PM_{ij}\text{-}MM_{ij})/n_i$ for each probe set *i* (original Affymetrix algorithm)
- Model-Based Expression Index
  - (Li and Wong, *PNAS,* 2001)
  - $MBEI_i = \theta_i$ estimated from $PM_{ij}\text{-}\ MM_{ij} = \theta_i\ \phi_j + \varepsilon_{ij} \Rightarrow$ weighted average difference
- New algorithms to address negative signals, etc.
  - "New" Affymetrix algorithm
  - PM only (no mismatch subtraction)
  - Average background adjustment (Irizarry *et al.,* 2002)

Outline

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

cDNA/2-color spotted arrays: Need for Normalization

- Unequal incorporation of labels
  - green better than red
- Unequal amounts of sample
- Unequal PMT voltage

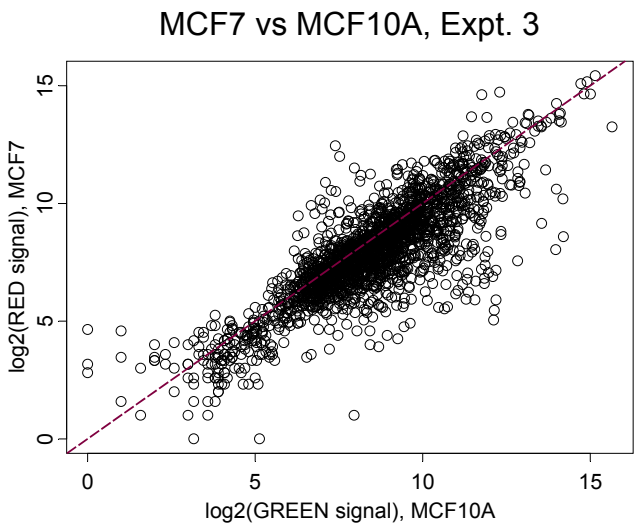
Normalization Methods for cDNA/2-Color Spotted Arrays

- Model-based methods
  - Normalization incorporated into model
- Ratio-based methods
  - Median (or Mean) Centering Method
  - Lowess Method
  - Multitude of other methods

    Chen *et al.*, *Journal of Biomedical Optics*, 1997

    Yang *et al.* (http://oz.berkeley.edu/users/terry/zarray)
    - Scaling factors, separately by printer pin, etc.

Median (or Mean) Centering

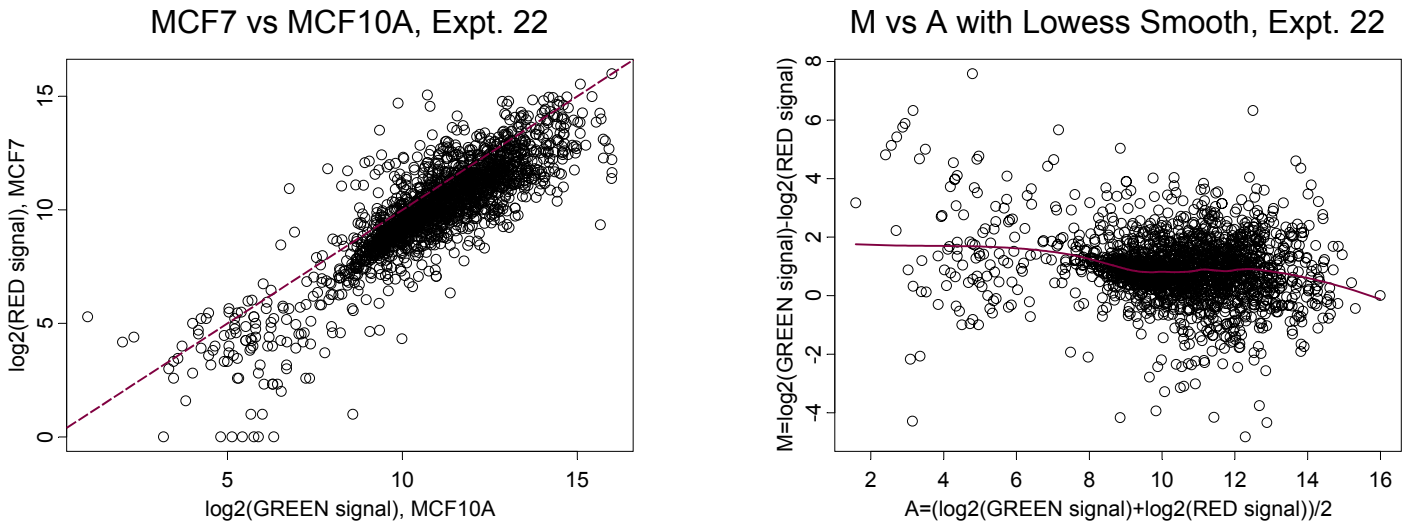


In plot of log(red signal) versus log(green signal), if point scatter is parallel to 45° line, adjust intercept to 0.

Subtract median or mean log-ratio (computed over all genes on the slide or only over housekeeping genes) from each log-ratio.

Lowess Normalization: M vs A plots

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



$$M = \log_2(\text{GREEN signal}) - \log_2(\text{RED signal})$$

$$A = (\log_2(\text{GREEN signal}) + \log_2(\text{RED signal}))/2$$

Normalization: Affymetrix Arrays

- Need
  - Variations in amount of sample or environmental conditions
  - Variations in chip, hybridization, scanning
- Methods
  - Median, lowess, quantile adjustments, . . .
  - Across probe cells or across genes summaries?
  - Adjust to fixed value or to “reference” array

Filtering Genes

- "Bad" values on too many arrays.
- Not differentially expressed across arrays.
  - Variance (assumes approx. normality)

    Let s^2_i = sample variance of gene *i* (log) measurements across *n* arrays.

    Exclude gene *i* if

    $(n\text{-}1)\, s^2_i < \chi^2(1-\alpha, n\text{-}1) \times \text{median}(s^2_1, s^2_2, \ldots, s^2_n)$.
  - Fold difference

    Examples: Max/Min < 3 or 4

    (95^{th} percentile/5^{th} percentile) < 2 or 3

Outline

1) Introduction: Technology
2) Data Quality & Image Processing
3) Normalization & Filtering
4) Study Objectives & Design Considerations
5) 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

- Sample selection
- Sample size planning
- Controls
- Sources of variability/levels of replication
- For cDNA/2-color spotted arrays:
  - Reverse fluor experiments
  - Allocation of samples to (cDNA) array experiments
    - Kerr and Churchill, *Biostatistics,* 2001
    - Dobbin and Simon, *Bioinformatics,* in press

Sample Selection

- Experimental Samples
  - A random sample from the population under investigation?
  - Broad versus narrow inclusion criteria?
- Reference Sample (cDNA array experiments using reference design)
  - In most cases, does not have to be biologically relevant.
    - Expression of most genes, but not too high.
    - Same for every array
  - Other situations exist (e.g., matched normal & cancer)

Sample Size Planning

- No comprehensive method for planning sample size exists for gene expression profiling studies.

