Statistical Analysis of Microarray Data

Eric Polley Lori Long

Biometric Research Branch
National Cancer Institute

CIT, January 2014

Class Overview

1st Half Lecture on statistical analysis of microarray data – Eric Polley

2nd Half Hands-on BRB ArrayTools workshop – Lori Long

- Slides available at: `http://ecpolley.github.com/CIT_Microarray_Course/index.html`

Outline

Gene Expression Microarrays

Permit simultaneous evaluation of expression levels of thousands of genes

Popular Platforms:

- Spotted cDNA arrays (2-color)
- Affymetrix GeneChip (1-color)
- Spotted Oligo arrays (1- or 2-color)
- Bead arrays (e.g. Illumina-DASL)

Spotted cDNA Arrays

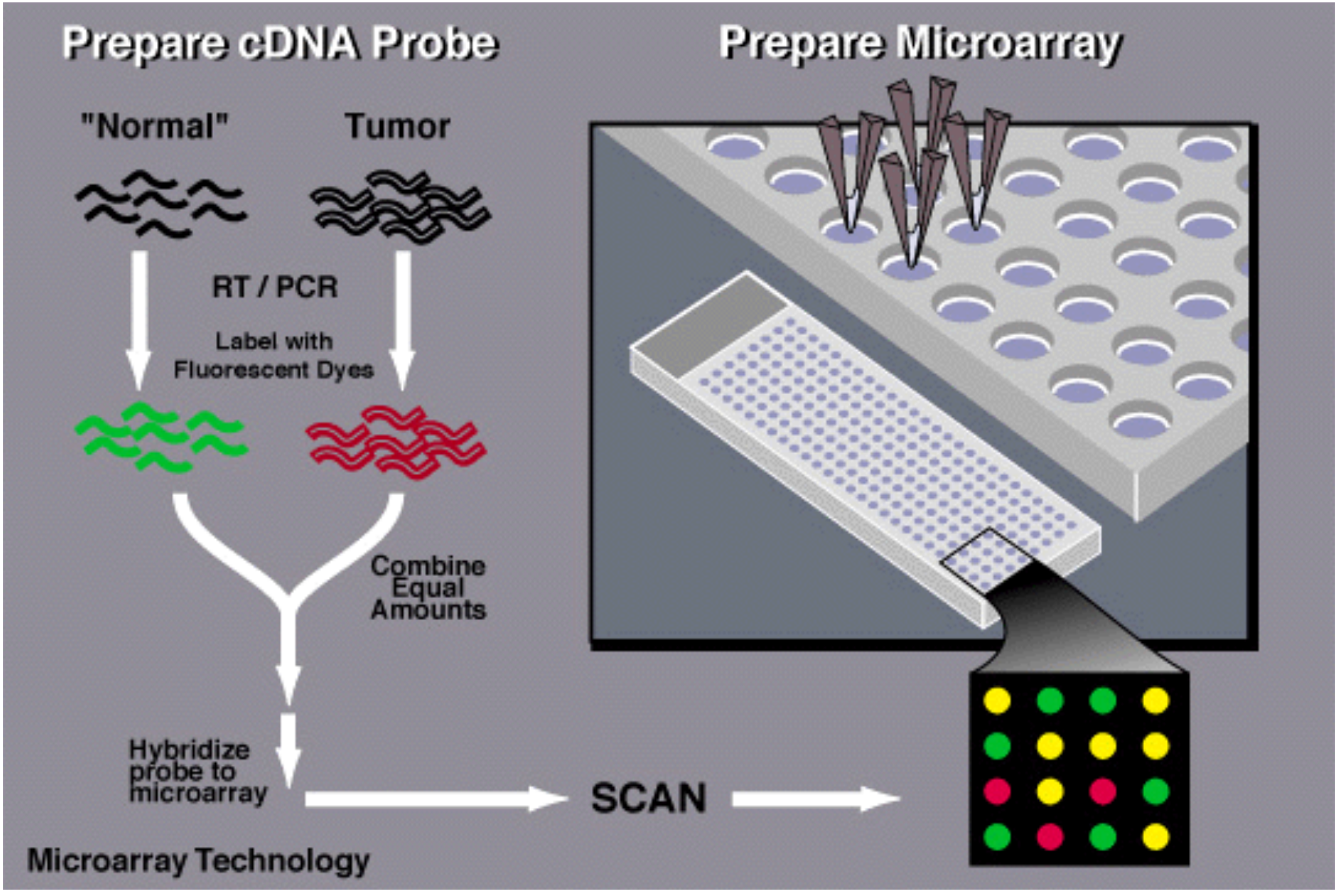
cDNA arrays: Schena *et al.*, *Science*, 1995.

Each spot corresponds to a gene. Sometimes multiple spots per gene.

Two-color (two-channel) system:

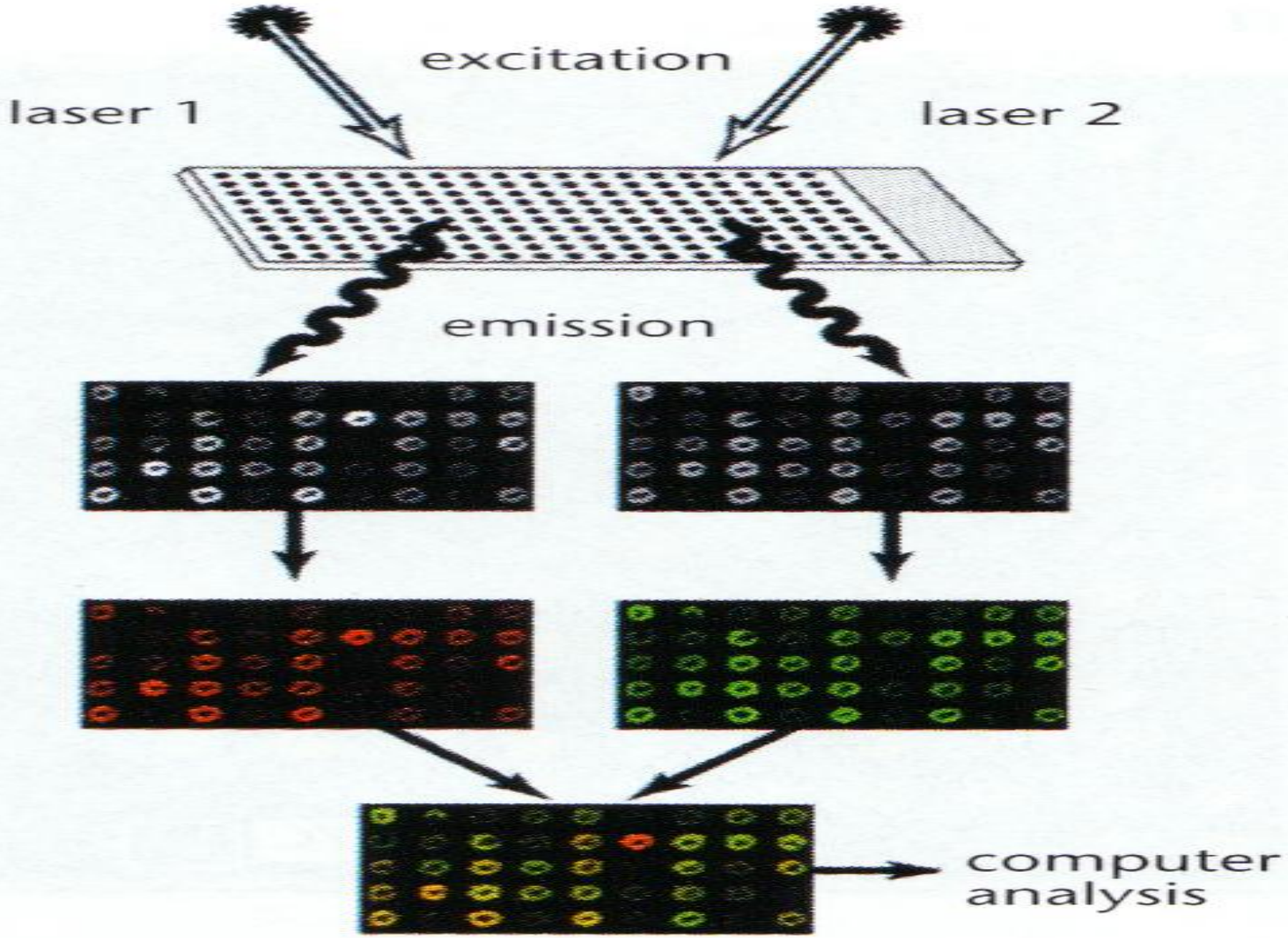
- Two colors represent the two samples competitively hybridized
- Each spot has “red” and “green” measurements associated with it.

cDNA Array



http://www.genome.gov/10000533/

cDNA Array



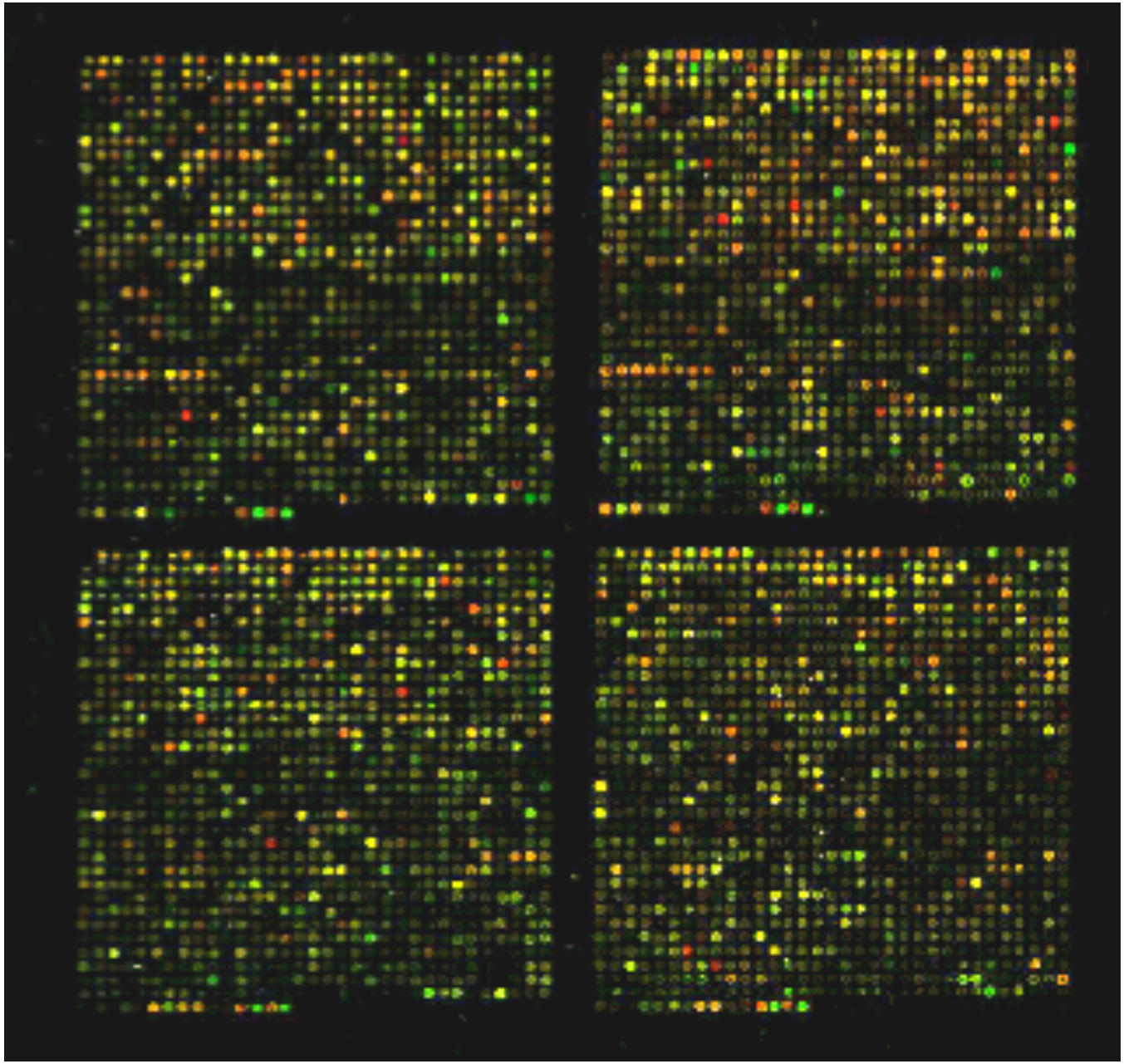


Figure: Overlaid “red” and “green” images for cDNA microarray

AFFYMETRIX GENECHIP

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 one or more probe sets:

- 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)

