Statistical Analysis of Gene Expression Microarray Data

Lisa M. McShane*
Joanna H. Shih
Kevin K. Dobbin

Biometric Research Branch
National Cancer Institute

*Lecturing on October 26, 2005

Class Overview

- Day 1: Discussion of statistical analysis of microarray data – Lisa M. McShane
- Day 2: Hands-on BRB ArrayTools workshop – Amy Lam & Supriya Menezes

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

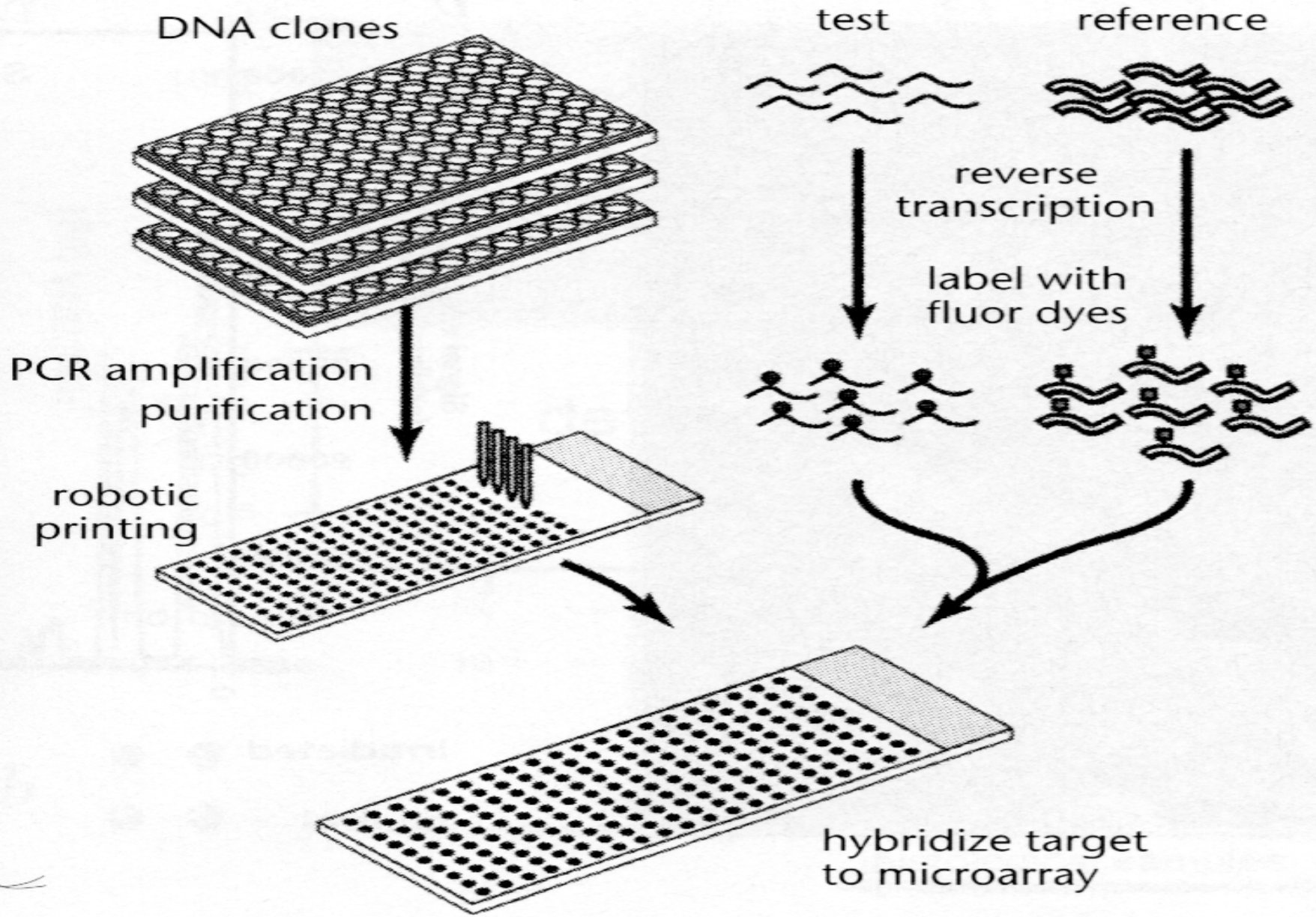
Gene Expression Microarrays

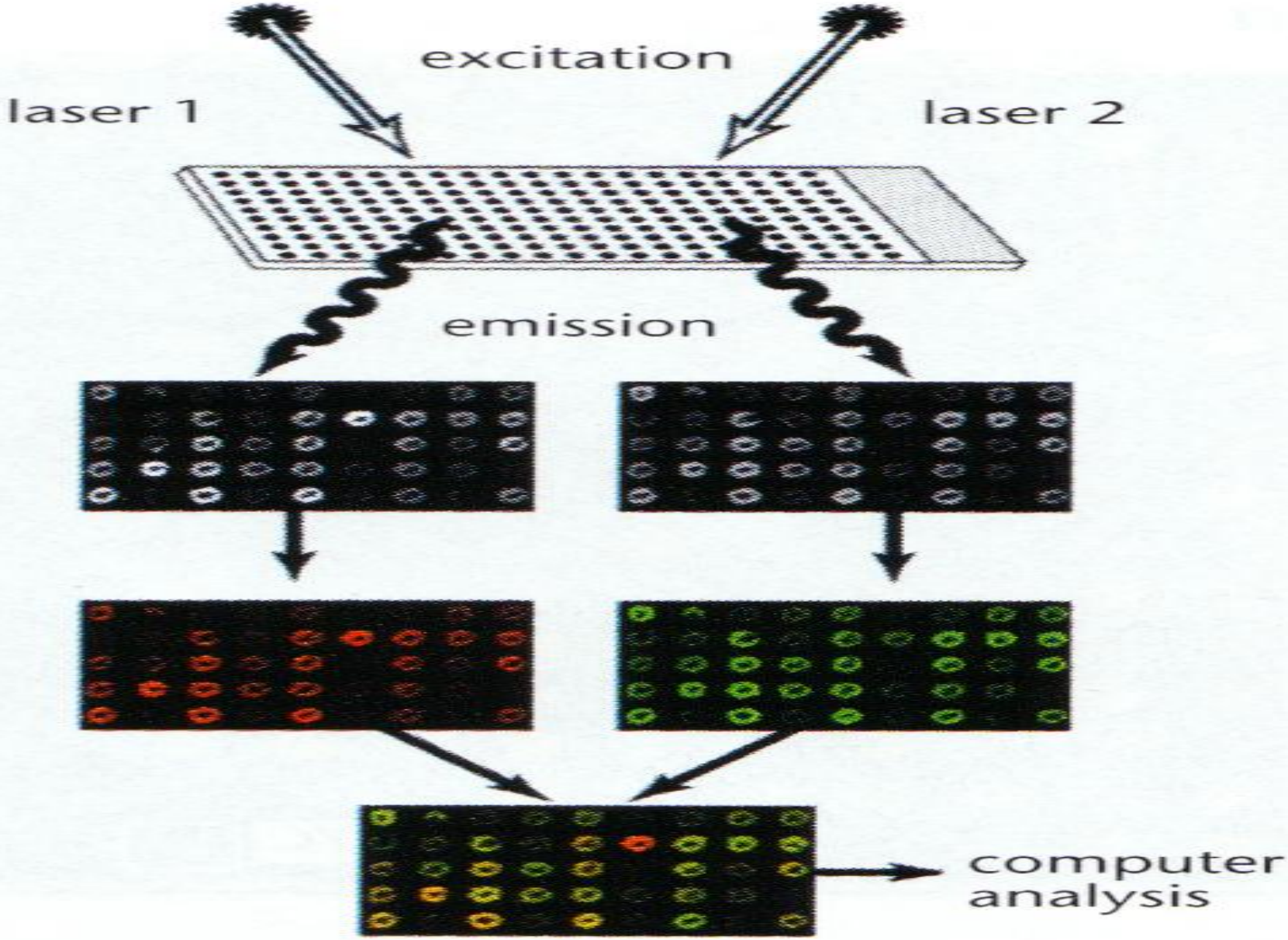
- Permit simultaneous evaluation of expression levels of thousands of genes
- Main platforms
  - Spotted cDNA arrays (2-color)
  - Affymetrix GeneChip (1-color)
  - Spotted oligo arrays (2-color or 1-color)
  - Nylon filter arrays

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

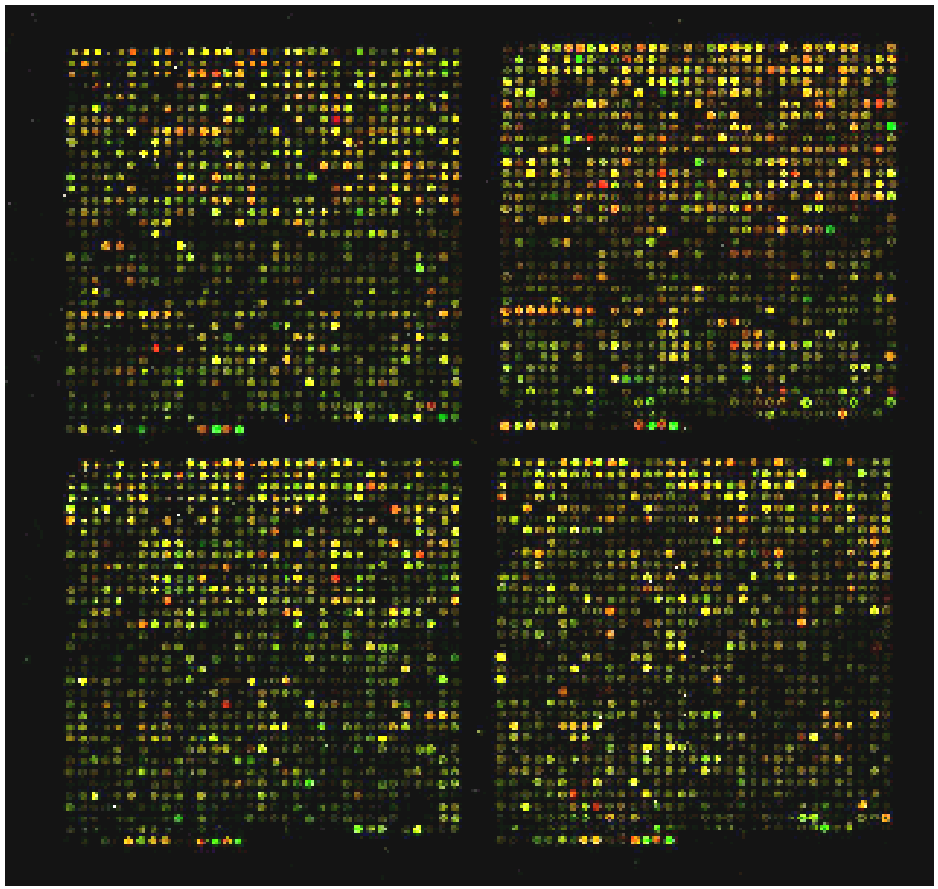
- cDNA arrays: Schena *et al.*, *Science*, 1995
- Each gene represented usually by one spot (occasionally multiple)
- Two-color (two-channel) system
  - Two colors represent the two samples competitively hybridized
  - Each spot has “red” and “green” measurements associated with it

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

cDNA Microarray Image
(overlaid “red” and “green” images)

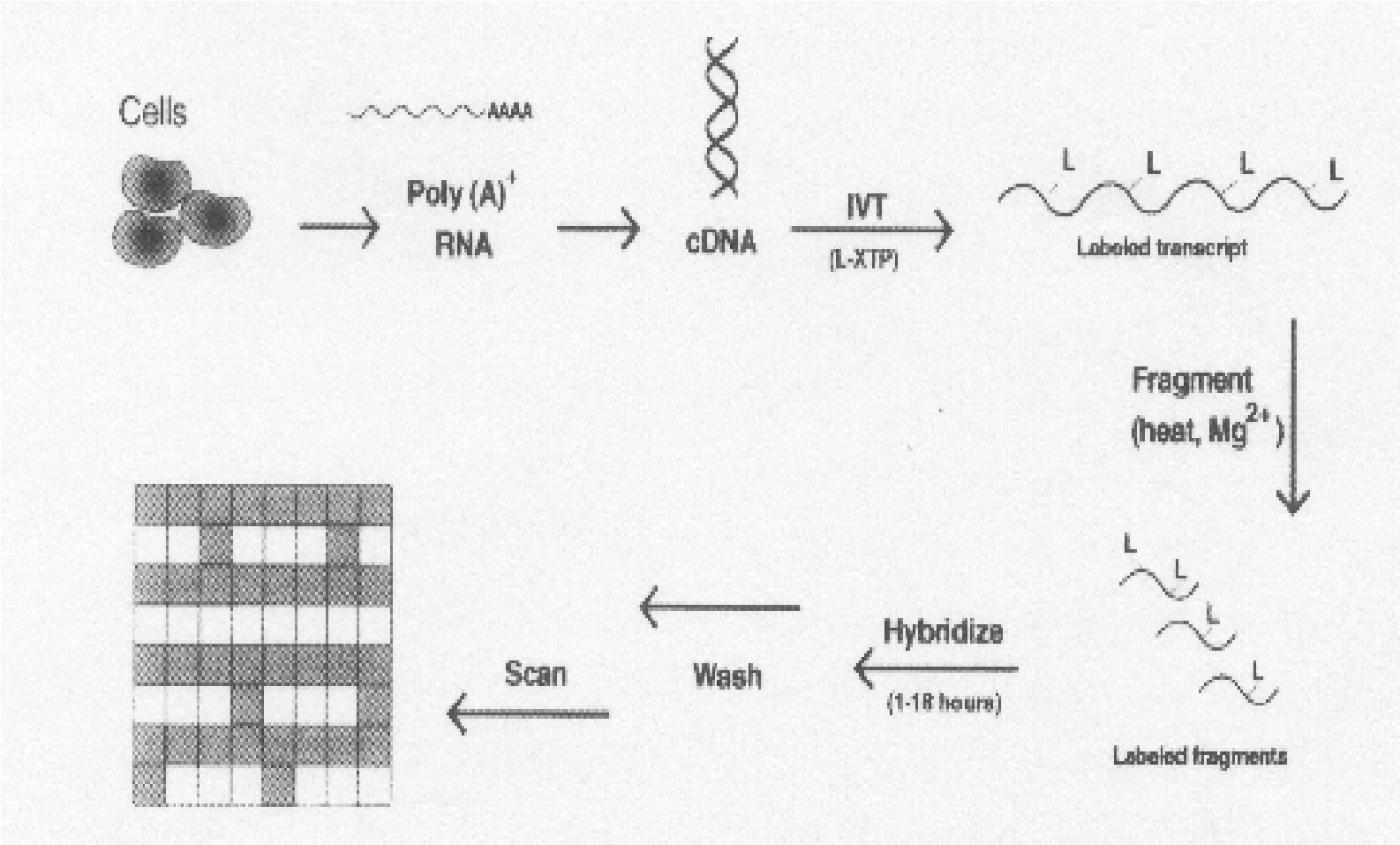


Affymetrix GeneChip Arrays

- 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 oligo probe is 25 nucleotides in length
  - 11-20 PM:MM pairs per probe set

