Design &Analysis of DNA Microarray Studies With BRB-ArrayTools

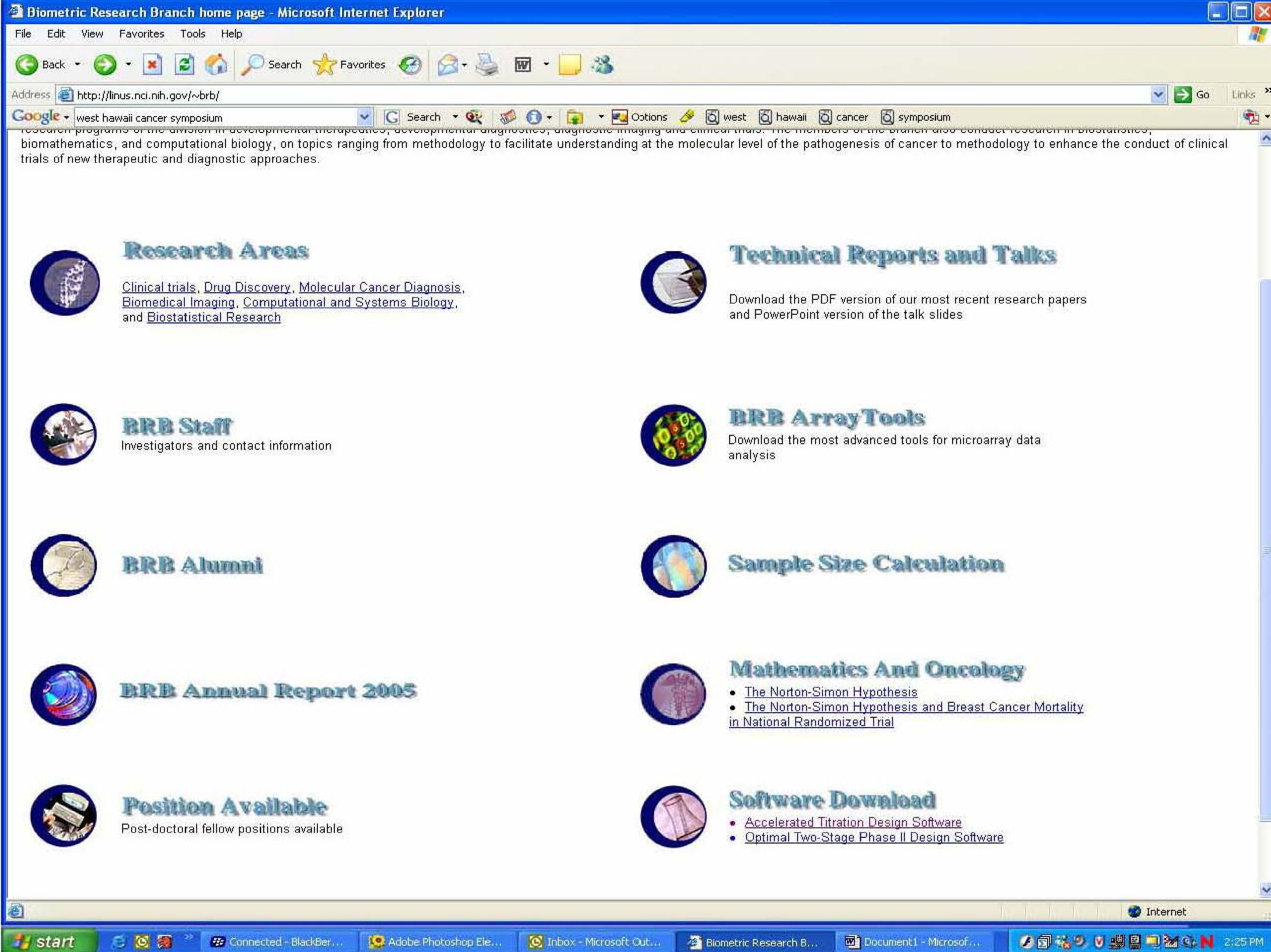
Dr. Richard Simon

rsimon@nih.gov

http://linus.nci.nih.gov/brb

http://linus.nci.nih.gov/brb

- http://linus.nci.nih.gov/brb
  - Powerpoint presentations
  - Reprints & Technical Reports
  - BRB-ArrayTools software
  - BRB-ArrayTools Data Archive
  - Sample Size Planning for Targeted Clinical Trials



research programs of the division in developmental therapeutics, developmental diagnostics, diagnostic imaging and clinical trials. The members of the branch also conduct research in biostatistics, biomathematics, and computational biology, on topics ranging from methodology to facilitate understanding at the molecular level of the pathogenesis of cancer to methodology to enhance the conduct of clinical trials of new therapeutic and diagnostic approaches.

Research Areas

Clinical trials, Drug Discovery, Molecular Cancer Diagnosis, Biomedical Imaging, Computational and Systems Biology, and Biostatistical Research

Technical Reports and Talks

Download the PDF version of our most recent research papers and PowerPoint version of the talk slides

BRB Staff

Investigators and contact information

BRB ArrayTools

Download the most advanced tools for microarray data analysis

BRB Alumni

Sample Size Calculation

BRB Annual Report 2005

Mathematics And Oncology

- The Norton-Simon Hypothesis
- The Norton-Simon Hypothesis and Breast Cancer Mortality in National Randomized Trial

Position Available

Post-doctoral fellow positions available

Software Download

- Accelerated Titration Design Software
- Optimal Two-Stage Phase II Design Software

Challenges in Effective Use of DNA Microarray Technology

- Design & Analysis are bigger challenges than data management.
  - Much greater opportunity for misleading yourselves and others than traditional single gene/protein studies
- Limited availability of experienced statistical collaborators
- Predominance of hype, mis-information, and dangerous methods promulgated by biomedical scientists as well as professional statistical/computational scientists
- Predominance of flashy software that encourages misleading analyses

Objectives of BRB-ArrayTools

- Provide biomedical scientists access to statistical expertise for the analysis of expression data
  - training in analysis of high dimensional data
  - access to critical assessment of methods published in a rapidly expanding literature

BRB-ArrayTools

- Integrated package
- Excel-based user interface
  - Doesn’t use Excel analyses
  - state-of-the art analysis methods programmed in R, Java & Fortran
  - Data not stored as worksheets
    - >1000 arrays and 65000 genes per project
- Based on continuing evaluation of validity and usefulness of published methods
  - Methods carefully selected by R Simon
  - Not a repository like Bioconductor
- Publicly available for non-commercial uses from BRB website:

BRB-ArrayTools

- Not tied to any database
  - Importer for common databases and platforms
    - MadB, GenePix, Agilent, MAS5/GCOS
    - Imports .cel files
    - Import wizzard for any files output by image analysis program
  - Import (collate)
    - Expression data (eg separate file for each array)
    - Spot (probeset) identifiers
    - Experiment descriptor worksheet
      - Rows correspond to arrays
      - Columns are user defined phenotypes to drive the analyses
        - Can be updated during analysis
  - Imported data saved as project folder containing project workbook and binary files
    - Project workbook can be re-opened in Excel at any time
    - Output saved in html files in output folder

BRB-ArrayTools

- Highly computationally efficient
  - Non-intensive analyses in R
  - Intensive analyses in FORTRAN
    - eg BRB-AT version of SAM is 9x + more efficient than Bioconductor or web based versions
      - And more accurate
- Extensive gene and pathway annotation features

BRB-ArrayTools

- Plug-in facility for user written R functions
- Message board and listserve
- Extensive built-in help facilities, tutorials, datasets, usersguide, data import and analysis wizzards, sample statistical analysis sections, …

BRB-ArrayTools Archive of Human Tumor Expression Data

- http://linus.nci.nih.gov/brb/DataArchive.html
- Archive of BRB-ArrayTools zipped project folders of expression profiles of human tumors and associated clinical/pathological descriptors
  - Published data
- Easy way to archive your data and to analyze someone else’s data
  - Download, unzip, open in Excel

- *Design and Analysis of DNA Microarray Investigations*
  - R Simon, EL Korn, MD Radmacher, L McShane, G Wright, Y Zhao. Springer (2003)

Brief Review of Microarray Technology

Microarray Expression Profiling

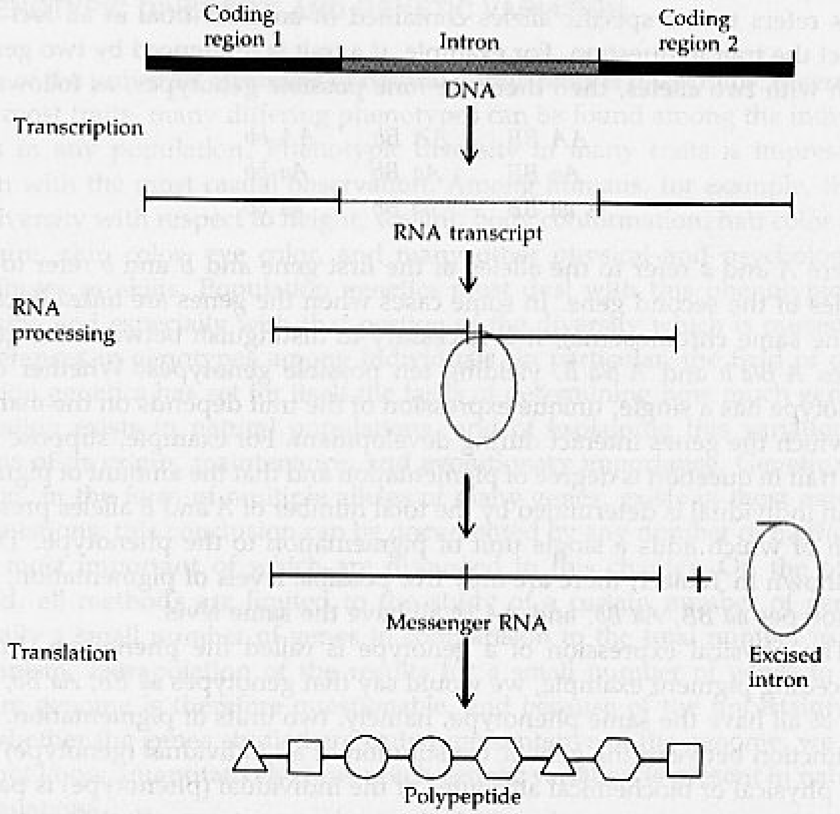
- Would like to know the concentration of each protein in a cell
  - Proteins do the work of cells
  - Proteins have many shapes and parallel assays for all proteins have not been developed

Microarray Expression Profiling

- One gene *transcription* produces one mRNA molecule produces one protein molecule
- # genes ≅ # mRNA types
- mRNA molecule can be reverse transcribed into DNA and will bind only to the gene from which it was originally transcribed (to which it is *homologous*)