Figure: Affymetrix Oligo GeneChip array

---

http://en.wikipedia.org/wiki/DNA_microarray

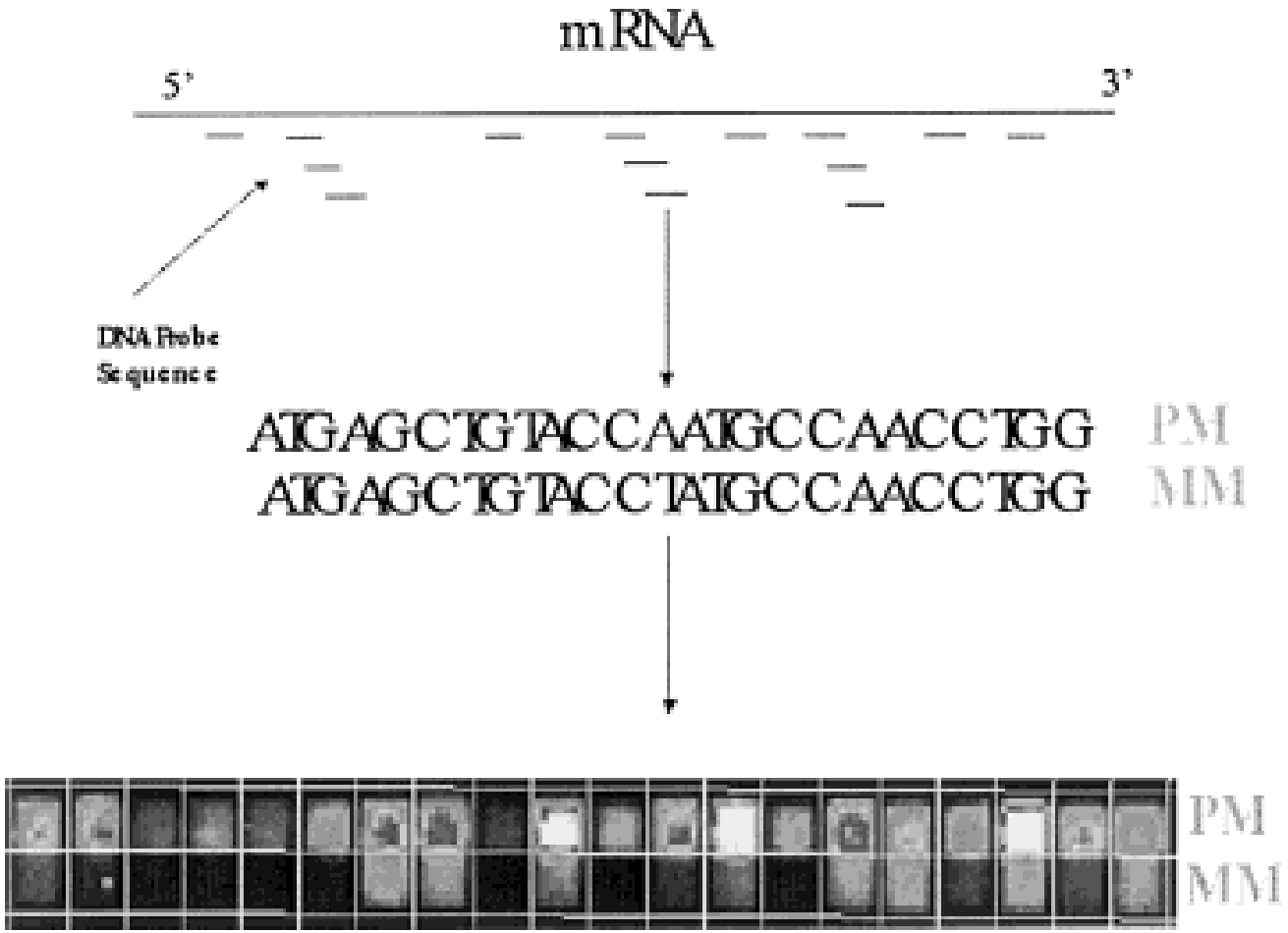


Figure: Perfect Match - Mismatch Probe Pair

From Schadt *et al.*, *Journal of Cellular Biochemistry*, 2001

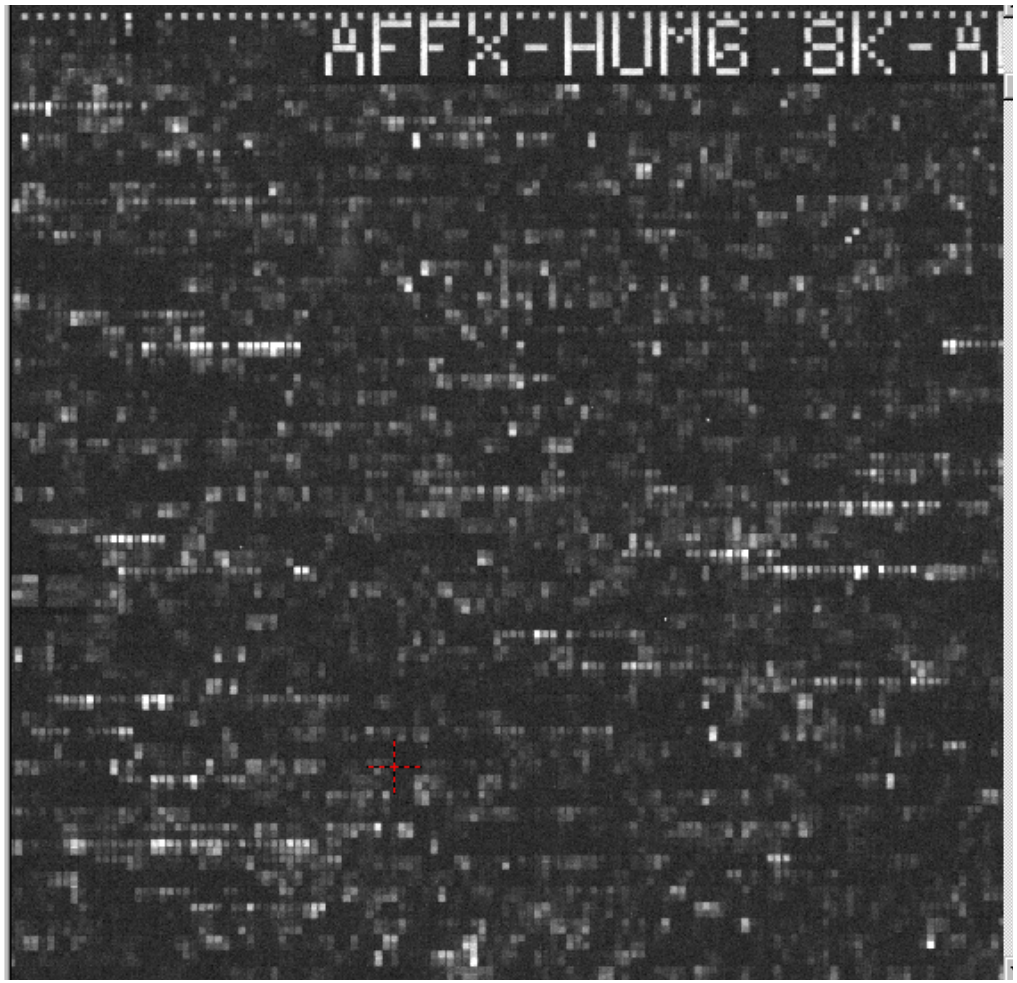


Figure: Image of scanned Affymetrix GeneChip

SPOTTED ARRAYS: QA

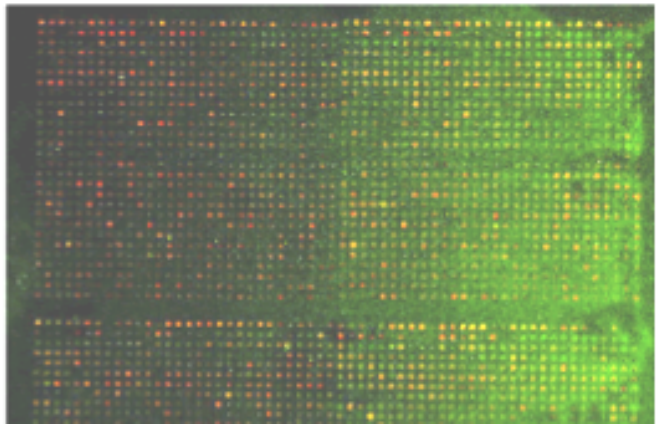


Figure: Background haze

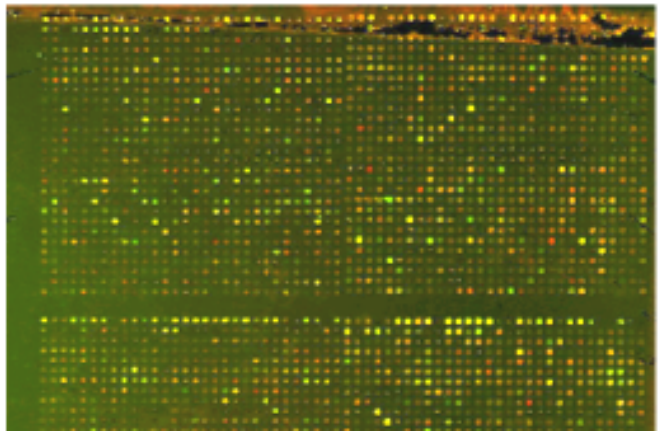


Figure: Edge effect

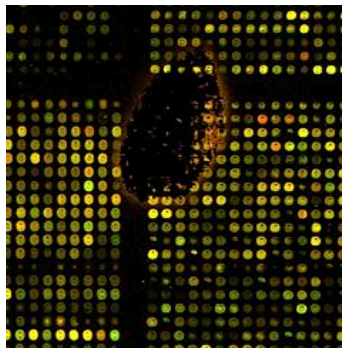


Figure: Bubble

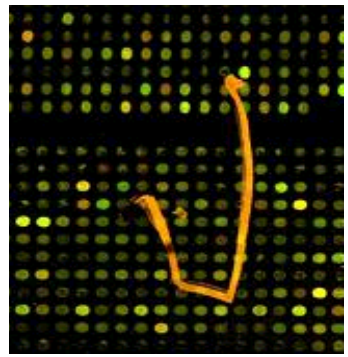


Figure: Fiber or Scratch

- Visual inspection of arrays advisable
- Danger:
  Garbage In $\Rightarrow$
  Garbage Out

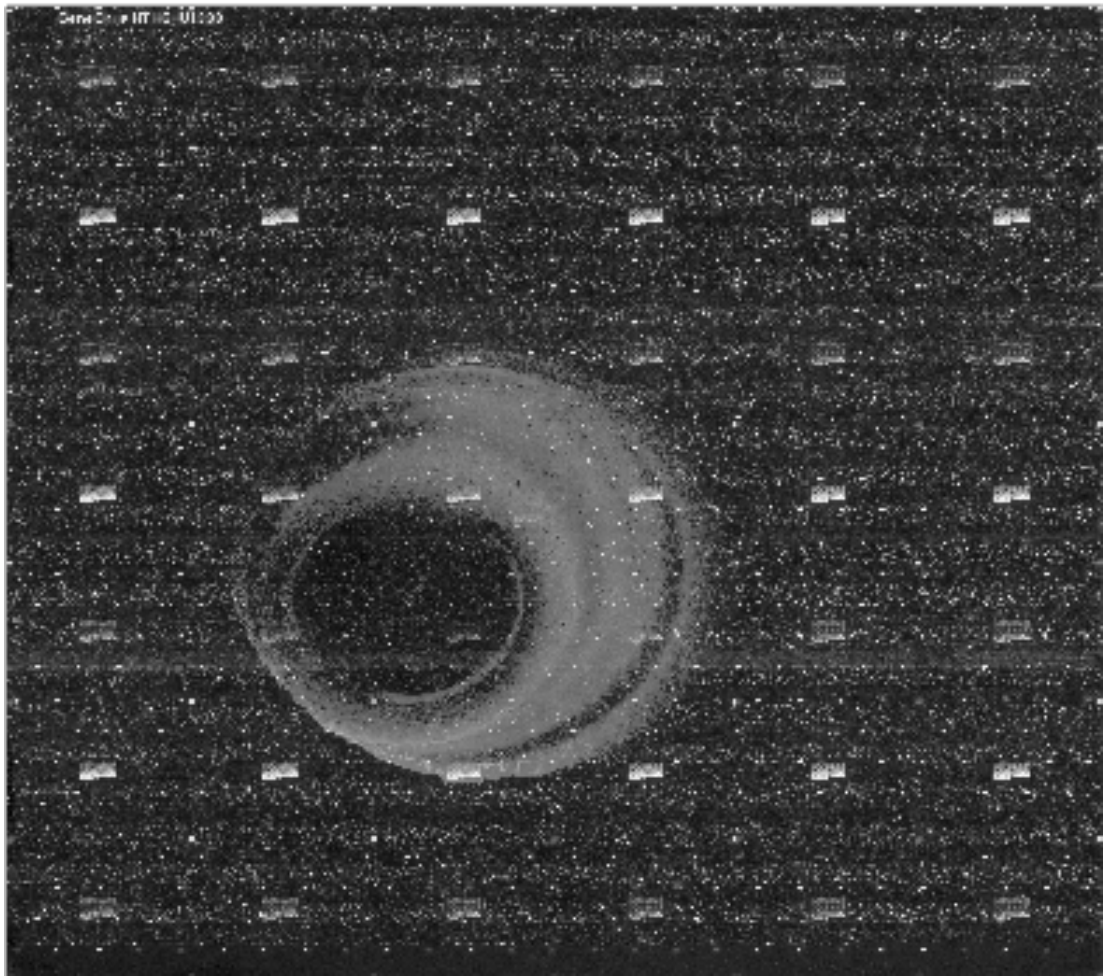


Figure: Affymetrix Arrays: Quality problems

Summarize

Which measurement to use as the gene expression value

Normalize

Remove unwanted variability (more on the sources of variability later)

2-Color Arrays: Gene Summary

Expression levels from the two colors are *(usually)* summarized by the log ratio for each spot:

$$X = \log_2 \left(\frac{\text{RED}}{\text{GREEN}} \right)$$

Affymetrix Arrays: Gene Summary

For Affymetrix arrays, more methods exist for summarizing the individual probes:

- AvDiff
- MAS5
- MBEI
- RMA
- GCRMA

Affymetrix Arrays: Gene Summary

Original Affymetrix algorithm (AvDiff):

$$Y_i = \sum_j \frac{1}{n_i} (PM_{ij} - MM_{ij})$$

Revised Affymetrix algorithm to address negative signals (MAS 5.0 series):

$$Y_i = \exp \{ave_T (\log(PM_{ij} - IM_{ij}))\}$$

where $ave_T(\cdot)$ is the Tukey biweight method and

$$IM = \begin{cases} MM & \text{if } MM < PM \\ PM - \delta & \text{if } MM > PM \end{cases}$$

AFFYMETRIX ARRAYS: GENE SUMMARY

Model based summaries from Li and Wong (*PNAS*, 2001; *Genome Biology*, 2001; incorporated into `dChip`)

- $MBEI_i = \theta_i$ estimated from:

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}$$

  where ϕ_j is the j^{th} probe sensitivity index and ε_{ij} is random error.

- $MBEI_i^* = \theta_i^*$ estimated from:

$$PM_{ij} = \nu_i + \theta_i^* \phi_j'$$

  where ϕ_j' is the j^{th} probe sensitivity index and ν_i is baseline response for i^{th} probe *pair*

AFFYMETRIX ARRAYS: GENE SUMMARY

- `RMA`: Irizarry *et al.* (*Nucleic Acids Research*, 2003; *Biostatistics*, 2003): $RMA_i = \mu_i$ estimated from

$$\mathrm{T}(PM_{ij}) = \mu_i + \alpha_j + \varepsilon_{ij}$$

  where $\mathrm{T}(PM_{ij})$ is the cross-hybridization corrected, (quantile-) normalized and log-transformed PM intensities.
- `GCRMA`: Wu *et al.* (*J. Amer. Stat. Assoc.*, 2004): Apply cross-hybridization correction that depends on G-C content of probe

Need for Normalization

For cDNA/2-color spotted arrays:

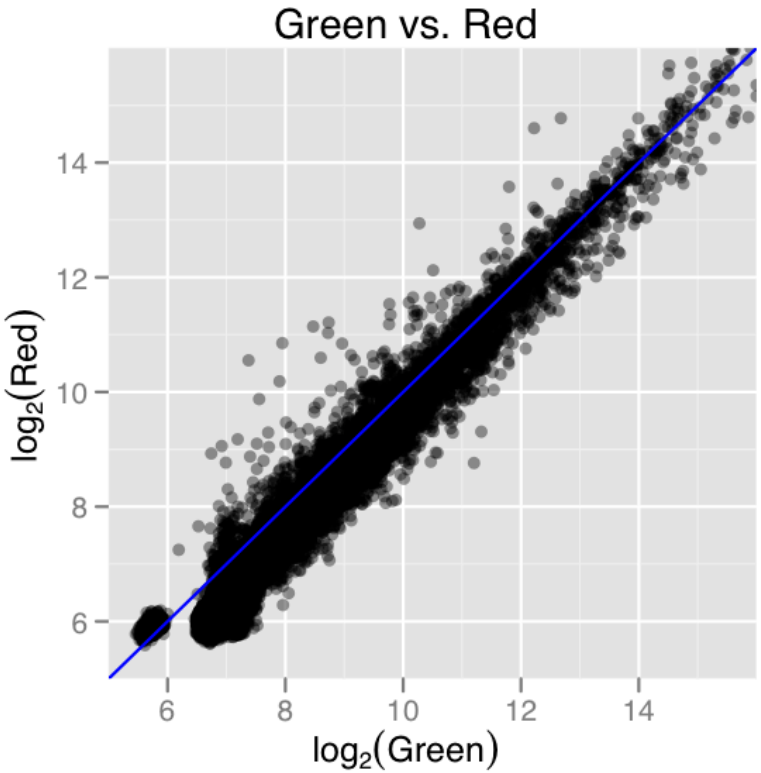
- Unequal incorporation of labels. Green brighter than red
- Unequal amounts of sample
- Unequal PMT voltage
- Autofluorescence greater at shorter scanning wavelength

2-Color Arrays: Normalization

Methods for ratio-based summaries:

- 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).