    (PM = perfect match, MM = mismatch)

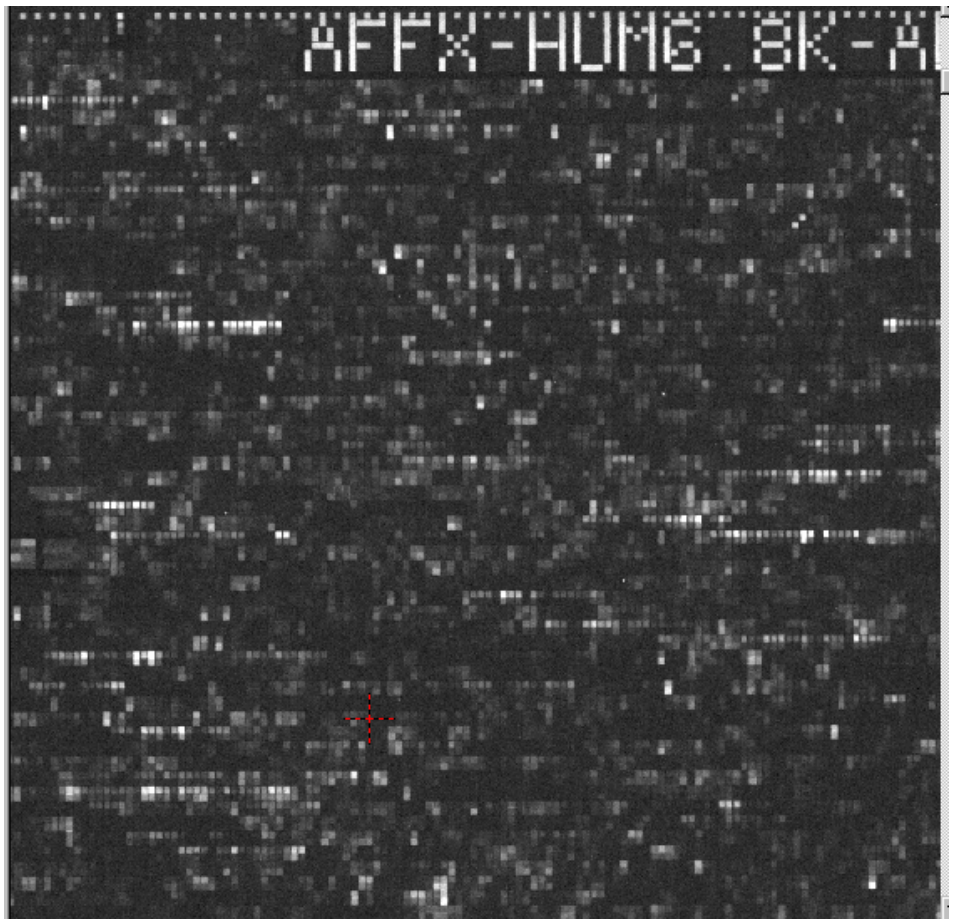
Affymetrix: Assay procedure



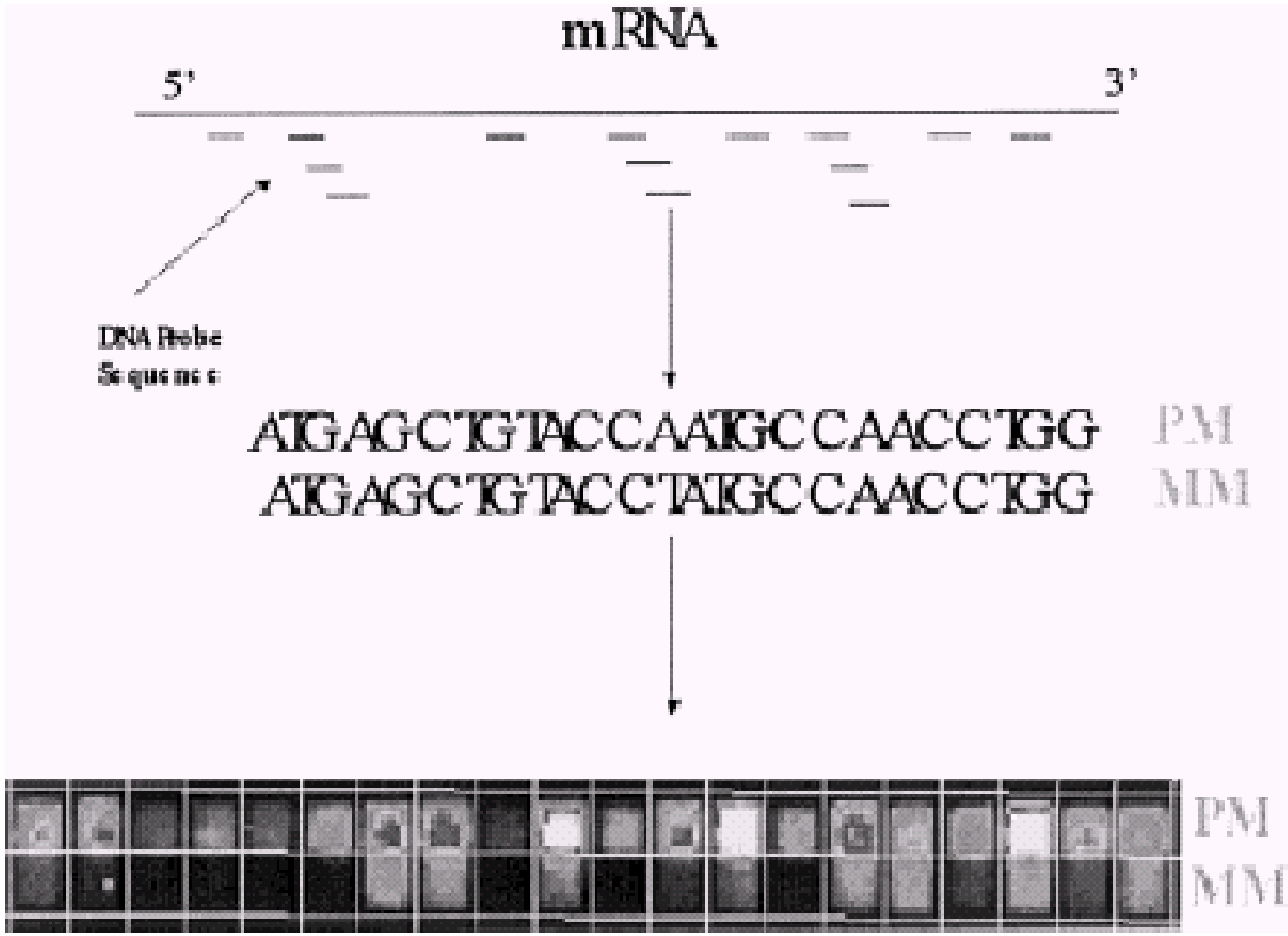
(Figure 1 from Lockhart *et al.*, *Nature Biotechnology*, 1996)

[Affymetrix] Hybridization Oligo “GeneChip” Array

Image of a Scanned Affymetrix GeneChip



Perfect Match - Mismatch Probe Pairs

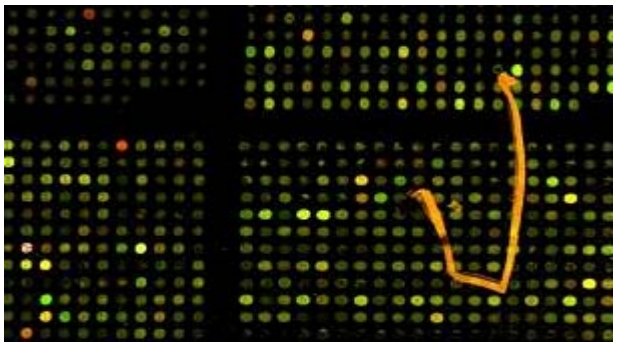


(Figure 2 from Schadt *et al.*, *Journal of Cellular Biochemistry*, 2001)

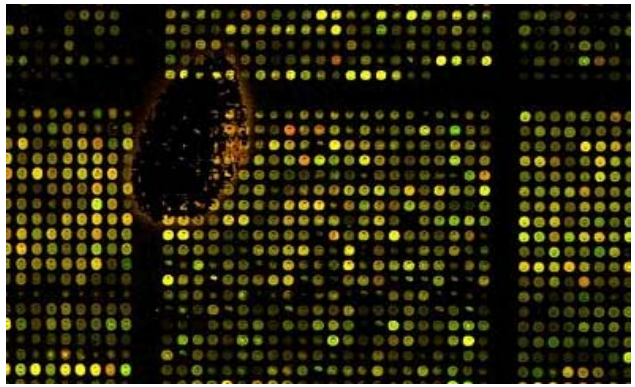
Outline

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

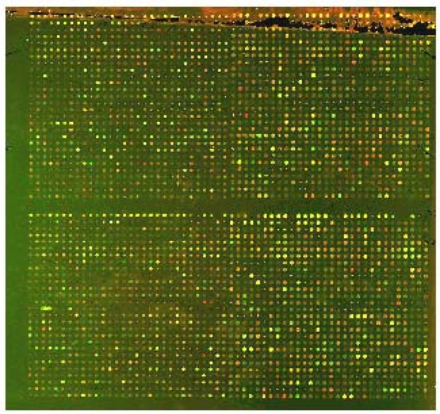
cDNA/Spotted Arrays: Slide Quality



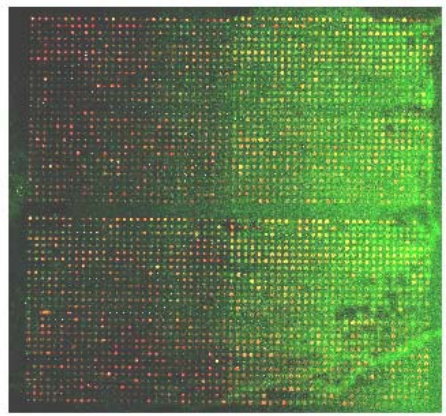
Fiber or scratch?



Bubble



Edge effect



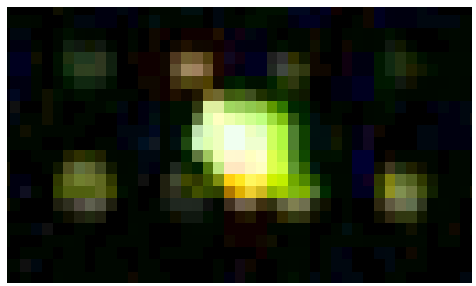
Background haze

cDNA/Spotted Arrays: Spot Quality

Poorly defined borders

Large holes

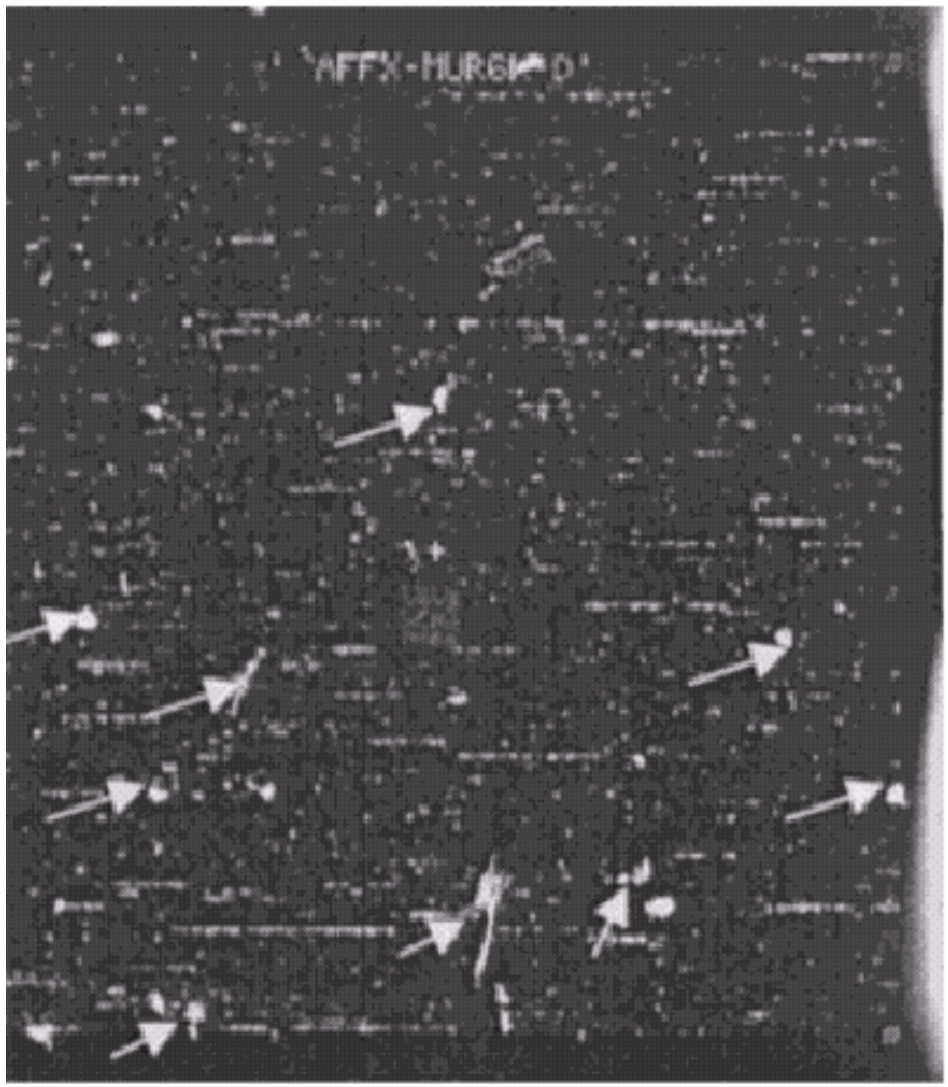
Saturated spot



Dust specs

Affymetrix Arrays: Quality Problems

(Figure 1 from Schadt *et al.*, *Journal of Cellular Biochemistry*, 2001)



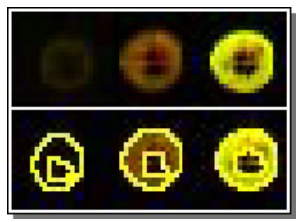
cDNA/2-color spotted arrays: Image Processing

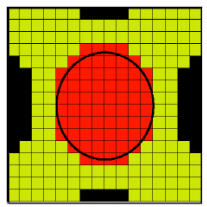
(Yang, *et al.*, *J. Computational and Graphical Statistics*, 2002)

- Gridding
- 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

- 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 in zone 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 Statistical Algorithms Description Document*

Affymetrix Arrays: Probe Set (Gene) Summaries

- $AvDiff_i = \Sigma(PM_{ij}-MM_{ij})/n_i$ for each probe set *i* (original Affymetrix algorithm)

- New Affymetrix algorithm to address negative signals (MAS 5.0, 5.1 & GCOS 1.0)
  - anti-log of a robust average (Tukey biweight) of the $\log(PM_{ij}-IM_{ij})$, where
    IM=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.

