Statistical Design, Analysis, and Interpretation of Gene Expression Microarray Data

Lisa M. McShane, Ph.D.

August 11, 2003

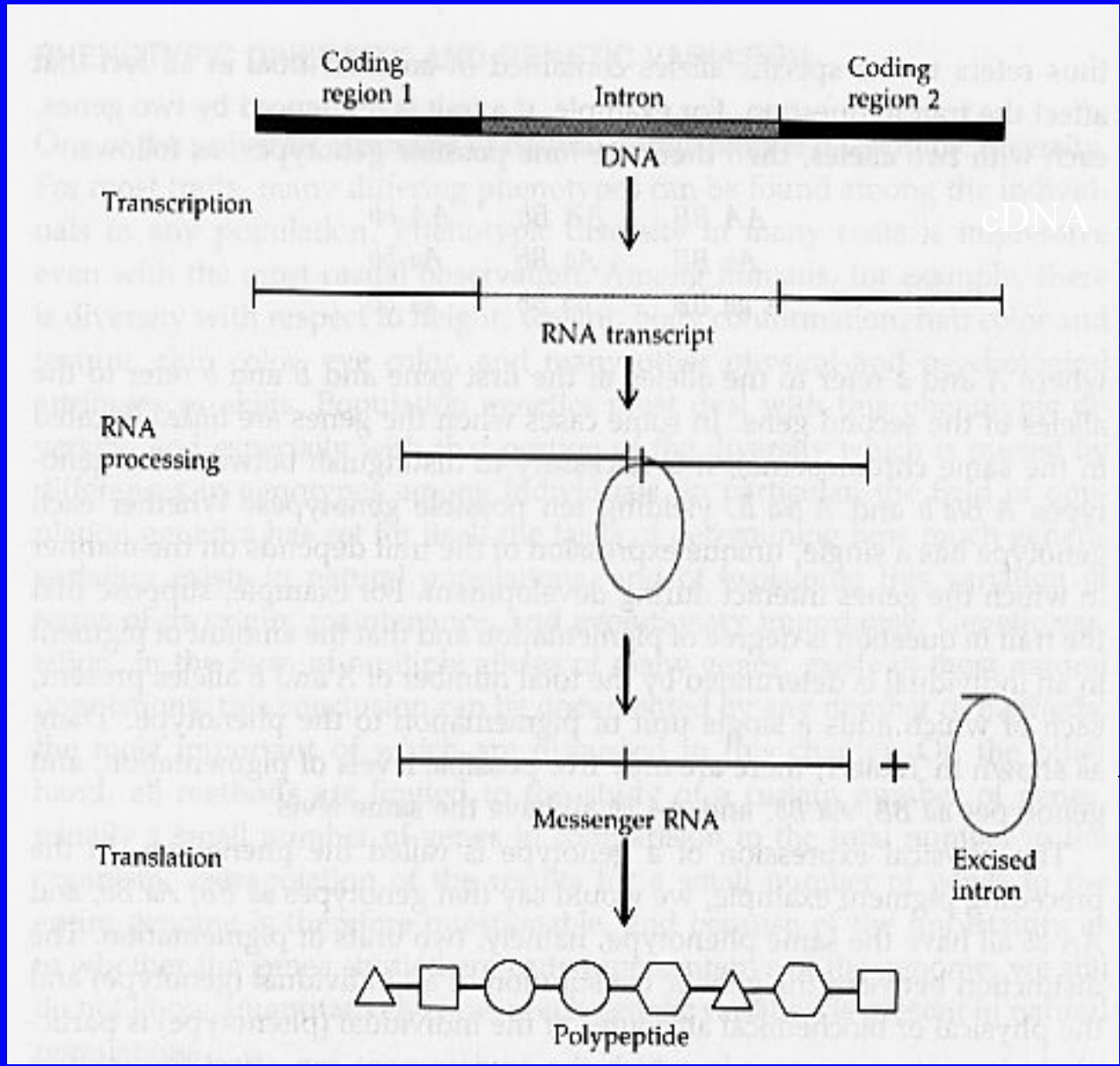
Biometric Research Branch
National Cancer Institute