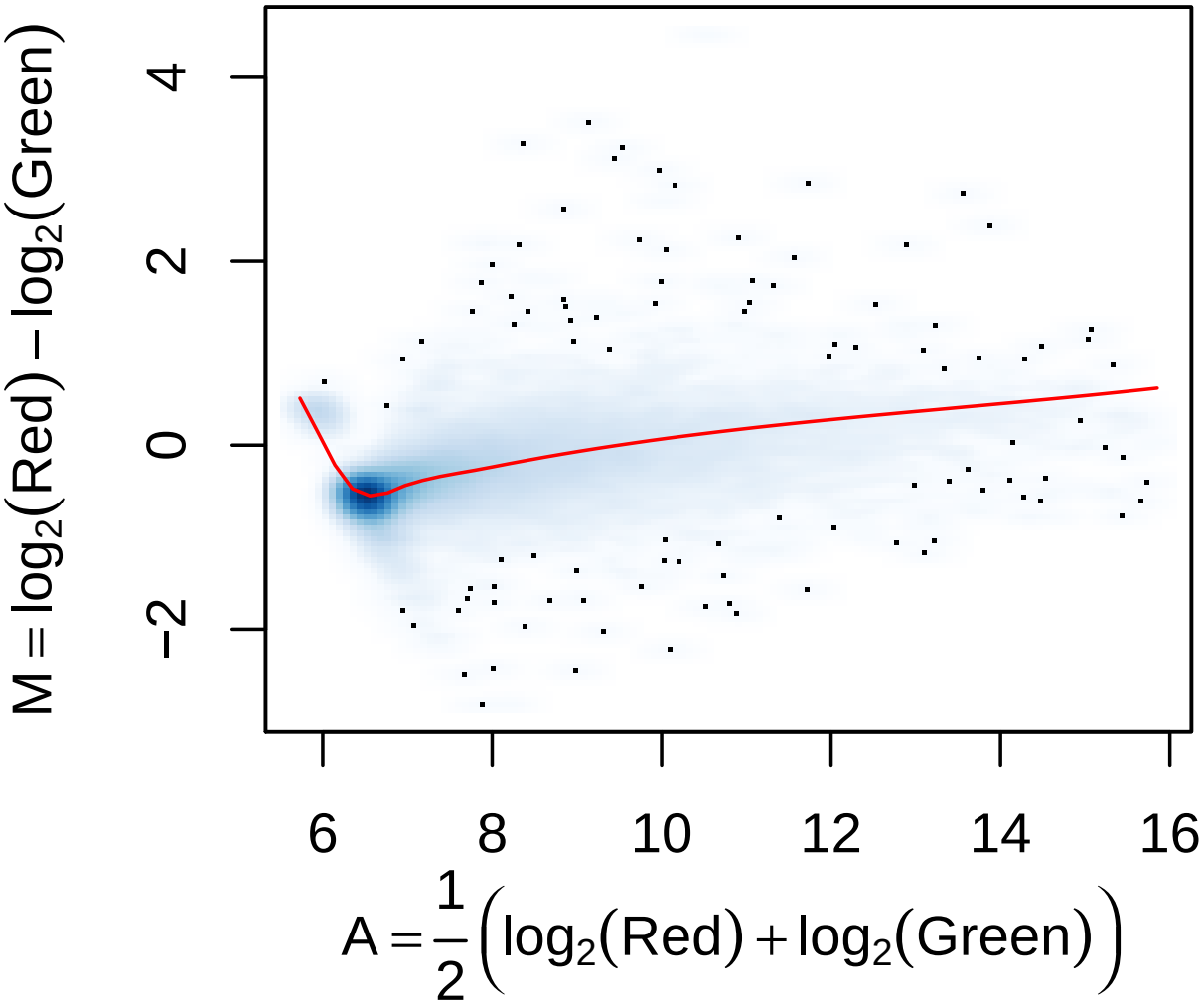
2-Color Arrays: Normalization



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

Data from Agilent *Rat Whole Genome* Array in `CC14` package

M vs A plot



Yang *et al.*, *Nucleic Acids Research*, 2002

2-Color Arrays: Normalization

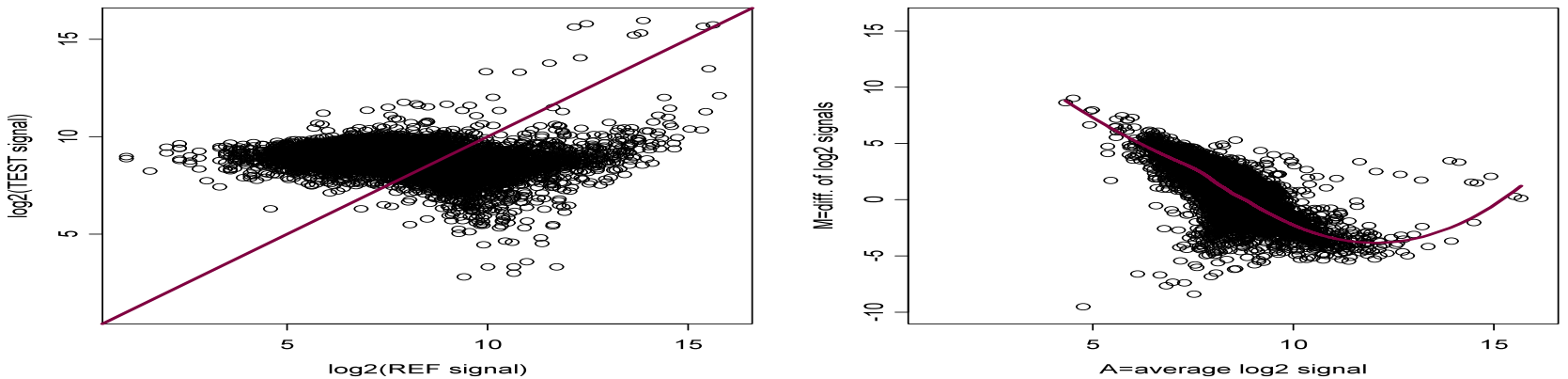


Figure: Bad Array Example

Affymetrix Normalization

- 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
- Built in to `dChip`, `RMA`, and `MAS 5.0 series` algorithms:
  - Li and Wong (*PNAS*, 2001; *Genome Biology*, 2001)
  - Irizarry *et al.* (*Nucleic Acids Research*, 2003; *Biostatistics*, 2003)
  - Bolstad *et al.* (*Bioinformatics*, 2003)

After QC and normalization, filtering genes can increase the statistical power for most study objectives.

Very important: The filtering needs to be done correctly to not adversely impact downstream analyses.
(Bourgon *et al.*, *PNAS* 2010).

- "Bad" or missing values on too many arrays
- Not differentially expressed across arrays (non-informative). Two ways to identify genes:

Variance

s_i^2 is the sample variance of (log) measurements of gene i ($i = 1, 2, \ldots, K$). Exclude gene i if:

$$(n-1)s_i^2 < \chi^2(\alpha, n-1) \times \text{median}(s_1^2, s_2^2, \ldots, s_K^2)$$

Fold Difference

Exclude gene i if:

$\max_i / \min_i < 3$ or 4; or $95^{th}\%/5^{th}\% < 2$ or 3.

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

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

Global tests

- Compare whole profiles
- Statistical significance based on permutation test

Gene-level analyses

- Model-based methods (e.g. multi-parameter)
- Test-based methods (e.g. t-tests, nonparametric tests)
- Hybrid variance methods

Global tests for differences in profiles between classes:

- Choice of summary measure of difference, for example:
  - Sum of squared univariate t-statistics
  - Number of genes univariately significant at α 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

Model-based methods

Multi-parameter modeling of channel-level data (e.g. Gaussian mixed models), hierarchical Bayesian models, etc. May borrow information across genes and use multiple comparison adjustments.

Test-based methods

t-test, F-test, or Wilcoxon tests for each gene. Multiple comparison adjustment commonly used.

Goal:

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

Before going into multiple testing, a review of single testing:

Let Y_1 be the expression levels for gene Y in group 1 and Y_2 the expression levels in group 2. The two-sample t-test is:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s is a measure of the within gene variability.

For testing, we set up the null and alternative hypothesis:

H_0 : The two groups have the same expression level for gene Y, i.e. NOT differentially expressed.

H_1 : The gene is differentially expressed between the two groups.

We define a cut-off value for the test statistic based on the sample size and the α value (usually $\alpha = 0.05$) and claim the gene is DE if:

$$|t| \geq t_\alpha$$

Why adjust for multiple testing?

If we had 10,000 genes and performed a t-test for each gene, even if none were DE, we would expect $10,000 \times 0.05 = 500$ genes to have a p-value less than 0.05.

We mostly describe two group comparisons with the t-test, but most of the material here is applicable to other settings such as:

- Paired samples (paired t-test)
- More than 2 groups (ANOVA)
- Time course studies
- eQTL
- Time-to-event data (Survival data)

	Don't Reject	Reject	
True Null	$\mathbf{U}$	$\mathbf{V}$	m_0
False Null	$\mathbf{T}$	$\mathbf{S}$	$m - m_0$
	$m - \mathbf{R}$	$\mathbf{R}$	m

We compute m tests where m_0 are true nulls and $m - m_0$ are false nulls (DE).

The test rejects $\mathbf{R}$ out of m hypotheses, with $\mathbf{S}$ correctly rejected. $\mathbf{V}$ represents a type I error and $\mathbf{T}$ represents a type II error.

Adapted from Benjamini and Hochberg, *JRSS-B*, 1995.

Different ways to control:

- Actual number of false discoveries: FD
- Expected number of false discoveries: $\mathrm{E}(\mathrm{FD})$
- Actual proportion of false discoveries: FDP
- Expected proportion of false discoveries: $\mathrm{E}(\mathrm{FDP})$ = false discovery rate (FDR)

$$\mathrm{FD} = \mathbf{V} \quad \text{and} \quad \mathrm{FDP} = \mathbf{V}/\mathbf{R}$$

Multiple Testing: Simple Procedures

Control expected number of false discoveries

- $\mathrm{E}(\mathrm{FD}) \leq u$
- conduct each of m tests at level u/m

Bonferroni control of family-wise error (FWE)

- Conduct each of m tests at level α/m
- At least $(1-\alpha)100\%$ confident that $\mathrm{FD} = 0$

Multiple Testing: Simple Procedures

Problems with the 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
  - Esimate quantities similar to FDR (old SAM) or control FDP (new SAM)
- Bayesian
  - Efron *et al.*, *JASA*, 2001
  - Manduchi *et al.*, *Bioinformatics*, 2000
  - Newton *et al.*, *J Comp Bio*, 2001
- Step-down permutation procedures
  - Westfall and Young, 1993 Wiley (FWE)
  - Korn *et al.*, *JSPI*, 2004 (FD and FDP control)

Korn *et al.* FD

$\mathrm{FD}(2)$: We are 95% confident that the actual number of false discoveries is not greater than 2

Korn *et al.* FDP

$\mathrm{FDP}(0.10)$: We are 95% confident that the actual proportion of false discoveries does not exceed 0.10

Tusher *et al.* SAM

$\mathrm{SAM}_{old}(0.10)$: On *average*, the false discovery proportion will be controlled at 0.10

Current SAM

$\mathrm{SAM}_{new}(0.10)$: Similar to Korn FDP procedure

Bayesian Methods

Higher posterior probability of differential expression

The step-down permutation procedure for FD and FDP control is available in BRB-ArrayTools for class comparison, survival analysis, and quantitative traits analysis.