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

- 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
- Wu, Irizarry, Gentleman, Murillo, Spencer (*J. Amer. Stat. Assoc., 2004)*
  - Apply background correction that depends on G-C content of probe.

Affymetrix Arrays: Comparison of Background Corrections

- MAS 5.0/GCOS: Estimate background using mismatch probes for the gene.
- RMA: Some target hybridizes to the MM probe; for high expressed genes MM is brighter than true background; use smaller BG estimate (approx. mode of the MM probes across all MM probe sets).

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 brighter than red
- Unequal amounts of sample
- Unequal PMT voltage

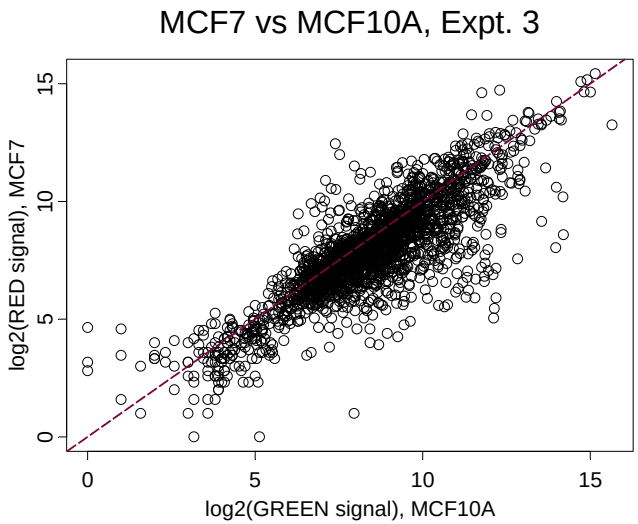
Normalization Methods for cDNA/2-Color Spotted Arrays

- 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

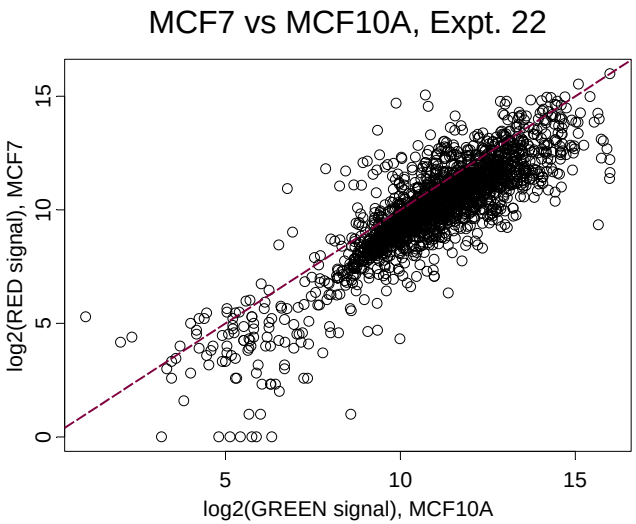


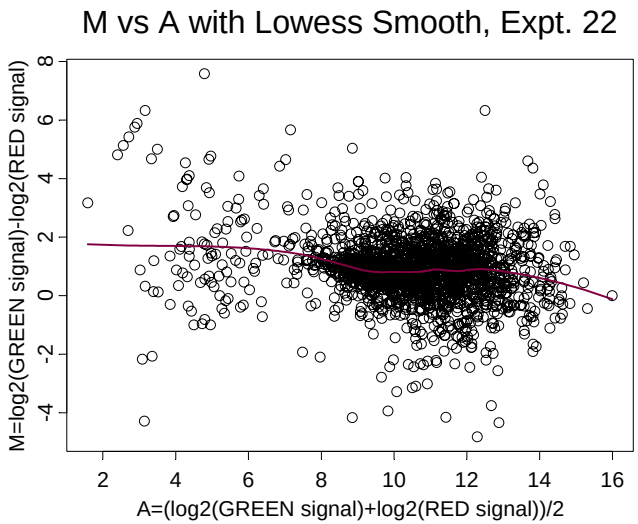
In plot of log(red signal) versus log(green signal), if point scatter is parallel to 45° line, adjust intercept to 0.

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

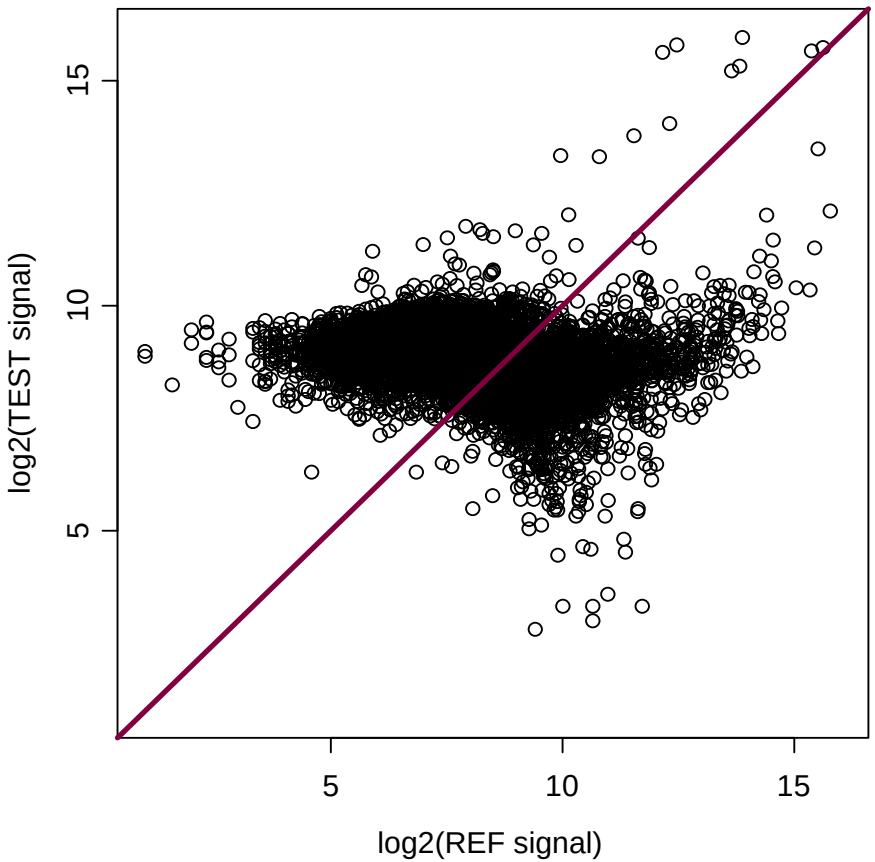


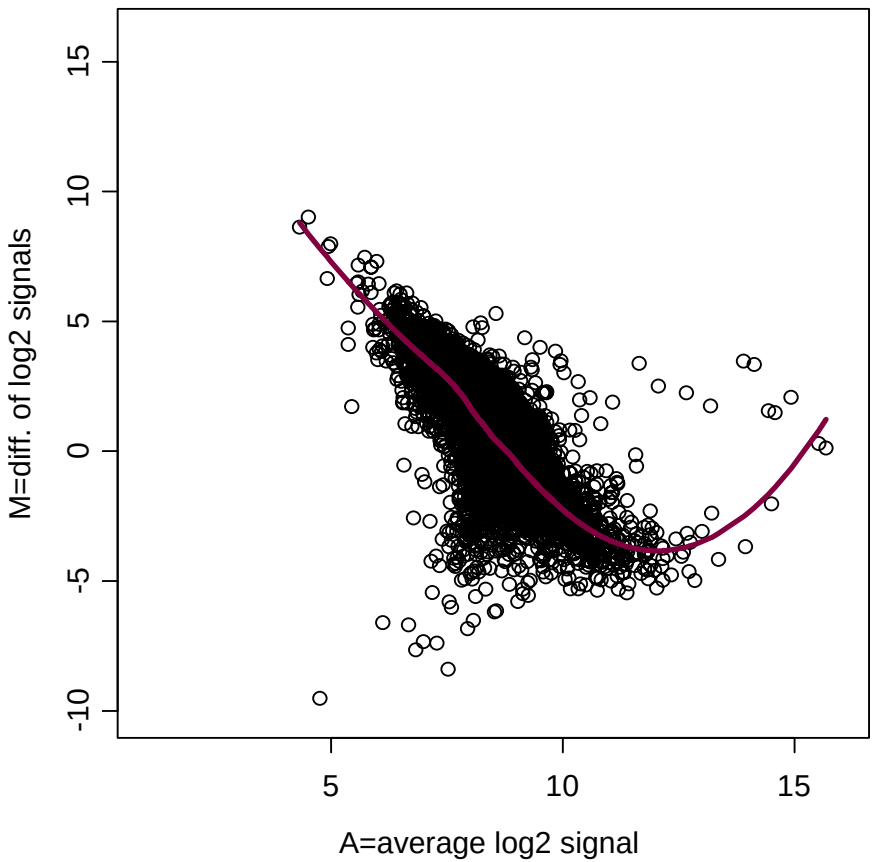


$M = \log_2(\text{GREEN signal}) - \log_2(\text{RED signal})$

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

Bad Array Example





Normalization: Affymetrix Arrays

- Variations due to sample, chip, hybridization, scanning
- Probe set-level vs probe-level
- Quantile normalization, intensity-dependent, etc.
- 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

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

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

    Exclude gene *i* if (gene has smaller var than median)

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

  - Fold difference

    Max/Min < 3 or 4, (95^{th} percentile/5^{th} percentile) < 2 or 3

    Filter if k% of genes have FC<2 or 3 relative to median

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.
- Cluster temporal gene expression patterns to get insight into genetic regulation in response to a drug or toxin.

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

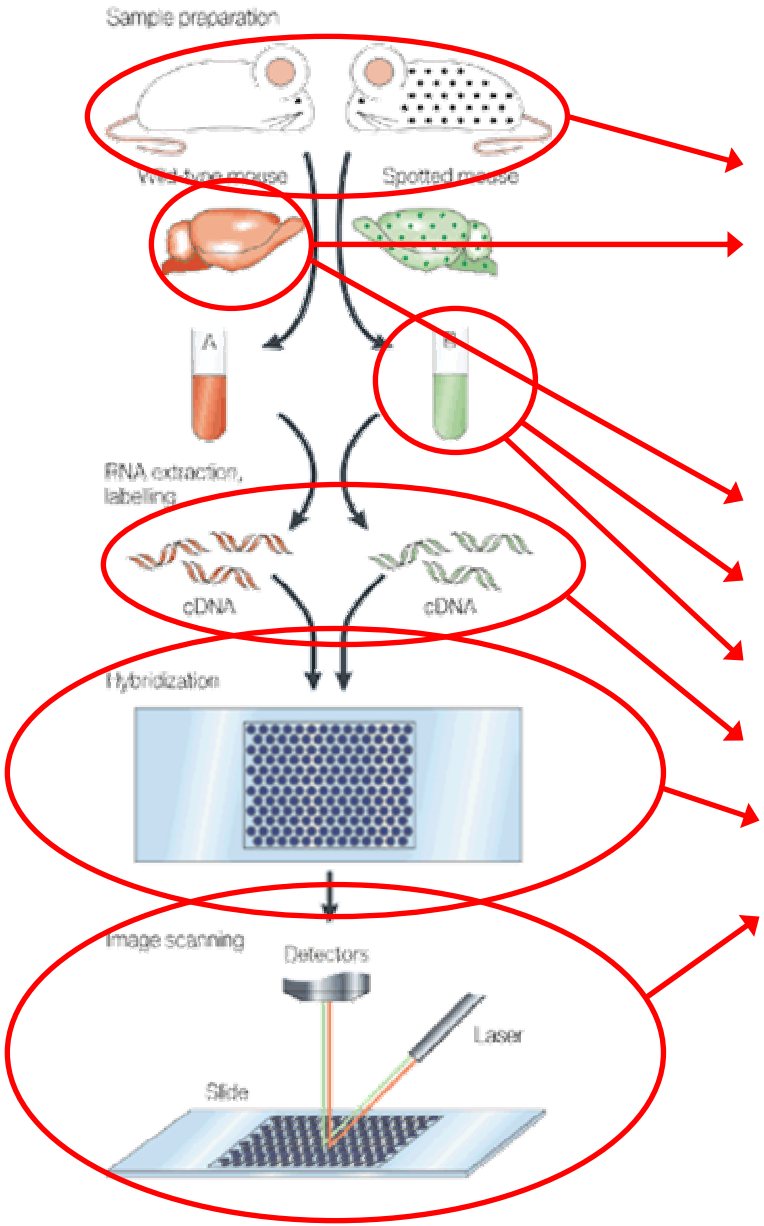
- Sample selection, including reference sample
- Sources of variability/levels of replication
- Pooling
- Sample size planning
- Controls
- For cDNA/2-color spotted arrays:
  - Reverse fluor experiments
    - Dobbin, Shih and Simon, *Bioinformatics*, 2003
  - Allocation of samples to (cDNA) array experiments
    - Kerr and Churchill, *Biostatistics*, 2001
    - Dobbin and Simon, *Bioinformatics*, 2002

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)

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)

Levels of replication

- Technical replicates
  - RNA sample divided into multiple aliquots and re-arrayed.
- Biological replicates
  - Use a different human/animal for each array.
  - In cell culture experiments, re-grow the cells under the same condition for each array (independent replication).