Outline

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

Outline

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



Reverse transcription

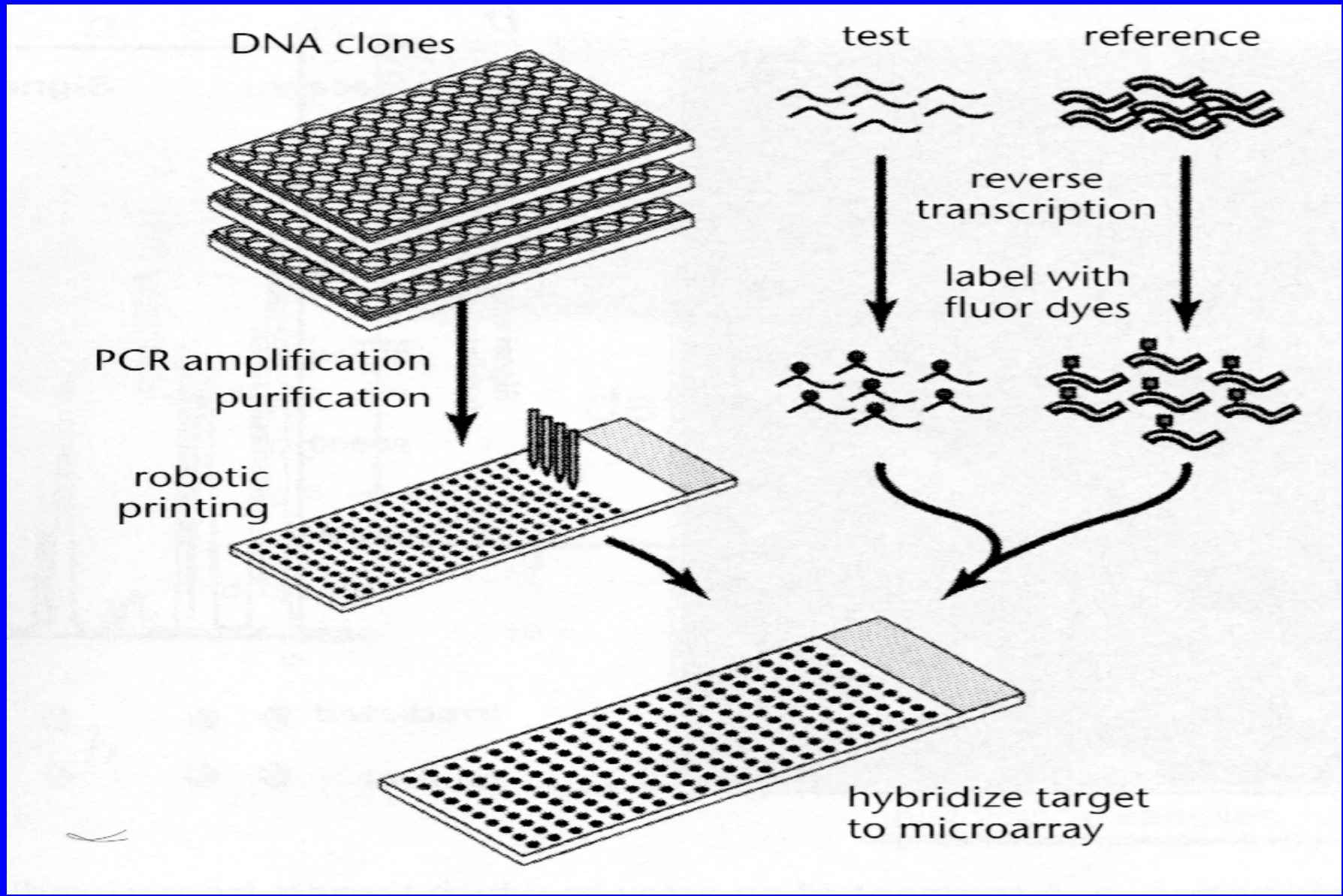
Gene Expression Microarrays

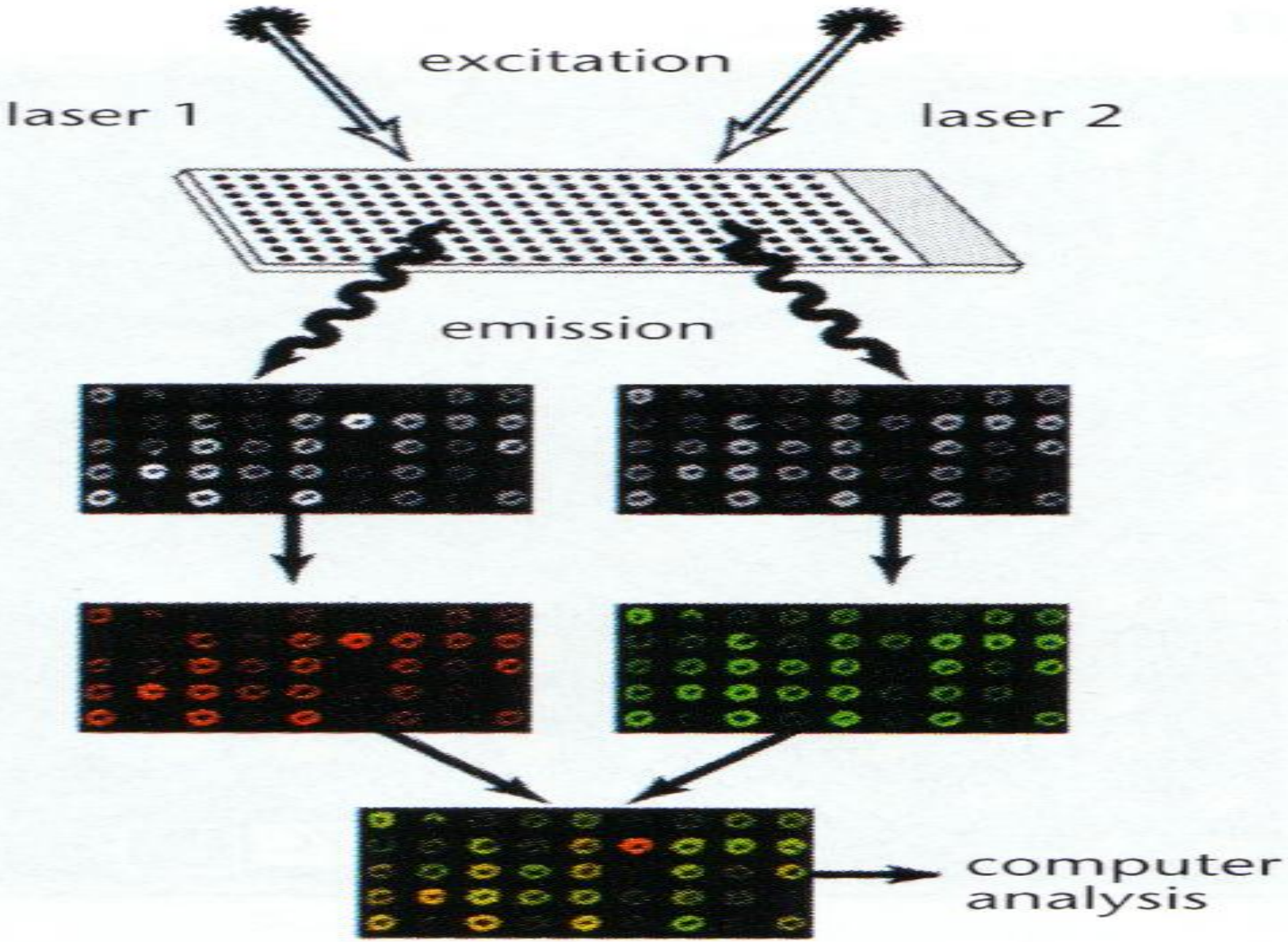
- Permit simultaneous evaluation of expression levels of thousands of genes
- Main platforms
  - Spotted cDNA arrays (2-color)
  - Affymetrix GeneChip (1-color)
  - Spotted oligo arrays (2-color or 1-color)
  - Nylon filter arrays

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

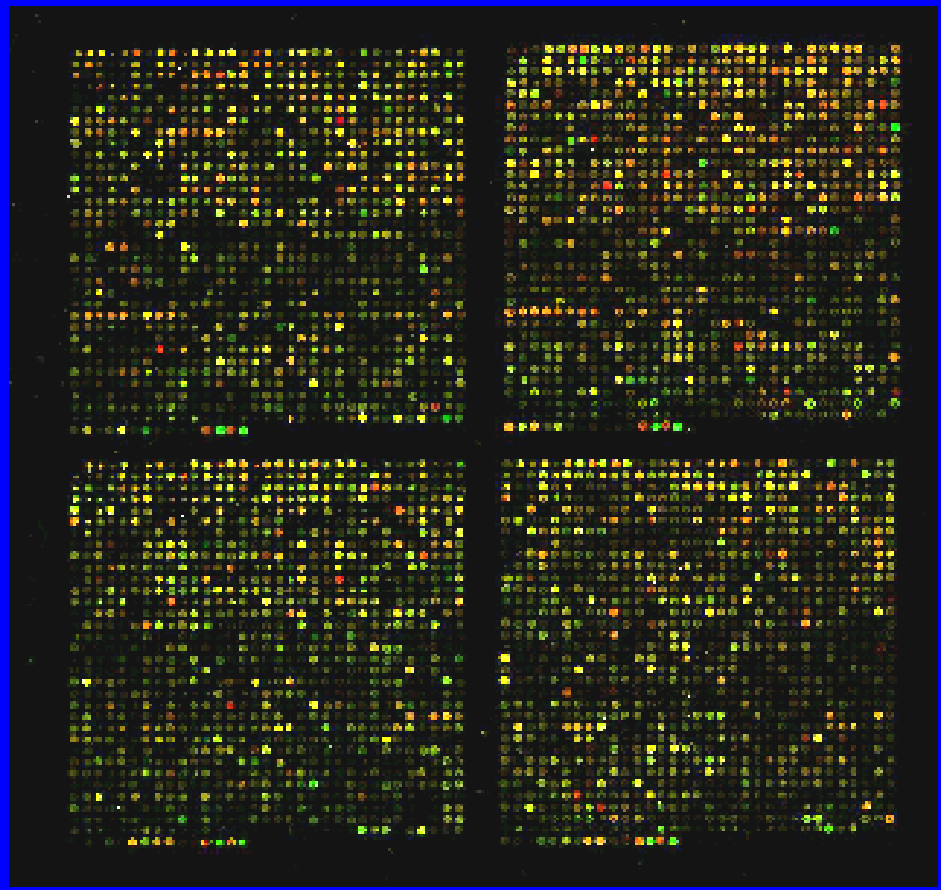
cDNA Arrays


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)

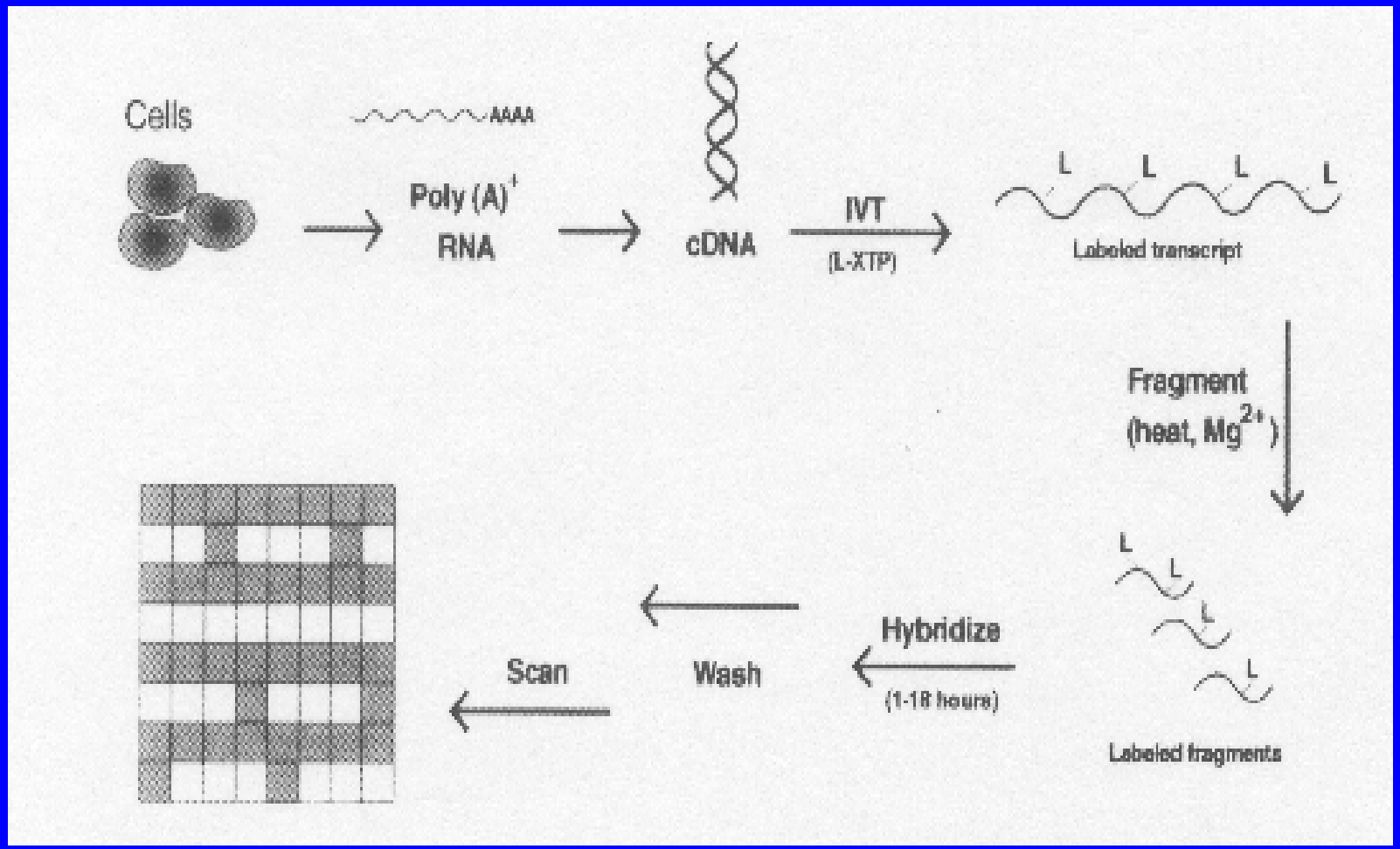


Affymetrix GeneChip Arrays

- 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)