Can also set number of permutations

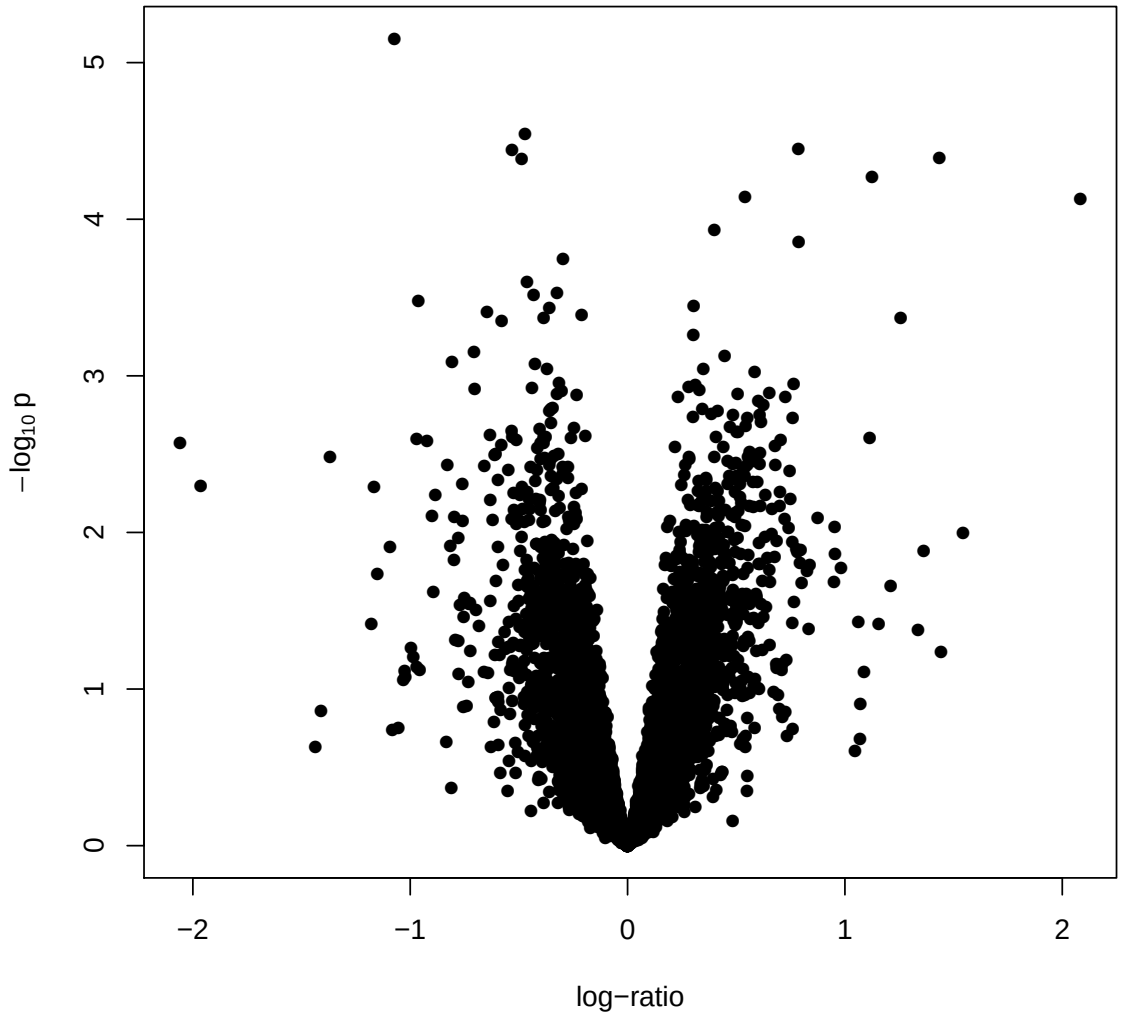


Figure: Volcano Plot showing fold change versus p-value

Gene Set Enrichment Analysis

The methods above test each gene individually, but this can be difficult for the following reasons:

- Large number of genes (need to control Type I error rate)
- Small sample size
- Large variability across samples

A complementary approach looks for class differences in *pre-defined* sets of genes:

- Pathways (KEGG, BioCarta)
- Chromosome
- Protein domains

Gene Set Enrichment Analysis (GSEA) was proposed by Subramanian *et al.* (*PNAS*, 2005) and has been extended by many others.

With a pre-defined set of (possibly over-lapping) gene sets, $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K$ and the statistics for each gene, $\mathbf{T} = \{t_1, t_2, \ldots, t_p\}$, define a gene-set score:

$$S_k(T) = KS(k, T), \quad \text{For } k = 1, \ldots, K$$

Where $KS(k, T)$ is the Kolmogorov-Smirnov test comparing the empirical distributions between

$$\{t_j : j \in \mathcal{S}_k\} \quad \text{and} \quad \{t_j : j \notin \mathcal{S}_k\}$$

The Kolmogorov-Smirnov test is not critical to the GSEA procedure. Another popular score in the Maxmean from Efron and Tibshirani (*Ann. App. Stat.*, 2007). Both are available in BRB-ArrayTools.

To test the significance of the gene set score, a permutation test is run.

BREAK

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

Cluster Analysis

- Hierarchical
- K-means
- Self-organizing maps
- Maximum likelihood/mixture models
- Many more...

Graphical Displays

- Dendrogram
- Heatmap
- Multidimensional scaling plot

Hierarchical Agglomerative Clustering Algorithm

Two approaches:

1. Cluster genes with respect to expression across specimens
2. Cluster specimens with respect to gene expression profiles

Often helpful to filter genes that show little variation across specimens and median/mean center genes.

Hierarchical Agglomerative Clustering Algorithm

Merge two "closest" observations into a cluster. Continue merging closest clusters/observations.

Two things to define:

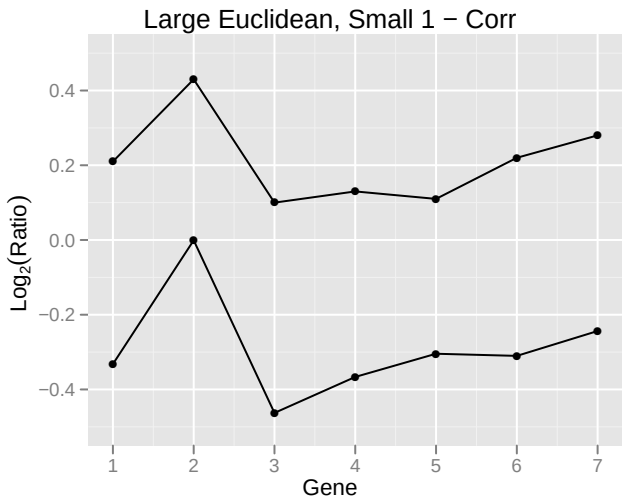
1. How is distance between individuals measured?
   - Euclidean
   - Maximum
   - Manhattan
   - 1 - Correlation
2. How is distance between clusters measured?
   - Average linkage
   - Complete linkage
   - Single linkage

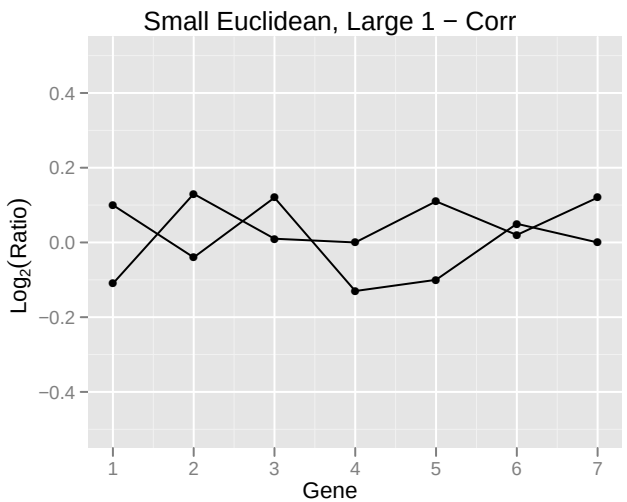
Euclidean distance:

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

1 - Correlation:

Large values reflect lack of linear association (pattern dissimilarity)





Average Linkage

Merge clusters whose average distance between all pairs is minimized. Particularly sensitive to distance metric

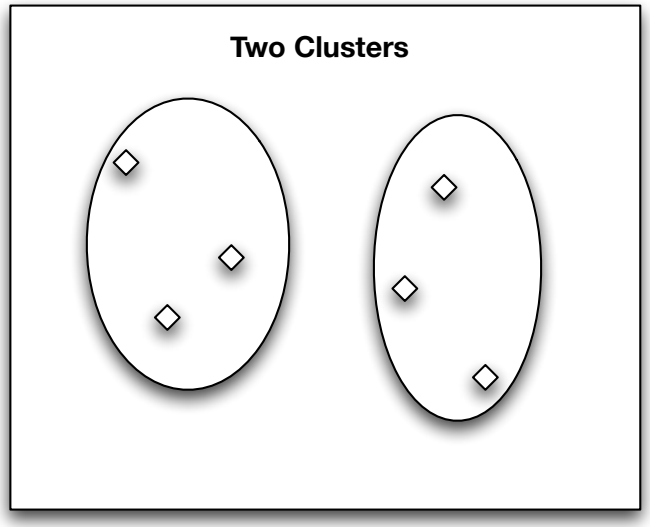
Complete Linkage

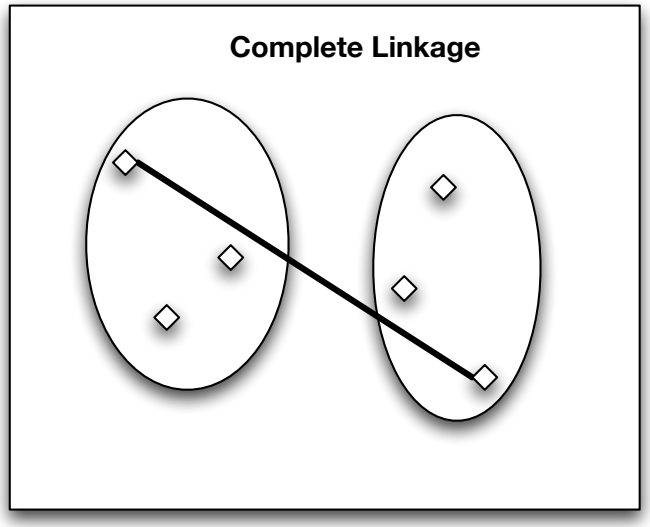
Merge clusters to minimize the maximum distance within any resulting cluster. Tends to produce compact clusters

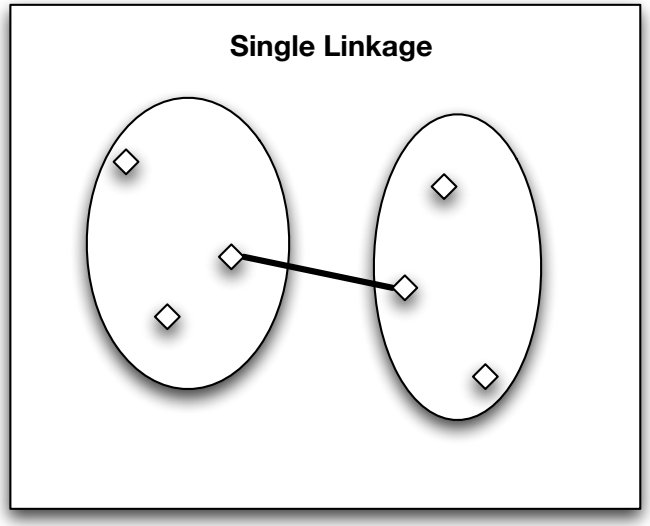
Single Linkage

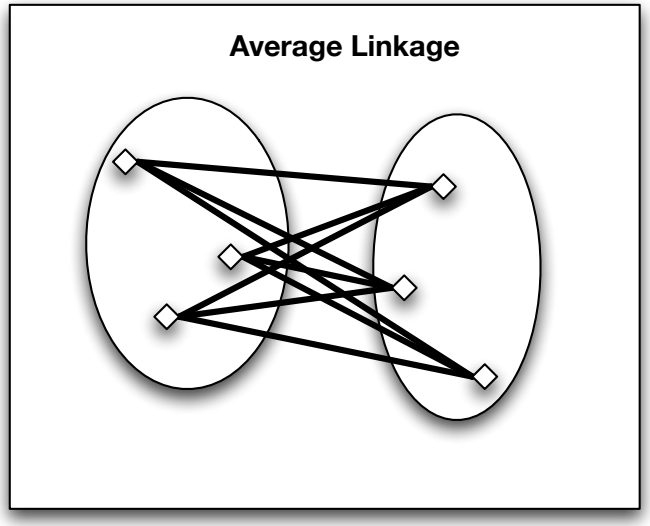
Merge clusters at minimum distance from one another. Prone to “chaining” and sensitive to noise