- In lieu of such a method…
  - Plan sample size based on comparisons of two classes involving a single gene.
  - Make adjustments for the number of genes that are examined.

Sample Size Planning

GOAL: Identify genes differentially expressed in a comparison of pre-defined classes of specimens.

- Total sample size when comparing two equal sized, independent groups:

$$n = 4\sigma^2(z_{\alpha/2} + z_{\beta})^2/\delta^2$$

where δ = mean difference between classes

σ = standard deviation

$z_{\alpha/2}$, z_{β} = standard normal percentiles

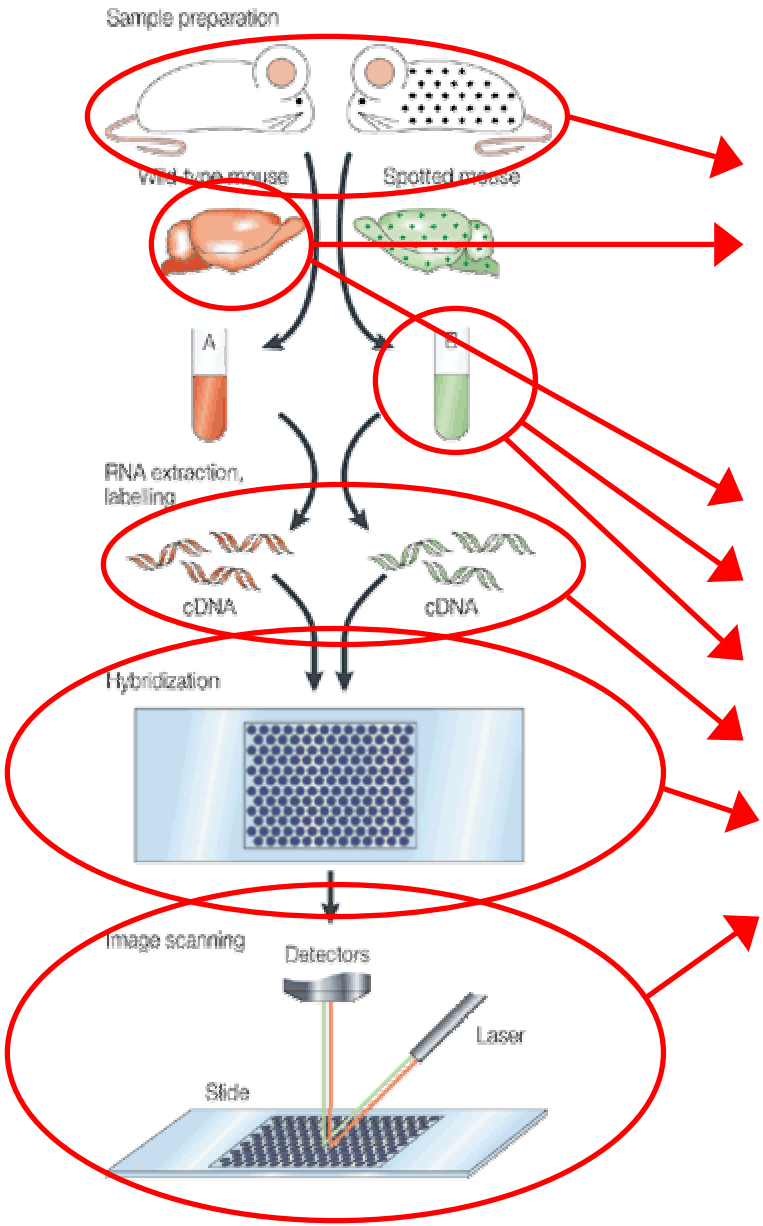
- Choose α small, e.g. $\alpha = .001$
- Alternative formulas for unequal, paired, or multiple groups

Controls

- Multiple clones (cDNA arrays) or probe sets (oligo arrays) for same gene spotted on array
- Spiked controls (e.g. yeast or *E. coli)*

Sources of Variability

(cDNA Array Example)



- Biological Heterogeneity in Population
- Specimen Collection/ Handling Effects
  - Tumor: surgical bx, FNA
  - Cell Line: culture condition, confluence level
- Biological Heterogeneity in Specimen
- RNA extraction
- RNA amplification
- Fluor labeling
- Hybridization
- Scanning
  - PMT voltage
  - laser power

(Geschwind, *Nature Reviews Neuroscience*, 2001)

Examples of Replication

- RNA sample divided into multiple aliquots
- Multiple RNA samples from a specimen
- Multiple subjects from population(s)

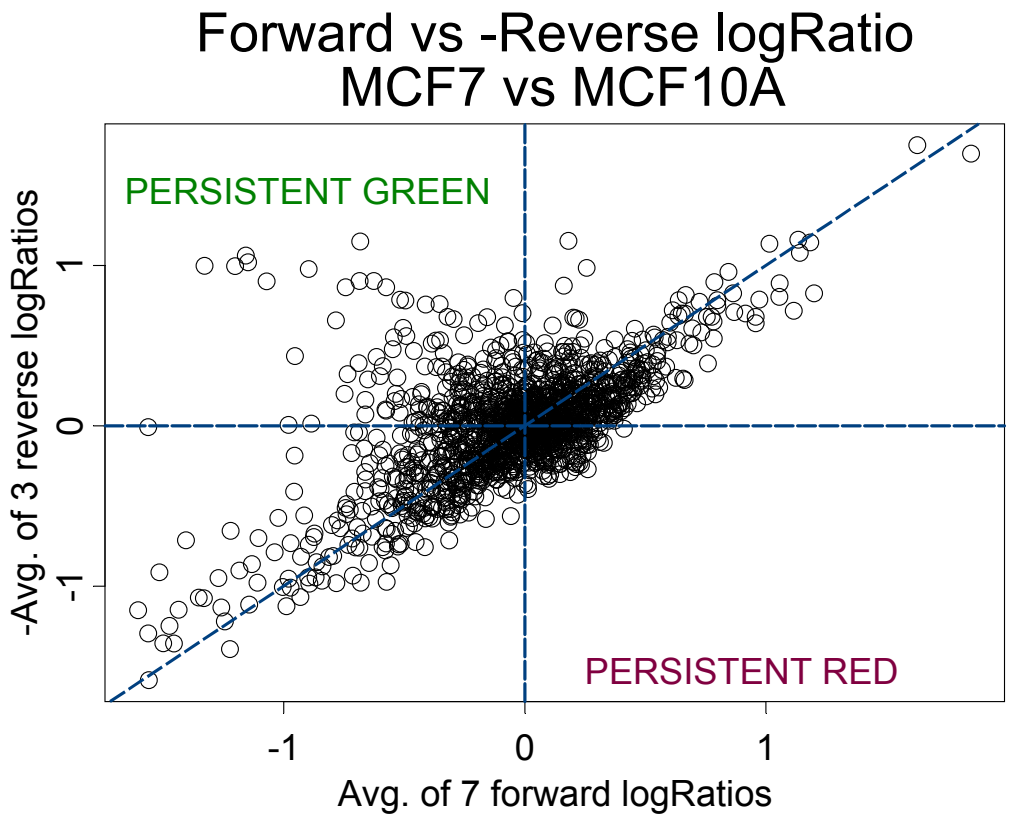
Level of Replication Determines Inference

- Replication of samples should generally be at the “subject” level because we want to make inference to the population of “subjects”, not to the population of sub-samples of a single biological specimen.