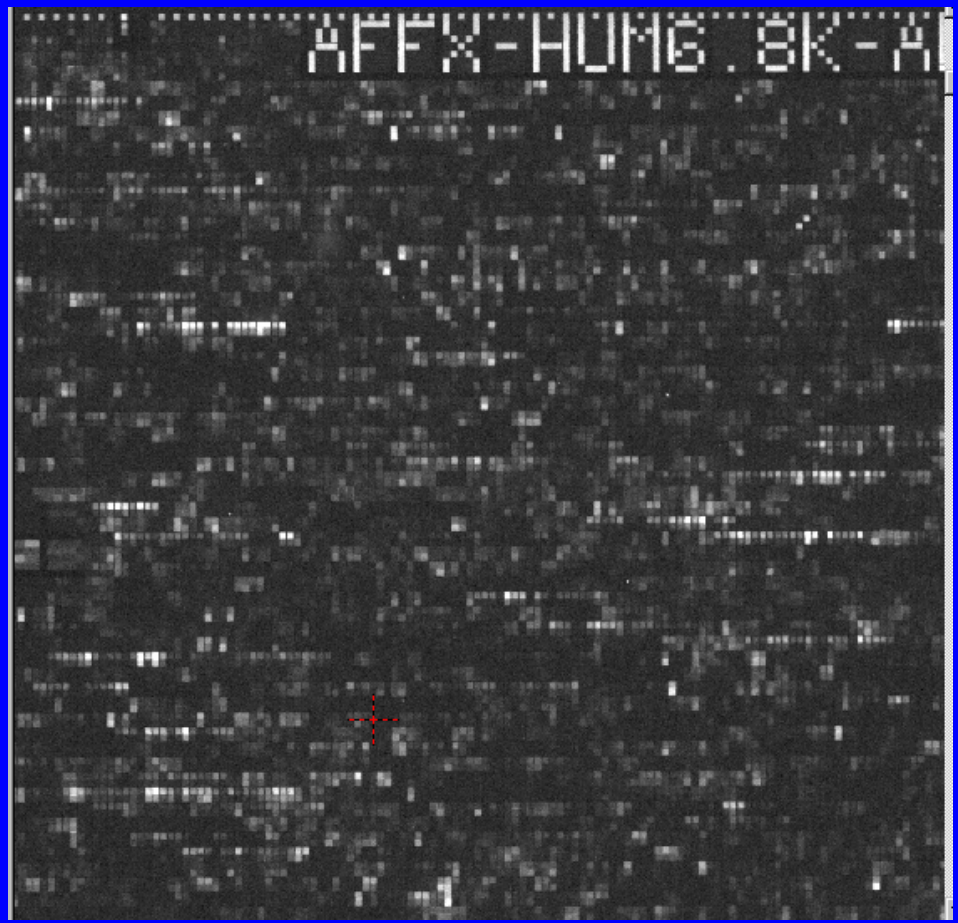
Affymetrix: Assay procedure



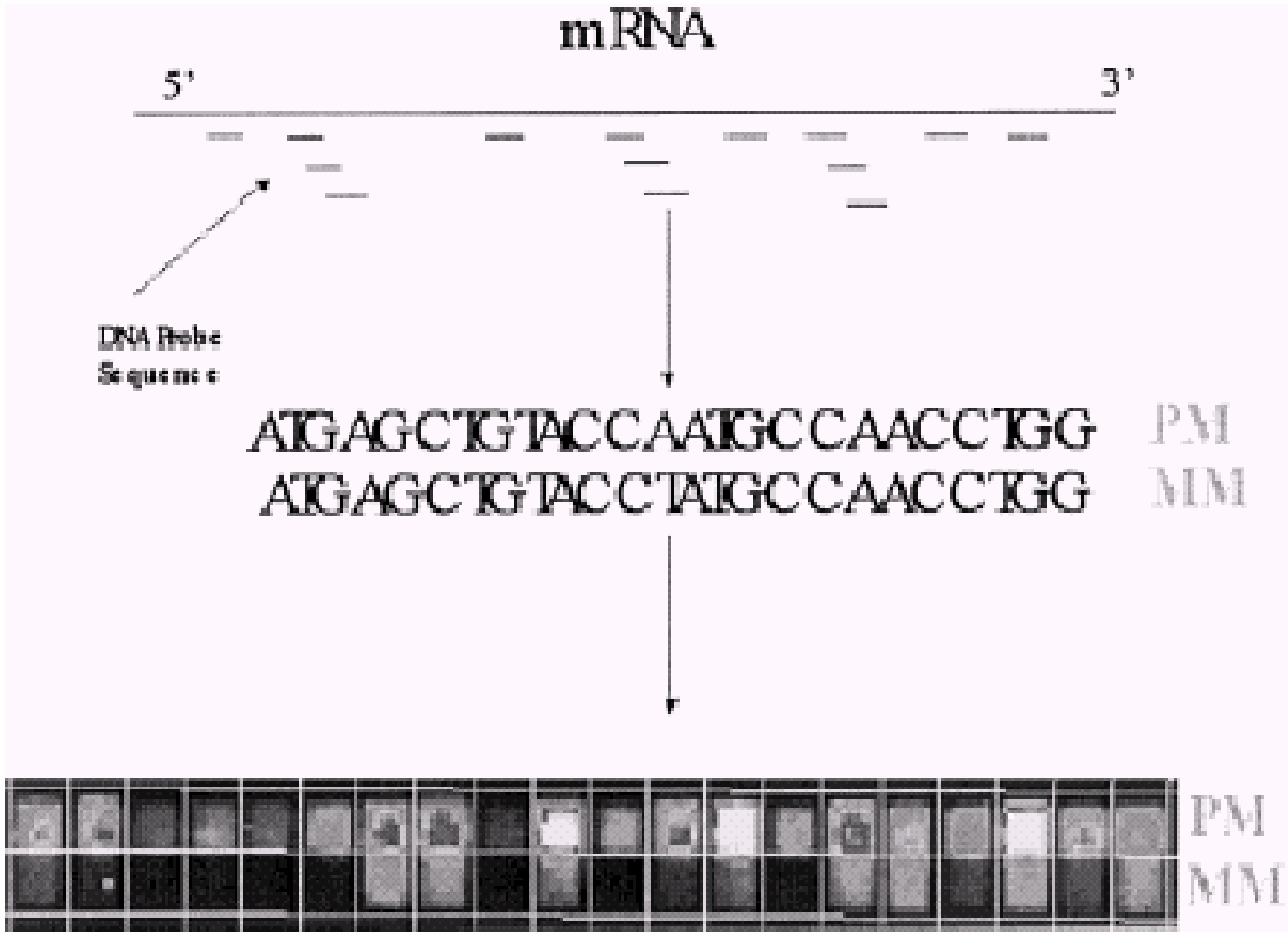
(Figure 1 from Lockhart *et al.*, *Nature Biotechnology*, 1996)

[Affymetrix] Hybridization Oligo “GeneChip” Array

Image of a Scanned Affymetrix GeneChip



Oligo Arrays: Perfect Match - Mismatch Probe Pairs

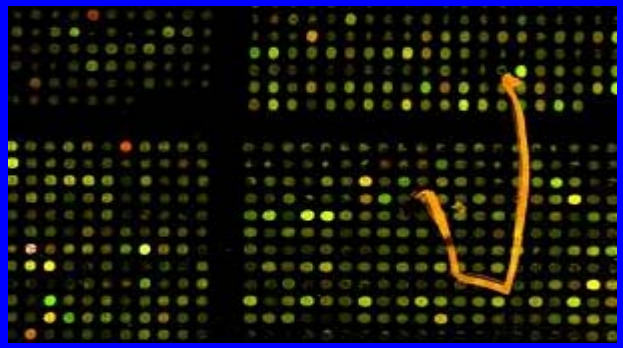


(Figure 2 from Schadt *et al.*, *Journal of Cellular Biochemistry*, 2000)

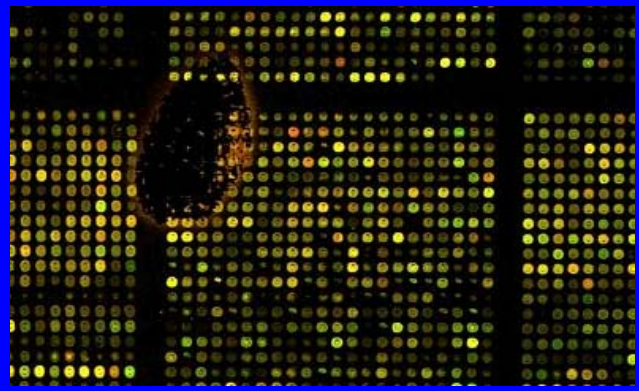
Outline

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

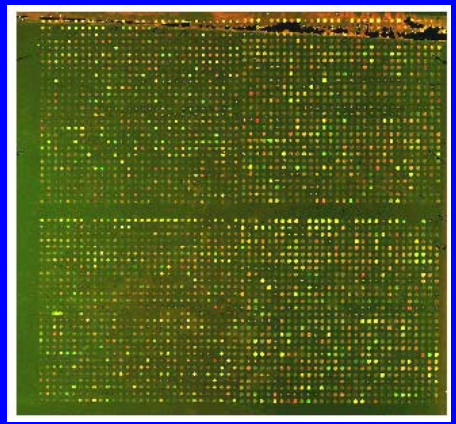
cDNA/Spotted Arrays: Slide Quality



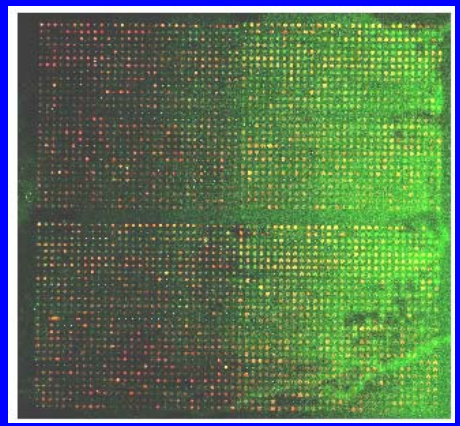
Fiber or scratch?