Microarray Expression Profiling

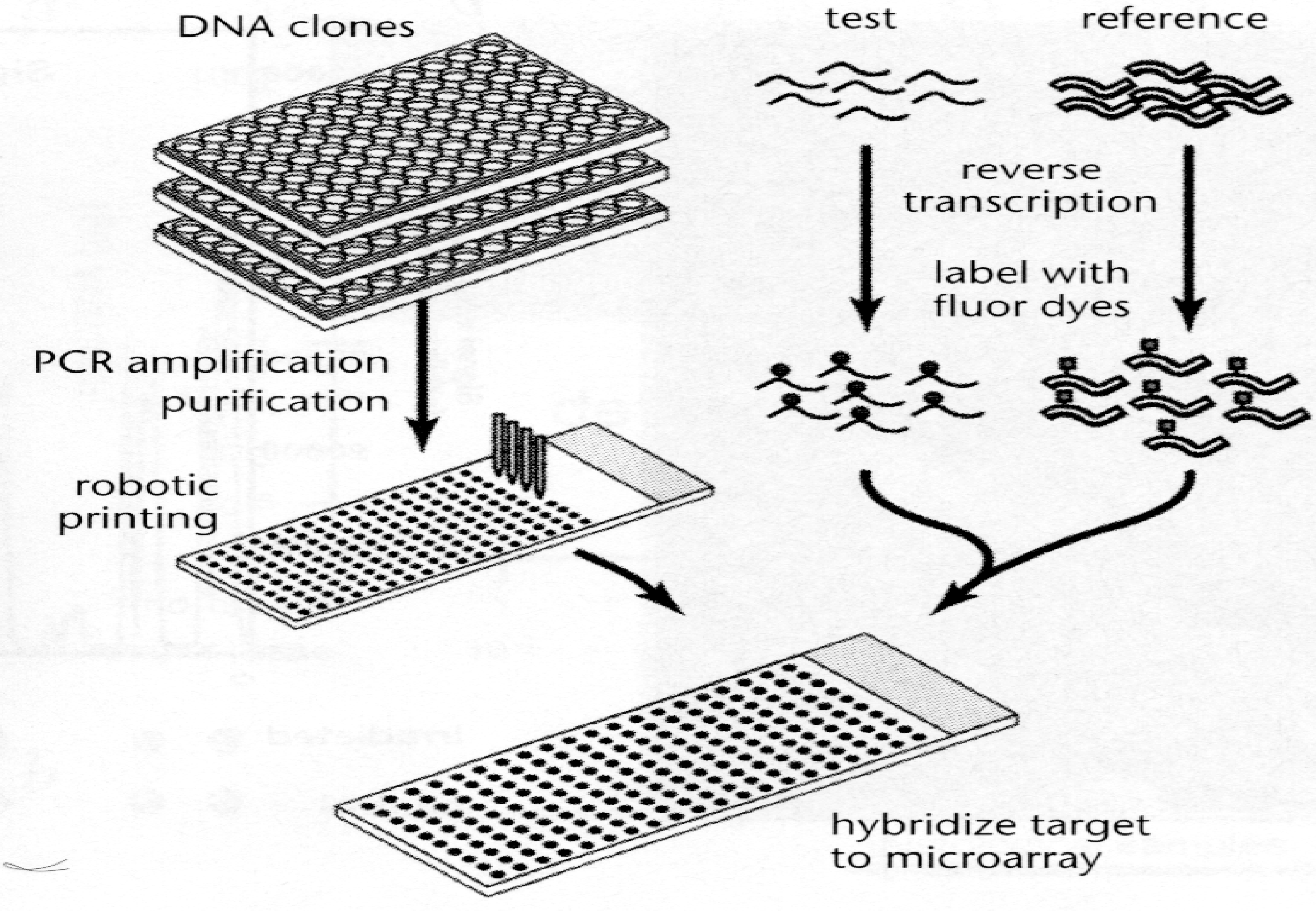
- Estimates abundance of mRNA molecules of each type present in cells
  - Assay not sensitive enough to analyze single cells so estimate is for average of sample of cells
- Microarray contains a spot of DNA corresponding to each gene
  - Spots are in known fixed positions
  - Spots contain fewer nucleotides that the full gene

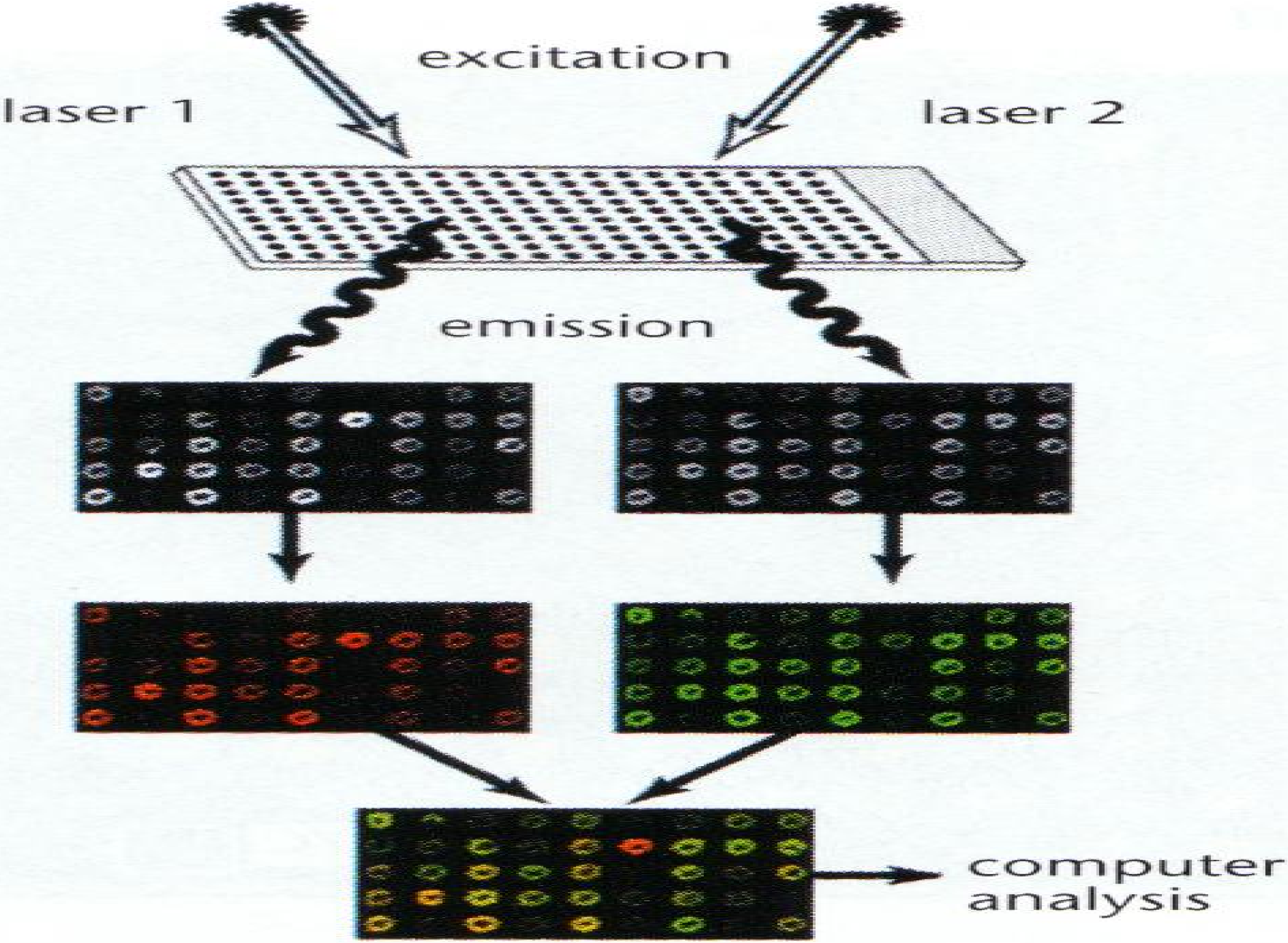

Coding region 1
Intron
Coding region 2
DNA
Transcription
RNA transcript
RNA processing
Messenger RNA
+
Excised intron
Translation
Polypeptide

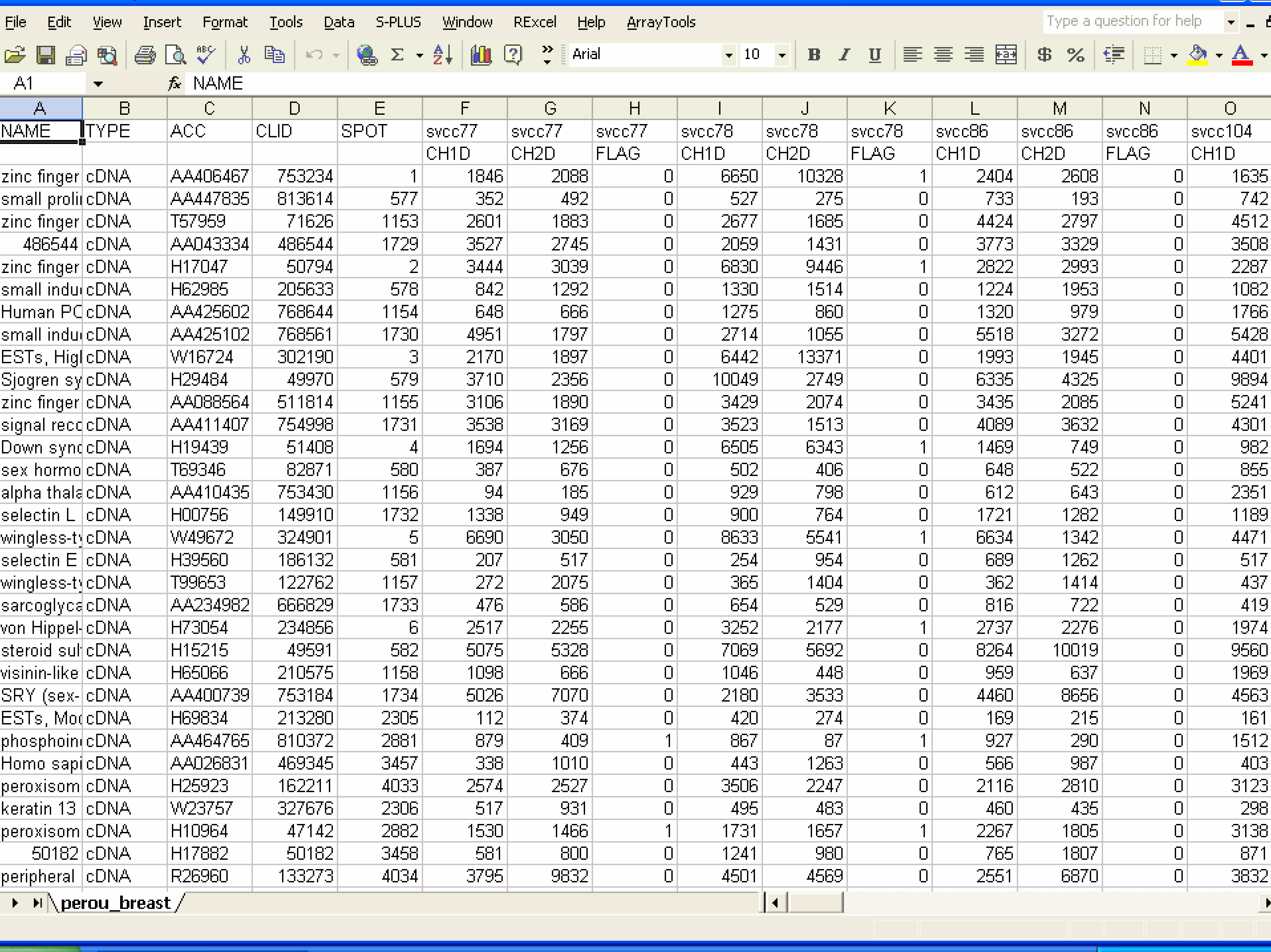
Gene Expression Microarrays

- Permit simultaneous evaluation of expression levels of thousands of genes
- Main platforms
  - cDNA printed on glass slides
  - Externally synthesized oligos printed on glass slides
  - Affymetrix GeneChips
  - Oligos in-situ synthesized on glass slides
  - cDNA printed on nylon filters