Linkage Methods

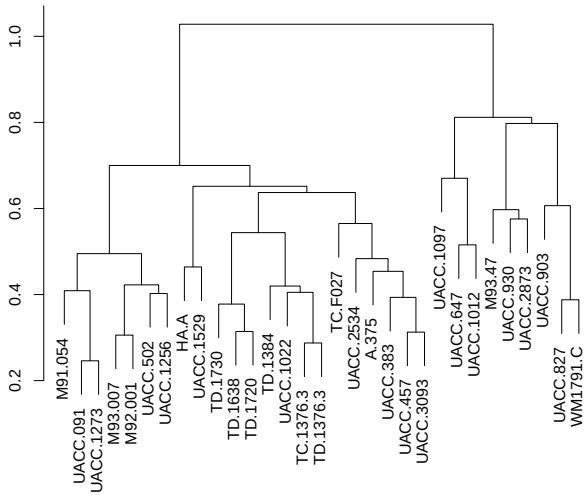




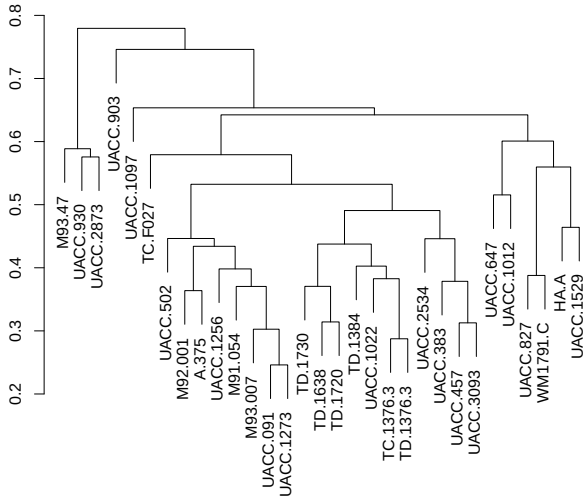




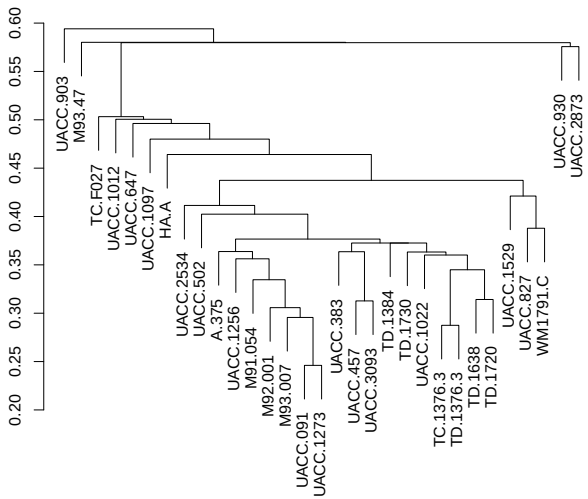
Complete



Average

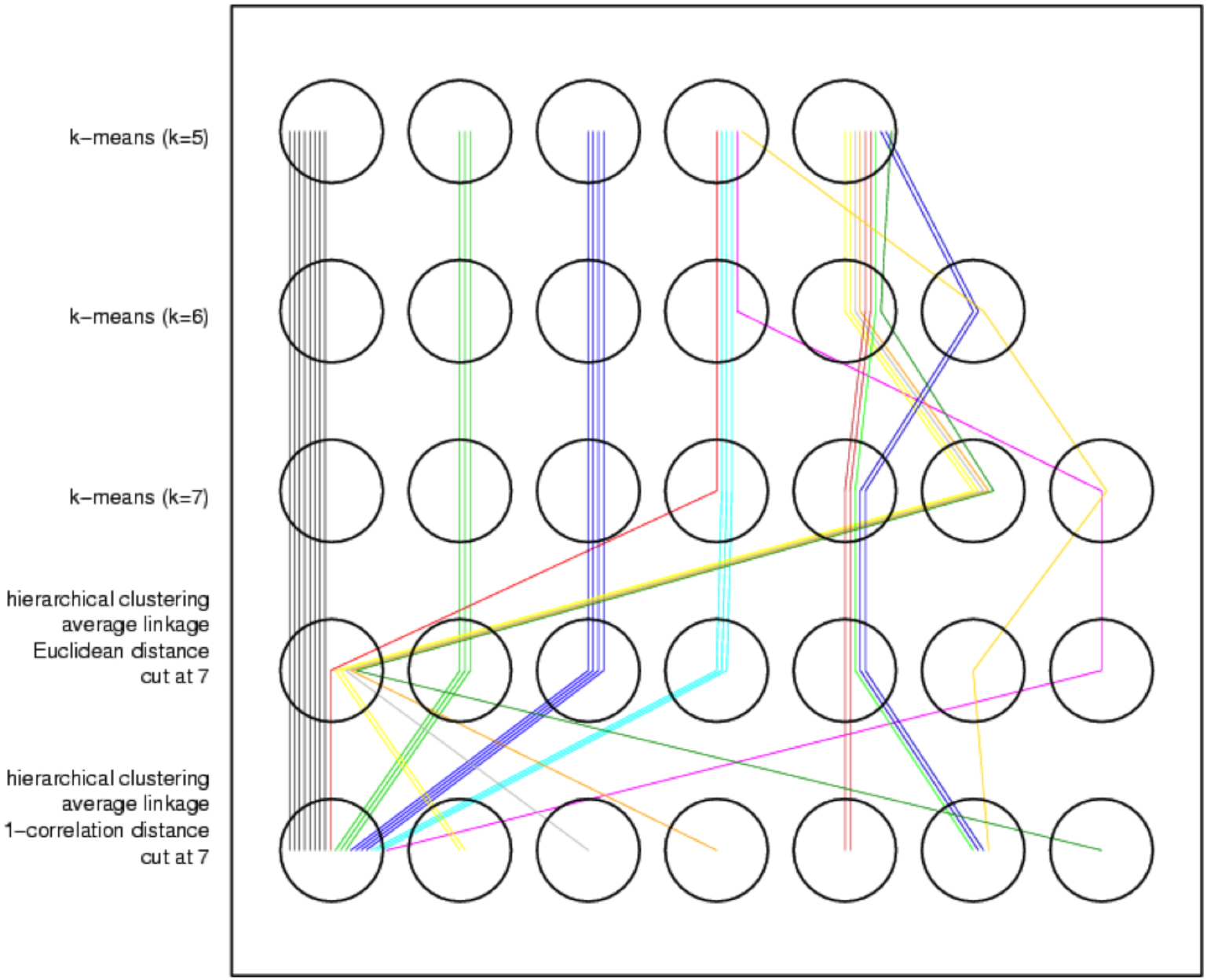


Single



Dendrograms using 3 different linkage methods, 1 - Corr distance (Data from Bittner *et al.*, *Nature*, 2000)

Does Cluster Method Matter?



How to interpret the 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 clusters
- Different clustering algorithms may find different structure using the same data

Global test of clustering

Based on inter-sample distances in transformed dimension-reduced space

Available as an option in BRB-ArrayTools for multidimensional scaling of samples

Assessment of reproducibility

Are the individual clusters using the selected cuts of the dendrogram in hierarchical clustering reproducible? (McShane *et al.*, *Bioinformatics*, 2002)

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:

- Gaussian Errors –McShane *et al.*, *Bioinformatics*, 2002
- Bootstrap residual errors–Kerr and Churchill, *PNAS*, 2001

Perturbation Method in BRB-ArrayTools

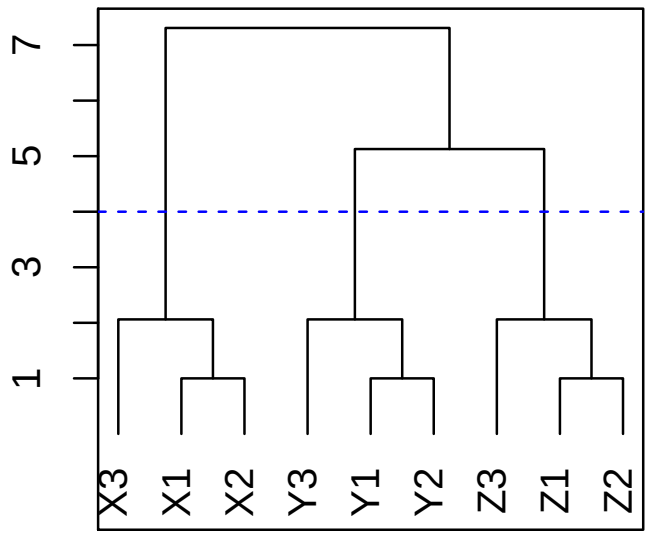
Perturb the log-gene measurements by adding Gaussian noise and then re-cluster. For each cluster:

1. Compute proportion of elements that occur together in the original cluster and remain together in perturbed data clustering when cutting dendrogram at the same level k
2. Average the cluster-specific proportions over many perturbed data sets to get an R-index for each cluster
3. the R-index may be obtained in BRB-ArrayTools for the hierarchical clustering of samples by selecting the 'Compute cluster reproducibility measures' options†
4. Hope for R-index ≥ 0.75

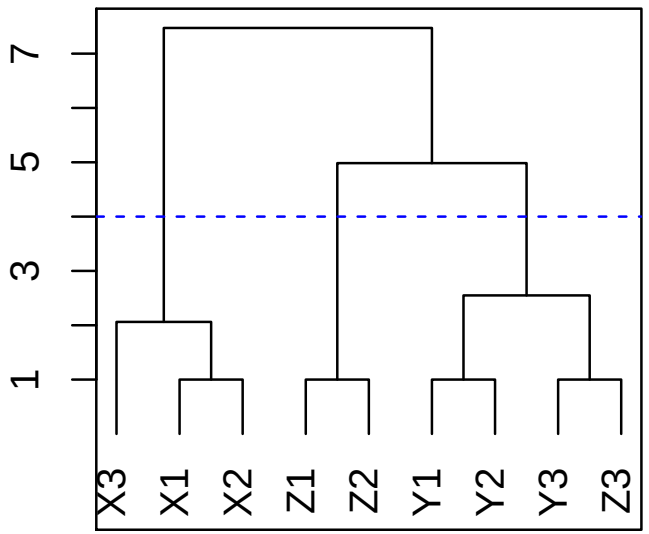
† R-index not available for gene clustering

R-INDEX EXAMPLE

Original Data



Perturbed Data



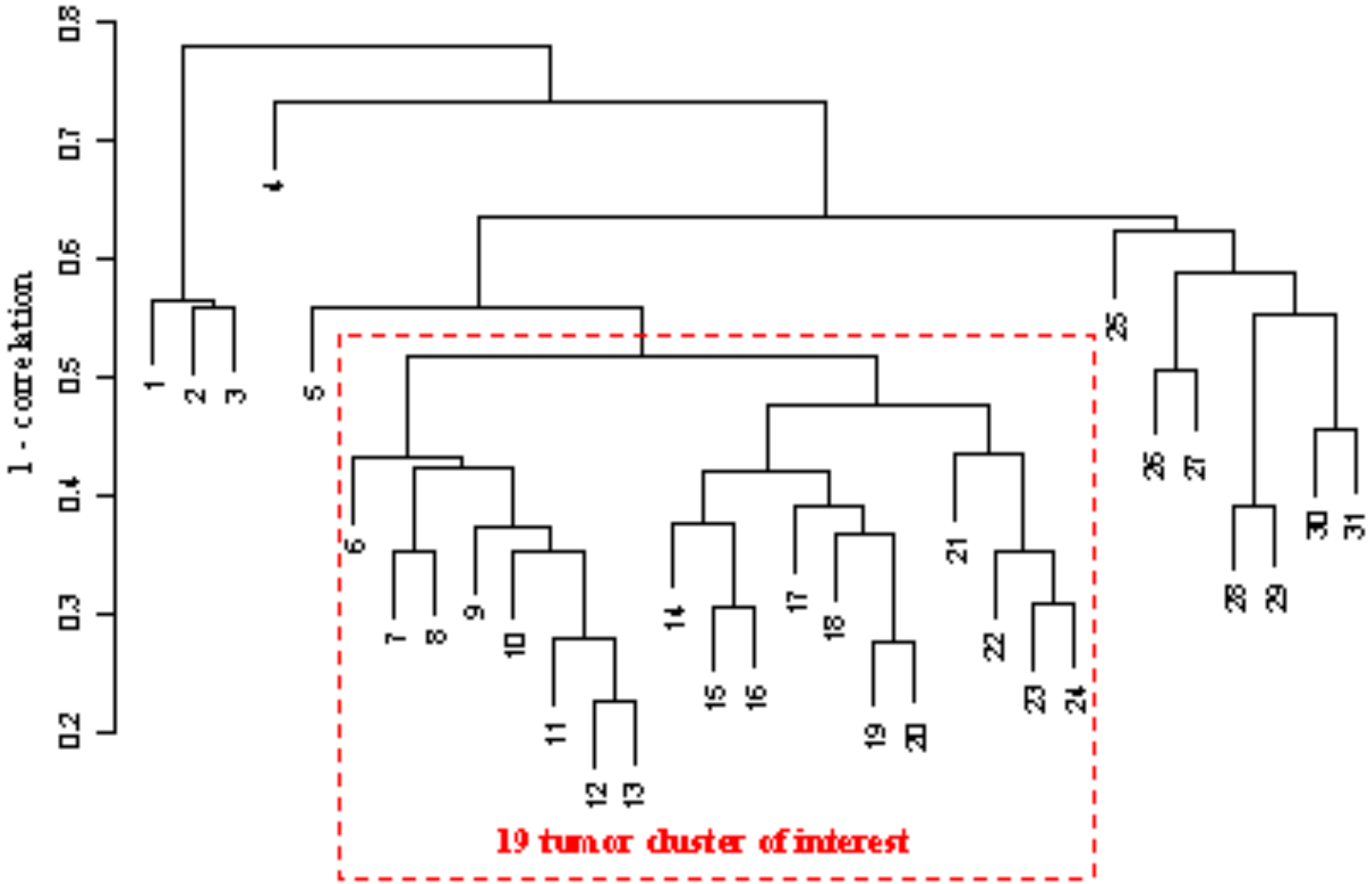
- 3 out or 3 pairs in X remain together
- 3 out of 3 pairs in Y remain together
- 1 out of 3 pairs in Z remain together
- $R = (3 + 3 + 1)/(3 + 3 + 3) = 0.78$

Cluster Reproducibility: Melanoma

From 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

CLUSTER REPRODUCIBILITY: MELANOMA



Cluster Reproducibility: Melanoma

For hierarchical clustering, the cluster of interest had R-index = 1.0 (highly reproducible)

Melanomas in the 19 element cluster tended to have:

- reduced invasiveness
- reduced mortality

Estimating Number of Clusters

- GAP statistic (Tibshirani *et al.*, *JRSS B*, 2002): detects too many false clusters (not recommended).
- Yeung *et al.* (*Bioinformatics, 2001*): jackknife method, estimate # of gene clusters.
- Dudoit *et al.* (*Genome Biology, 2002*): prediction-based resampling.
- Comparison of methods for estimating number of clusters (Milligan and Cooper, *Psychometrika*, 1985): uncertain performance in high dimensions.

Heat Map

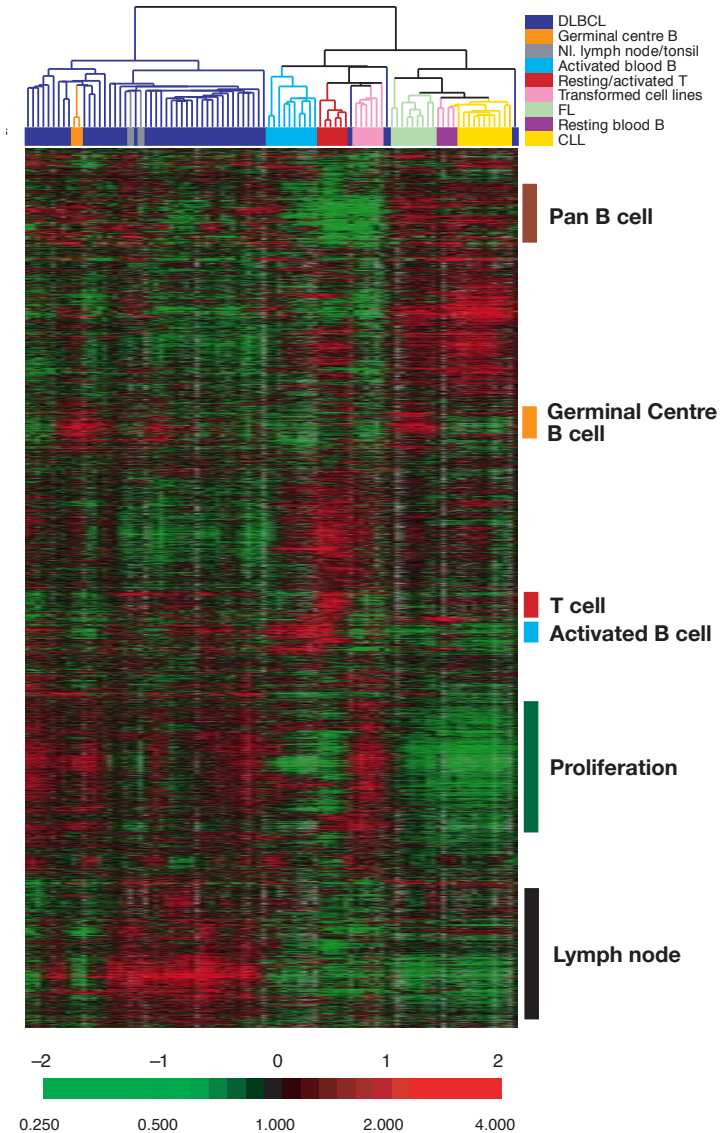


Figure: Lymphoma data (Alizadeh *et al.*, *Nature*, 2000)

Multidimensional Scaling (MDS)

High-dimensional data points are represented in a lower-dimensional space (e.g. 3D):

- Principal components or optimization methods
- Depends only on pairwise distances between points
- “Relationships” need not be well-separated clusters