Summary: Replication levels

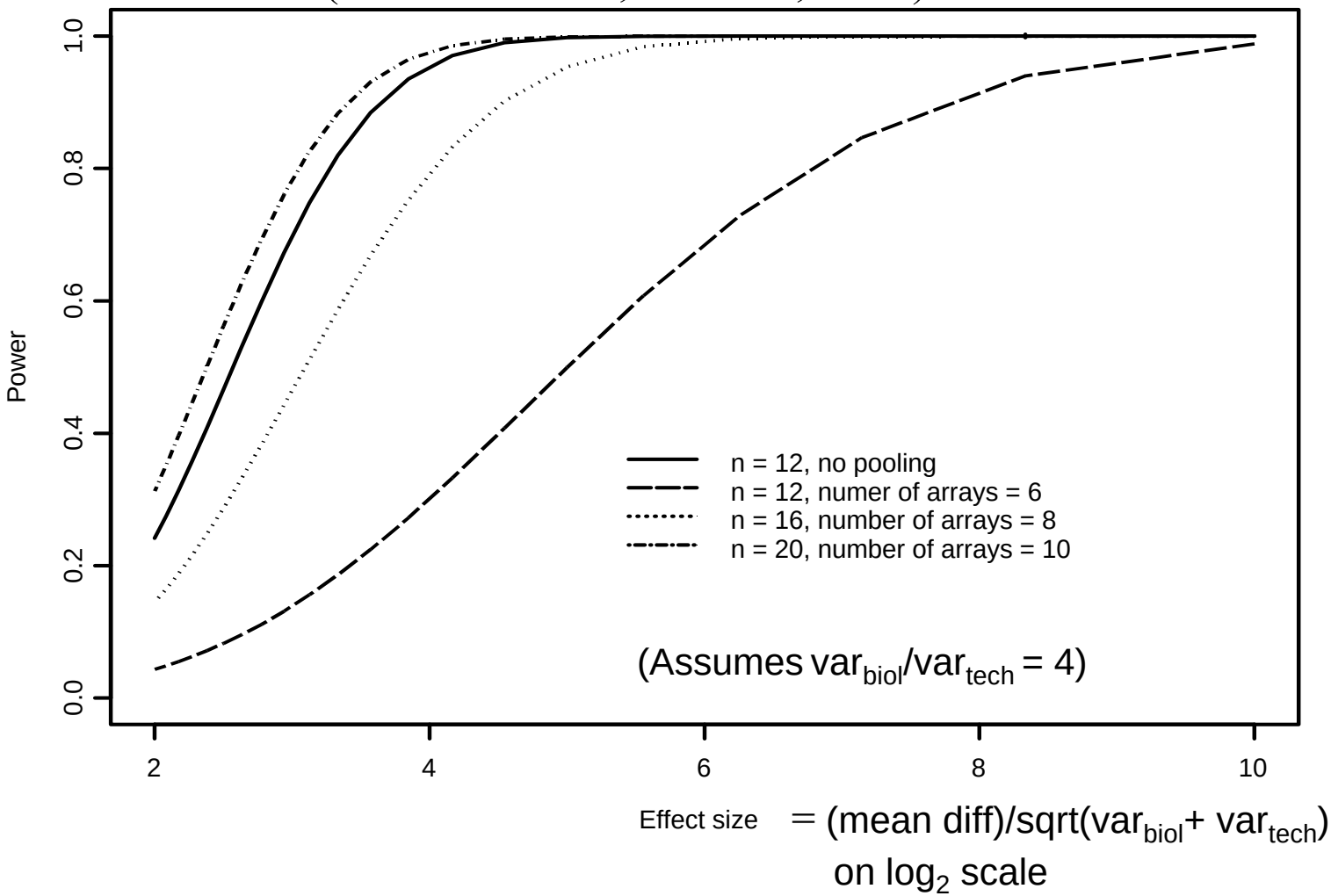
- Independent biological replicates are required for valid statistical inference.
- Maximizing biological replicates usually results in the best power for class comparisons.
- Technical replicates can be informative, e.g., for QC issues.
- But, systematic technical replication usually results in a less efficient experiment.

Is Pooling Advantageous?

- If RNA samples tiny, pooling is an alternative to amplification.
- If RNA samples big enough, then there is not usually an advantage unless arrays are very expensive and samples very cheap.
- NO FREE LUNCH: Pooling samples for each array can reduce the number of arrays needed to achieve desired precision and power, but this will come at the COST of requiring that a larger number of biologically distinct samples be used.
- Single pool with many aliquots hybridized to arrays is NOT smart! Inference requires independent replication.

Power to detect specified expression difference between 2 group

(McShane et al., *JMGBN*, 2003)



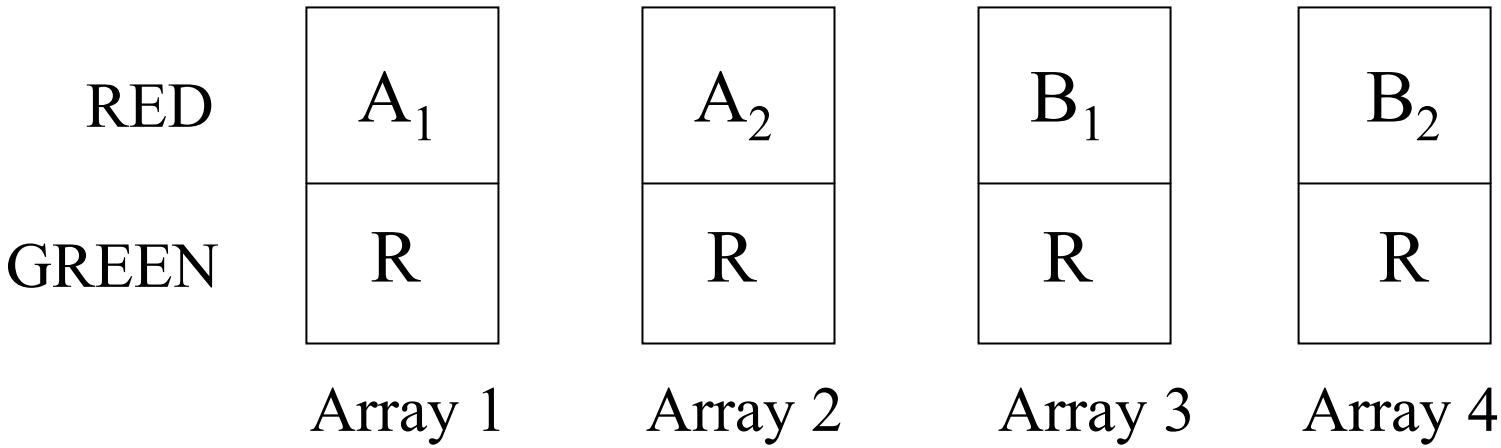
Kendziorski et al., *Biostatistics*, 2003

Shih et al., *Bioinformatics*, 2004

Class Comparison: Allocation of Specimens in cDNA Array Experiments

- Reference Design (traditional)
- Balanced Block Design
- Others
  - All pairs design
  - Loop Design (Kerr and Churchill, *Biostatistics*, 2001)
  - Variations on loop designs

Reference Design



A_i = *i*th specimen from class A

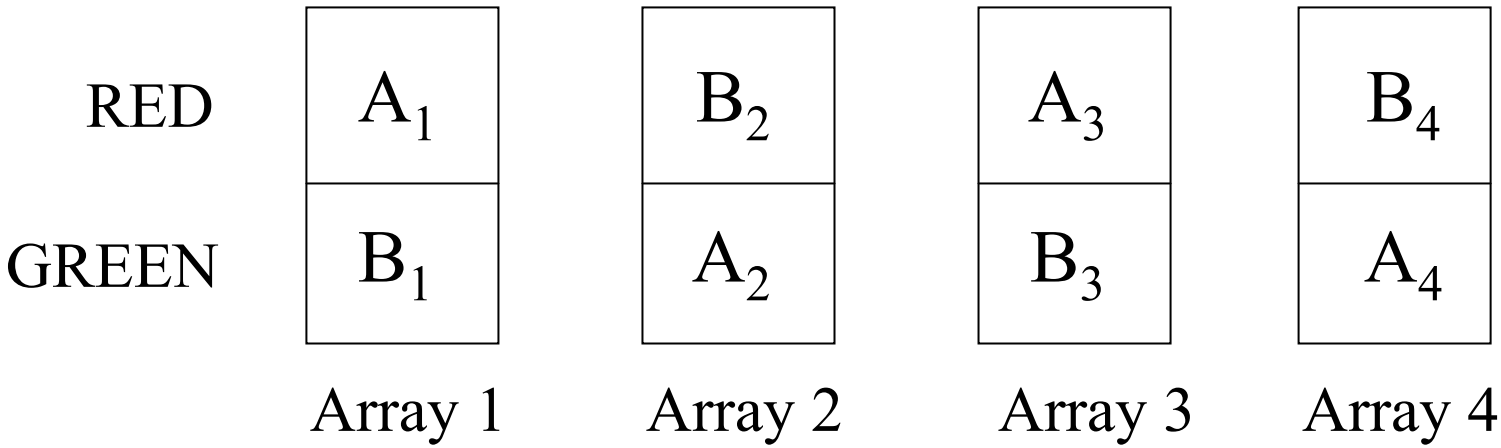
B_i = *i*th specimen from class B

R = aliquot from reference pool

Reference Design

- If the reference sample is not biologically relevant to the test samples, the class comparison is done between groups of arrays.
- If the comparison between the reference sample and test samples is biologically meaningful (e.g. reference sample is a mixture of normal samples, test samples are types of tumor samples), the class comparison is done between green and red channels – some reverse fluor experiments are required to adjust for potential dye bias.

Balanced Block Design



A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

Class comparison: optimal design
(Dobbin and Simon, 2002)

- Balanced block design most powerful/efficient
- Drawbacks of BB
  - Less flexible, some analyses may not be possible.
  - No basis for combination with other array data.
  - Cluster analysis of samples may be distorted if array effects large relative to between-sample correlations.

Number of classes being compared	Relative efficiency* (BB/R)
2	2.4
3	1.8

*Depends on ratio of biological to technical variation

Example: If comparing two classes, a common reference design (R) will require 2.4 times as many arrays as a balanced block (BB) design to obtain an equally precise estimate of mean class difference.

Class discovery: optimal design

- Class discovery by clustering samples
  - Reference design appears better than alternatives that have been proposed.
  - It's possible that the reference design could be improved upon.
- Class discovery by clustering genes
  - Example: Cluster genes observed in time series
  - A reference design will work, but may be inferior to a block design that pairs samples from different timepoints.

Summary Recommendations for Sample Allocation Schemes

- For 2-group comparison, block design is most efficient but precludes clustering.
- For cluster analysis or comparison of many groups, loop design is particularly bad and reference design is preferable.
- Reference design permits easiest analysis, allows greatest flexibility in making comparisons within and between experiments (using same reference), and is most robust to technical difficulties.
- The BRB-ArrayTools software performs class comparison between "groups of arrays" (e.g. reference designs) or between "red and green channels" (e.g. block designs), but currently not for loop designs.

Sample Size Planning

for 2-group comparisons with cDNA arrays using common reference design or with Affymetrix arrays

- 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

- Approx. total sample size required to compare 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

(δ and σ on log scale)

- More accurate iterative formulas recommended if n is approximately 60 or less

Sample Size Planning
Choosing α and β

Let K = # of genes on array

M= # of genes truly differentially expressed at a fold difference of $\theta = 2^{\delta}$

Expected number of false positives:

$EFP \leq (K-M)\times\alpha$ (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

Sample Size Planning: Effect of α and β on FDR

- False Discovery Rate (FDR) is the expected proportion of false-positive genes on the gene list.

$$\text{FDR} = \frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\beta)\pi}$$

π=proportion of differentially expressed genes

π	α	1-ß	FDR
.005	.01	.95	68%
.005	.01	.80	71%
.005	.001	.95	17%
.005	.001	.80	20%
.05	.001	.95	2%

Sample Size Planning

Choosing σ and δ

Value of σ will be determined by biology and experimental variation

- Within a *single class*, what SD is expected for expression measure?
  - For $\log_2$ ratios, σ in range 0.25 – 1.0 (typically smallest for animal model and cell line experiments)

Value of δ is the size of mean difference ($\log_2$ scale) you want to be able to detect

- 2-fold: $\delta = \log_2(2) = 1$
- 3-fold: $\delta = \log_2(3) = 1.59$

Example Sample Size Calculation

K = 10,000 genes on array

M = 100 genes differentially expressed 2-fold

Specify $\alpha = 0.001$, $\beta = 0.05$

($z_{\alpha/2} = 3.291$, $z_{\beta} = 1.645$)

$\sigma = 0.75$

$\delta = 1$ (2-fold)