Bubble



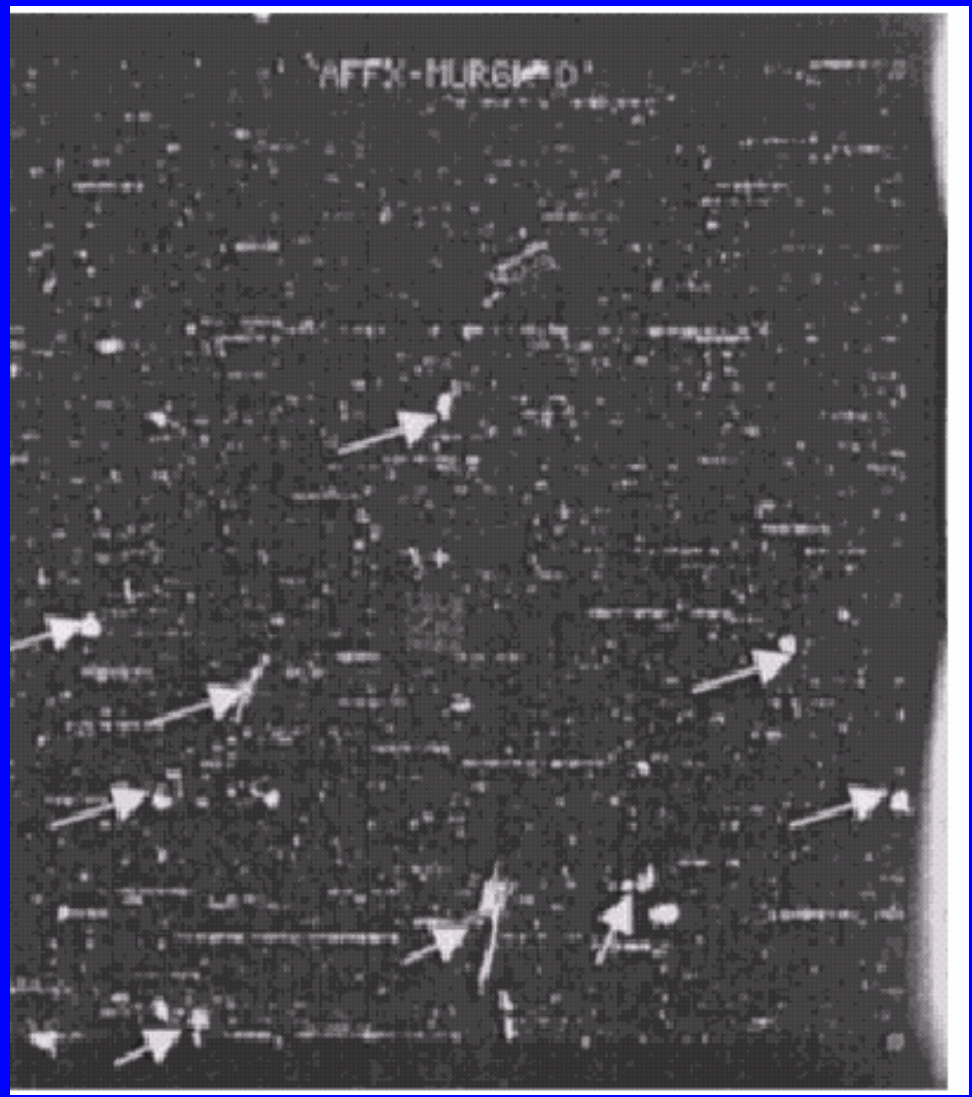
Edge effect



Background haze

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)

- Begin with image consisting of millions of pixels
- Segmentation (F vs B)
- Background correction & signal calculation
- Spot flagging criteria
- Gene(spot)-level summaries
  - Signal ratio (Red/Green)
  - 2 channel signals (Red, Green)

Affymetrix Arrays: Image Processing

- .DAT image files → .CEL files
- Each probe cell: 10x10 pixels
- Grid alignment to probe cells
- Signals:
  - Remove outer 36 pixels → 8x8 pixels
  - The probe cell signal, PM or MM, is the 75th percentile of the 8x8 pixel values
- Background correction: Average of the lowest 2% probe cell values is taken as the background value and subtracted
- Summarize over probe pairs 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

- $AvDiff_i = \Sigma(PM_{ij}-MM_{ij})/n_i$ for each probe set *i*
  (OLD Affymetrix algorithm, MAS 4.0 and earlier)

- New Affymetrix algorithm to address negative signals (MAS 5.0)
  - anti-log of a robust average (Tukey biweight) of the $\log(PM_{ij}-CT_{ij})$, where
    CT=MM, if MM < PM
    = adjusted to be less than PM, if MM≥PM

Affymetrix Arrays:
Model-based Probe Set (Gene) Summaries

- Li and Wong (*PNAS*, 2000; *Genome Biology*, 2001)
  - $MBEI_i = \theta_i$ estimated from $PM_{ij}\text{-}MM_{ij} = \theta_i \varphi_j + \varepsilon_{ij}$ => weighted average difference
  - $MBEI_i^* = \theta_i^*$ estimated from $PM_{ij} = \nu_i + \theta_i^* \varphi_j'$: probe set summaries are based on PM signals only.
- Irizarry *et al.* (*Nucleic Acids Research*, 2003; *Biostatistics*, 2003)
  - $RMA_i = e_i$ estimated from $T(PM_{ij}) = e_i + a_j + \varepsilon_{ij}$, where T(PM) represents the PM intensities which have been background corrected, normalized and log-transformed

cDNA/2-color spotted arrays: Need for Normalization

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

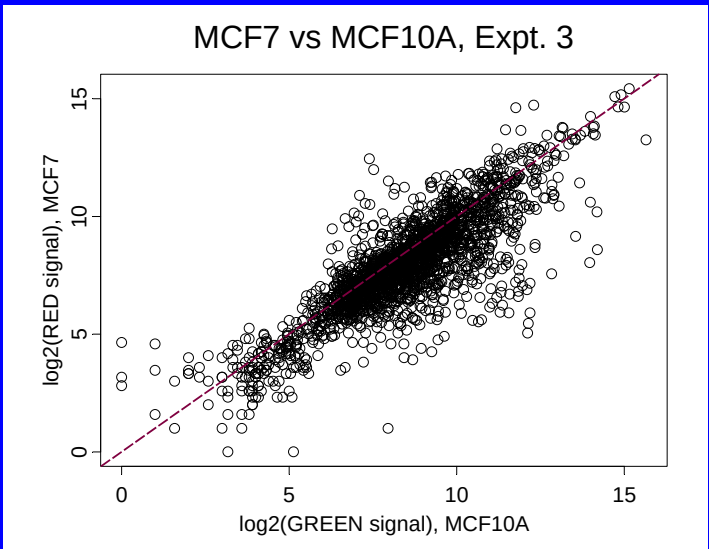
Normalization Methods for cDNA/2-Color Spotted Arrays

- 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.*, *Nucleic Acids Research*, 2002
    - Scaling factors, separately by printer pin, etc.