Does it pay to replicate arrays?

- When the subject-to-subject variability is greater than the experimental variability, it is more efficient to assay specimens from additional subjects than it is to perform replicate arrays for the same subjects.
- Some replicates may be helpful as quality checks, but can be misleading if replication covers only some of the sources of variability.

cDNA/2-Color Spotted Arrays: Reverse Fluor Experiments



cDNA/2-color spotted arrays:

Should reverse fluor “replicates” be performed for each array?

cDNA/2-color spotted arrays: Reverse Fluors

- When interested in interpreting an individual ratio . . .
  - If gene-specific dye bias is relatively constant across specimens, dye bias can be accounted for by performing some reverse fluor experiments.

cDNA/2-color spotted arrays: Reverse Fluors

- When interested in class comparisons . . .
  - If gene-specific dye bias is constant across arrays, it will “cancel out”.
  - If average dye bias within each group is the same for all groups, the average bias will “cancel out”.

cDNA/2-color spotted arrays: Reverse Fluors

- When interested in class discovery . . .
  - Usefulness of reverse fluor experiments and replicates will depend on nature and magnitude of both dye bias and experimental variability relative to between subject variability.
  - For some clustering methods, constant dye biases will “wash out”.
  - Some reverse fluors and replicates may be useful as informal quality checks.

cDNA/2-color spotted arrays: Reverse Fluors

- When interested in class prediction . . .
  - Considerations of replicates and reverse fluor experiments are similar to those for the case of class comparisons.

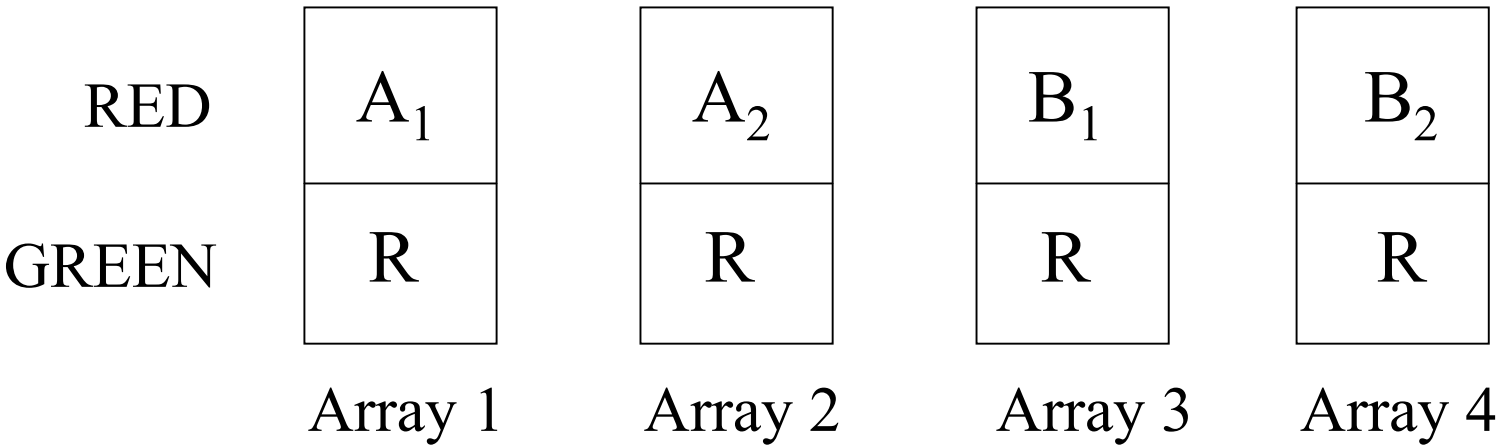
How best to allocate effort?

In many cases it may make a lot of sense to first identify a list of potentially interesting genes, and then verify their expression by other more accurate methods rather than trying to eliminate the biases and noise in the microarray-based measurements by performing many replicate arrays or reverse fluor experiments (2-color arrays).

Class Comparison: Allocation of Specimens to cDNA Array Experiments

- Reference Design
- Loop Design
  - Kerr and Churchill, *Biostatistics,* 2001
- Block Design