cDNA Array


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

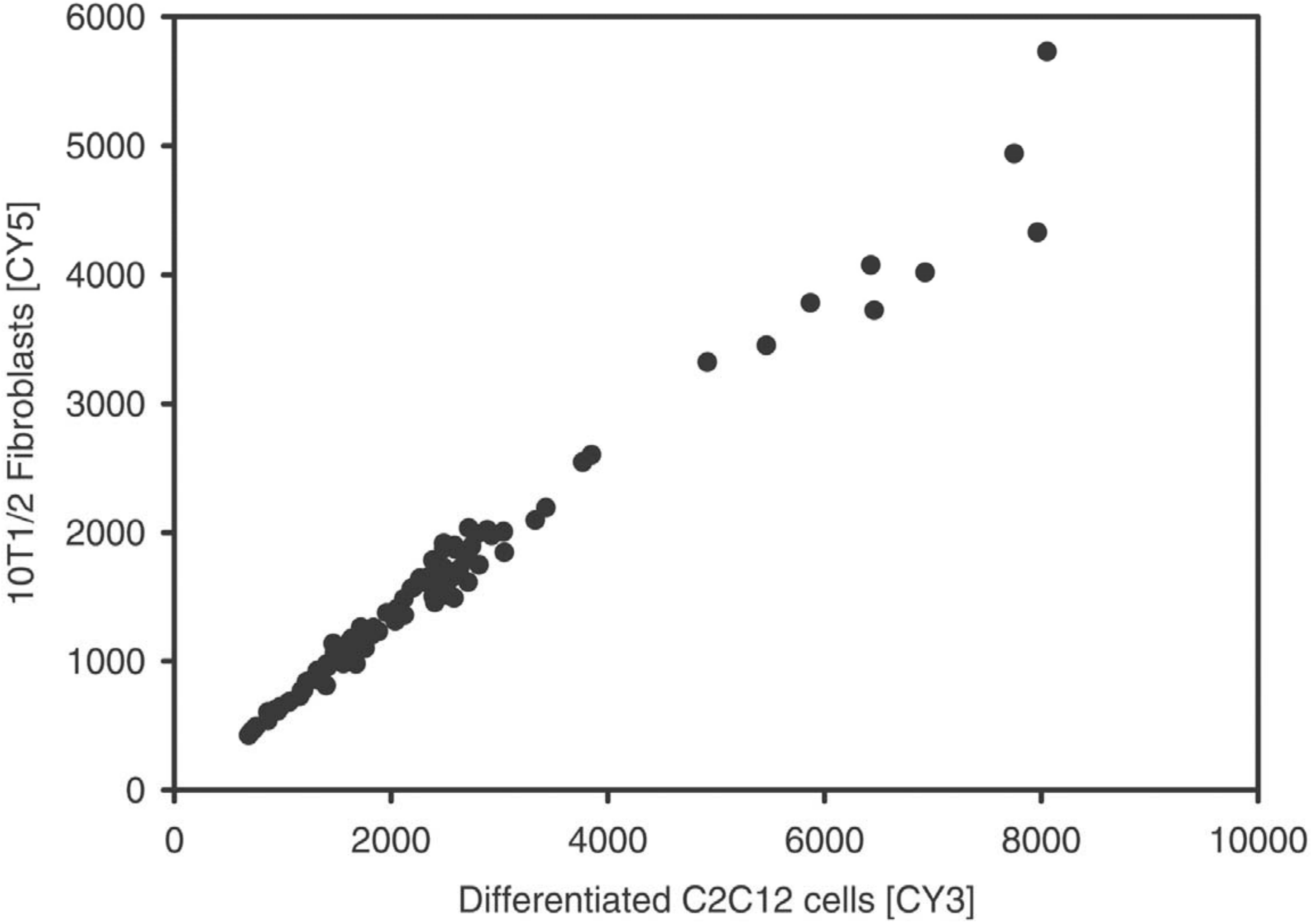

excitation
laser 1
laser 2
emission
computer analysis



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
NAME	TYPE	ACC	CLID	SPOT	svcc77	svcc77	svcc77	svcc78	svcc78	svcc78	svcc86	svcc86	svcc86	svcc104
					CH1D	CH2D	FLAG	CH1D	CH2D	FLAG	CH1D	CH2D	FLAG	CH1D
zinc finger	cDNA	AA406467	753234	1	1846	2088	0	6650	10328	1	2404	2608	0	1635
small proli	cDNA	AA447835	813614	577	352	492	0	527	275	0	733	193	0	742
zinc finger	cDNA	T57959	71626	1153	2601	1883	0	2677	1685	0	4424	2797	0	4512
486544	cDNA	AA043334	486544	1729	3527	2745	0	2059	1431	0	3773	3329	0	3508
zinc finger	cDNA	H17047	50794	2	3444	3039	0	6830	9446	1	2822	2993	0	2287
small indu	cDNA	H62985	205633	578	842	1292	0	1330	1514	0	1224	1953	0	1082
Human PC	cDNA	AA425602	768644	1154	648	666	0	1275	860	0	1320	979	0	1766
small indu	cDNA	AA425102	768561	1730	4951	1797	0	2714	1055	0	5518	3272	0	5428
ESTs, Higl	cDNA	W16724	302190	3	2170	1897	0	6442	13371	0	1993	1945	0	4401
Sjogren sy	cDNA	H29484	49970	579	3710	2356	0	10049	2749	0	6335	4325	0	9894
zinc finger	cDNA	AA088564	511814	1155	3106	1890	0	3429	2074	0	3435	2085	0	5241
signal recc	cDNA	AA411407	754998	1731	3538	3169	0	3523	1513	0	4089	3632	0	4301
Down sync	cDNA	H19439	51408	4	1694	1256	0	6505	6343	1	1469	749	0	982
sex hormo	cDNA	T69346	82871	580	387	676	0	502	406	0	648	522	0	855
alpha thala	cDNA	AA410435	753430	1156	94	185	0	929	798	0	612	643	0	2351
selectin L	cDNA	H00756	149910	1732	1338	949	0	900	764	0	1721	1282	0	1189
wingless-ty	cDNA	W49672	324901	5	6690	3050	0	8633	5541	1	6634	1342	0	4471
selectin E	cDNA	H39560	186132	581	207	517	0	254	954	0	689	1262	0	517
wingless-ty	cDNA	T99653	122762	1157	272	2075	0	365	1404	0	362	1414	0	437
sarcoglyca	cDNA	AA234982	666829	1733	476	586	0	654	529	0	816	722	0	419
von Hippel-	cDNA	H73054	234856	6	2517	2255	0	3252	2177	1	2737	2276	0	1974
steroid sul	cDNA	H15215	49591	582	5075	5328	0	7069	5692	0	8264	10019	0	9560
visinin-like	cDNA	H65066	210575	1158	1098	666	0	1046	448	0	959	637	0	1969
SRY (sex-	cDNA	AA400739	753184	1734	5026	7070	0	2180	3533	0	4460	8656	0	4563
ESTs, Mod	cDNA	H69834	213280	2305	112	374	0	420	274	0	169	215	0	161
phosphoin	cDNA	AA464765	810372	2881	879	409	1	867	87	1	927	290	0	1512
Homo sapi	cDNA	AA026831	469345	3457	338	1010	0	443	1263	0	566	987	0	403
peroxisom	cDNA	H25923	162211	4033	2574	2527	0	3506	2247	0	2116	2810	0	3123
keratin 13	cDNA	W23757	327676	2306	517	931	0	495	483	0	460	435	0	298
peroxisom	cDNA	H10964	47142	2882	1530	1466	1	1731	1657	1	2267	1805	0	3138
50182	cDNA	H17882	50182	3458	581	800	0	1241	980	0	765	1807	0	871
peripheral	cDNA	R26960	133273	4034	3795	9832	0	4501	4569	0	2551	6870	0	3832

cDNA & Printed Oligo Arrays

- Each gene represented 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


6000
5000
4000
3000
2000
1000
0
10T1/2 Fibroblasts [CY5]
0
2000
4000
6000
8000
10000
Differentiated C2C12 cells [CY3]