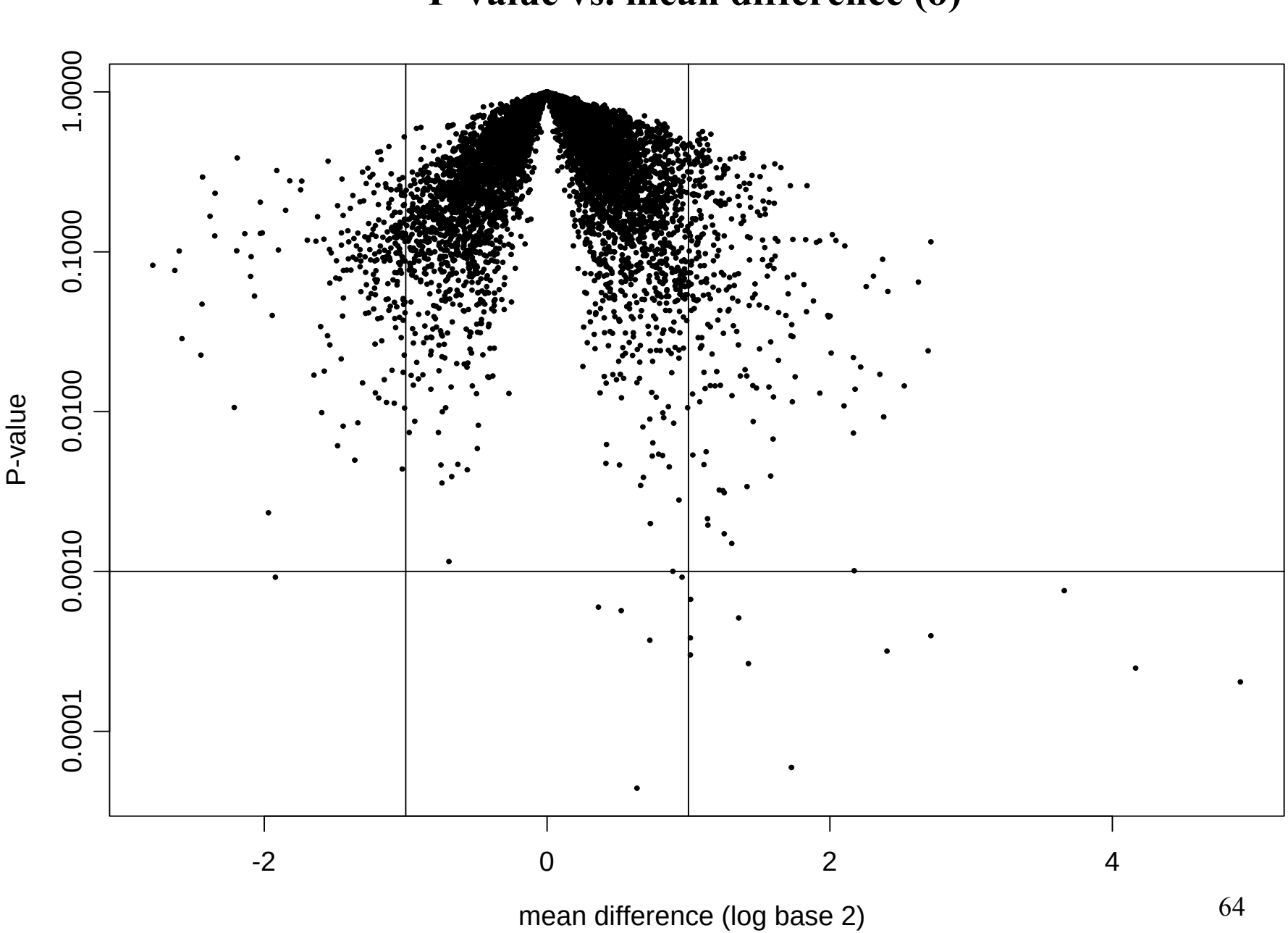
NEED n = 55 (approximately 28 per group)

Expect $\leq$ 10 false positives

Expect to miss approximately 5/100 2-fold genes

Sample Size Examples
($\alpha = .001$)

σ	δ	Fold-difference (2^{δ})	n per group	Power(%)
.25	1	2	6	95
.5	1	2	14	95
.25	1	2	5	82
.5	1	2	5	14
.25	1.20	2.29	5	95
.5	2.39	5.24	5	95


P-value vs. mean difference (δ)
P-value
1.0000
0.1000
0.0100
0.0010
0.0001
-2
0
2
4
mean difference (log base 2)

Further sample sizes references

- Technical replicates for comparing 2 samples
  - Lee et al., PNAS, 2000
  - Black and Doerge, Bioinformatics, 2002
- Sample sizes for pooled RNA designs
  - Shih et al., Bioinformatics, 2004
- Sample sizes for balanced block designs, paired data, dye swaps, technical replicates, etc.
  - Dobbin et al., Bioinformatics, 2003
  - Dobbin and Simon, Biostatistics, 2005

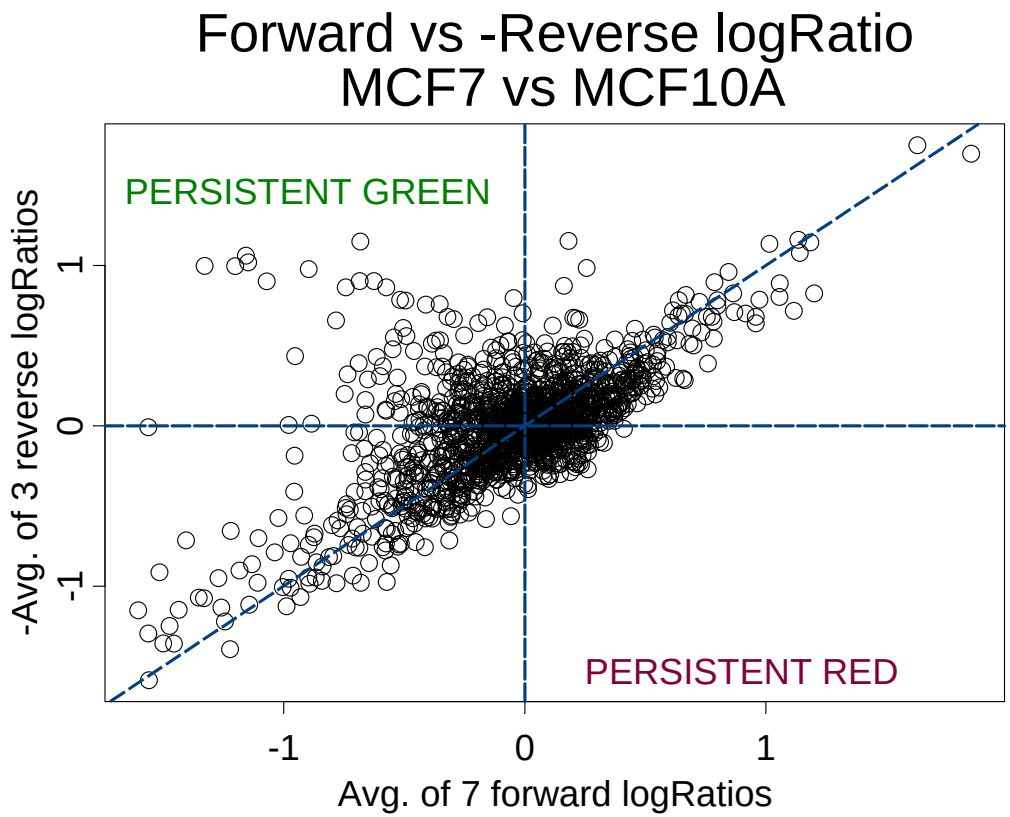
How best to allocate effort?

- Microarrays can serve as a good high-throughput screening tool to identify potentially interesting genes.
- Verification of results via a different, more accurate, assay often preferable to running many arrays or technical replicates.
- Gene IDs associated with sequences can change over time, so periodic verification is advisable.

Controls

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

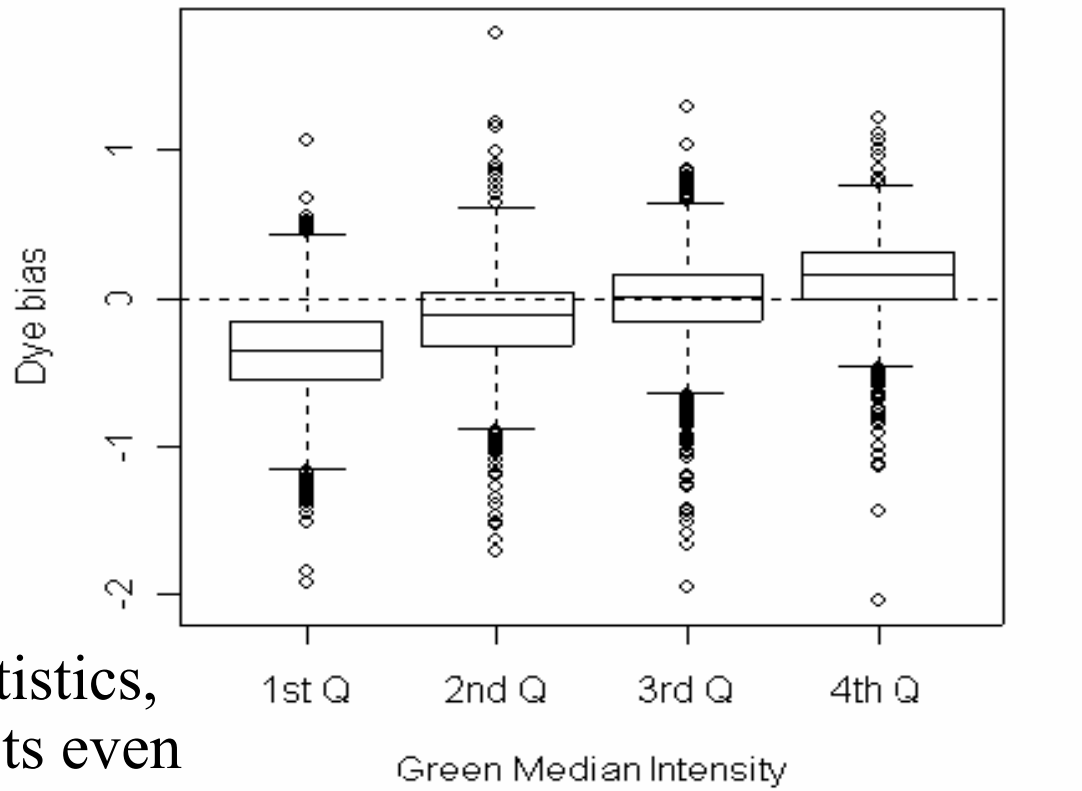
cDNA/2-Color Spotted Arrays: Reverse Fluor Experiments



Does dye bias exist?

- Direct labeling? Yes.
- Indirect labeling? Yes
- Does normalization (e.g., loess) fix it? No.

Indirect labeling
Log base 2 scale dye bias estimates



Dobbin and Simon (Biostatistics, 2005) show dye bias persists even after normalization, and may depend on intensity.

cDNA/2-color spotted arrays under common reference design:

Should reverse fluor "replicates" be performed for every array?

Usually NO!

See Dobbin, Shih and Simon, *Bioinformatics*, 2003 for a comprehensive discussion of reverse fluor replication

Reverse Fluors: cDNA/2-color spotted arrays with common reference design

- When interested in interpreting individual ratios . . .
  - If gene-specific dye bias depends on gene sequence and not sample characteristics, dye bias can be adjusted for by performing *some* reverse fluor experiments.
  - If dye bias depends on both the gene and the sample, dye swaps won't help (Dobbin, Shih and Simon, 2005)!

- In BRB-ArrayTools reverse fluor arrays must be specified during the data importation (collation) step.

Reverse Fluors: cDNA/2-color spotted arrays with common reference design

- When interested in class comparisons and using common reference design. . .
  - When comparing classes of non-reference samples tagged with the same dye, the dye bias should cancel out.
  - Reverse fluors are not required.

Reverse Fluors: cDNA/2-color spotted array with balanced block design

- For each class, half the samples should be tagged with Cy3 and half with Cy5.
- When comparing different classes, dye bias will cancel out of the class comparisons.
- No reverse fluors are required.

Reverse Fluors: cDNA/2-color spotted arrays with common reference design

- 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 (Euclidean distance), constant dye biases should "wash out".
  - Some reverse fluors and replicates may be useful as informal quality checks.

Reverse Fluors: cDNA/2-color spotted arrays with common reference design

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

Within and between platform variation

- Within-platform reproducibility depends on
  - Quality of assay protocol
  - Experience of lab technician
  - Quality and standardization of chips and reagents
- With good quality and standardization, within-platform reproducibility can be quite high, e.g., correlation > 93% within and between labs on same RNA.
- Between-platform reproducibility much more problematic. Often 50% or lower gene list agreement. Pathway agreement can be better.

References: Dobbin et al. (Clin Cancer Res, 2005); Petersen et al., (BMC Genomics, 2005); Irizarry et al. (Nature Methods, 2005).

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

- Global tests
  - Compare whole profiles
  - Permutation tests
- Gene-level analyses
  - Model-based methods
  - Non-model-based methods
  - Hybrid variance methods

Global Tests for Differences in Profiles 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
- BRB-ArrayTools uses the number of univariately significant genes as a summary measure for the global test for differences between profiles.