Reference Design

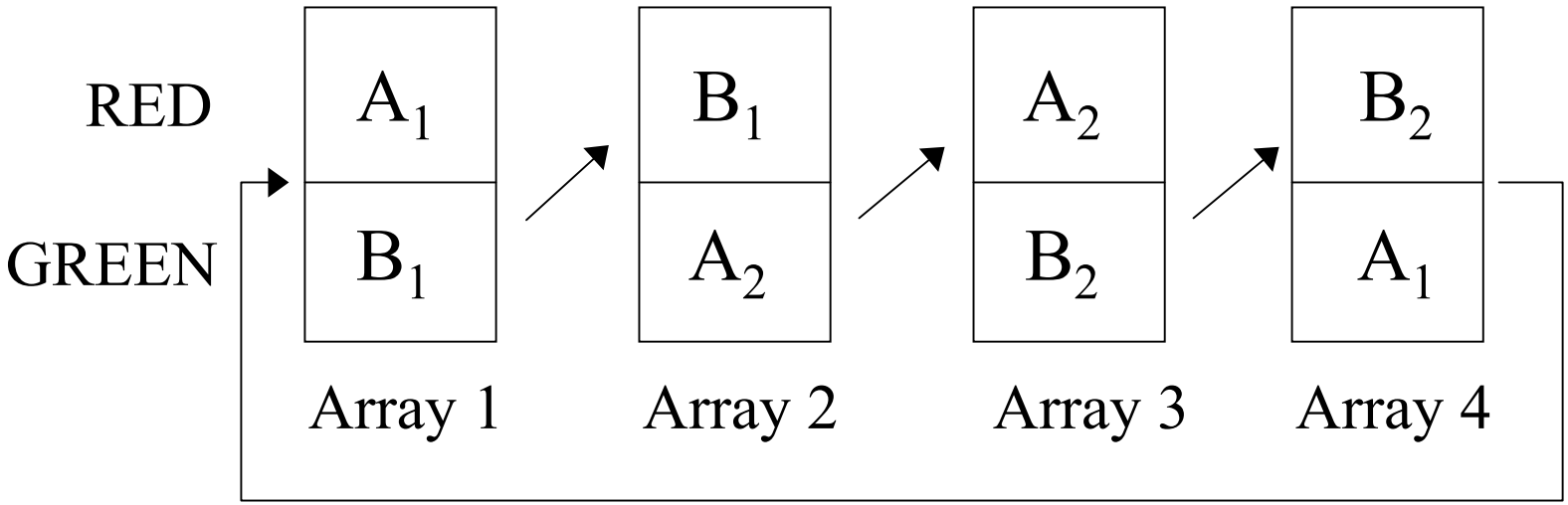


A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

R = aliquot from reference pool

Loop Design

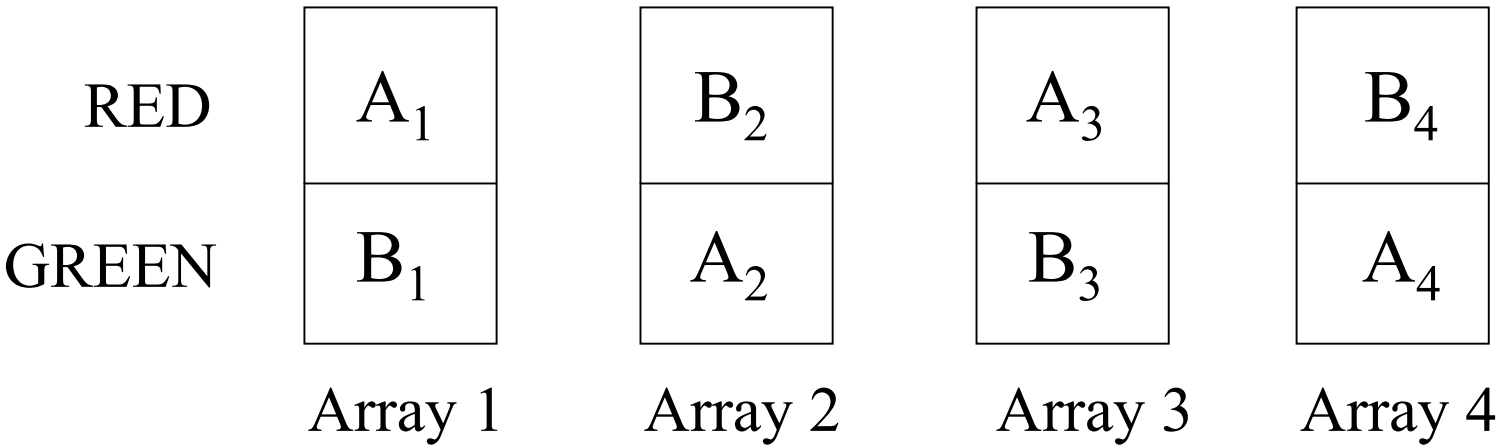


A_i = aliquot from *i*th specimen from class A

B_i = aliquot from *i*th specimen from class B

(Requires two aliquots per specimen)

Block Design



A_i = ith specimen from class A

B_i = ith specimen from class B

Comparison of Designs

- For class discovery, a **Reference** design is preferable because of large gains in cluster performance.
- For class comparisons . . .
  - With a fixed number of arrays, **Block** design is more efficient than **Loop** or **Reference** design, but **Block** design precludes clustering.
  - With a fixed number of specimens, **Reference** design is more efficient than **Loop** or **Block** design when intra-class variance is “large” relative to experimental variation.

Outline

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

Analysis Strategies for Class Comparisons

- Model-based methods
- Global tests
- Multiple testing procedures to identify differentially expressed genes

Model-based Methods for cDNA Arrays

- 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
- Dobbin & Simon, *Bioinformatics, 2002,* in press

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

Multiple Testing Procedures

Identification of differentially expressed genes while controlling for false discoveries (genes declared to be differentially expressed that in truth are not).

- *Actual Number* of False Discoveries: FD
- *Expected Number* of False Discoveries: E(FD)

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

Simple Procedures

- Control expected number of false discoveries
  - $E(FD) \leq u$
  - Conduct each of k tests at level u/k
- Bonferroni control of familywise error (FWE) rate at level α
  - Conduct each of k tests at level α/k
  - At least $(1-\alpha)100\%$ confident that FD = 0

Problems With Simple Procedures

- Bonferroni control of FWE is very conservative
- 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
  - Statistical properties unclear
- 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

- Cluster genes with respect to expression across specimens

- Cluster specimens with respect to gene expression profiles
  - Filter genes that show little variation across specimens
  - Median or mean center genes

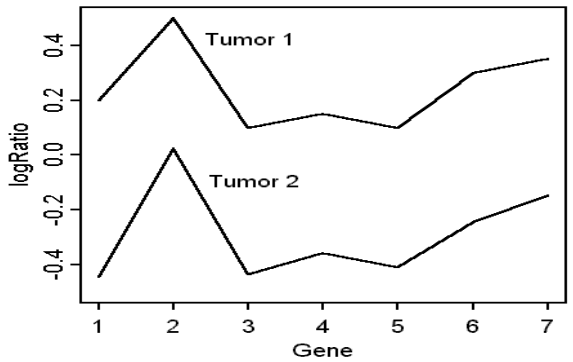
Hierarchical Agglomerative Clustering Algorithm

- 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

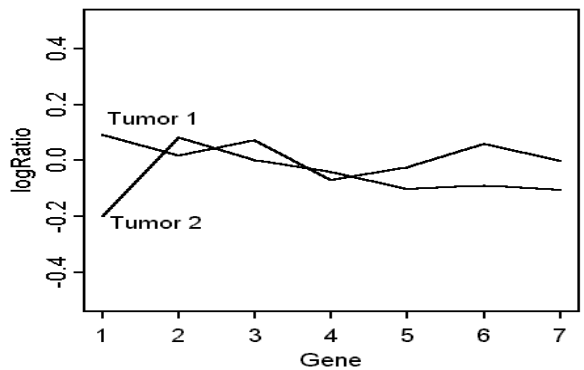
Common Distance Metrics for Hierarchical Clustering

- Euclidean distance
  - Measures absolute distance (square root of sum of squared differences)
- 1-Correlation
  - Large values reflect lack of linear association (pattern dissimilarity)