[Affymetrix] Hybridization Oligo Array

Affymetrix GeneChips

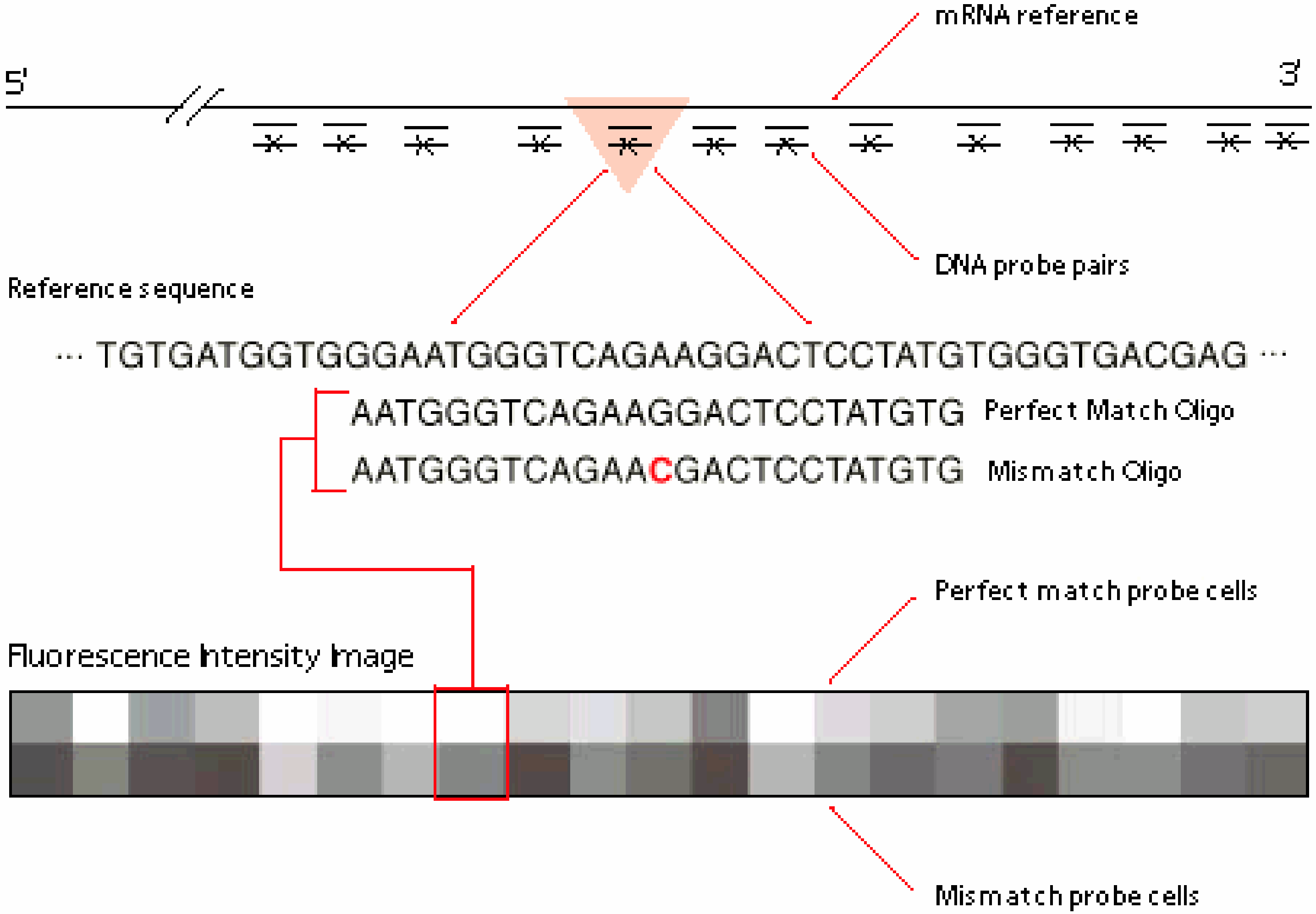
- Contain multiple probes (spots) per gene
- Probes corresponding to the same gene must be processed to give a probe-set summary intensity for each gene
- Single label system
  - Higher reproducibility makes use of dual-labels unnecessary

Affymetrix Arrays

- Single sample hybridized to each array
- Each gene represented by a “probe set”
  - One probe type per array “cell”
  - Typical probe is a 25-mer oligo
  - 11-16 PM:MM pairs per probe set

    (PM = perfect match, MM = mismatch)

GeneChip Expression Array Design



mRNA reference

5'

3'

DNA probe pairs

Reference sequence

··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAG ···

AATGGGTCAGAAGGACTCCTATGTG Perfect Match Oligo

AATGGGTCAGAACGACTCCTATGTG Mismatch Oligo

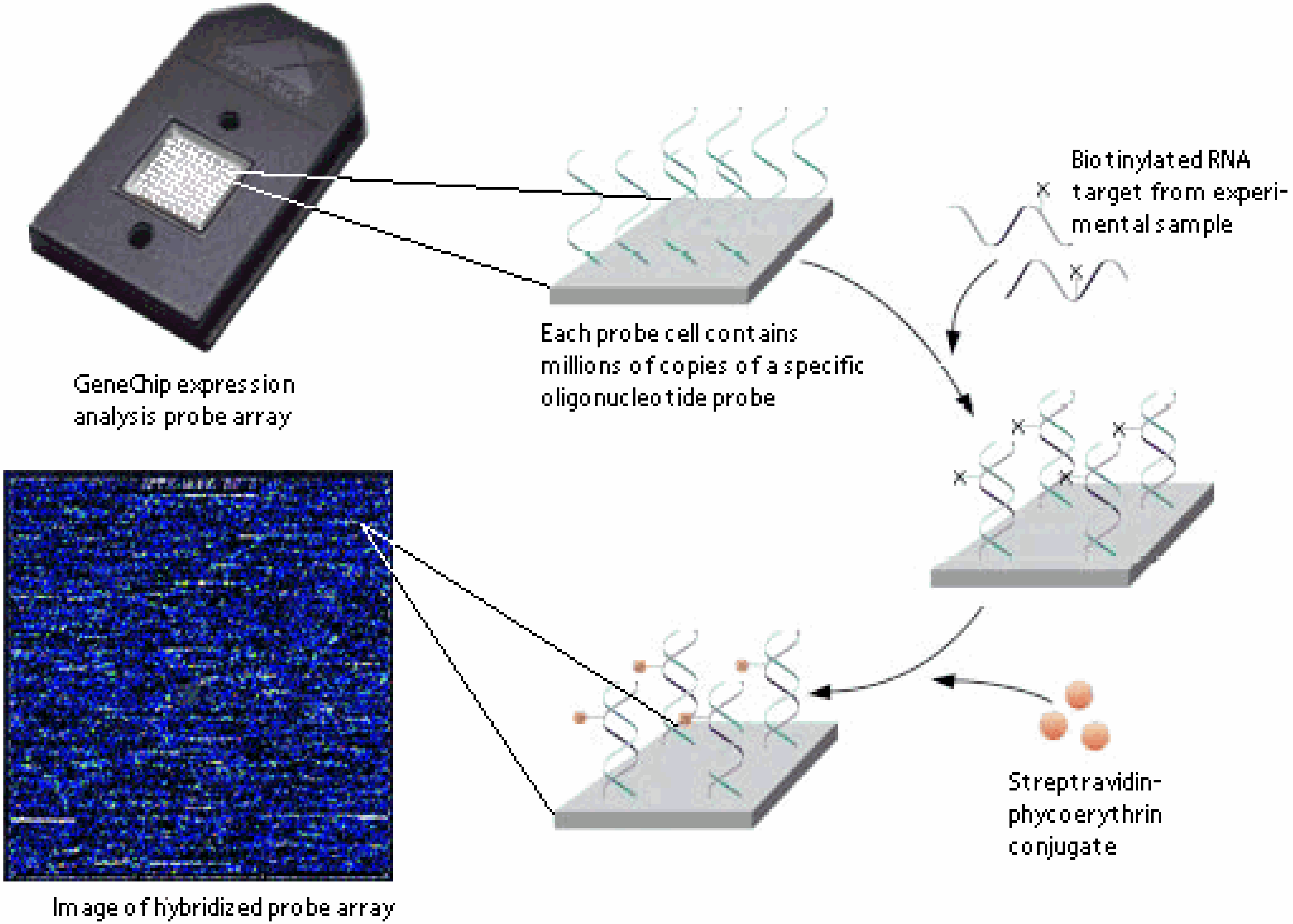
Perfect match probe cells

Fluorescence Intensity Image

Mismatch probe cells

Source: Affymetrix website

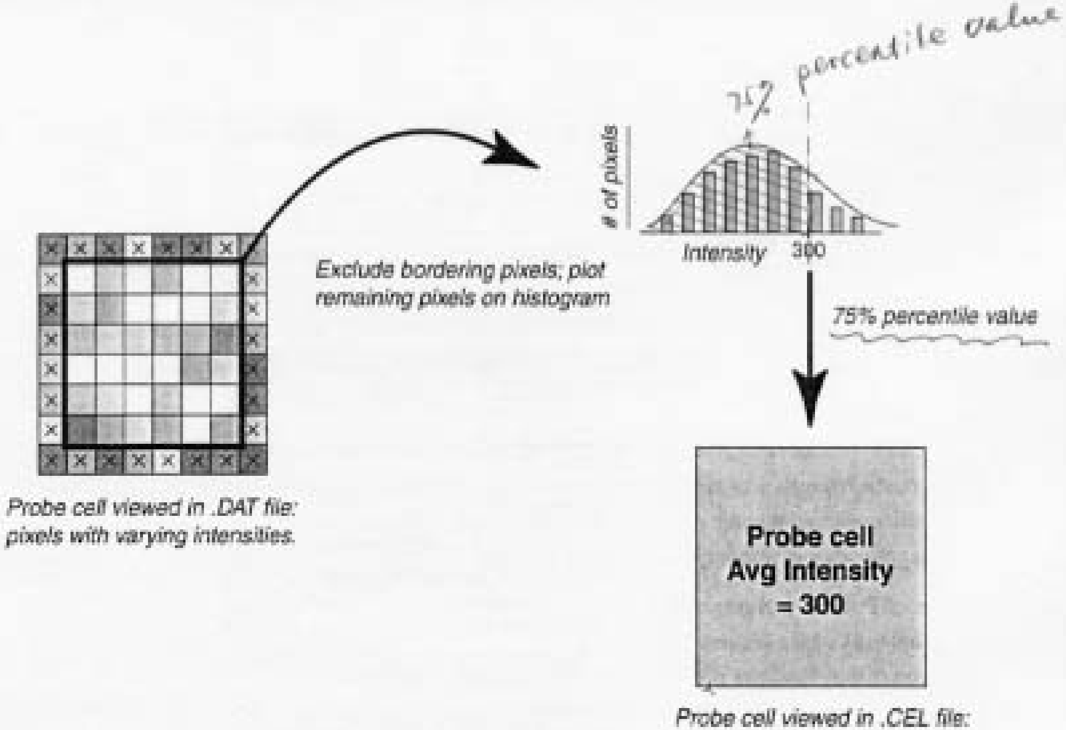
GeneChip® Expression Analysis Process


Biotinylated RNA target from experimental sample
GeneChip expression analysis probe array
Each probe cell contains millions of copies of a specific oligonucleotide probe
Streptravidin-phycoerythrin conjugate
Image of hybridized probe array