Summary of Results:

Number of genes significant at 0.001 level of the univariate test: 52

Probability of getting at least 52 genes significant by chance (at the 0.001 level) if there are no real differences between the classes: 0.001

Gene-Level Analyses

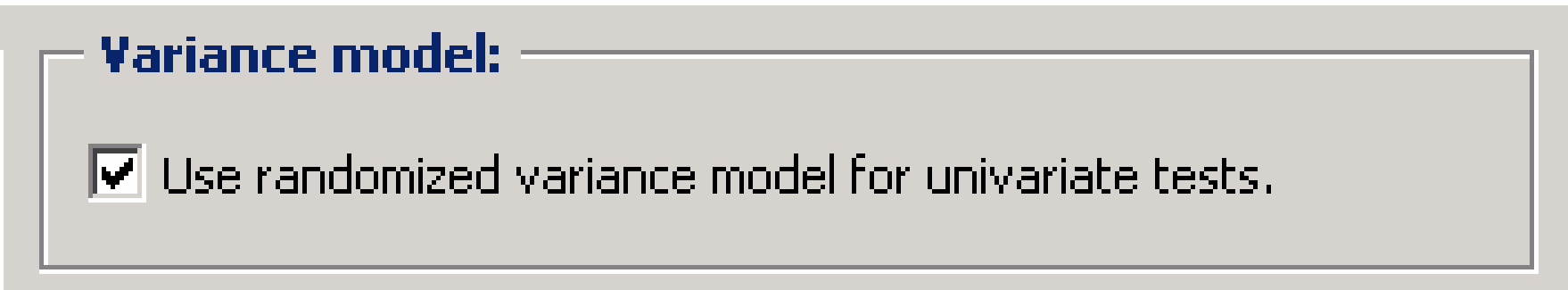
- Model-based methods
  - Multi-parameter modeling of channel-level data (e.g., Gaussian mixed or ANOVA models), hierarchical Bayesian models, etc.
  - May borrow information across genes
  - May use multiple comparison adjustments
- Non-model-based methods
  - Log ratios or signal (e.g., Affymetrix)
  - T-test, F-test, or nonparametric counterparts (e.g., Wilcoxon)
  - Multiple comparison adjustment commonly used
- Random variance methods
  - Variance estimates borrow across genes

Model-based Gene-by-Gene Methods for cDNA Arrays

- Non-Bayesian Examples
  - 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
- Bayesian & Empirical Bayes Examples
  - Tadesse *et al.*, *Biometrics*, 2003
  - Ibrahim *et al.*, *JASA*, 2002
  - Efron *et al.*, *JASA*, 2001; *Stanford Tech Rep*, 2001
  - Newton *et al.*, *J Comp Biology* 2001
  - Manduchi *et al.*, *Bioinformatics* 2000

Hybrid Variance Methods for Small Sample Gene-Level Analyses

- Bayesian:
  - Baldi and Long, *Bioinformatics*, 2001
- Frequentist:
  - Wright and Simon, *Bioinformatics*, 2003
    - Available as the 'Random variance' option in BRB-ArrayTools for Class Comparison and Class Prediction analyses:



Multiple Testing Procedures for Gene-Level Analyses

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

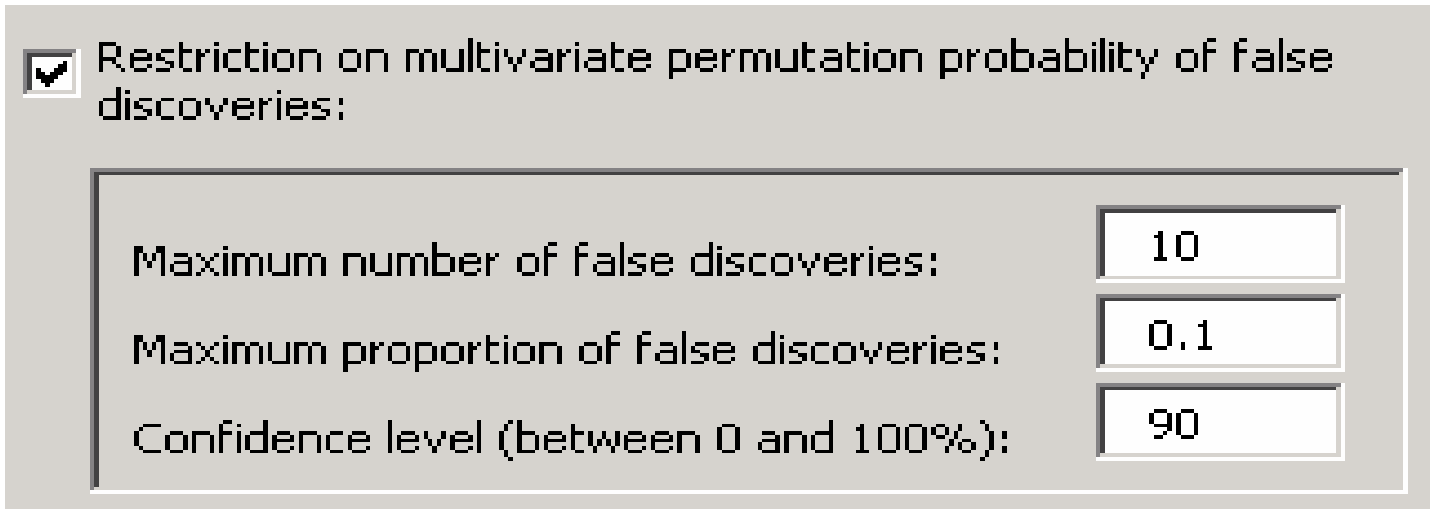
- Review by Dudoit *et al*. (*Statistical Science*, 2003)
- "SAM" - Significance Analysis of Microarrays
  - Tusher *et al*., *PNAS*, 2001 and relatives
  - Estimate quantities similar to FDR (old SAM) or control FDP (newer versions of SAM)
- Bayesian
  - Efron *et al*., *JASA*, 2001; *Stanford Tech 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*, 2004 (FD and FDP control)

Examples of Types of Control

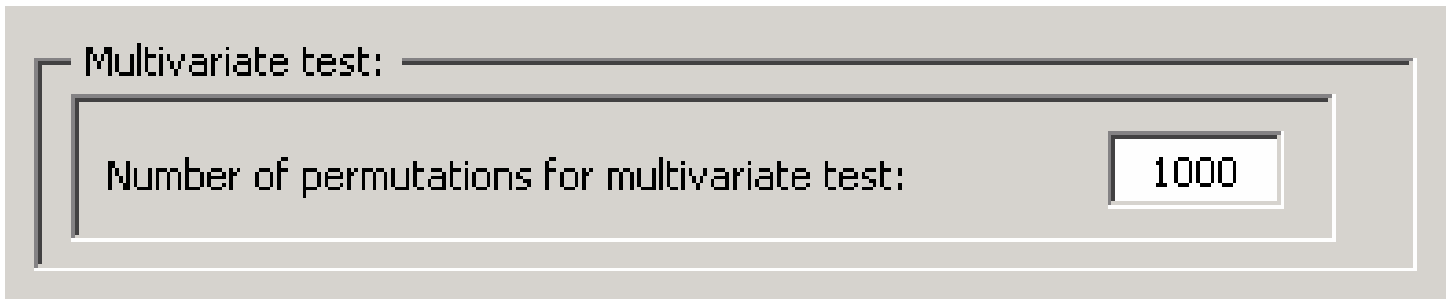
- Korn *et al.* FD procedure: "We are 95% confident that the (actual) number of false discoveries is no greater than 2."
- Korn *et al.* FDP procedure: "We are 95% confident that the (actual) proportion of false discoveries does not exceed approximately 0.10."
- Tusher *et al.* SAM: "On *average*, the false discovery proportion will be controlled at approximately 10%."
- Current SAM – more similar to Korn FDP procedure
- Bayesian methods: "High posterior probability of differential expression"

Multiple Testing Procedure Available in BRB-ArrayTools

- The step-down permutation procedure for FD and FDP control (Korn, *et al*.) is available in BRB-ArrayTools for Class Comparison, Survival Analysis, and Quantitative Traits Analysis (finding genes significantly correlated with a quantitative variable). The following screenshot from the analysis dialog shows the default options:



- The number of permutations may be specified on the 'Options' page:



Class Discovery

- Cluster analysis algorithms

  (Gordon, 1999, Chapman Hall)
  - Hierarchical
  - K-means
  - Self-Organizing Maps
  - Maximum likelihood/mixture models
  - Multitude of others
- Graphical displays
  - Hierarchical clustering
    - Dendrogram
    - “Ordered” color image plot (heatmap)
  - 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

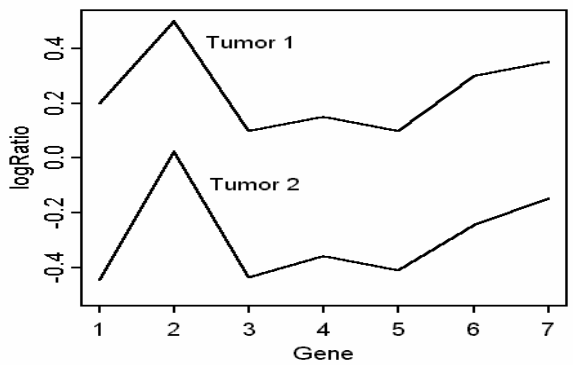
Hierarchical Agglomerative Clustering Algorithm

- 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

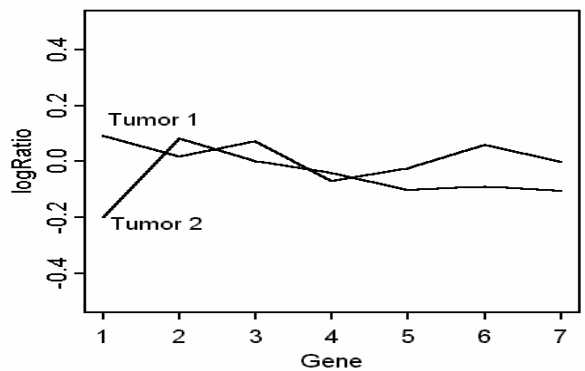
Common Distance Metrics for Hierarchical Clustering

- 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

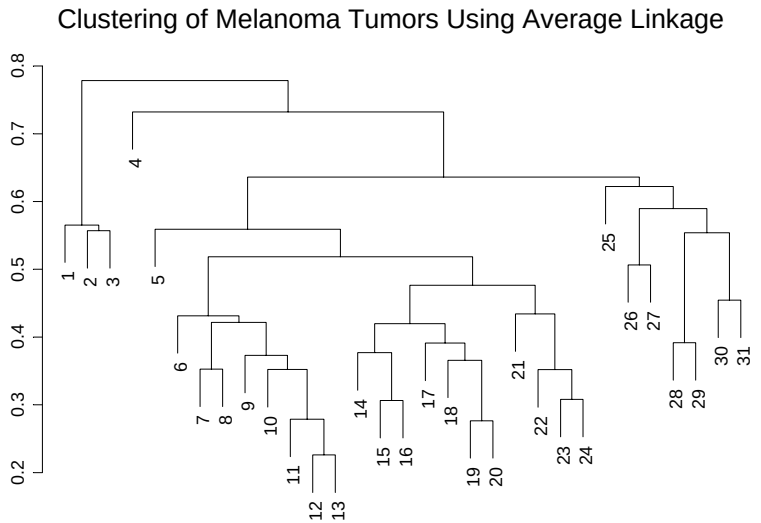


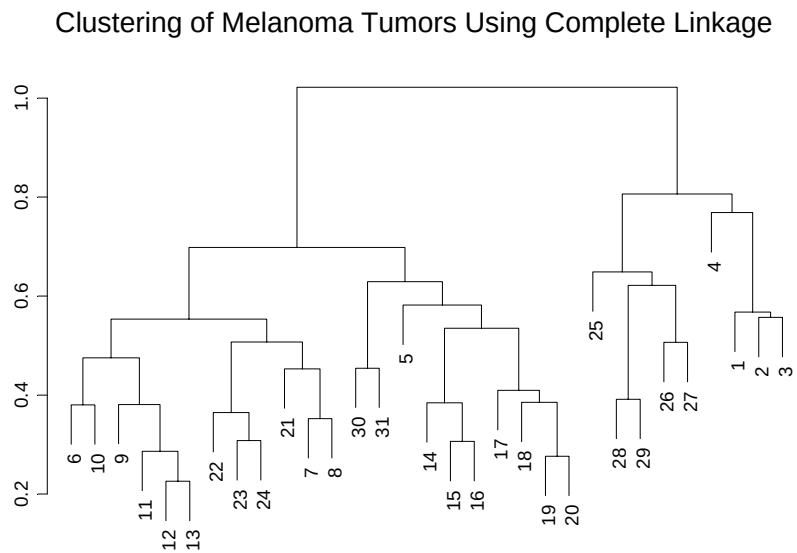
Euclidean distance small, 1-Correlation large

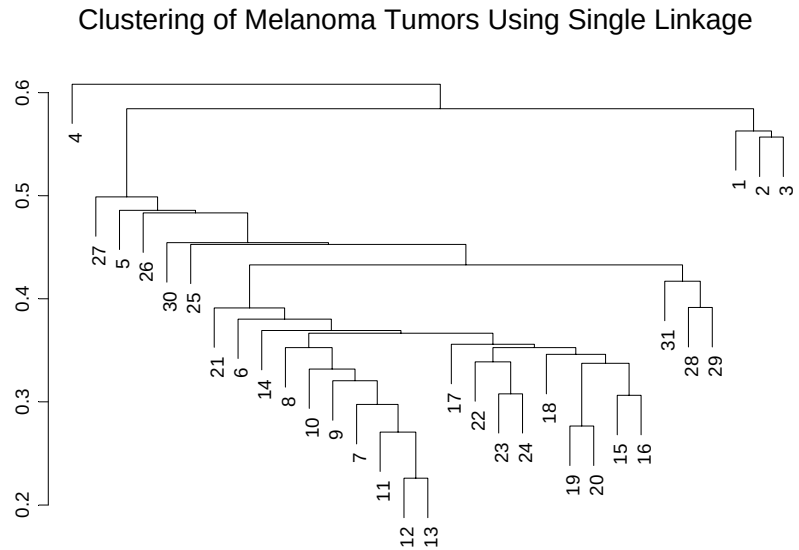


Linkage Methods

- 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



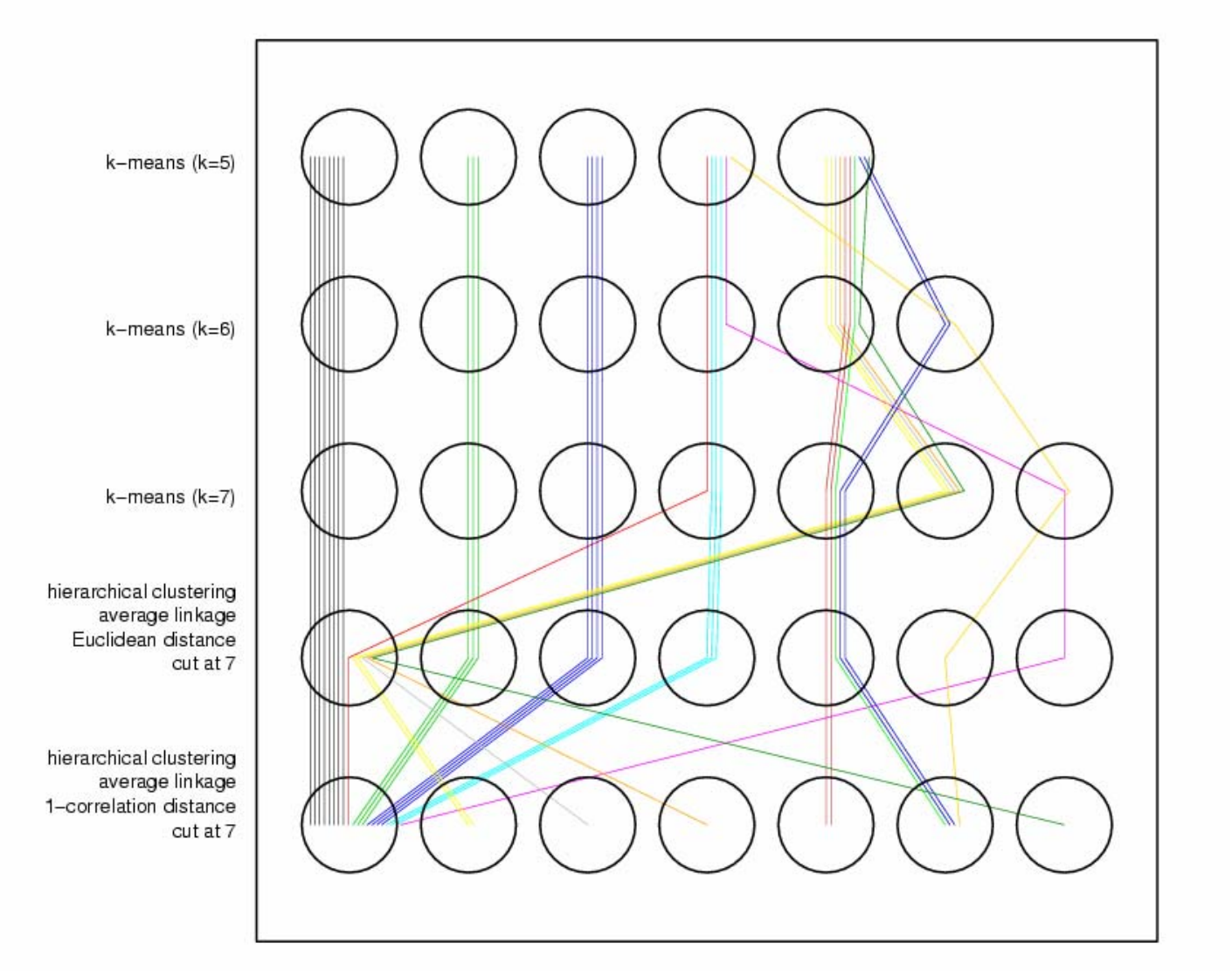




Dendrograms using 3 different linkage methods,
distance = 1-correlation

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

Does clustering method matter?
One set of specimens clustered by different methods