Euclidean distance large, 1-Correlation small

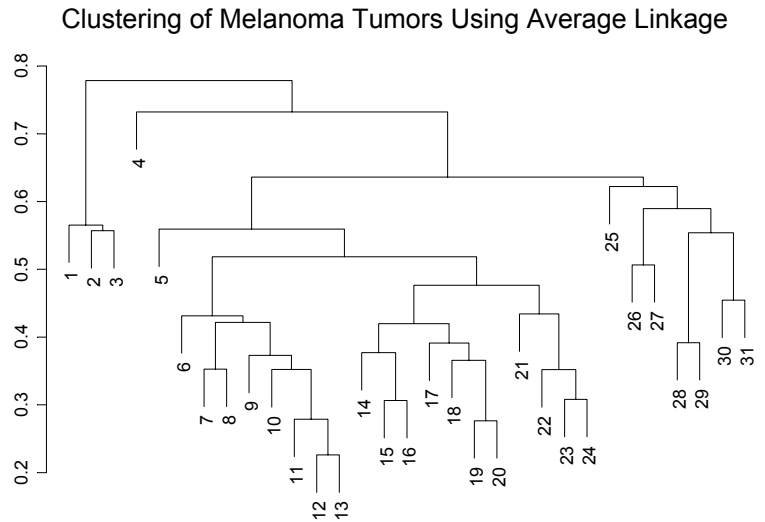


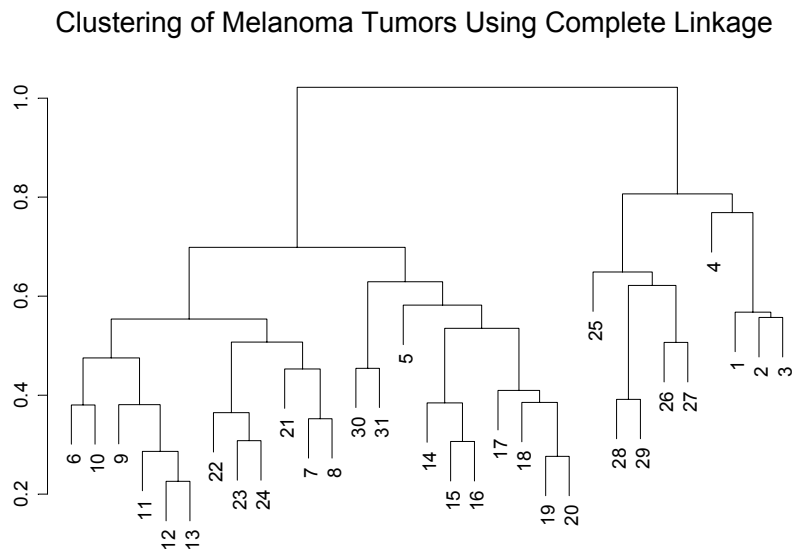
Euclidean distance small, 1-Correlation large

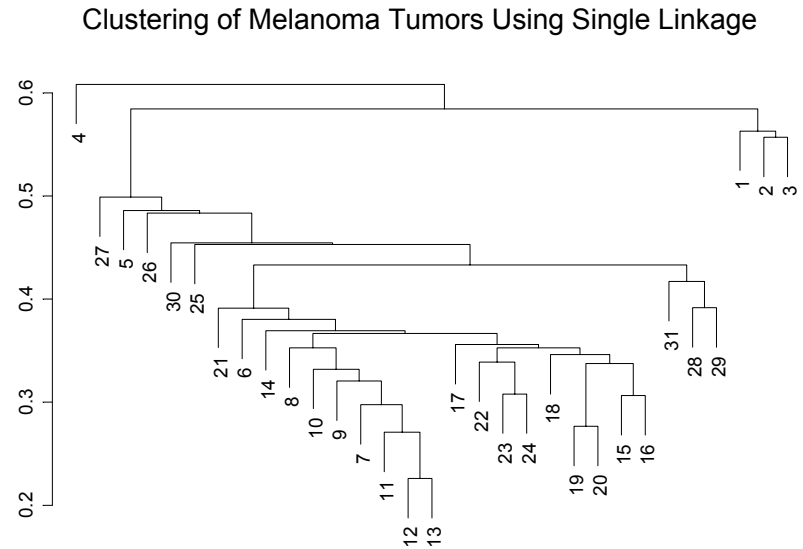


Linkage Methods

- 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
  - Prone to “chaining” and sensitive to noise







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
  - Where to “cut” the dendrogram?
- Different clustering algorithms may find different structure using the same data.
- Which clusters do we believe?
  - Reproducible between methods
  - Reproducible within a method

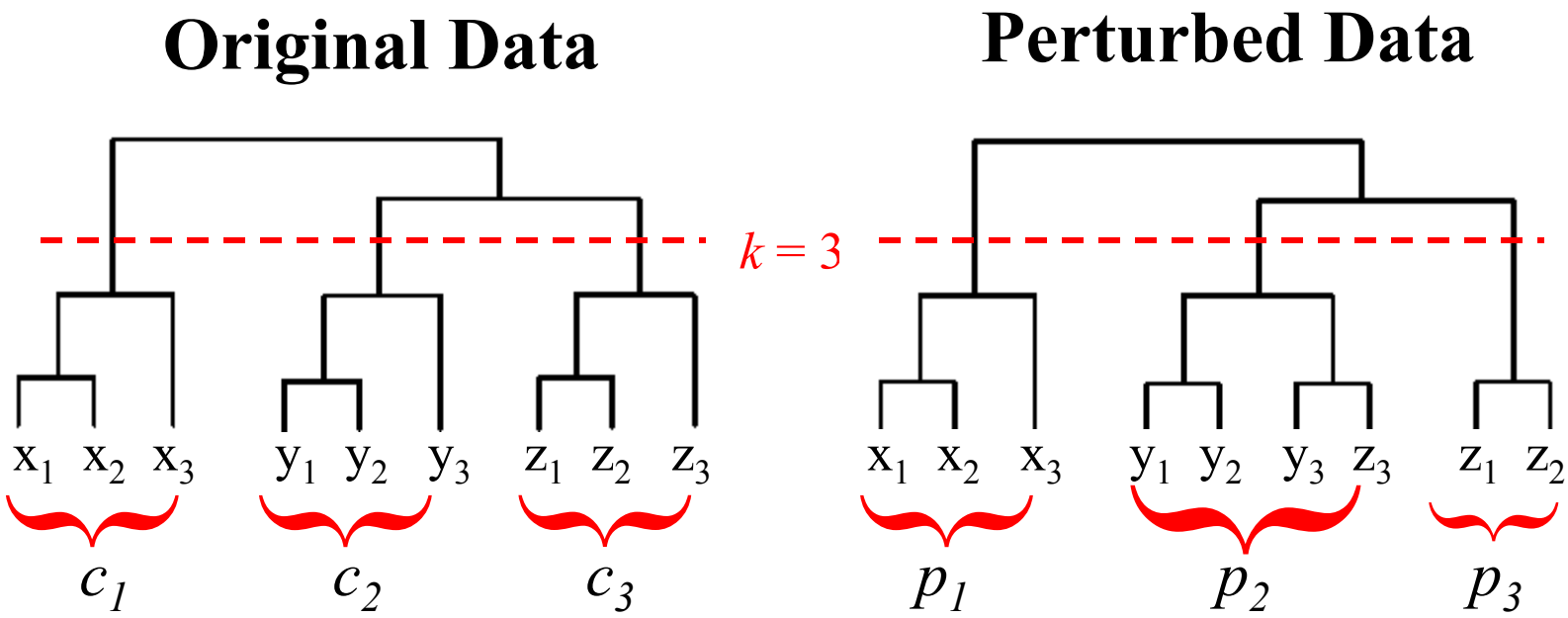
Assessing Cluster Reproducibility: Data Perturbation Methods