Multidimensional Scaling

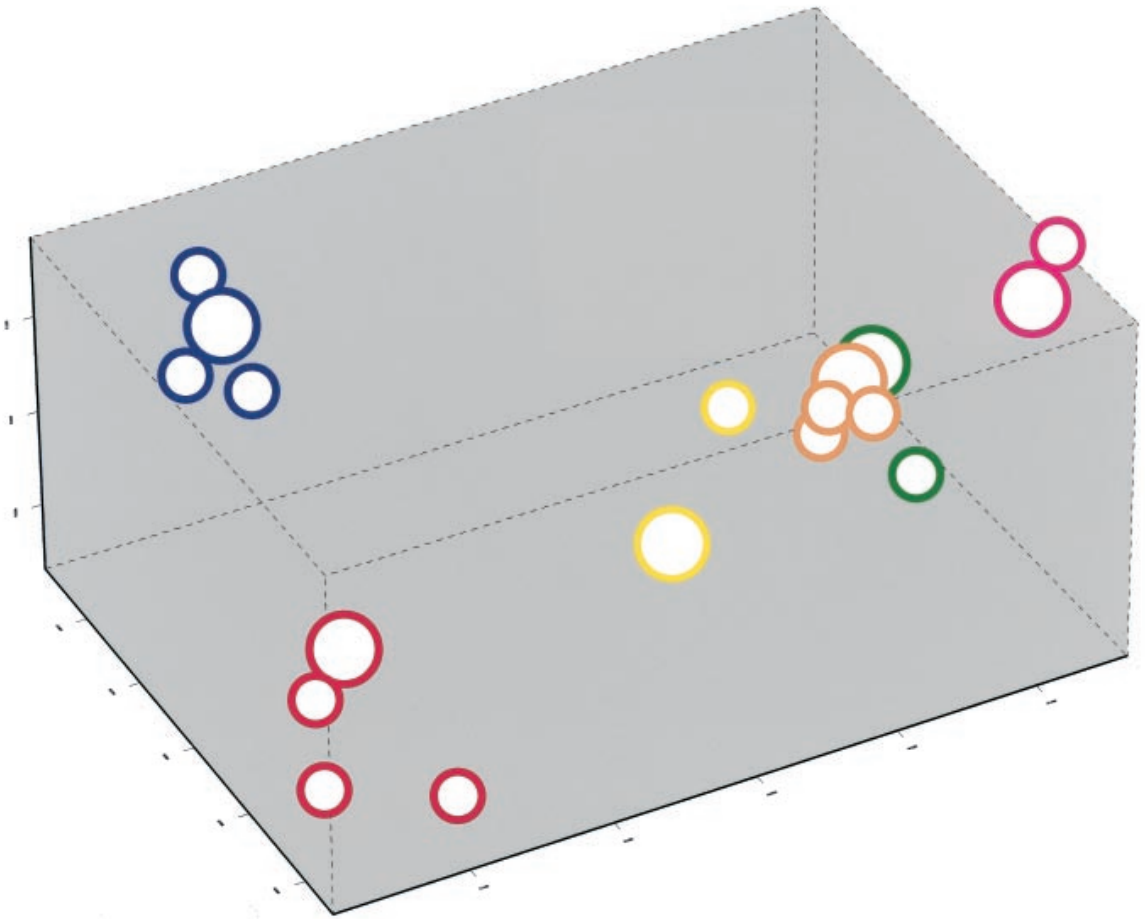


Figure: Color = patient, large circle = tumor, small circle = FNA. Assersohn *et al.*, Clinical Cancer Research, 2002.

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

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 actue 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. Bio.*, 2002).
- Support vector machines: classified ovarian tissue as normal or cancerous (Furey *et al.*, *Bioinformatics*, 2000).
- And many more...

COMPOUND COVARIATE PREDICTOR (CCP)

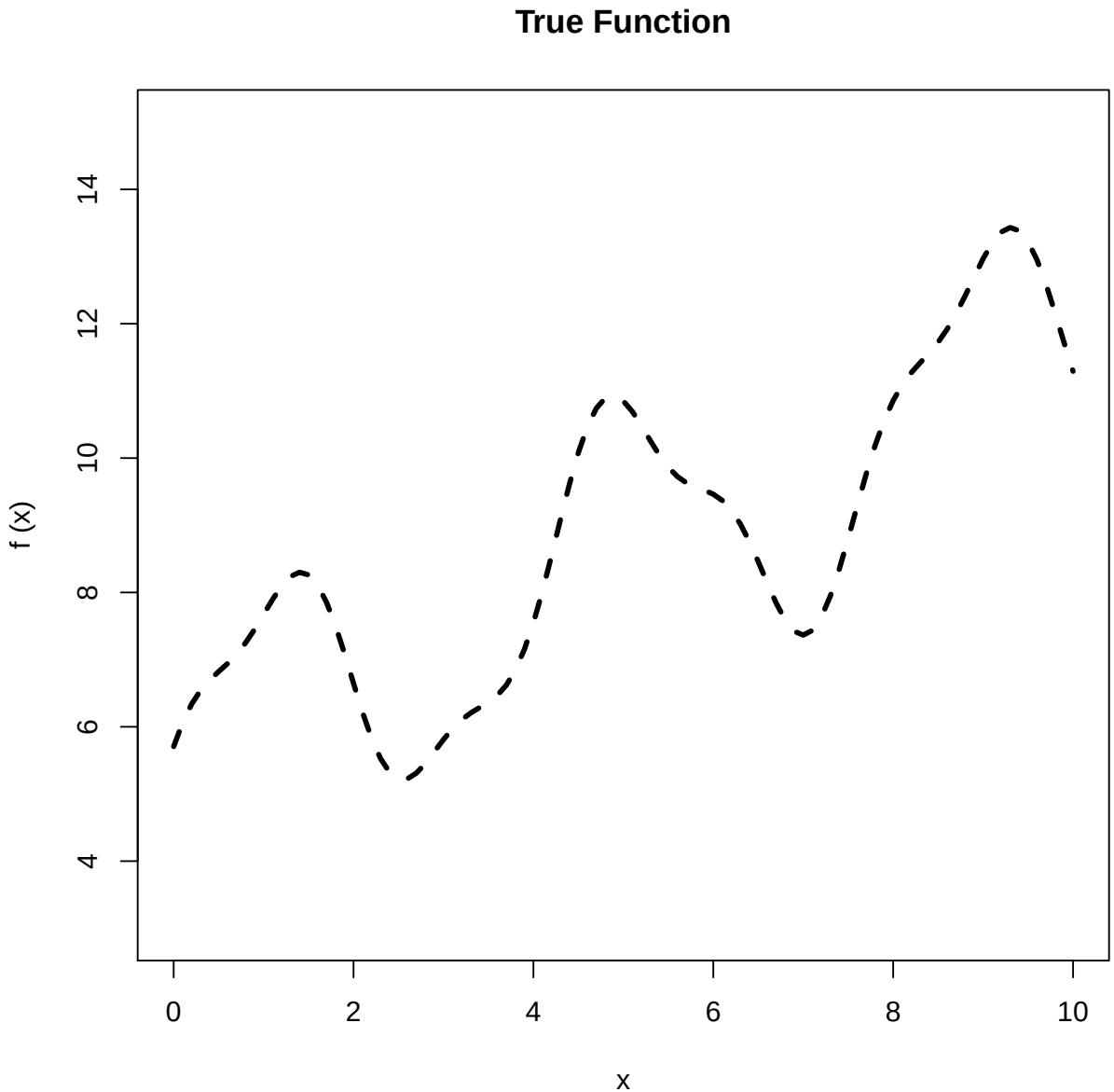
- Select "differentially expressed" genes by two-sample t-test with small α.

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

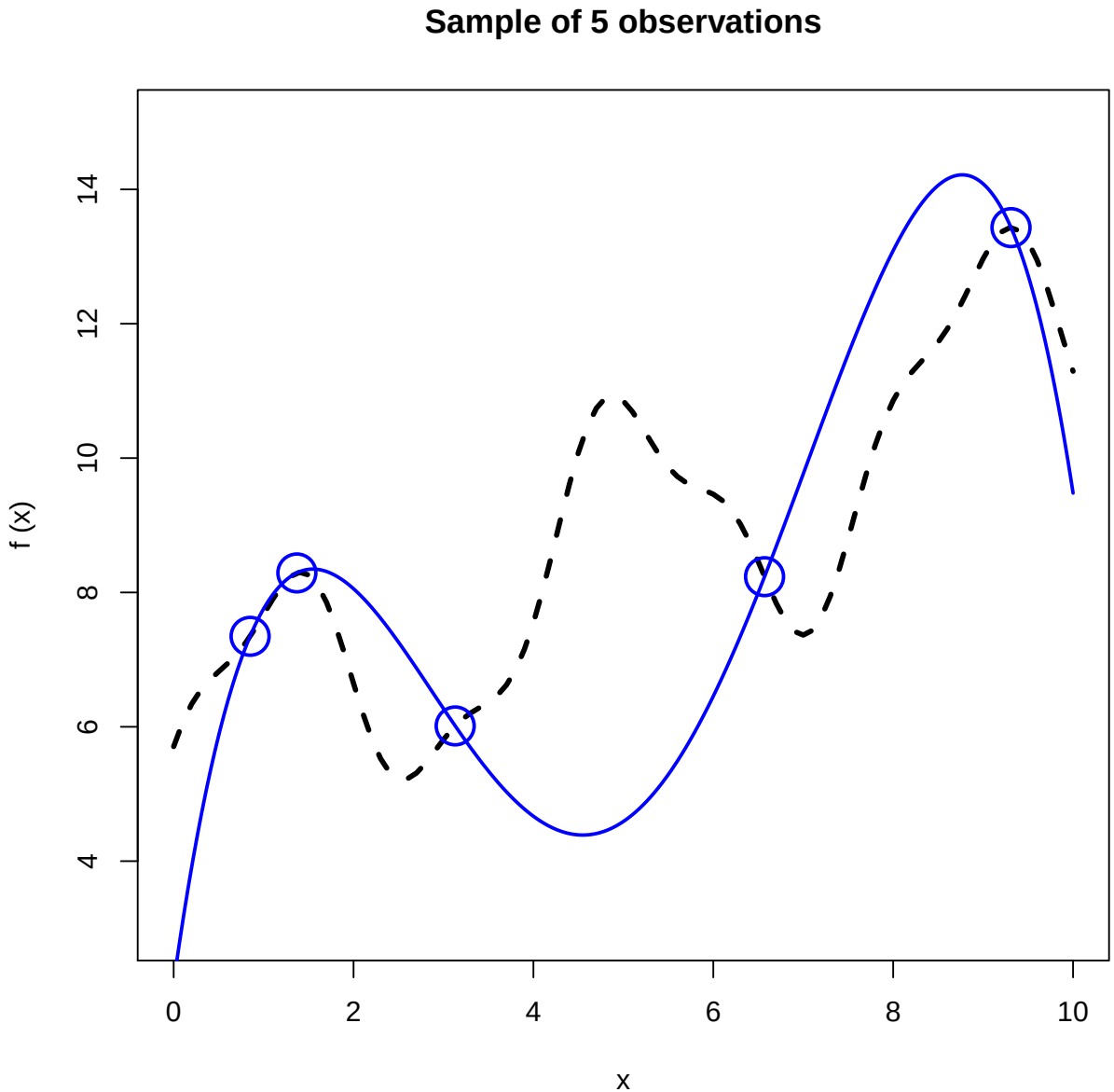
  t_j is the t-statistic for gene j,
  x_{ij} is the log expression measure for gene j in sample i,
  d is the number of differentially expressed genes (at level α).
- Threshold of classification: midpoint of the CCP means for the two classes.
- *Ref:* Tukey. *Controlled Clinical Trials*, 1993; Radmacher *et al.*, *J. Comp. Bio.*, 2002.

CLASSIFICATION PITFALLS

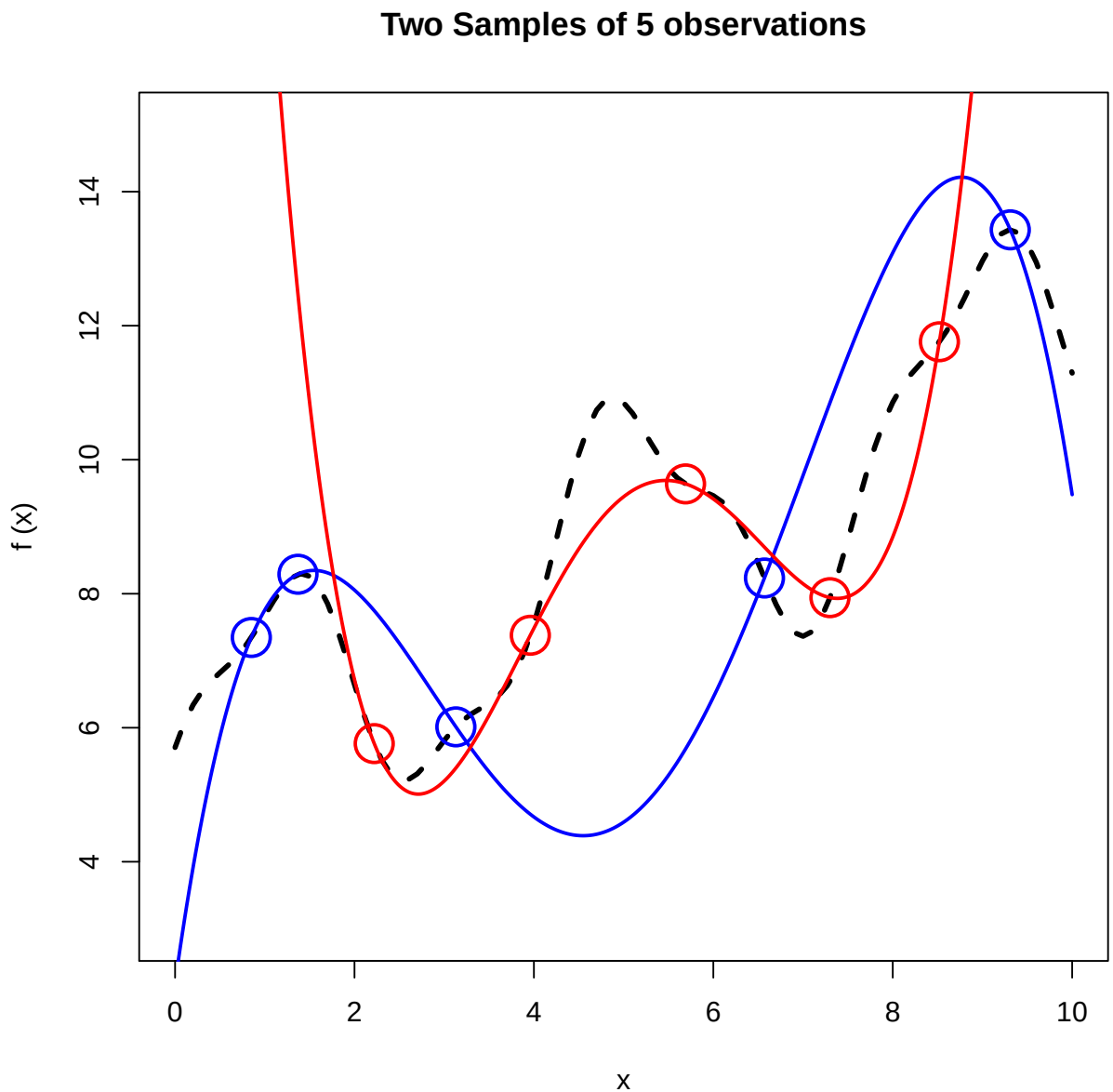
- When number of potential features is much larger than the number of cases ($p >> n$), one can always fit a predictor to have 100% accuracy on the data set used to build it.
- If applied naively, more complex modeling methods are more prone to over-fitting.
- Estimating accuracy by "plugging in" data used to build a predictor results in highly biased estimates of performance (re-substitution estimate).
- Internal and external validation of predictors are essential.
- *Ref:* Simon *et al.*, *JNCI*, 2003; Radmacher *et al.*, *J. Comp. Bio.*, 2002.