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

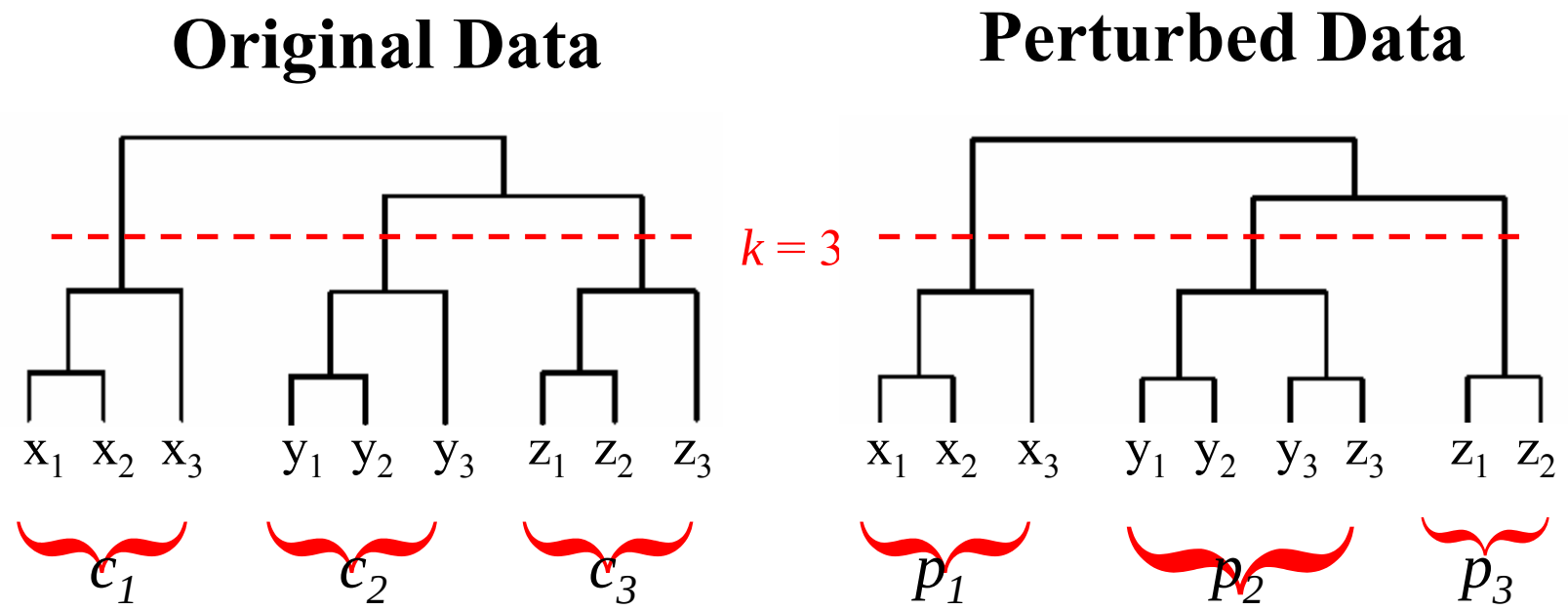
Assessing Cluster Reproducibility: Data Perturbation Methods

- 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*, *2002* – 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.
  - The *R-index* **may be obtained in BRB-ArrayTools** for the hierarchical clustering of samples by selecting the 'Compute cluster reproducibility measures' option. The *R-index* option is not implemented for the hierarchical clustering of genes.

R-index Example

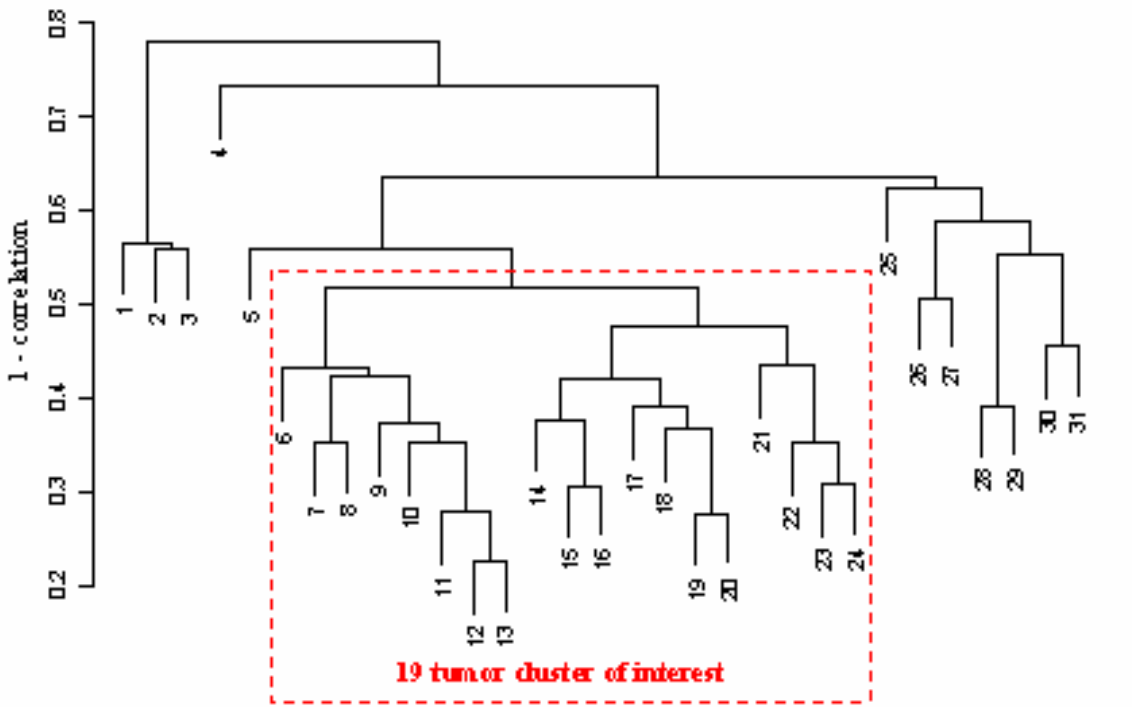


- 3 out of 3 pairs in c_1 remain together in perturbed clustering.
- 3 out of 3 in c_2 remain together.
- 1 out of 3 in c_3 remain together.
- *R*-index = (3 + 3 + 1)/(3 + 3 + 3) = 0.78

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.

$\Rightarrow$ highly reproducible

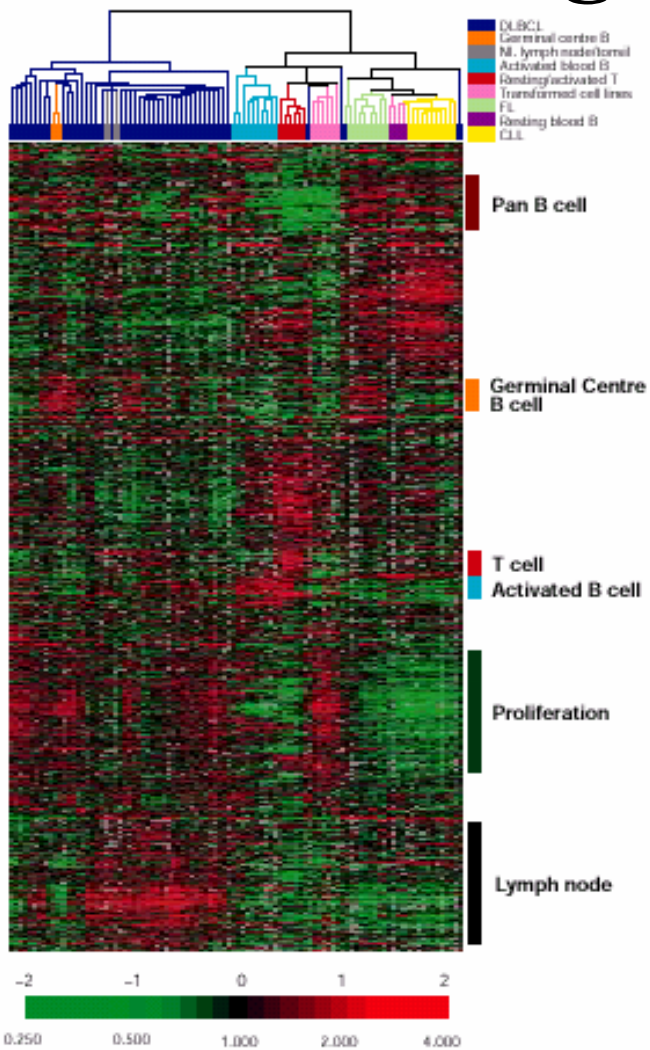
Melanomas in the 19 element cluster tended to have:

- 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*, 2002)
  - A global test of "no clustering" is available as an option in BRB-ArrayTools for the multidimensional scaling of samples. For computational purposes, the global test of clustering is applied to the multidimensional scaling coordinates (the dimension-reduced data) rather than to the original data.
- Gap Statistic (Tibshirani *et al.*, *JRSS B*, 2002) – estimate number of clusters (Does not work!)
- Comparisons of methods for estimating number of clusters in small dimension cases (Milligan and Cooper, *Psychometrika*, 1985)

Graphical Displays: Ordered Color Image Plot



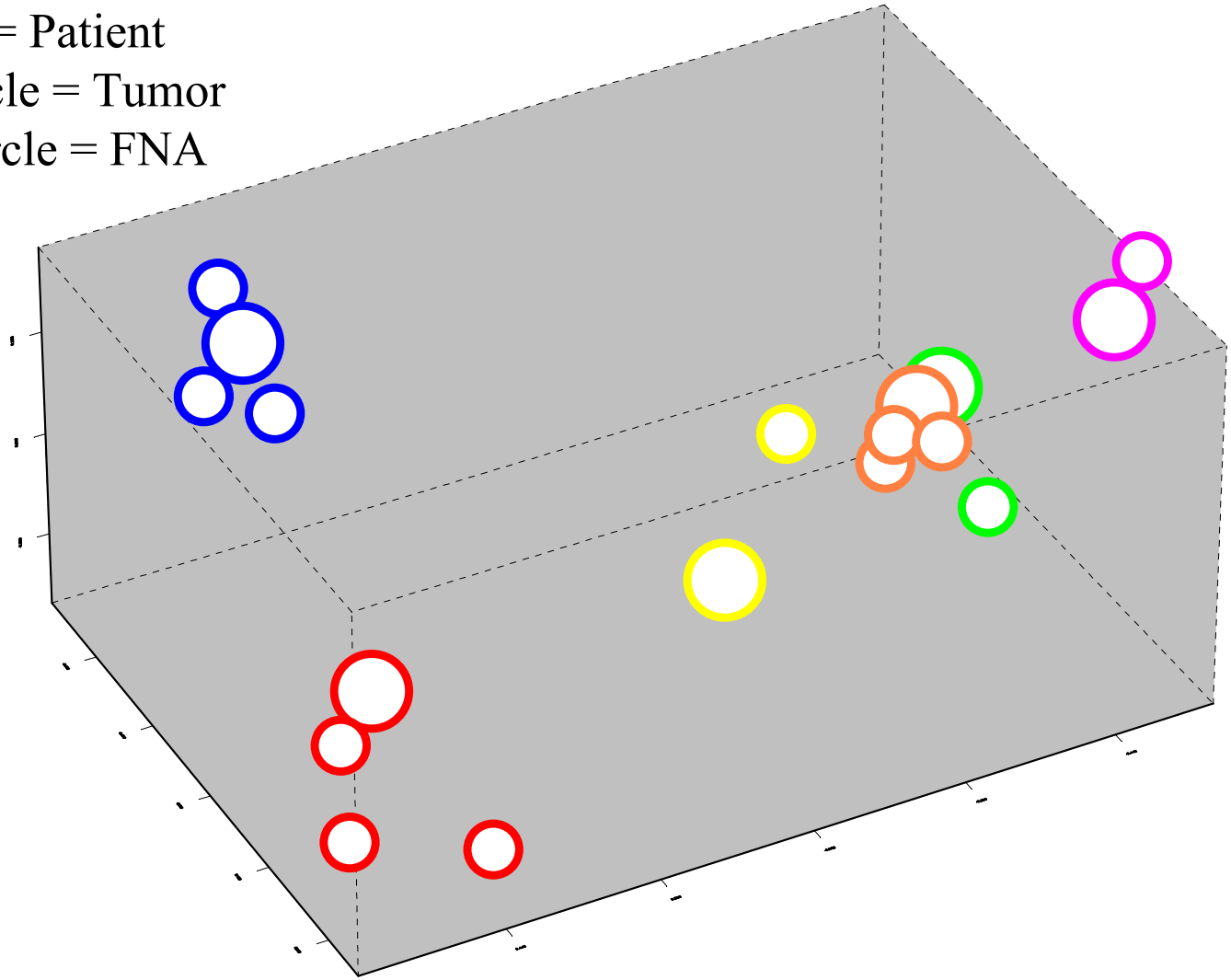
Hierarchical Clustering of Lymphoma Data (Alizadeh *et al.*, *Nature*, 2000)