- Most believable clusters are those that persist given small perturbations of the data.
  - Perturbations represent an anticipated level of noise in gene expression measurements.
  - Perturbed data sets are generated by adding random errors to each original data point.
    - McShane *et al*., *Bioinformatics,* in press – Gaussian errors
    - Kerr and Churchill, *PNAS*, 2001 – Bootstrap residual errors

Assessing Cluster Reproducibility: Data Perturbation Methods

- Perturb the log-gene measurements and re-cluster.
- For each original cluster:
  - Compute the proportion of elements that occur together in the original cluster and remain together in the perturbed data clustering when cutting dendrogram at the same level *k*.
  - Average the cluster-specific proportions over many perturbed data sets to get an *R-index* for each cluster.

R-index Example

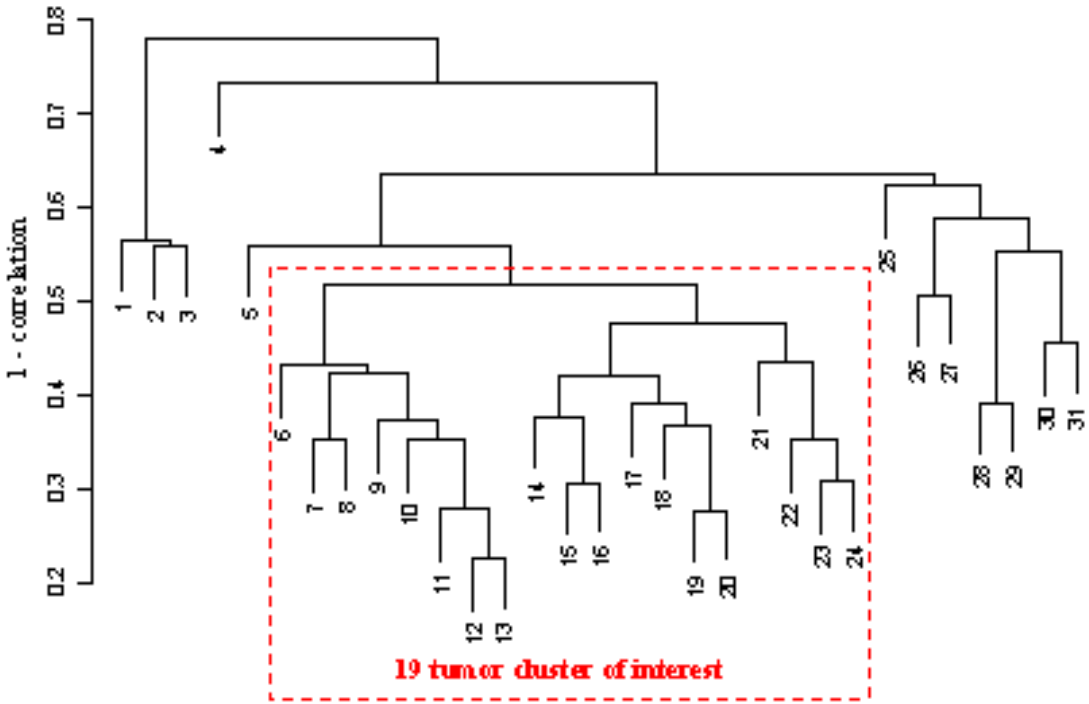


- 3 out of 3 pairs in c_1 remain together in perturbed clustering.
- 3 out of 3 in c_2 remain together.
- 1 out of 3 in c_3 remain together.
- *R*-index = (3 + 3 + 1)/(3 + 3 + 3) = 0.78

Cluster Reproducibility: Melanoma

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

Expression profiles of 31 melanomas were examined with a variety of class discovery methods. A group of 19 melanomas consistently clustered together.



For hierarchical clustering, the cluster of interest had an *R-index* = 1.0.

fi highly reproducible

Melanomas in the 19 element cluster tended to have:

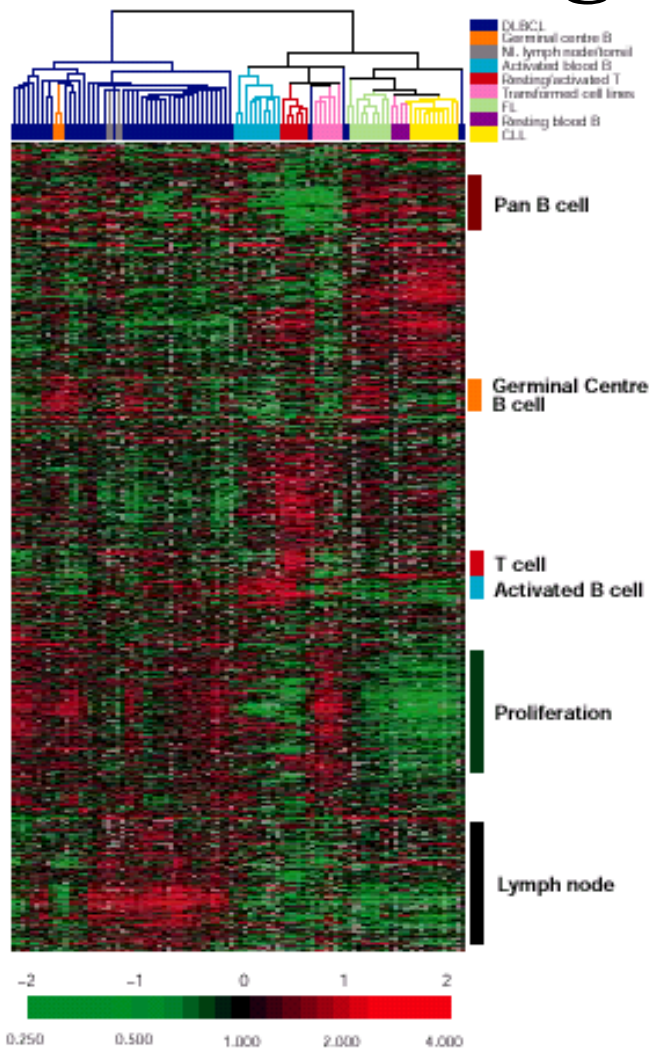
- reduced invasiveness
- reduced motility