Source: Affymetrix website

Affymetrix GeneChips

.dat file: a huge image file
.cel file: cell intensity file

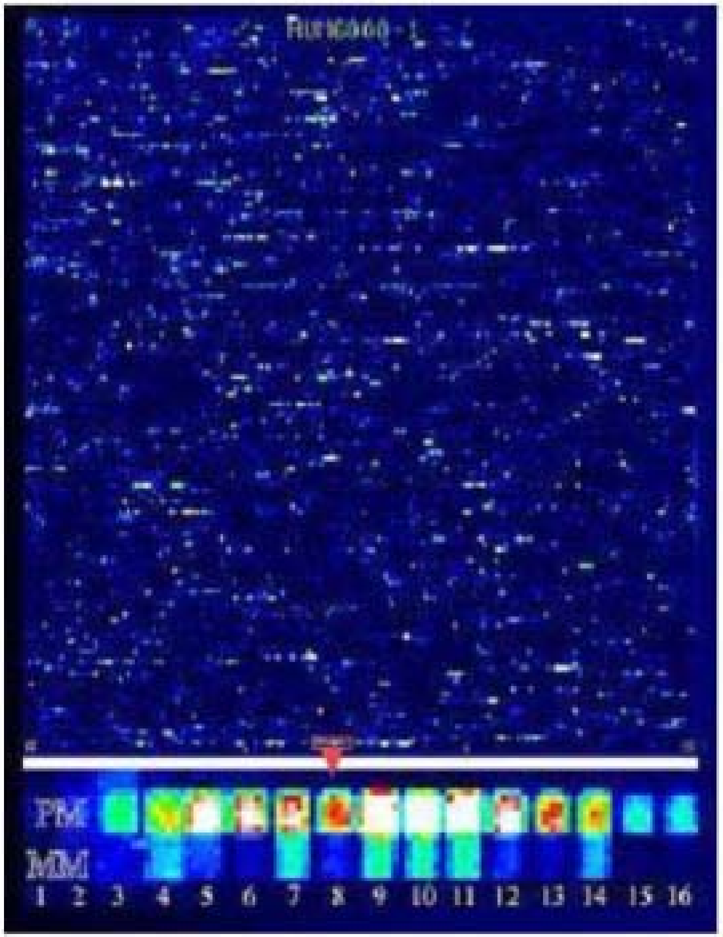


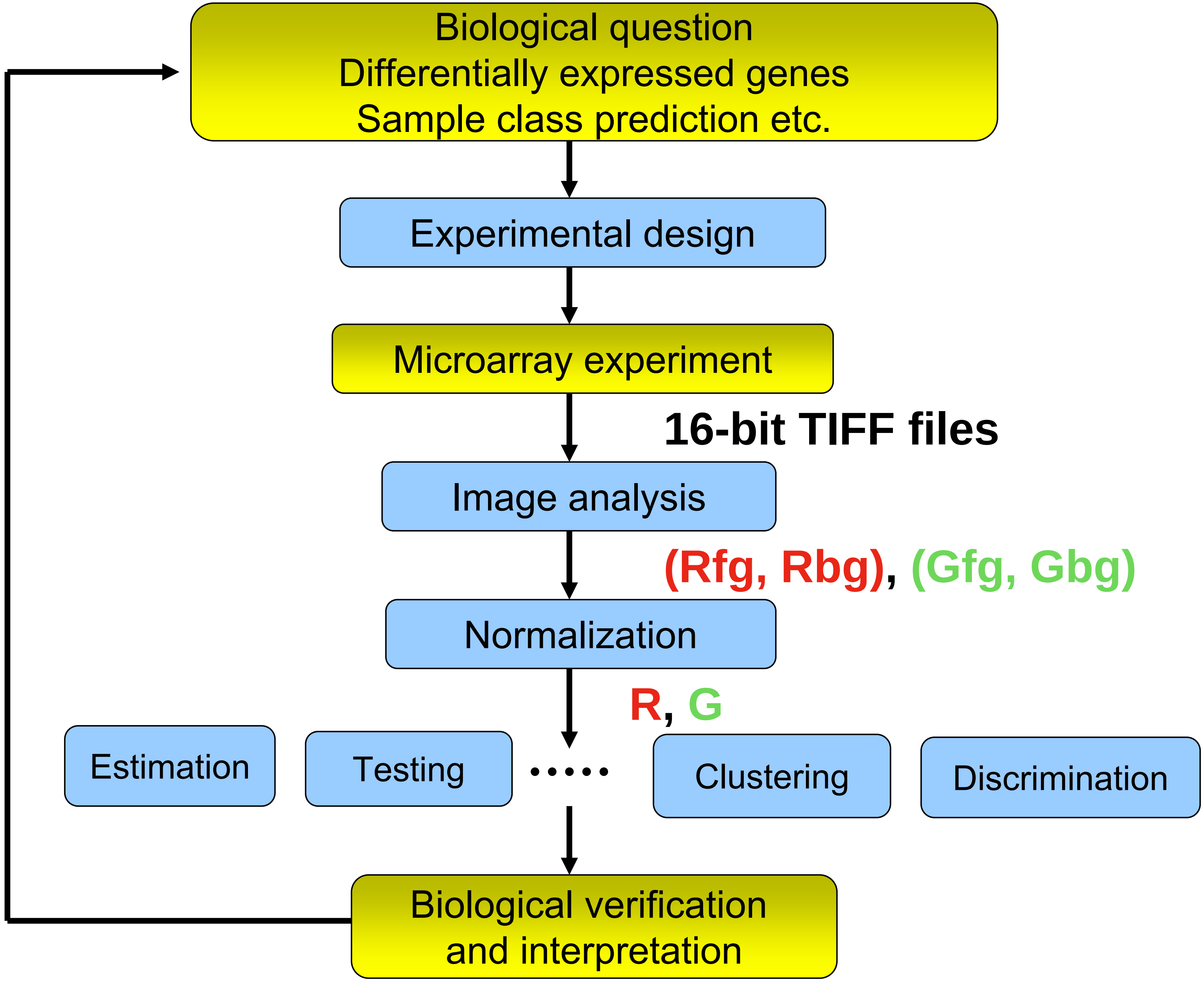
Exclude bordering pixels; plot remaining pixels on histogram

75% percentile value

Probe cell viewed in .DAT file: pixels with varying intensities.

Probe cell viewed in .CEL file: One intensity value.





Biological question
Differentially expressed genes
Sample class prediction etc.

Experimental design

Microarray experiment

16-bit TIFF files

Image analysis

(Rfg, Rbg), (Gfg, Gbg)

Normalization

R, G

Estimation

Testing

……

Clustering

Discrimination

Biological verification
and interpretation

Image Analysis

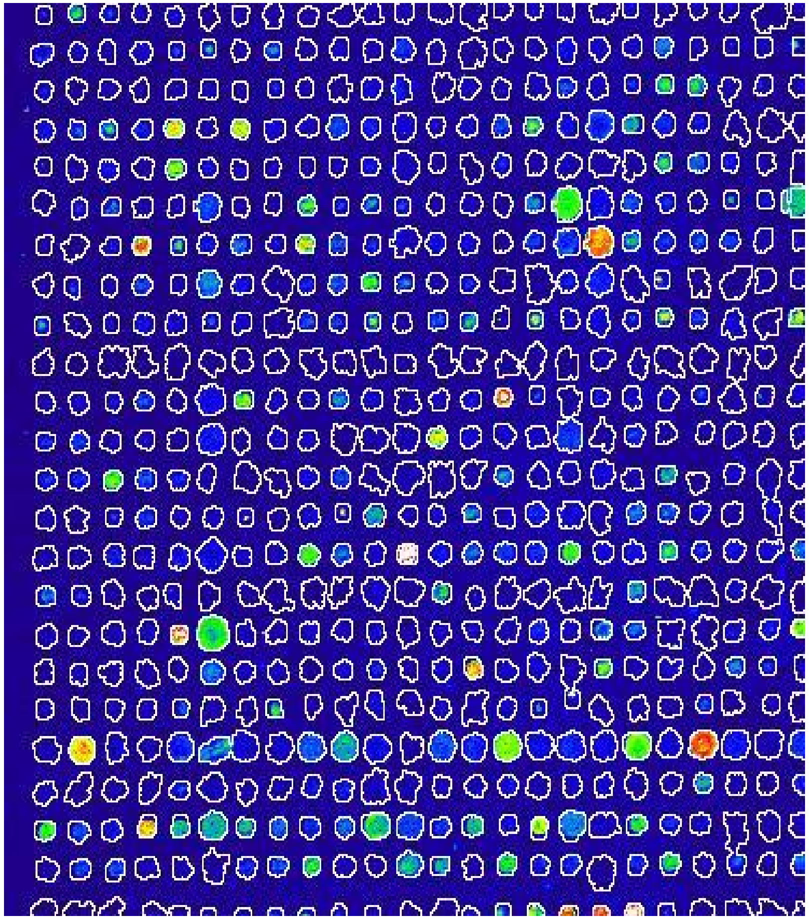
- Intensity is measured at fixed set of locations (pixels) arranged in rectangular patterns on the solid surface
- The distance between pixels is much less than the distance between probes
- The scanning microscope doesn’t know where the probes are; it just measures intensities at a fine grid of pixels

Image Analysis

1. **Gridding: isolate probes**

2. Segmentation:
classification of pixels either as signal or background.

3. Information extraction:
calculate signal intensity background and quality Measures for each channel at each probe



Need for Normalization for Dual-Channel Array Data

- Unequal incorporation of labels
  - green better than red
- Unequal amounts of sample
- Unequal signal detection

- Dual-channel arrays are normalized separately to adjust for dye bias

- Affymetrix arrays are normalized relative to each other to equalize intensities

What Genes To Use For Normalization?

- Constantly expressed genes (house-keeping)
- All genes on the array

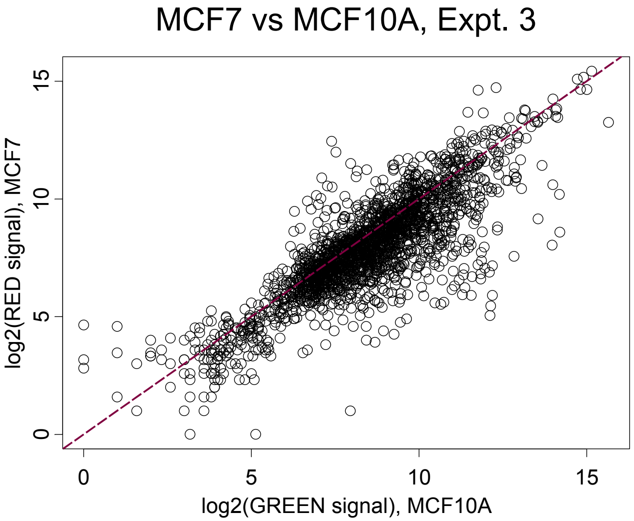
Global Normalization for Dual-Channel Arrays