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

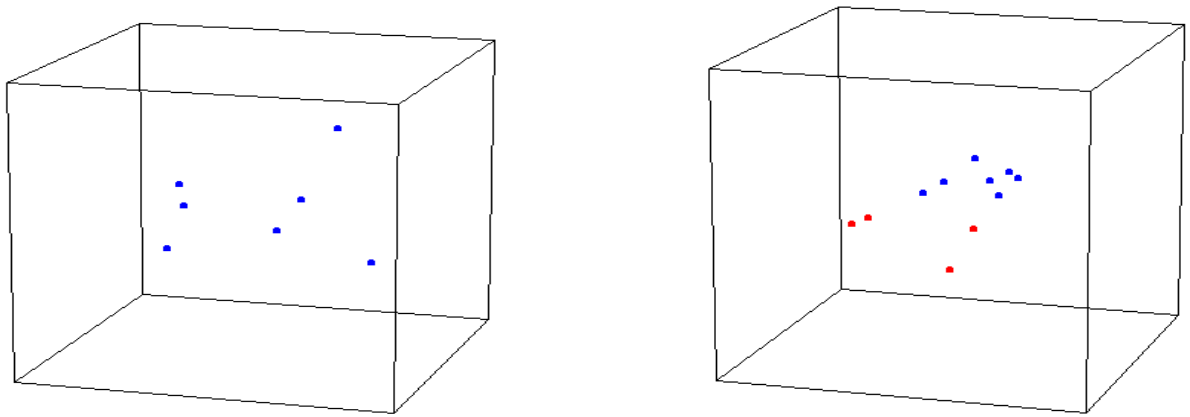
Color = Patient
Large circle = Tumor
Small circle = FNA



(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*, 2002)

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

(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

The Compound Covariate Predictor (CCP)

(Tukey, *Controlled Clinical Trials*, 1993)

- Select "differentially expressed" genes by two-sample *t*-test with small α.

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

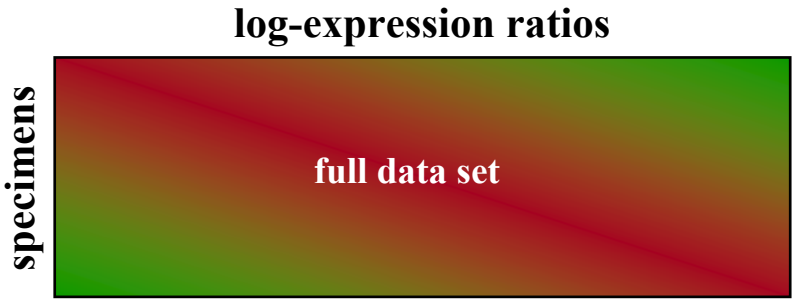
t_j is the two-sample *t*-statistic for gene *j*.

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

Sum is over all *d* differentially expressed genes.

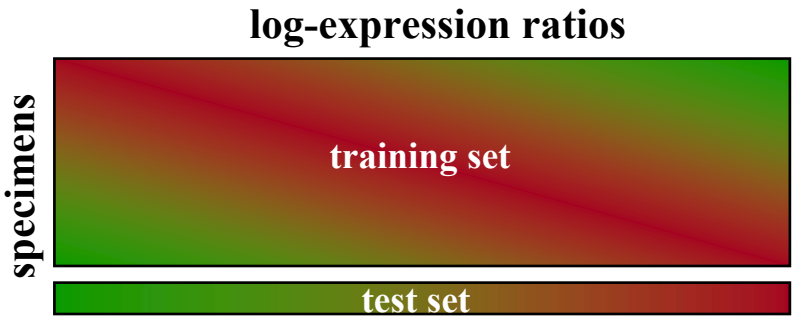
- Threshold of classification: midpoint of the CCP means for the two classes.

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.

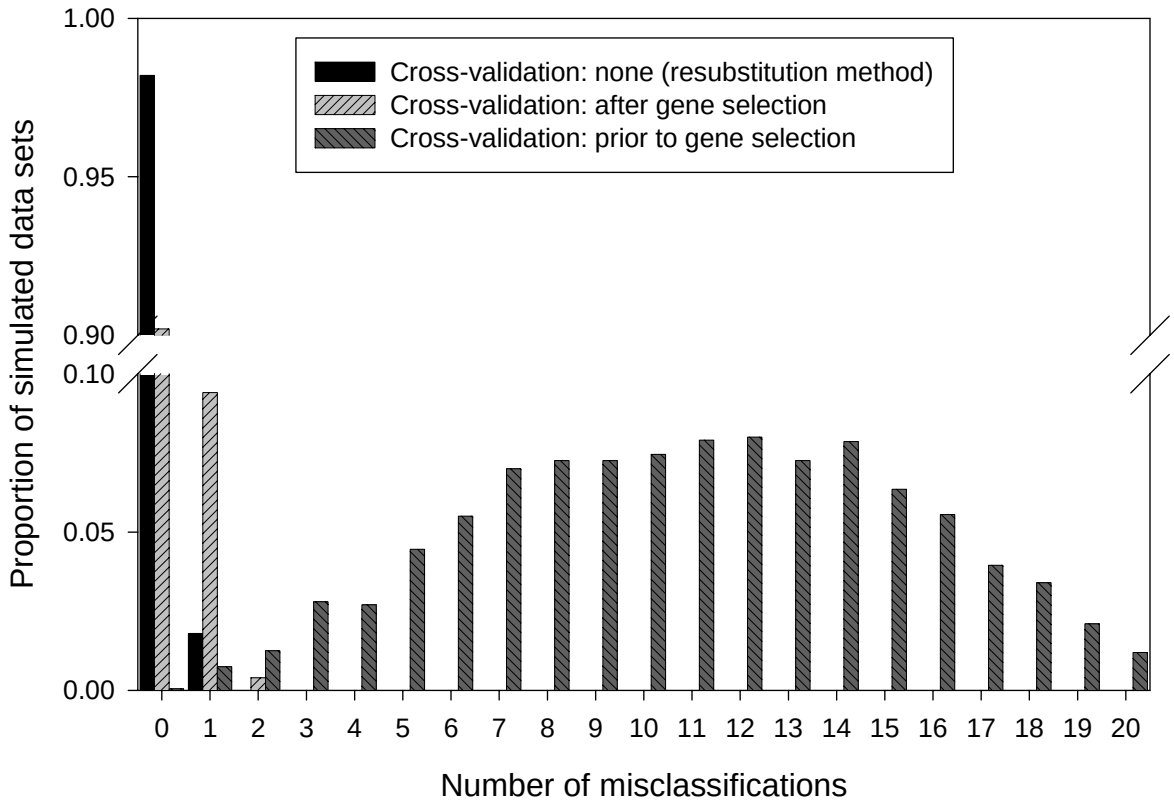
Prediction on Simulated Null Data

Generation of Gene Expression Profiles

- 20 specimens (P_i is the expression profile for specimen i)
- Log-ratio measurements on 6000 genes
- $P_i \sim \text{MVN}(\mathbf{0}, \mathbf{I}_{6000})$
- 10000 simulation repetitions
- 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.

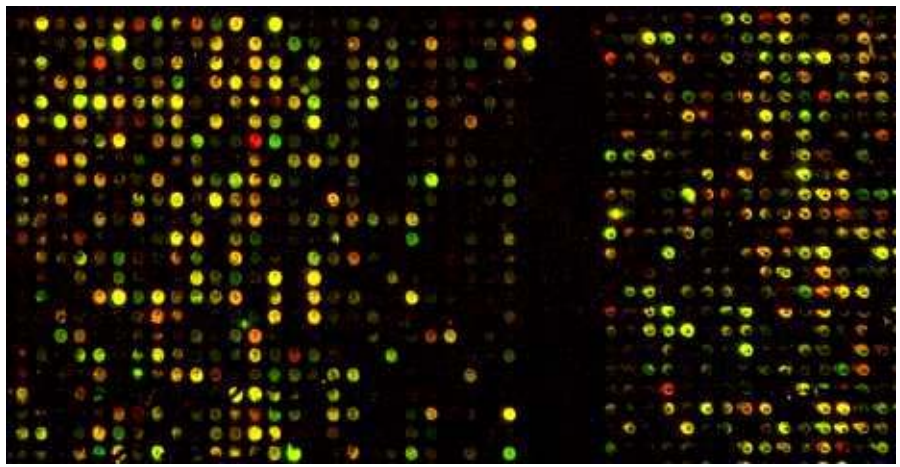

Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
Proportion of simulated data sets
1.00
0.95
0.90
0.10
0.05
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of misclassifications

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

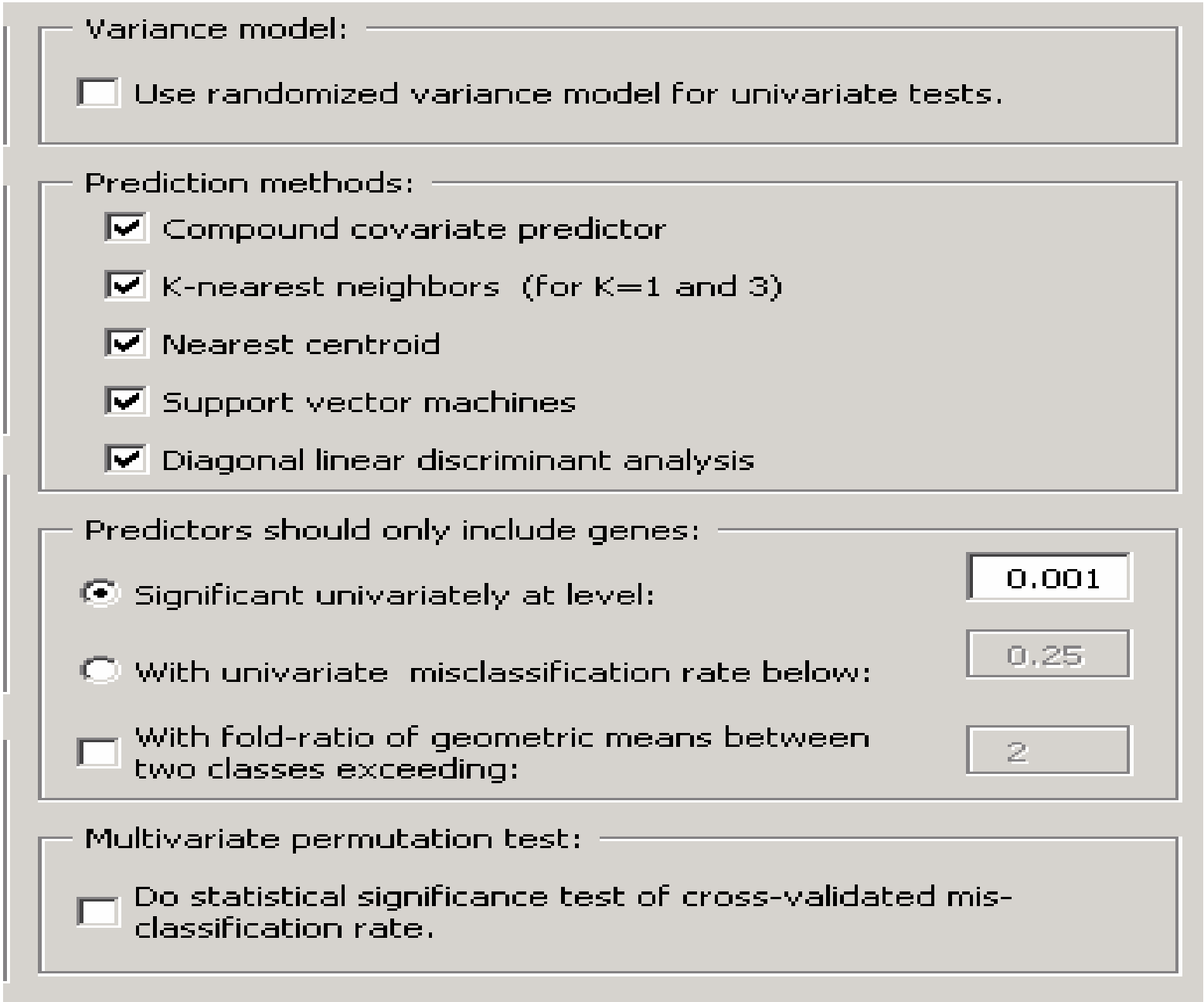
Can we distinguish *BRCA1*+ from *BRCA1*– cancers and *BRCA2*+ from *BRCA2*– cancers based solely on their gene expression profiles?

Classification of hereditary breast cancers with compound covariate predictor

Class labels	Number of differentially expressed genes (full data set, $\alpha = 0.0001$)	*m* = number of misclassifications using leave-one-out cross-validation	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

Class Prediction in BRB-ArrayTools

- Class prediction using a variety of prediction methods may be performed using BRB-ArrayTools. The predictors are automatically cross-validated, and a significance test may be performed on the cross-validated mis-classification rate. Independent test samples may also be classified using the predictors formed on the training set. The screenshot on the next page shows the available methods and options.


Variance model:
Use randomized variance model for univariate tests.
Prediction methods:
Compound covariate predictor
K-nearest neighbors (for K=1 and 3)
Nearest centroid
Support vector machines
Diagonal linear discriminant analysis
Predictors should only include genes:
Significant univariately at level:
0.001
With univariate misclassification rate below:
0.25
With fold-ratio of geometric means between two classes exceeding:
2
Multivariate permutation test:
Do statistical significance test of cross-validated mis-classification rate.

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, reference to book written by BRB members
  - BRB-ArrayTools software
  - .pdf of these talk slides: ftp://linus.nci.nih.gov/pub/techreport/CIT_course.pdf
- Berkeley: http://www.stat.berkeley.edu/users/terry/Group/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
- Whitehead Institute (Cancer Genomics Group): http://www.broad.mit.edu/cancer/index.html

Acknowledgements

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