Median (or Mean) Centering

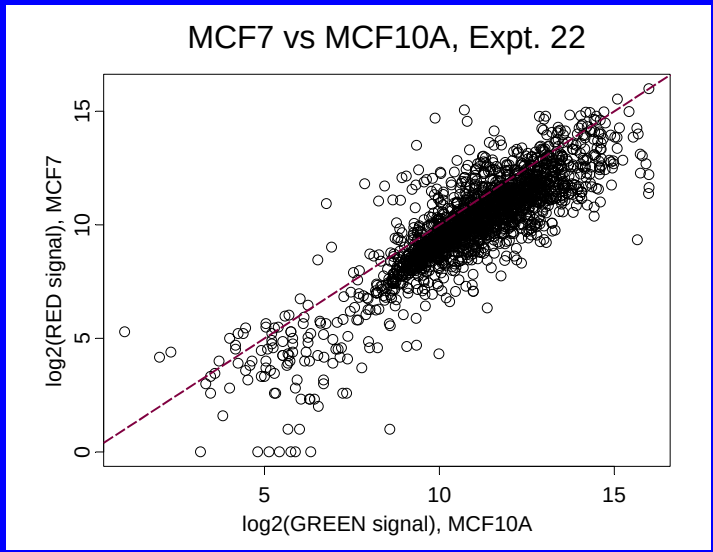


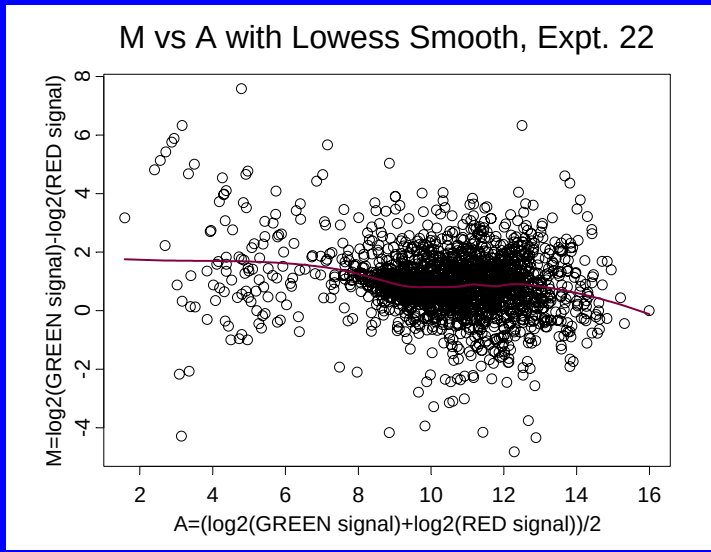
In plot of log(red signal) versus log(green signal), if point scatter is parallel to 45° line, adjust intercept to 0.

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.*, *Nucleic Acids Research*, 2002)

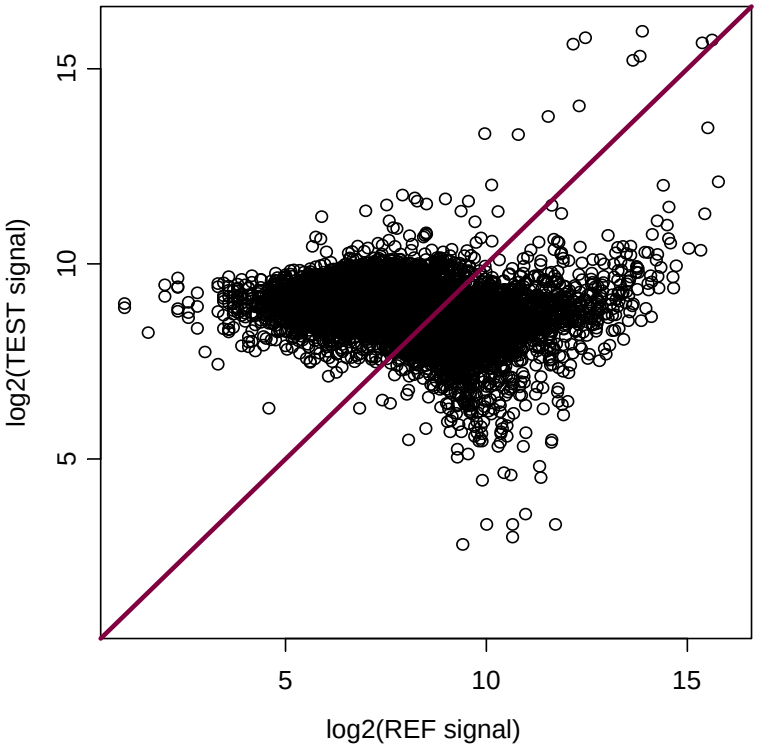


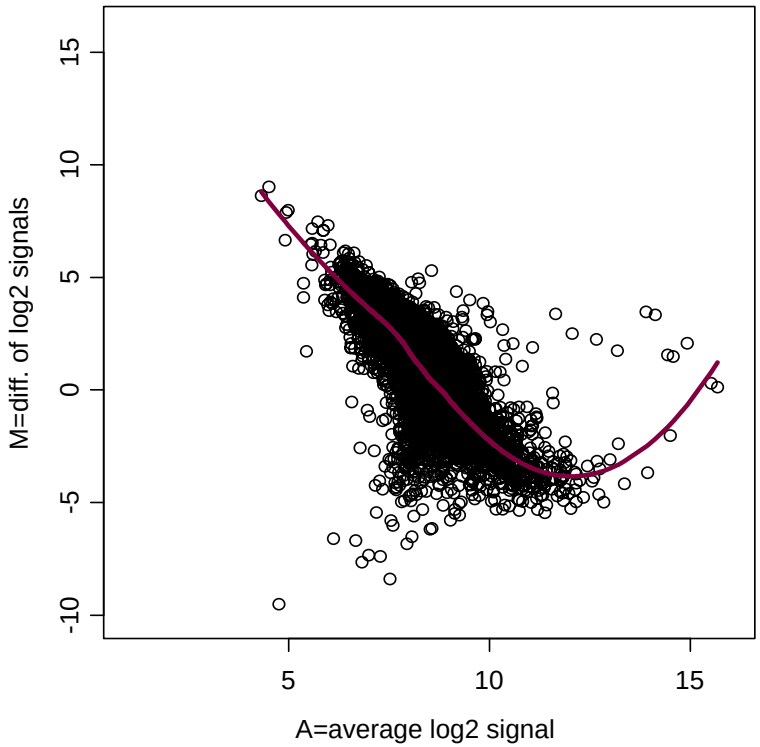


$M = \log_2(\text{GREEN signal}) - \log_2(\text{RED signal})$

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

Bad Arrays and Blind Normalizaton





Normalization: Affymetrix Arrays

- Variations due to sample, chip, hybridization, scanning
- Probe set-level vs probe-level
- Scale factor, intensity dependent, quantile-based, model-based
- Normalize across all arrays or pairwise
- PM-MM vs PM only
- References:
  - Li and Wong (*PNAS*, 2000; *Genome Biology*, 2001)
  - Irizarry *et al*. (*Nucleic Acids Research*, 2003; *Biostatistics*, 2003)
  - Bolstad *et al*. (*Bioinformatics*, 2003)

Filtering Genes

- Prior to cluster analyses only?
- "Bad" gene measurements on too many arrays.
- Not differentially expressed across arrays.
  - Gene variance < threshold
  - Fold change
    - Max/Min < 3 or 4
    - (95^{th} percentile/5^{th} percentile) < 2 or 3
    - Fold change relative to median

Outline

1) Introduction: Biology & Technology
2) Data Quality, Image Processing, Normalization & Filtering
3) Study Objectives & Design Considerations
4) Analysis Strategies Geared to 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
- Pooling
- For cDNA/2-color spotted arrays:
  - Allocation of samples to (cDNA) array experiments
    - Kerr and Churchill, *Biostatistics*, 2001
    - Dobbin and Simon, *Bioinformatics*, 2002
  - Reverse fluor experiments
    - Dobbin, Shih and Simon, *Bioinformatics*, 2003

Design Consideration Highlights

(Dobbin, Shih and Simon, *in press JNCI - Q&As*)

- Many types of "replicates" – biological vs technical replicates & implications for inference
- Reverse fluor "replicates" – might need fewer than you think and sometimes none at all
- Pooling
  - Optimal pooling design (Kendziorski, *Biostatistics*, 2003) depends on
    - Array vs biological sample cost
    - Technical vs biological variability
  - One pool, many aliquots – not smart

Does it pay to replicate arrays?

- It is generally more efficient to assay specimens from additional subjects than it is to perform replicate arrays for the same subjects because one's goal is usually to draw inferences to the population of subjects.
- Some replicates may be helpful as quality checks, but can be misleading if replication covers only some of the sources of variability. Also, poor quality RNA often yields highly reproducible array results!

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.
  - Specify size of difference to detect (2-fold, 3-fold, etc.)
  - Estimates of within-group variance
    - Small for cell lines or genetically identical animals
    - Large for human tumors
  - Set error rates accounting for number of genes examined

Sample Size Planning
Choosing α and β

Let K = # of genes on array

M= # of genes truly differentially expressed at specified fold difference

Expected number of false positives:

$EFP \leq K \times \alpha$ (α = significance level)

Expected number of false negatives for θ-fold genes:

$EFN_\theta = M \times \beta$ ($1-\beta$ = power)