True Function
f (x)
x
4
6
8
10
12
14
0
2
4
6
8
10

Over-Fit



Over-Fit



Over-Fit

- Models in high dimension are usually complex (not necessarily for the individual gene, but as a whole the model has a large space to live in).
- Sample sizes are always too small for precise estimation of the true model.
- Look for simpler models that provide reasonable approximations.

In almost every experiment, we are interested in the performance of the predictor on future samples (Generalization Error) and not the performance of the predictor on the current data (Resubstitution Error).

The difference between the generalization error and the resubstitution error is one measure of the over-fit.

Unless you have strong evidence you are not over-fitting, don't report the resubstitution error and assume it is an estimate of the generalization error.

Internal Validation

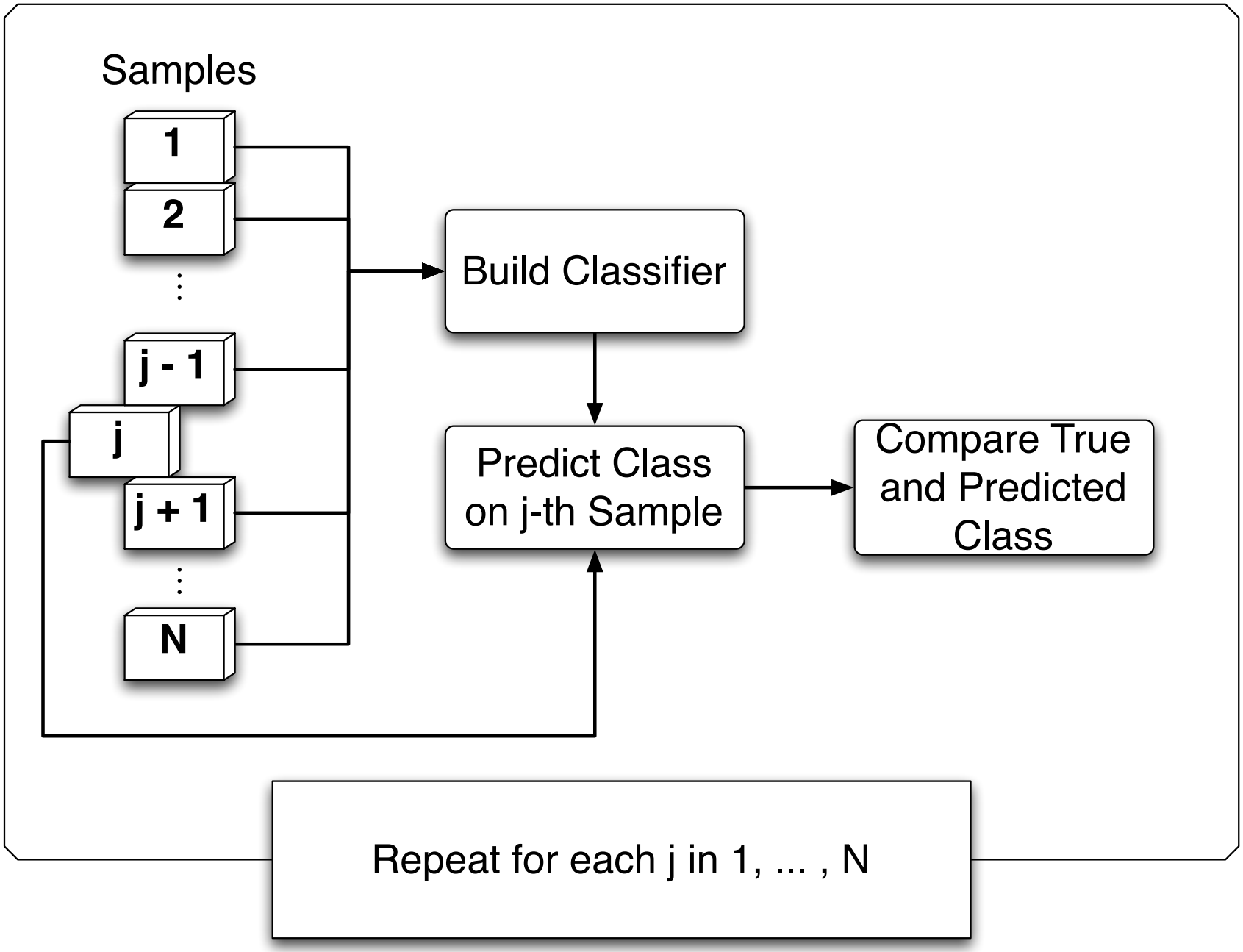
Within-sample Validation:

- Cross-validation (many flavors: leave-one-out, split-sample, k-fold, etc.)
- Bootstrap and other resampling methods
- See Molinaro *et al.* (*Bioinformatics*, 2005)

External Validation

Independent-sample validation

LEAVE-ONE-OUT CV (LOOCV)



Limitations of within-sample validation:

- Frequently performed incorrectly:
  - Improper CV (e.g. not including feature selection)
  - Special statistical inference procedures required (Lusa *et al.*, *Statistics in Medicine*, 2007; Jiang *et al.*, *Stat. Appl. Gen. and Mol. Bio.*, 2008).
- Large variance in estimated accuracy and effect sizes,
- Doesn't protect against biases due to selective inclusion/exclusion of samples.
- Built-in biases possible (e.g. lab batch, specimen handling).

Generation of Gene Expression Profiles

- 100 specimens (P_i, $i = 1, \ldots, 100$)
- Log-ratio measurements on 6000 genes
- $P_i \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}_{6000})$
- 1000 simulation repetitions
- Can we distinguish between the first 50 speciments (class 1) and the last 50 (class 2)? The class distinction is artificial here since all 100 were generated from the same distribution.

Prediction Method

Linear Discriminant Analysis (LDA) prediction using significant DE genes ($\alpha = 0.001$).

Resubstitution Method

1. Build LDA from all data.
2. For $i = 1, \ldots, 100$, apply LDA to sample i.
3. Compare predicted class to actual class.

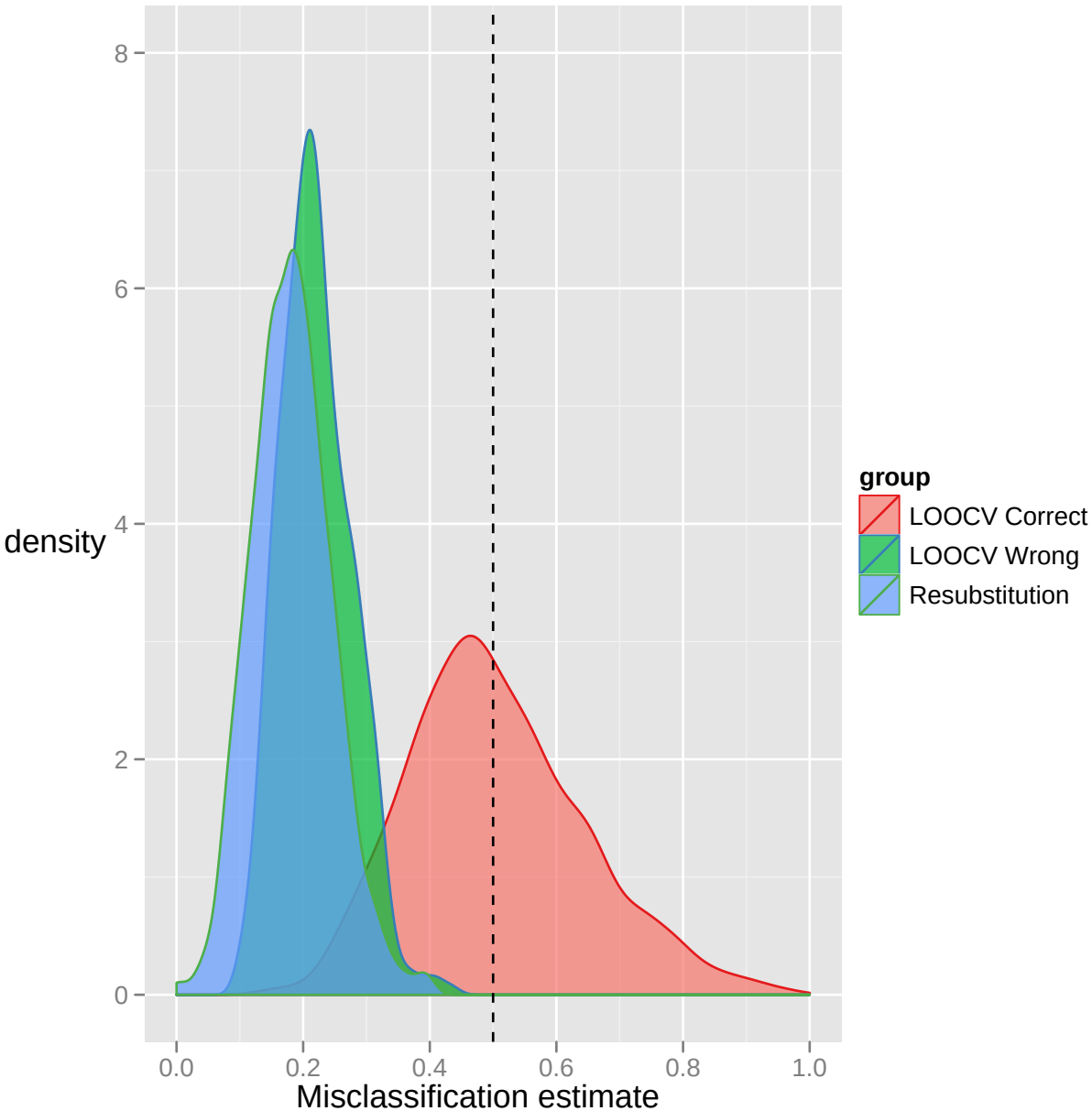
LOOCV Without Gene Selection

1. Select DE genes for LDA using all 100 samples.
2. For $i = 1, \ldots, 100$:
   1. Leave out sample i.
   2. Build $LDA(i)$ on other 99 samples.
   3. Apply $LDA(i)$ to sample i.
3. Compare predicted class to actual class

LOOCV with Gene Selection (Correct)

1. For $i = 1, \ldots, 100$:
    1. Leave out sample i.
    2. Select DE genes based on other 99 samples.
    3. Build $LDA(i)$ on other 99 samples.
    4. Apply $LDA(i)$ to sample i.
2. Compare predicted class to actual class

Prediction Simulation



Breast Cancer Example

Gene-Expression Profiles in Hereditary Breast Cancer (Hedenfalk *et al.*, *NEJM*, 2001).

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

Research questions

Can we distinguich *BRCA1+* from *BRCA1-* cancers and *BRCA2+* from *BRCA2-* cancers — based solely on their gene expression profiles?

Classification with Compound covariate predictor:

Class	# genes †	# misclass (m) ‡	proportion §
BRCA1$+/-$	9	1	0.004
BRCA2$+/-$	11	4	0.043

† $\alpha = 0.0001$ on the full data set

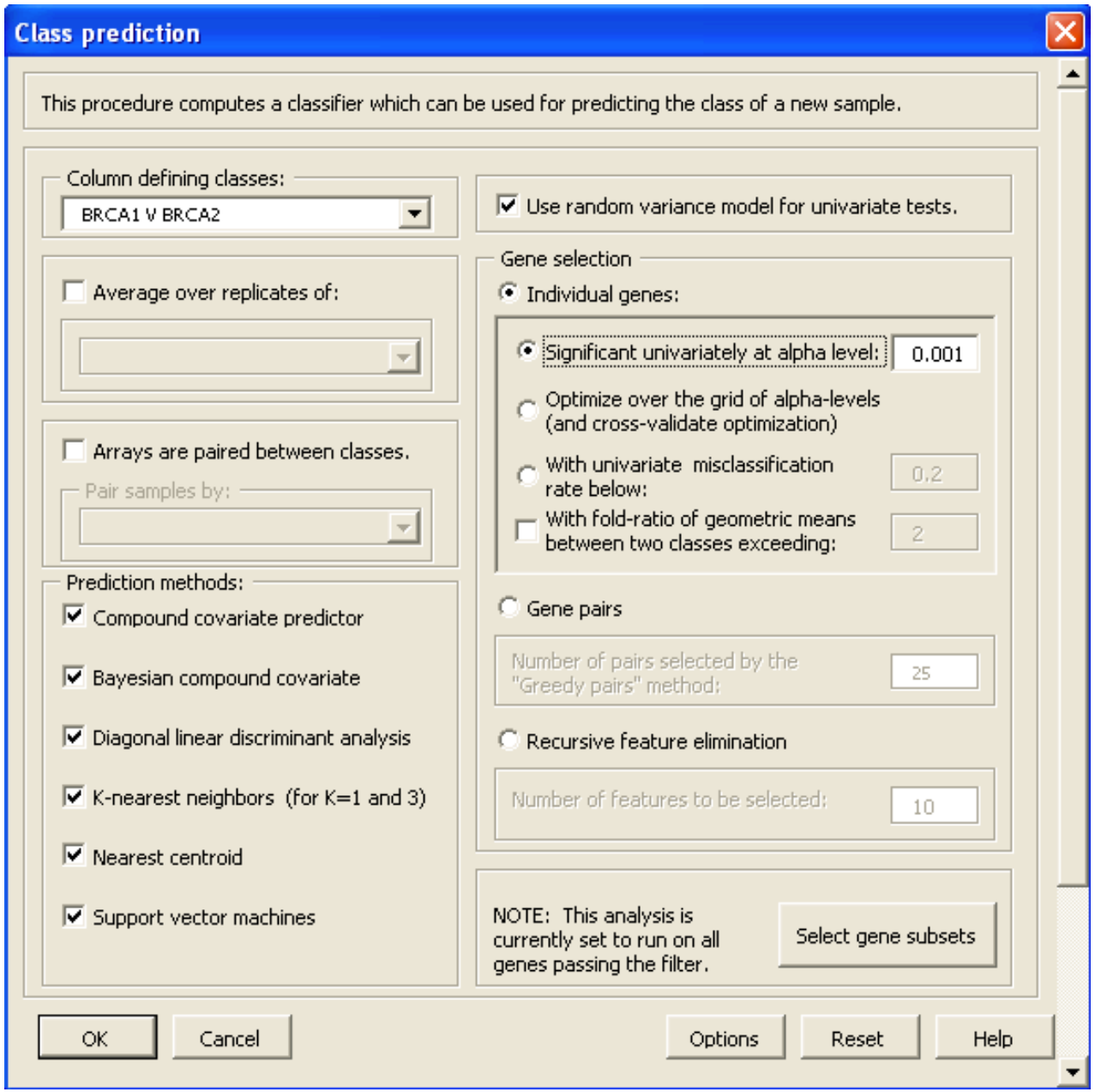
‡ Using LOOCV

§ Proportion of permutations with m or fewer misclassifications

CLASS PREDICTION IN BRB-ARRAYTOOLS

- Variety of prediction methods available
- 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.