- Assume $R_i \sim k\ G_i$

for all genes i in the normalization set

- Median-centered estimate
  - $k = \text{median}\{R_i/G_i\}$
  - $R_i' = R_i/k$

After Median Centering

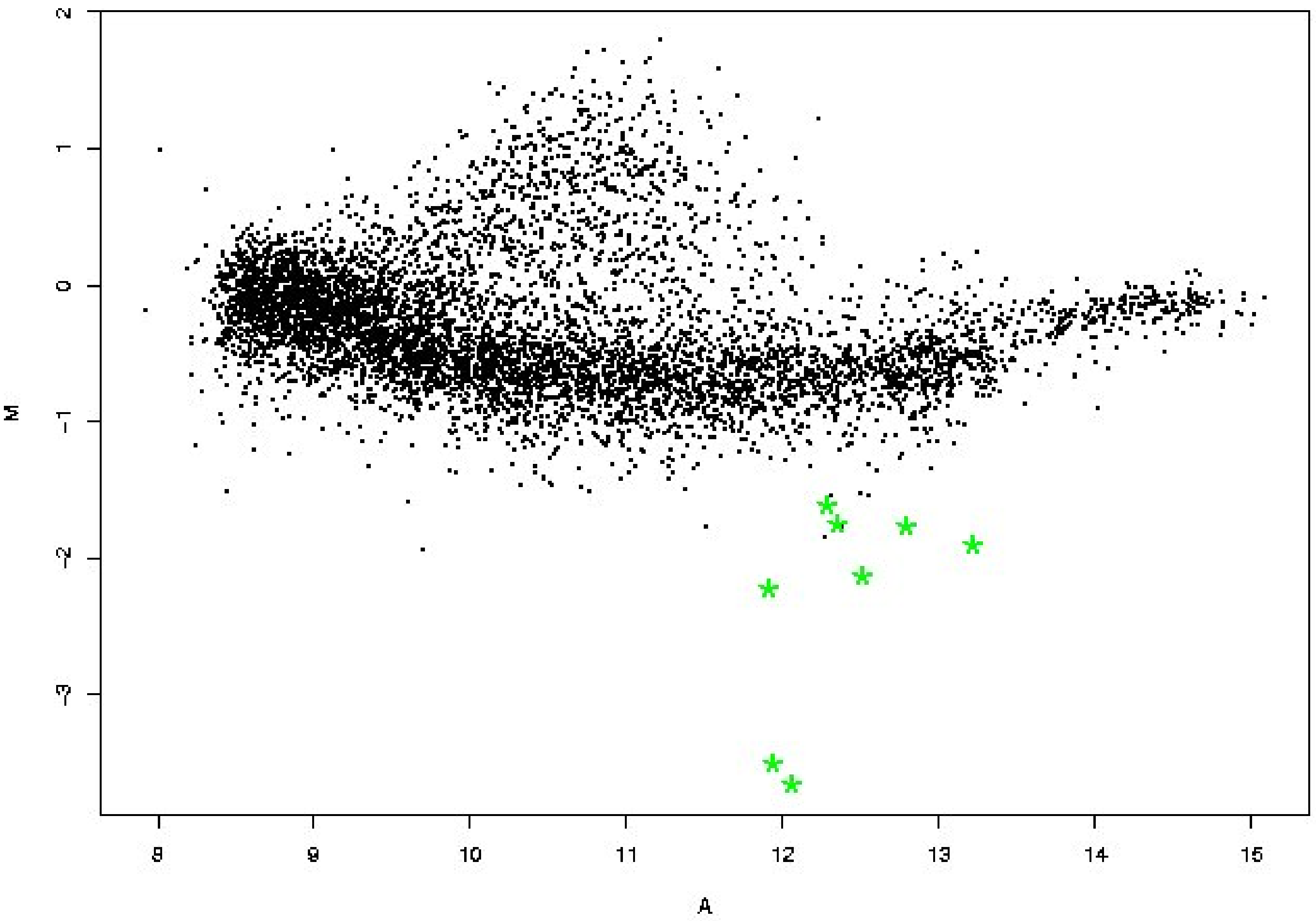


In plot of log(red signal) versus log(green signal), if point scatter is parallel to 45° line, adjust intercept to 0.