Evaluating the Number of Clusters

- Global test of "no clustering" followed by comparison of *R-index* and *D-index* over many cuts in the original dendrogram to assess how many clusters are reproducible (McShane *et al., Bioinformatics,* in press)

- Gap Statistic (Tibshirani *et al*., *JRSS B*, 2002) – estimate number of clusters
- Comparisons of methods for estimating number of clusters in small dimension cases (Milligan and Cooper*, Psychometrika,* 1985)

Graphical Displays: Ordered Color Image Plot



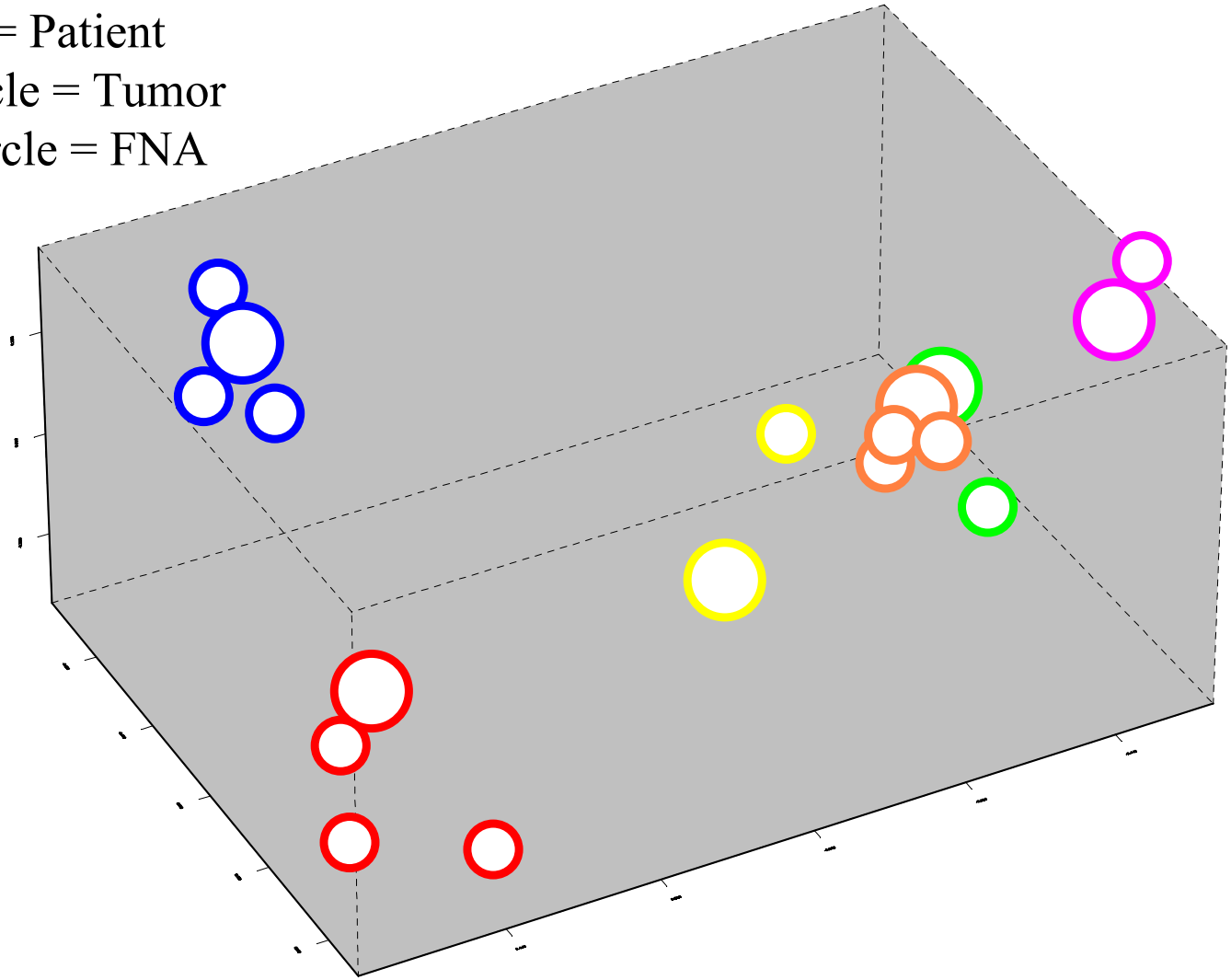
Hierarchical Clustering of Lymphoma Data (Alizadeh *et al., Nature*, 2000)

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 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

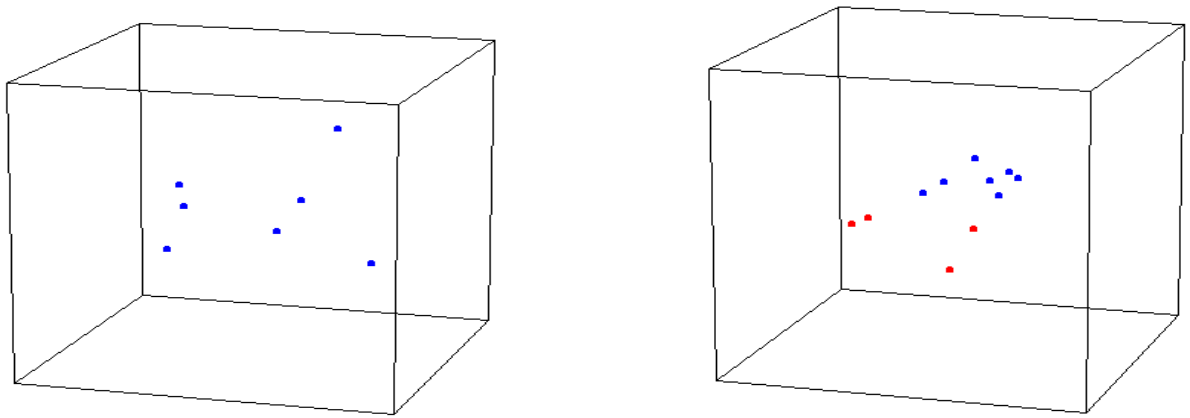
Color = Patient
Large circle = Tumor
Small circle = FNA



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

MDS Representation of Total and Amplified RNA Samples from Same Cell Line

(Fang *et al.*, unpublished)



- There appears to be a difference between total and amplified samples.
- Variability among amplified samples appears larger than variability among total samples.