CLASS PREDICTION IN BRB-ARRAYTOOLS



CLASS PREDICTION IN BRB-ARRAYTOOLS

Additional prediction plug-ins:

- Adaboost: Freund and Schapire, *In Proceedings of the Thirteenth Internal Conference on Machine Learning*, 1996.
- Prediction Analysis of Microarrays (PAM): Tibshirani *et al.*, *PNAS*, 2002.
- Random Forests: Breiman, *Machine Learning*, 2001.
- Top-scoring pairs: Geman *et al.*, *SAGMB*, 2004.

Design Considerations

- Sample selection, including reference sample
- Sources of variability/levels of replication
- Sample size planning
- Controls

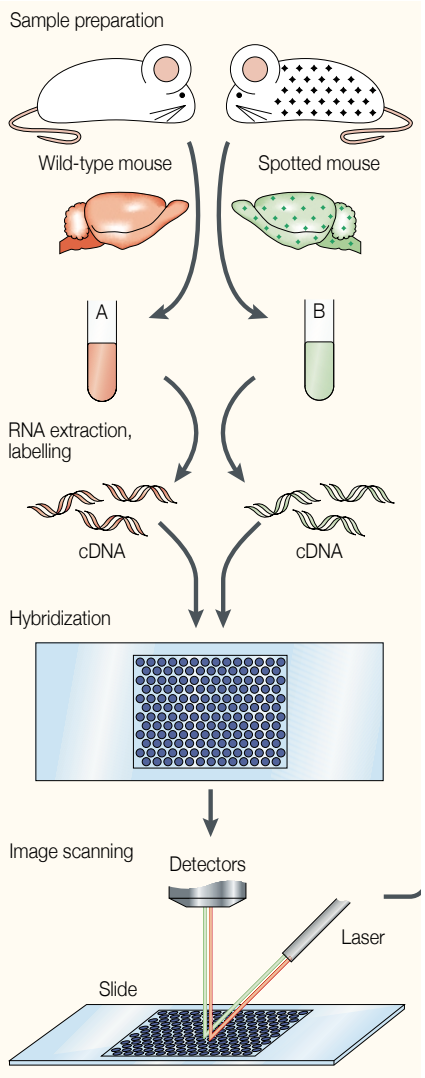
Experimental Samples

- A random sample from a population under investigation?
- Broad versus narrow inclusion criteria?

Reference Samples (cDNA)

- 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



- Biological heterogeneity in population
- Specimen Collection/handling effects
- Biological heterogeneity in specimen
- RNA extraction
- RNA amplification
- Fluor labeling
- Hybridization
- Scanning

Geschwind, *Nature Reviews Neuroscience*, 2001

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:

- 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.

Sample Size Planning

For 2-group comparisons with cDNA arrays using 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.

Approximate total sample size required to compare two equal sized, independent groups:

$$n = \frac{4\sigma^2 \left(Z_{\alpha/2} + Z_{\beta}\right)}{\delta^2}$$

Where:

$$\begin{aligned} \delta &= \text{mean diff. between classes (log scale)} \\ \sigma &= \text{standard deviation (log scale)} \\ Z_{\alpha/2}, Z_{\beta} &= \text{standard normal percentiles} \end{aligned}$$

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

How to choose α and β

$K =$ # of genes on array, $M =$ # of genes DE at $\theta = 2^{\delta}$

Expected number of false positives:

$$\text{EFP} \leq (K - M) \times \alpha$$

Expected number of false negatives:

$$\text{EFN}_{\theta} = M \times \beta$$

Popular choices for α and β:

$$\alpha = 0.001 \qquad \beta = 0.05 \text{ or } 0.10$$

$1 - \beta$ = Power

Effect of α and β on FDR

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

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

where π is the proportion of DE genes

π	α	$1-\beta$	FDR
0.005	0.01	0.95	68%
0.005	0.01	0.80	71%
0.005	0.001	0.95	17%
0.005	0.001	0.80	20%
0.05	0.001	0.95	2%

Choosing σ and δ

Value of δ 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 (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: $\delta = \log_2(3) = 1.59$

Sample Size Example

$K =$ 10,000 genes on array

$M =$ 100 genes DE

$\alpha = 0.001\,(Z_{\alpha/2} = 3.291)$

$\beta = 0.05\,(Z_{\beta} = 1.645)$

$\sigma = 0.75$

$\delta = 1$ (2-fold)

Need n = 55 (~ 28 per group).

Expect ≤ 10 false positives and miss $\approx 5/100$ 2-fold genes.

Sample Size Examples ($\alpha = 0.001$)

σ	δ	2^{δ}	n	Power(%)
0.25	1.00	2.00	6	95
0.50	1.00	2.00	14	95
0.25	1.00	2.00	5	82
0.50	1.00	2.00	5	14
0.25	1.20	2.29	5	95
0.50	2.39	5.24	5	95

n is per group

SAMPLE SIZE FOR CLASS PREDICTION

Raises unique issues:

- The classes may mostly overlap, even in the high dimensional space.
- There may be no good classifier.
- There will be an upper limit optimal performance that no classifier can exceed.

Solution: Determine sample size big enough to get "close to optimal" performance:

- Dobbin and Simon, *Biostatistics*, 2007; Dobbin, Zhao and Simon, *Clin Cancer Res*, 2008.
- `http://brb.nci.nih.gov`

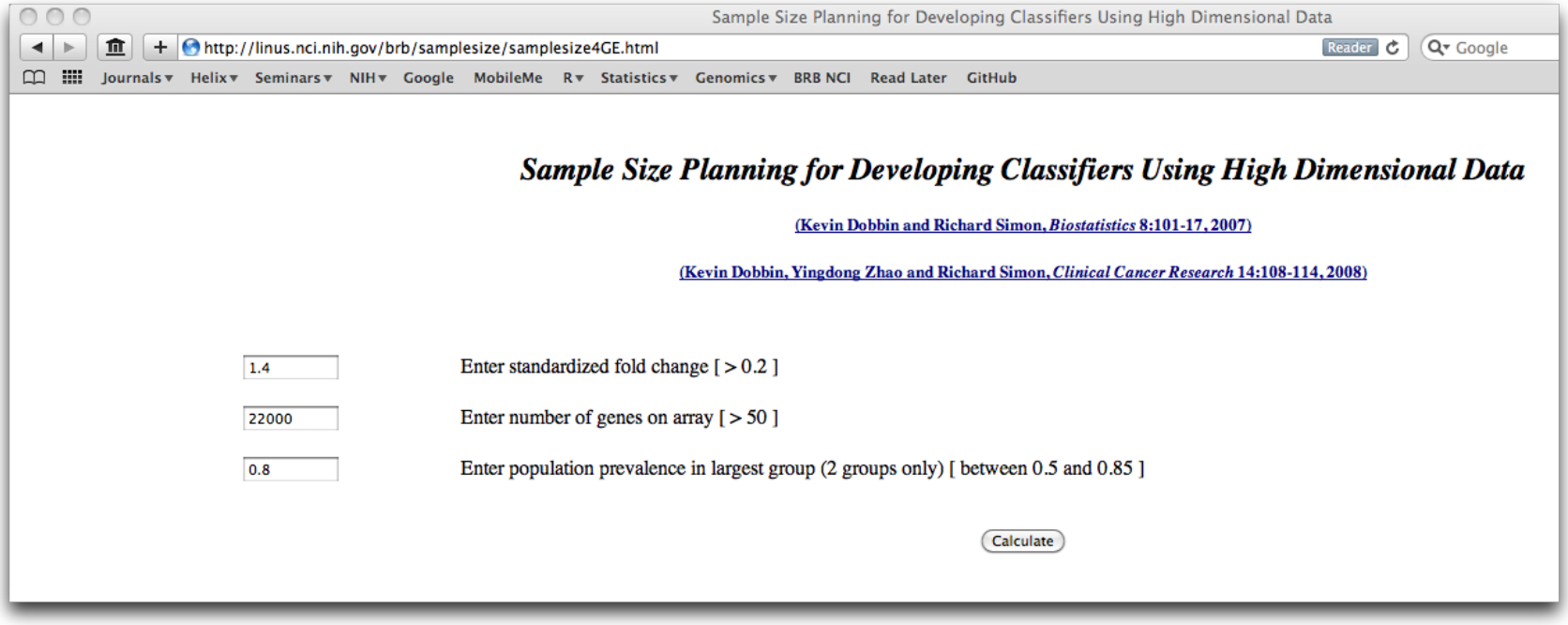
Sample Size for Class Prediction

3 essential inputs for sample size calculation with two classes:

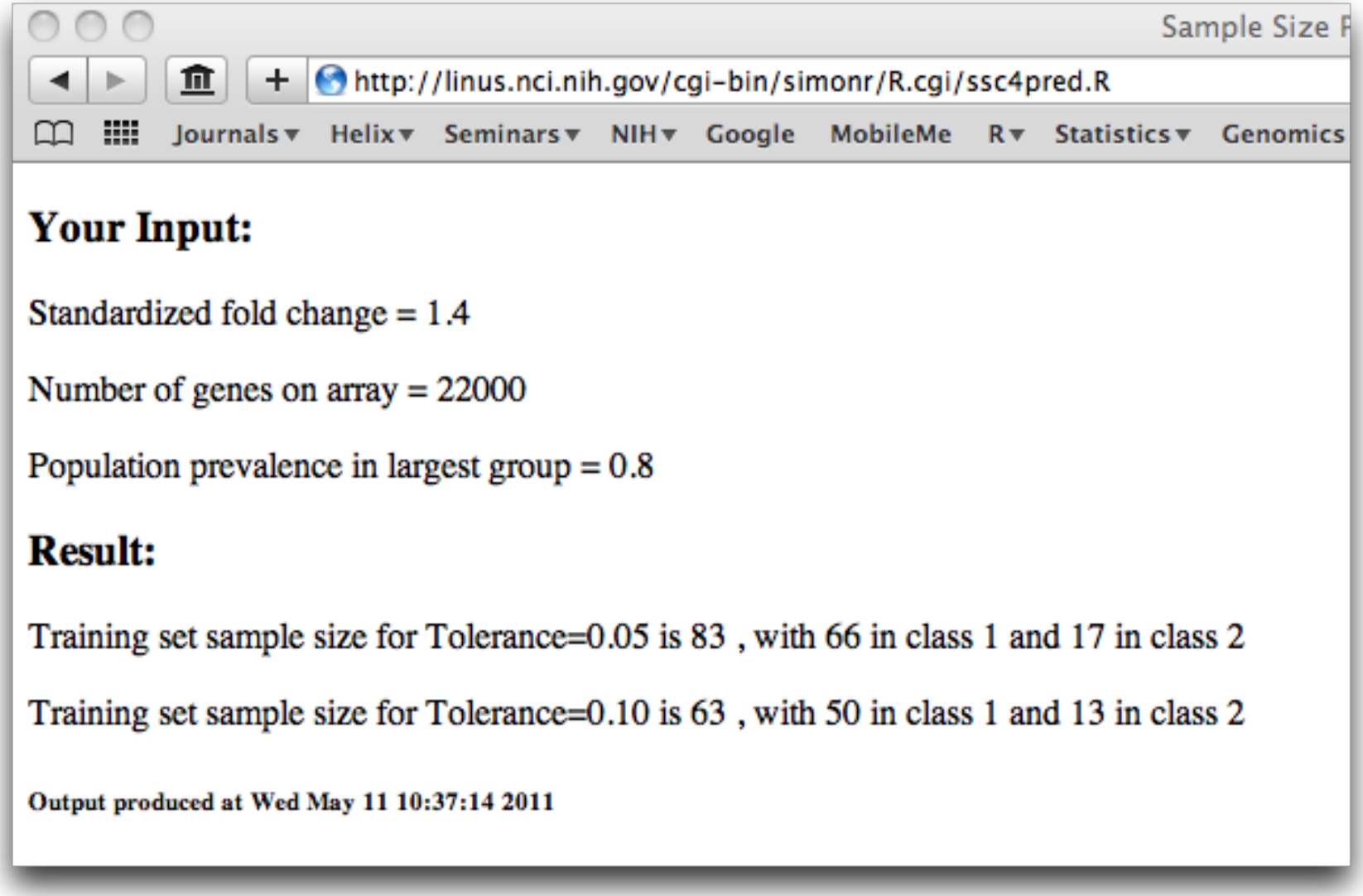
1. Number of genes on the array
2. The prevalence in each class
3. The fold-change for informative genes (difference in class means divided by within class SD, on the $\log_2$ scale)

For example, $\sim 22,000$ features on the Affymetrix U113A array, 20% respond to drug, so prevalence is 20% vs. 80%, and the fold change of 1.4.

SAMPLE SIZE FOR CLASS PREDICTION



SAMPLE SIZE FOR CLASS PREDICTION



Sample Size References

Technical replicates for 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.

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*)

Summary I

- 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 methods may be required.
- Chances for spurious findings are enormous, and validation of any findings on larger independent collections of specimens will be essential.

Summary II

- 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`
- BRB-ArrayTools: `http://linus.nci.nih.gov/BRB-ArrayTools.html`
- BRB textbook: `http://linus.nci.nih.gov/~brb/book.html`
- PDF of this talk: `http://linus.nci.nih.gov/~brb/presentations.htm`
- Bioconductor: `http://www.bioconductor.org/`

Acknowledgements

- Richard Simon
- Lisa McShane
- Joanna Shih
- Kevin Dobbin
- Michael Radmacher
- Members of the NCI Biometric Research Branch
- Our NCI collaborators and students in the NIH microarray classes
- BRB-ArrayTools development team