M vs. A

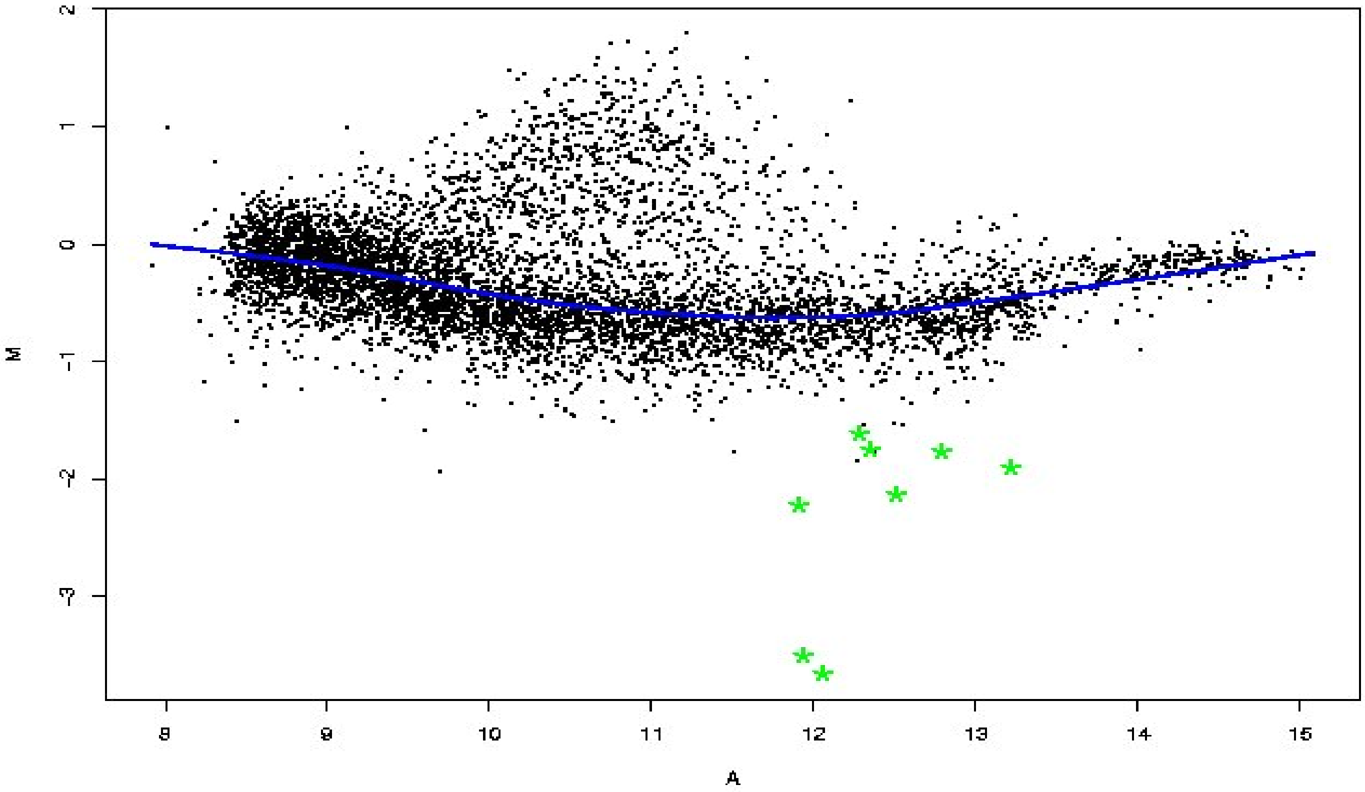
$M = \log_2(R / G)$

$A = \log_2(R*G) / 2$

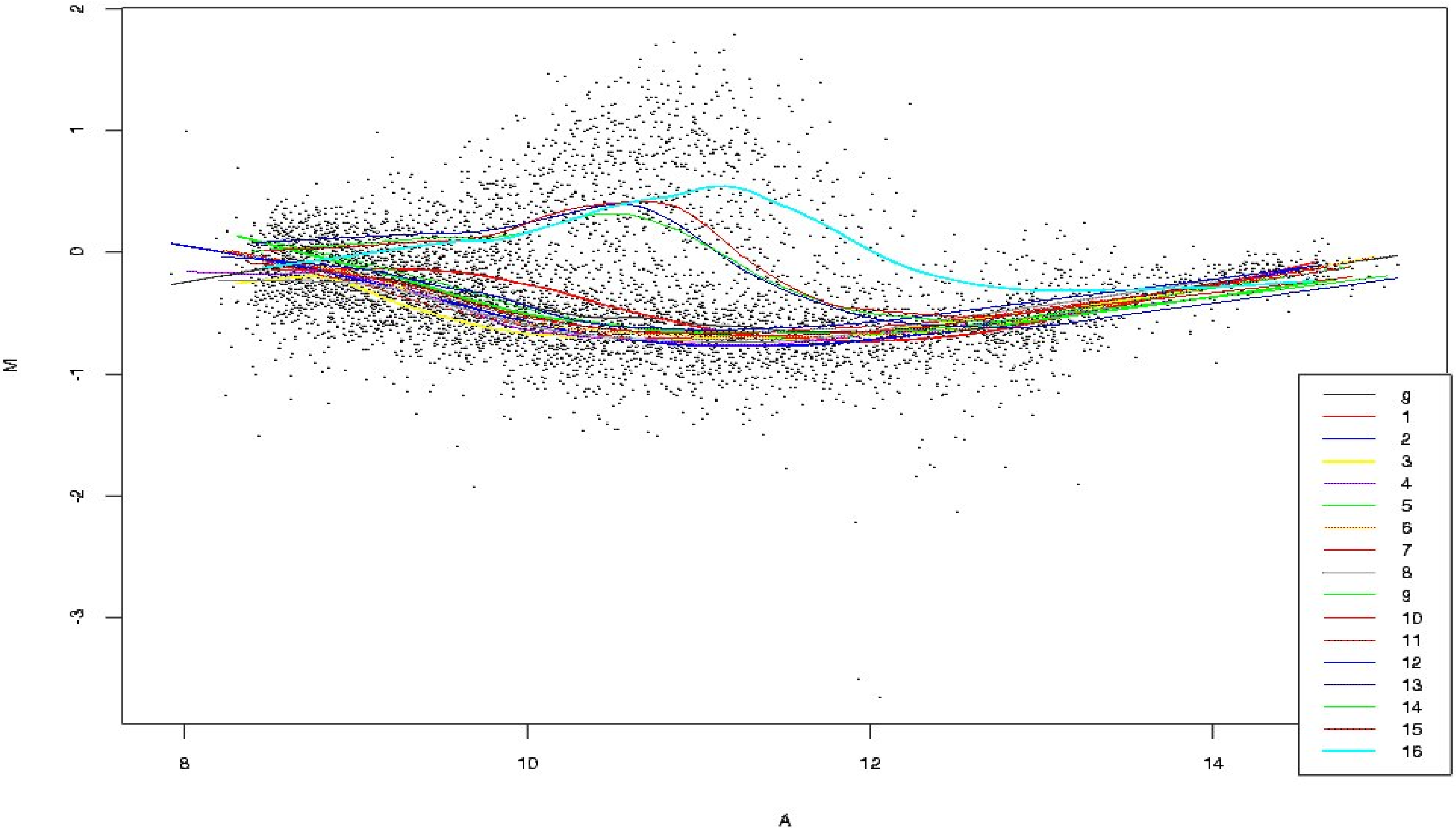


Normalization - lowess

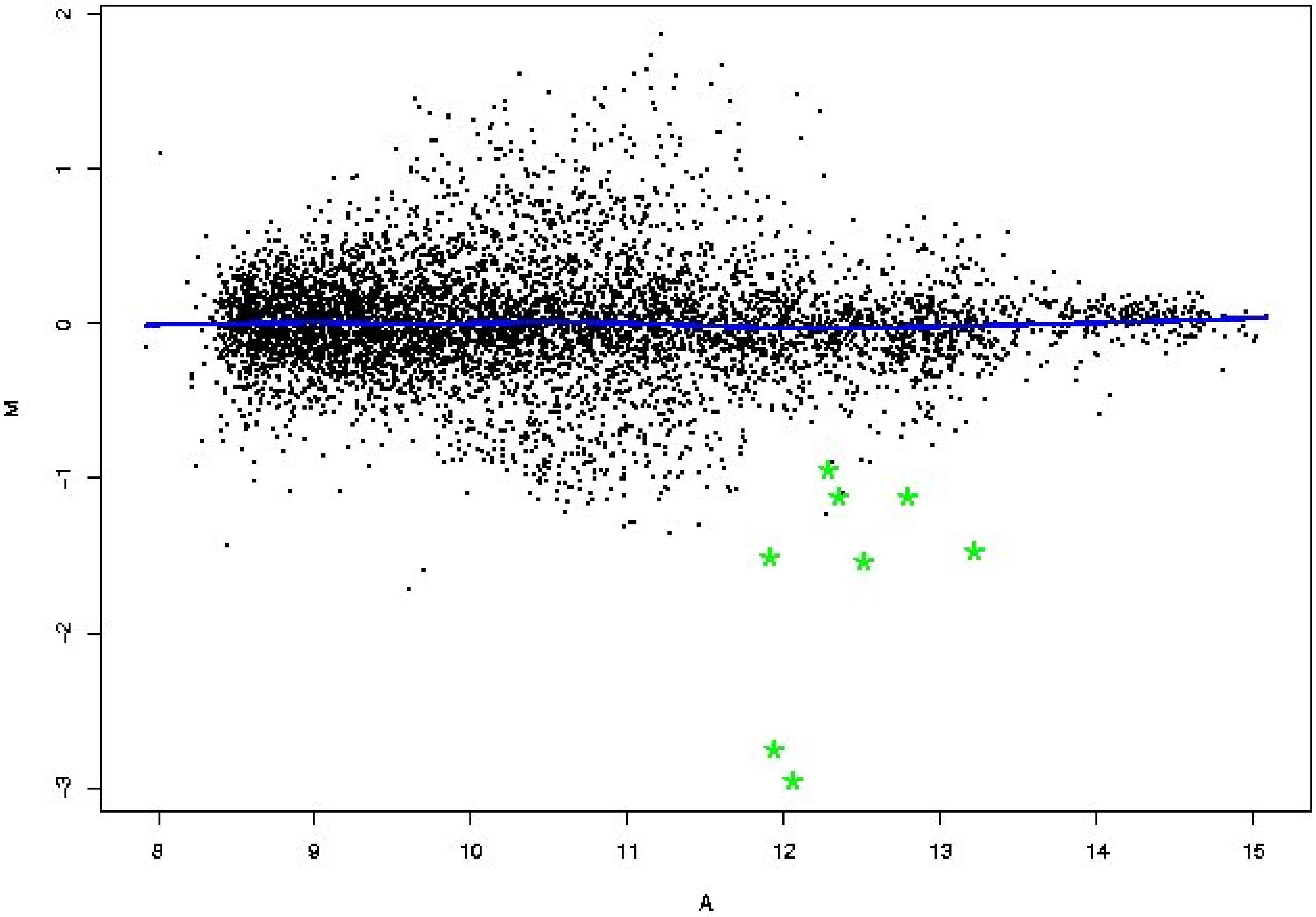
- Global lowess



Normalisation - print-tip-group



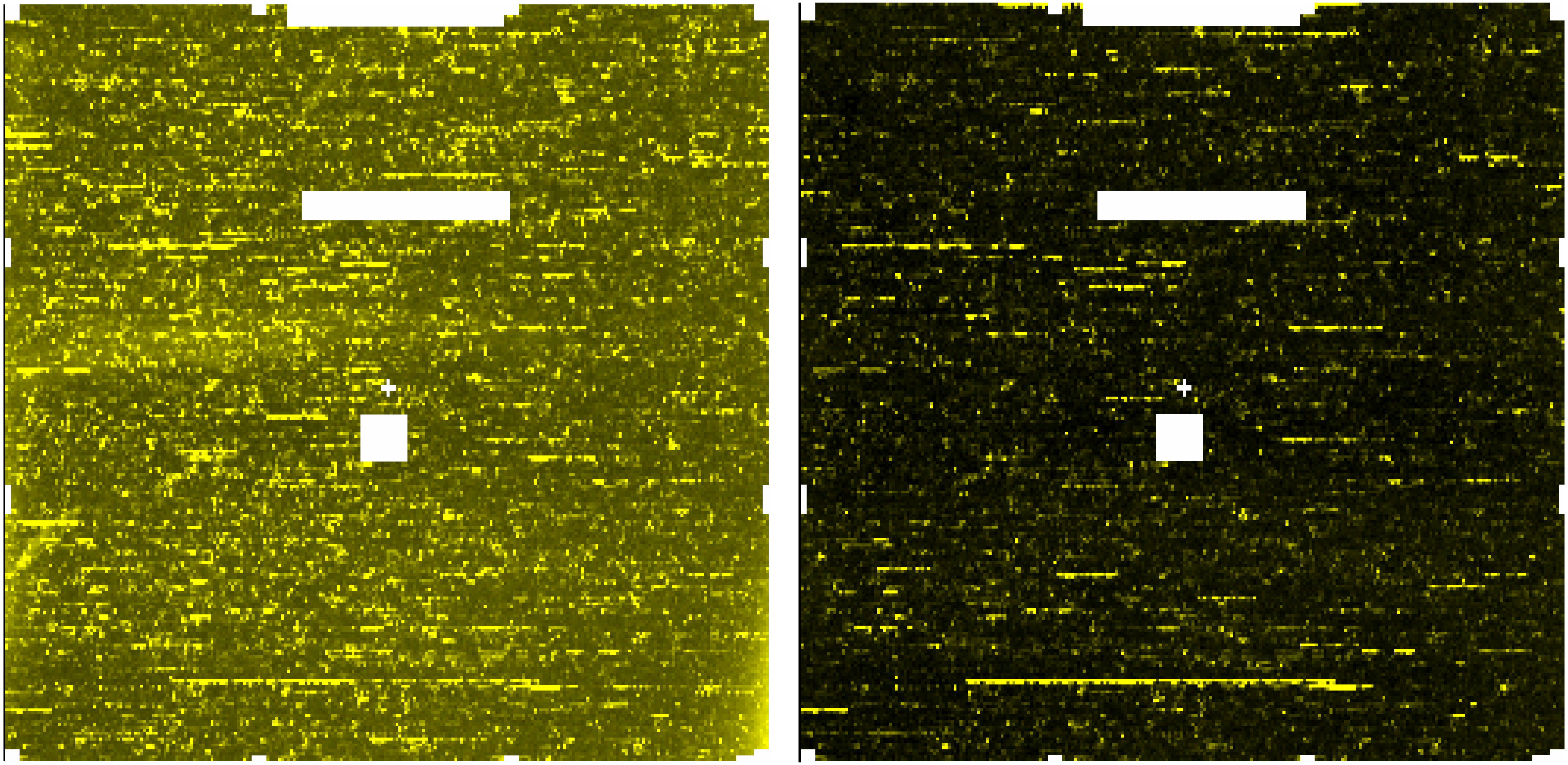
M vs. A - after print-tip-group normalization



Normalization for Affymetrix Arrays

- Need
  - Variations in amount of sample or environmental conditions
  - Variations in chip, hybridization, scanning

Normalization is needed to minimize non-biological variation between arrays



Normalization for Affymetrix Arrays

- Genes used
  - Affymetrix identifies housekeeping genes for some of their new arrays
- Methods
  - Scale each array so that it’s median signal equals a target value
  - Scale each array so that it’s median signal equals the median for a reference array
  - Intensity dependent normalization using lowess smoother based on ratios relative to a reference array

Spot Filtering Strategies

- Exclude if Signal < threshold in either channel

- Exclude if Signal < threshold in both channels

- If Min(R,G) < threshold
  - and Max(R,G) < threshold then exclude
  - Otherwise replace Min(R,G) by threshold

Gene Filtering Strategies

- "Bad" values on too many arrays.
- Not differentially expressed across arrays.
  - Proportion of arrays < 1.5 fold different from median for gene <20%

Affymetrix Arrays: Probe Set Summaries

MAS 4 Algorithm

- $\text{AvDiff}_i = \Sigma_j (\text{PM}_{ij}\text{-MM}_{ij}) / n_i$

for each probe set i

Summation over n_i =16-20 probes in probe set i

Excludes probe pairs that are more than 3 standard deviations from the average difference

Affymetrix Arrays:
Probe Set (Gene) Summaries
MAS5 Algorithm

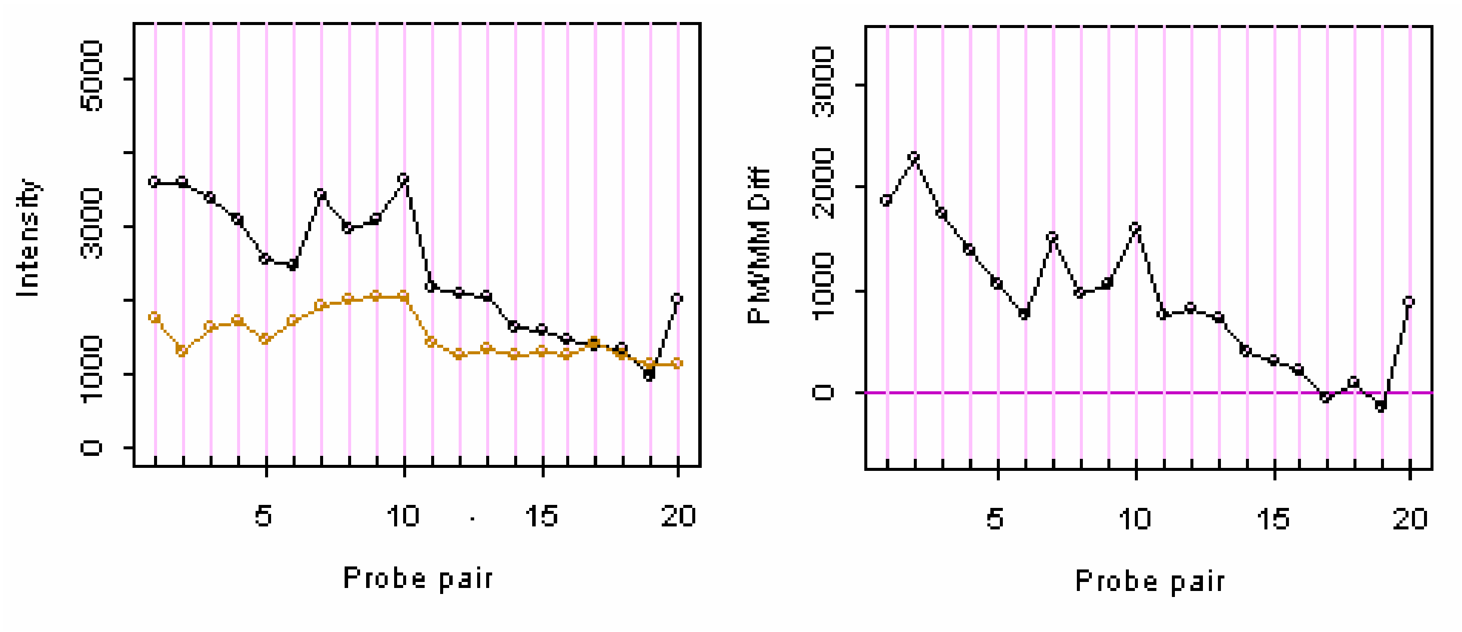
- Signal = $\Sigma w_{ij} (PM_{ij}-MM_{ij})^{+}$