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; Radmacher *et al., J. Comp. Biology*, in press)

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

(Simon, *et al. JNCI*, in press)

"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 α.

$$CCP_i = t_1 x_{i1} + t_2 x_{i2} + \ldots + t_d x_{id}$$

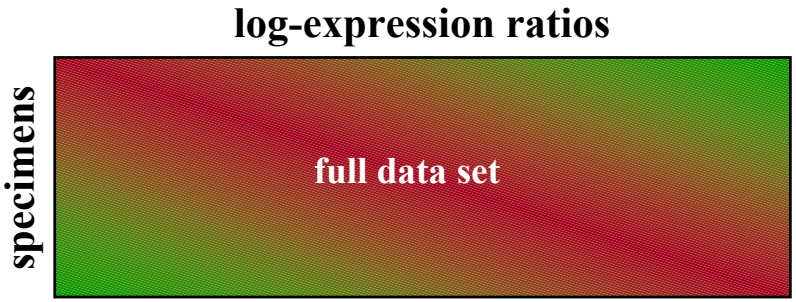
t_j is the two-sample *t*-statistic for gene *j*.

x_{ij} is the log expression measure for gene *j* in sample *i*.

Sum is over all *d* differentially expressed genes.

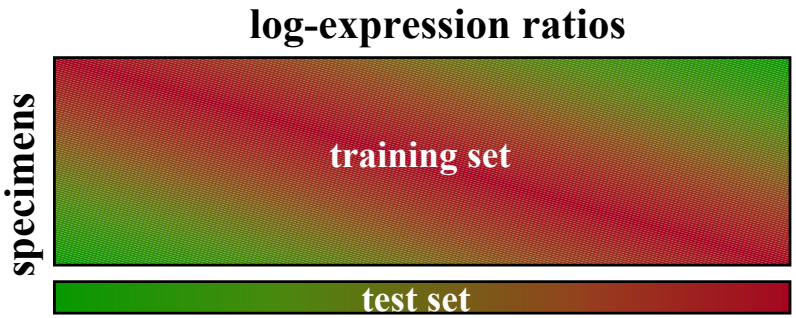
- Threshold of classification: midpoint of the CCP means for the two classes.

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.

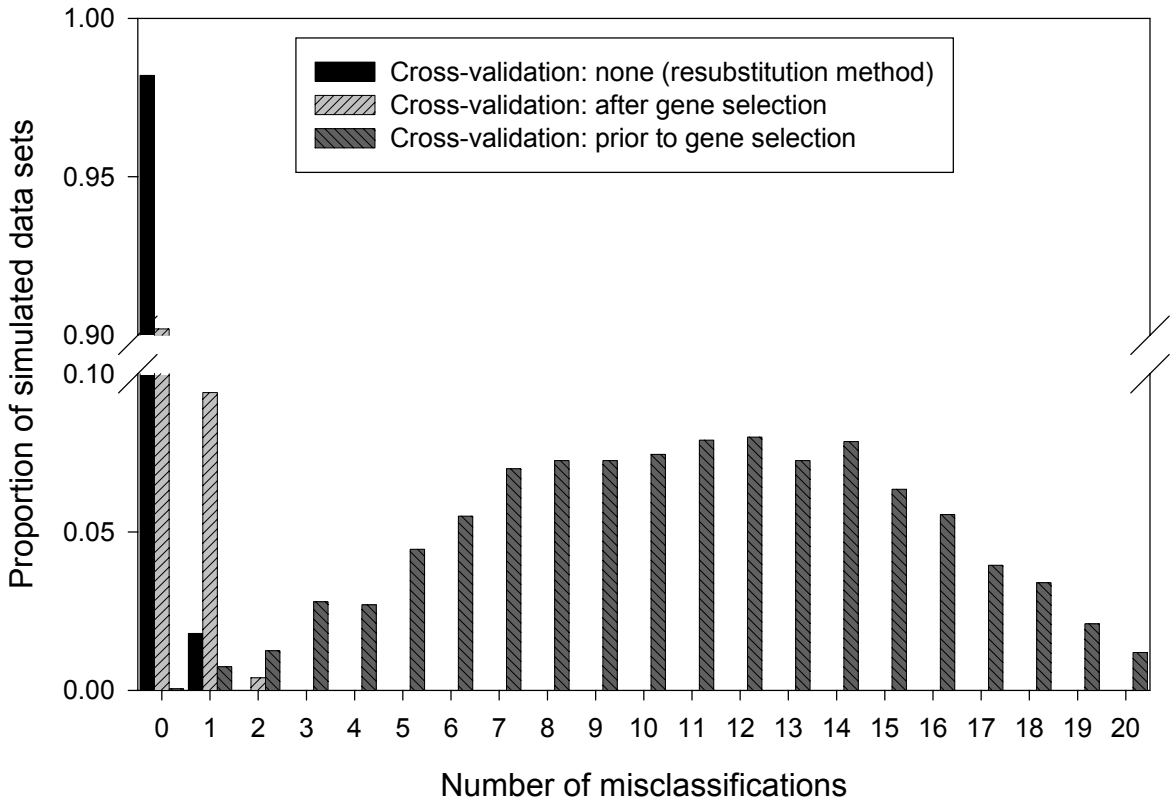
Prediction on Simulated Null Data

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.

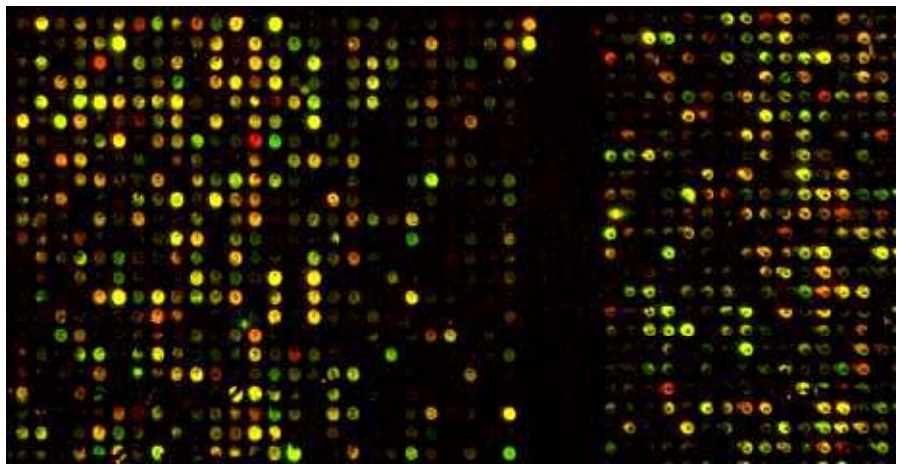

Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
Proportion of simulated data sets
1.00
0.95
0.90
0.10
0.05
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of misclassifications

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?

Classification of hereditary breast cancers with the compound covariate predictor

Class labels	Number of differentially expressed genes [1]	*m* = number of misclassifications [2]	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

[1]Using full data set and significance level $\alpha = .0001$

[2]Using leave-one-out cross-validation.

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)