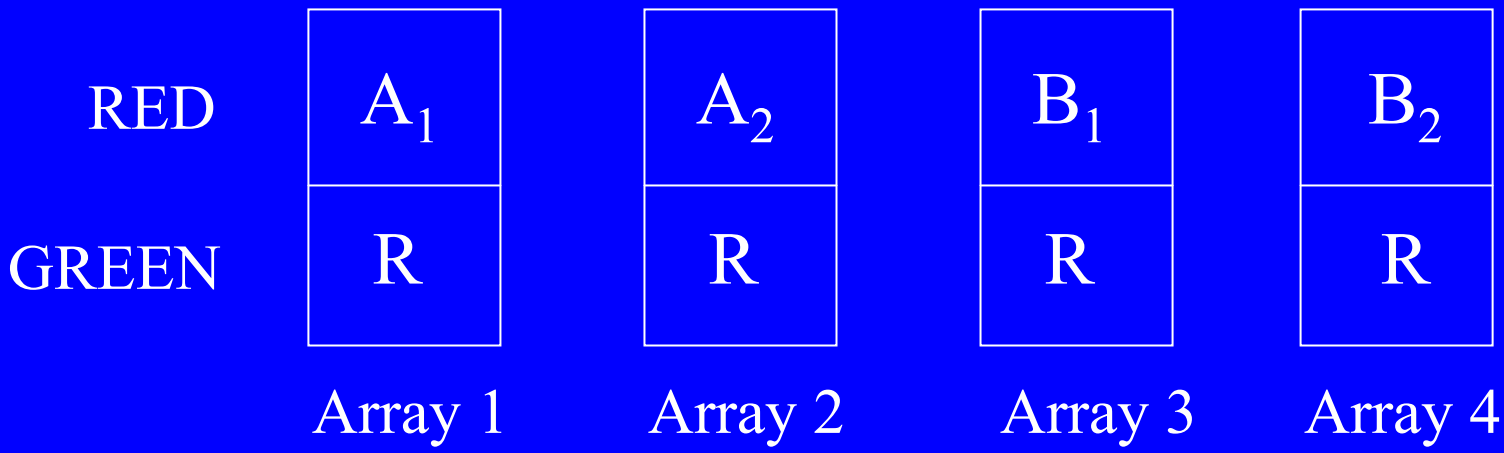
Popular choices for α and β:

$\alpha = 0.001$ $\beta = 0.05$ or 0.10

Class Comparison: Allocation of Specimens to cDNA Array Experiments

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

Reference Design

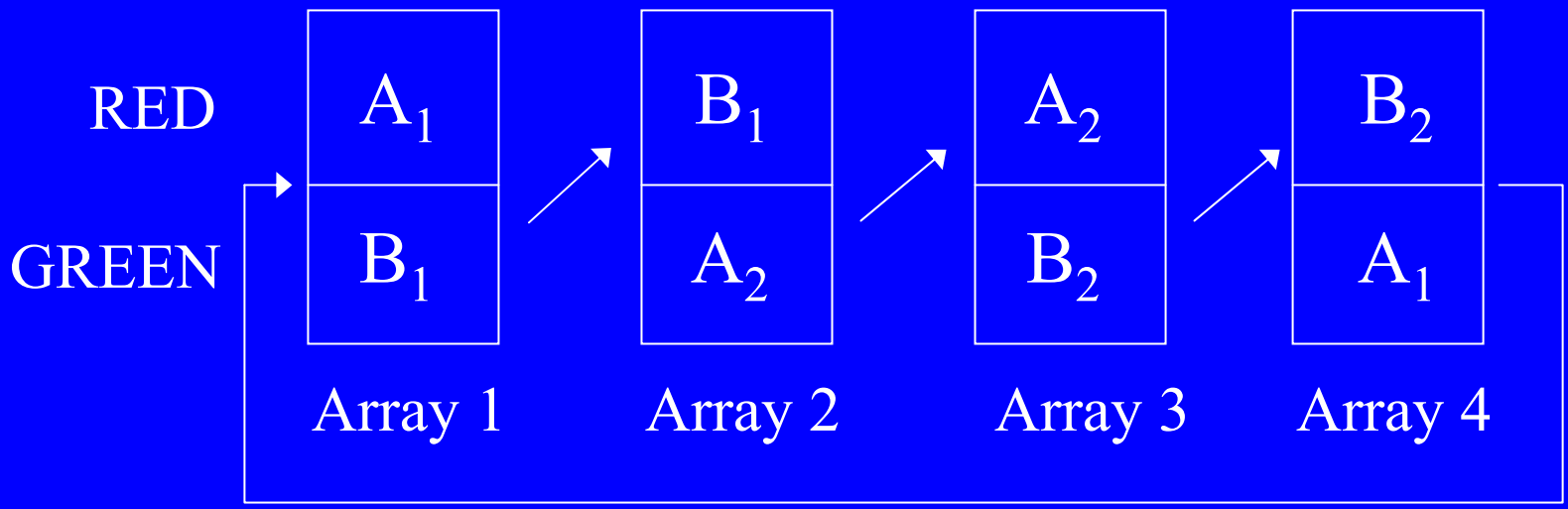


A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

R = aliquot from reference pool

Loop Design

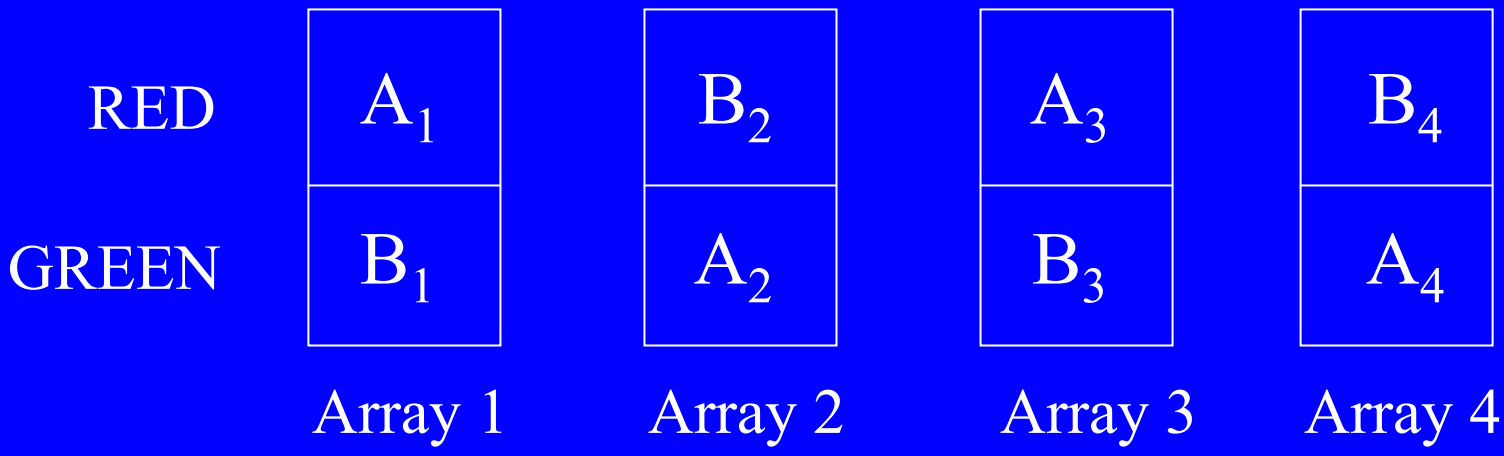


A_i = aliquot from *i*th specimen from class A

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

(Requires two aliquots per specimen)

Block Design



A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

Comparison of Designs

(Dobbin and Simon, *Bioinformatics*, 2002)

- 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 inter-sample variance is "large" relative to intra-sample variation.

Outline

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

Class Comparison Analysis

- Goal
  - Global tests
  - Multiple testing - identify differentially expressed genes
- Methods
  - Non-model-based
    - Log ratios
    - T-tests, F-tests, nonparametric tests, resampling methods
    - Usually reference design
  - Model-based
    - Individual channel data (ANOVA models)
    - Bayesian

Model-based ANOVA 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*

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

- Review by Dudoit *et al*. (*Statistical Science*, 2003)
- "SAM" - Significance Analysis of Microarrays
  - Tusher *et al*., *PNAS*, 2001 and relatives
  - Estimate "FDR-like" quantities
  - Algorithm is a moving target – user beware
- Bayesian
  - Efron *et al*., *JASA*, 2001; *Stanford Tech Rep*, 2001
  - Manduchi *et al*., *Bioinformatics* 2000
  - Newton *et al*., *J Comp Biology* 2001
- Step-down permutation procedures
  - Westfall and Young, 1993 Wiley (FWE)
  - Korn *et al*., *JSPI*, in press (FD and FDP control)

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

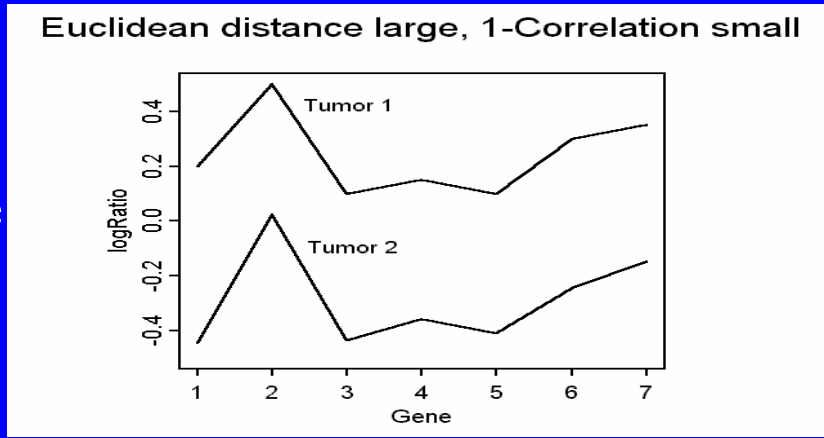
- Cluster genes with respect to expression across specimens
- Cluster specimens with respect to gene expression profiles
  - Filter genes that show little variation across specimens
  - Median or mean center genes

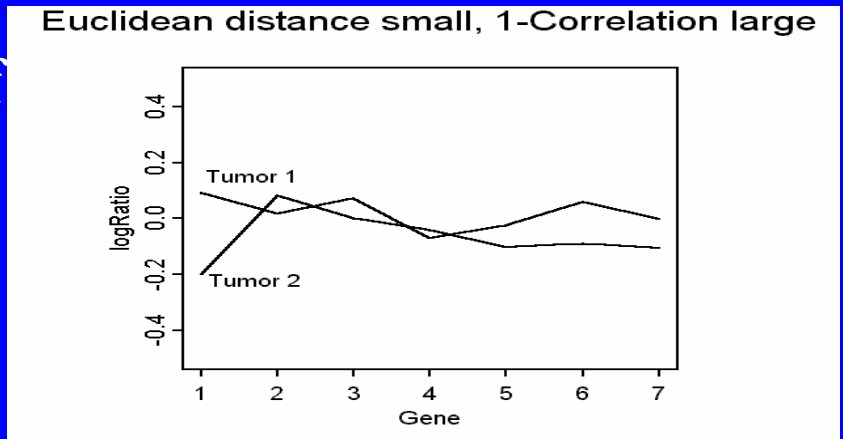
Hierarchical Agglomerative Clustering Algorithm

- 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

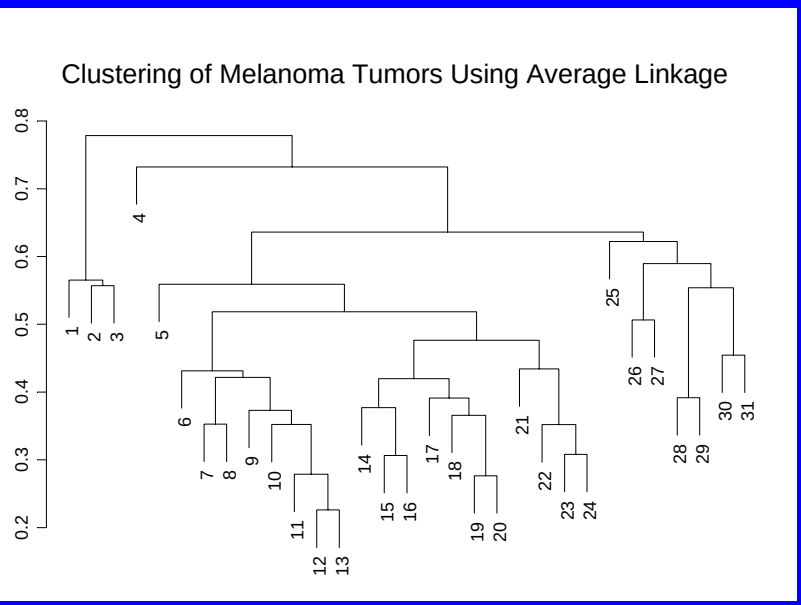
- Euclidean distance
  - Measures absolute distance (square root of sum of squared differences)
- 1-Correlation
  - Large values reflect lack of linear association (pattern dissimilarity)

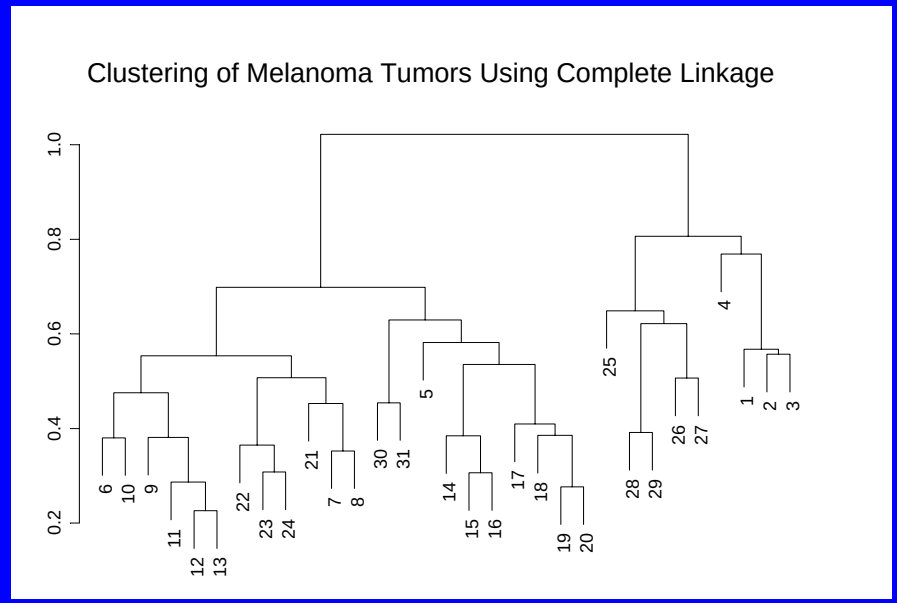


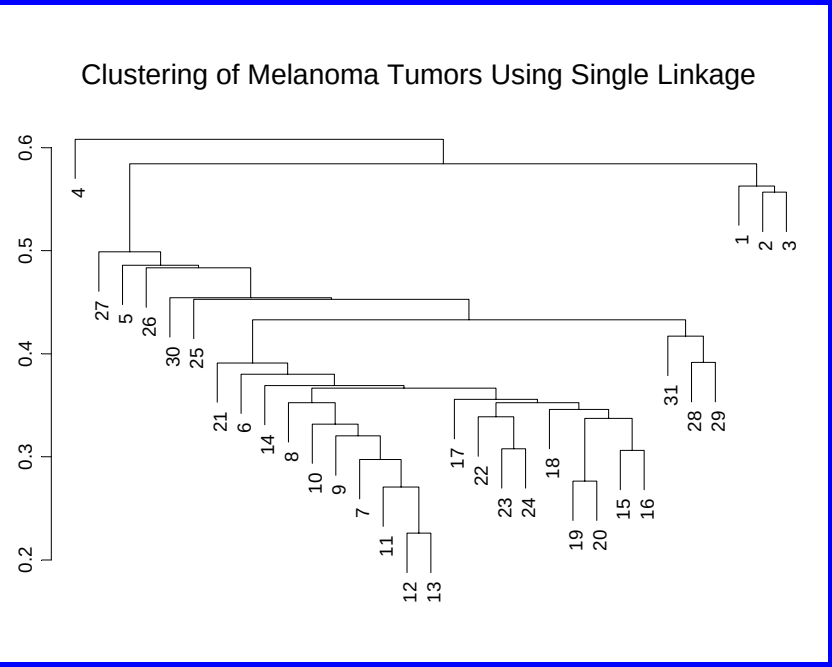


Linkage Methods

- 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





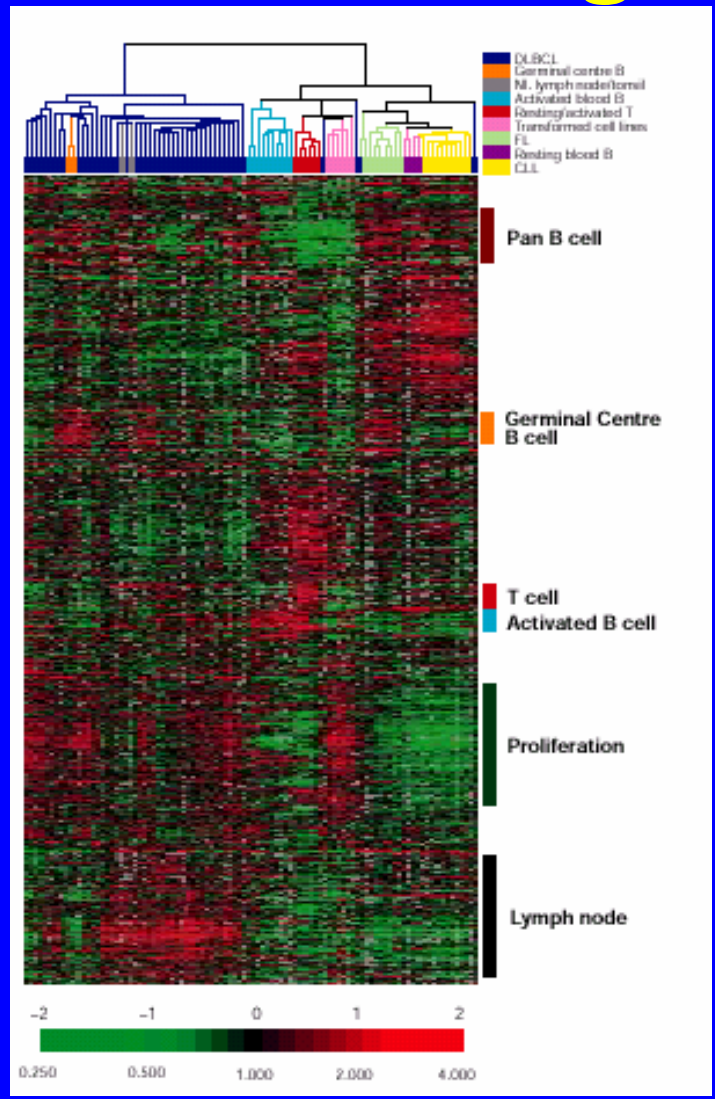


Dendrograms using 3 different linkage methods,
distance = 1-correlation

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

Graphical Displays: Ordered Color Image Plot

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



Interpretation of Cluster Analysis Results

- Cluster analyses always produce cluster structure
  - Where to "cut" the dendrogram?
  - Which clusters do we believe?
- Circular reasoning
  - Clustering using only genes found significantly different between two classes
  - "Validating" clusters by testing for differences between subgroups observed to segregate in cluster analysis
- Different clustering algorithms may find different structure using the same data

Assessing Clustering Results

- Data perturbation methods
  - McShane *et al.*, *Bioinformatics*, 2002 – Gaussian errors (global test + cluster-specific assessment)
  - Kerr and Churchill, *PNAS*, 2001 – Bootstrap residual errors
- Estimating the number of clusters
  - GAP statistic (Tibshirani *et al.*, *JRSS B*, 2002) – DOES NOT WORK!
  - Yueng *et al.* (*Bioinformatics*, 2001) – jackknife method, estimate # of genes clusters
  - Dudoit *et al.* (*Genome Biology*, 2002) – prediction-based resampling

Class Prediction Methods

Comparison of linear discriminant analysis, NN classifiers, classification trees, bagging, and boosting (Dudoit, *et al.*, *JASA*, 2002)

Weighted voting method (Golub *et al.*, *Science*, 1999)

Compound covariate prediction (Hedenfalk *et al.*, *NEJM*, 2001; Radmacher *et al.*, *J. Comp. Biology*, 2002)

Support vector machines (Furey *et al.*, *Bioinformatics*, 2000)

Neural Networks (Khan *et al.*, *Nature Medicine*, 2001)

Nearest Shrunken Centroids (Tibshirani *et al.*, *Statistical Science*, 2003)

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.

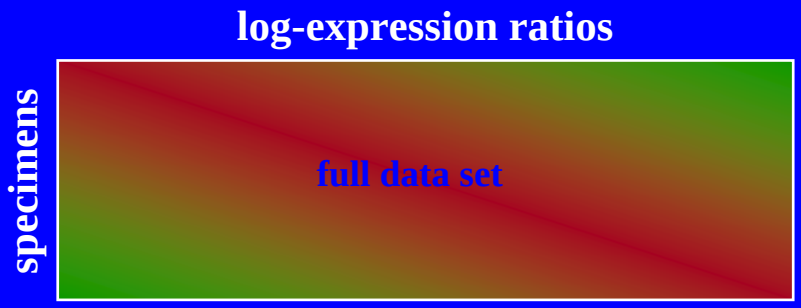
Pitfalls in Class Prediction for Microarray Data

(Radmacher *et al.*, *J Comp Biology*, 2002; Simon *et al.*, *JNCI*, 2003)

- Highly complex models prone to overfitting to data
- Internal validation performed improperly
  - Must include re-selection of features (genes)
  - Cross-validated predictions are not independent (can't treat cross-validated error rate as a binomial proportion)
- Lack of appropriate and sufficiently large independent (external) "validation" sets
  - Free of hidden biases

Assessing Prediction Accuracy

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.

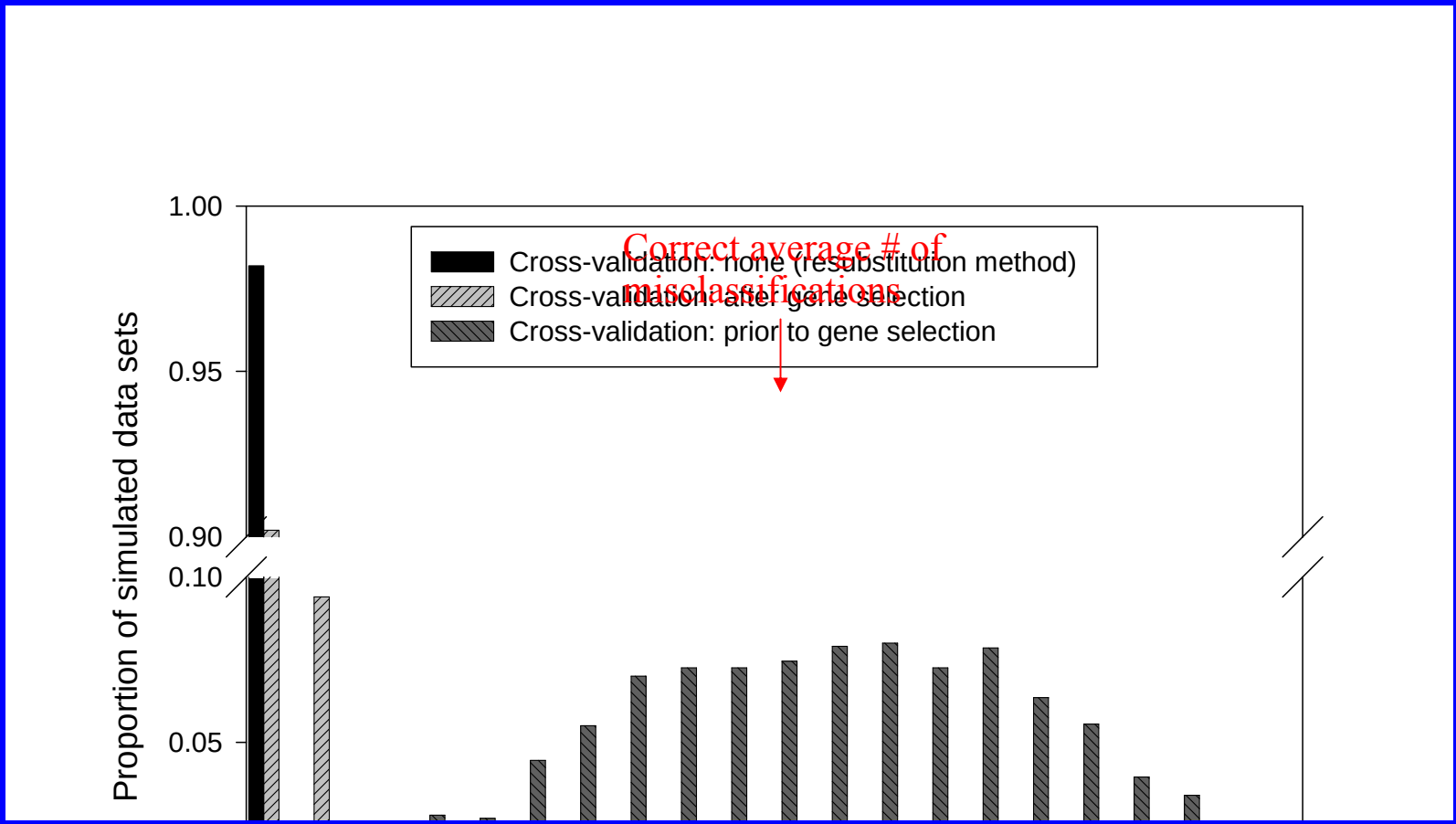
Prediction on Simulated Null Data

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.

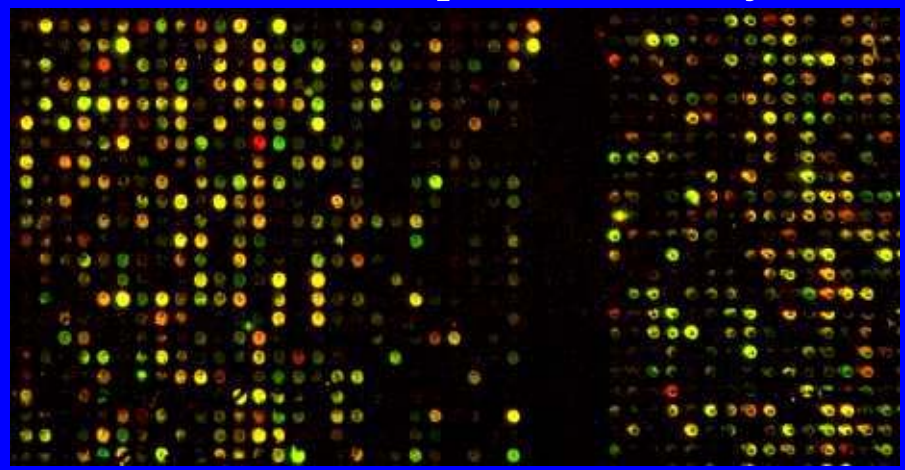

1.00
0.95
0.90
0.10
0.05
Proportion of simulated data sets
Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
Correct average # of misclassifications

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?

Hereditary Breast Cancers: Class Prediction Example

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.

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 can't compensate for poorly designed experiments.
- Fancy analysis tools don't necessarily outperform simple ones.
- Even the best analysis tools, if applied inappropriately, can produce incorrect or misleading results.

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/staff/churchill/labsite/
- Stanford:
  - http://genome-www5.stanford.edu/MicroArray/SMD/restech.html
  - http://www-stat.stanford.edu/~tibs/ (R. Tibshirani)
- Bioconductor: http://www.bioconductor.org/
  - R-based, open source pre-processing and analysis tools

Acknowledgements

- Richard Simon
- Joanna Shih
- Kevin Dobbin
- Michael Radmacher
- Other Members of the NCI Biometric Research Branch
- My NCI collaborators and students in the NIH microarray classes