Uses Tukey *biweights* that continuously down-weights probe pairs whose difference is far from the average difference

Negative probe pair differences are modified to make them non-negative

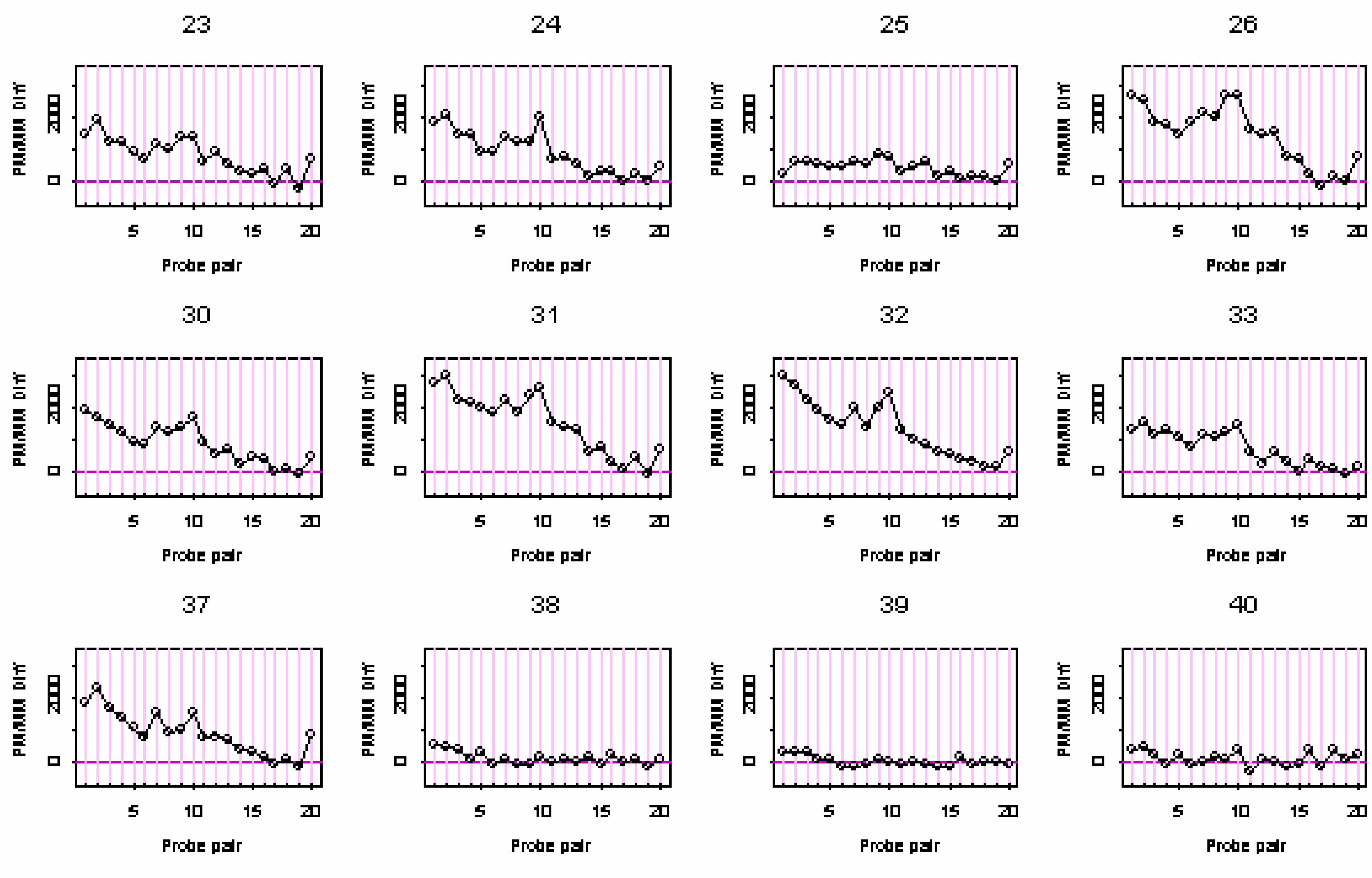
$(PM_{ij}-MM_{ij})^{+} = \max\{ 0, PM_{ij}-MM_{ij}\}$

Data for one probe set, one array



Use of PM/MM differences attempts to eliminate background and cross-hybridization signals

Data for one gene in many arrays



Li-Wong Model

A multiplicative model for each gene:

$$PM_{jk} - MM_{jk} = \theta_k \phi_j + \varepsilon_{jk}$$

θ_k: summary expression index for probe set on array k ,

ϕ_j : probe sensitivity index for probe pair j

ε_{jk}: random normal errors

$$\hat{\theta}_\mathrm{k} = \sum_j \hat{\phi}_j \left(PM_{jk} - MM_{jk} \right)$$

RMA (Irizarry et al.)

$$\log_2(PM'_{jk}) = \theta_k + \phi_j + \varepsilon_{jk}$$

θ_k: summary expression index for probe set on array k ,

ϕ_j : probe sensitivity index for probe pair j

ε_{jk}: random normal errors

$PM'_{jk} =$ globally background adjusted PM_{jk}

$PM'_{jk} = E(S_{jk} \mid PM_{jk})$ where $PM_{jk} = S_{jk} + bg_{jk}$

$$bg_{\mathrm{jk}} \sim N(\mu_k, \sigma_k^2)$$

$$S_{jk} \sim \mathrm{exponential}(\lambda_k)$$

RMA

- Estimate the background parameters globally for each array
- Estimate expression summaries θ_k for each probe set and each array k using Tukey’s median polish algorithm

Affymetrix Present/Absent Calls

- Based on Mann-Whitney rank test of the hypothesis that the probe specific PM-MM differences are independent observations with median value zero

Design of Microarray Studies

Myth

- That microarray investigations should be unstructured data-mining adventures without clear objectives

Myth

- That the greatest challenge is managing the mass of microarray data
- Greater challenges are:
  - Effectively designing and properly analyzing experiments that utilize microarray technology
    - Distinguishing hype and misinformation from sound methodology
    - Avoiding software developed by individuals with no qualifications for determining valid methodology
  - Organizing and facilitating effective interdisciplinary collaboration with statisticians, clinicians & biologists

Myth

- That data mining is an appropriate paradigm for analysis of microarray data
  - find interesting patterns that give clear answers to questions that were never asked
- That planning microarray investigations does not require “hypotheses” or clear objectives

- Good microarray studies have clear objectives, but not generally gene specific mechanistic hypotheses
- Design and analysis methods should be tailored to study objectives

Good Microarray Studies Have Clear Objectives

- Class Comparison
  - Find genes whose expression differs among predetermined classes
- Class Prediction
  - Prediction of predetermined class (phenotype) using information from gene expression profile
- Class Discovery
  - Discover clusters of specimens having similar expression profiles
  - Discover clusters of genes having similar expression profiles

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 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

Class Discovery Examples

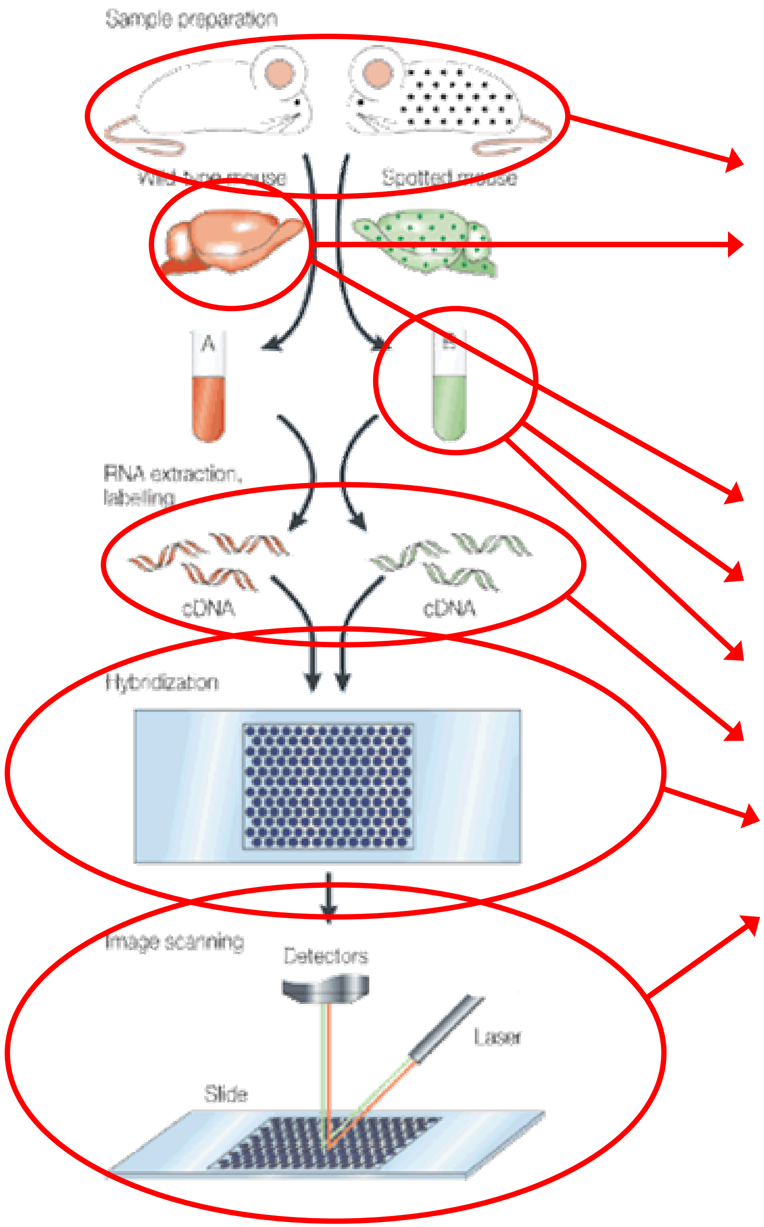
- Discover previously unrecognized subtypes of lymphoma
- Identify co-regulated genes

Design Considerations

- Sample and control selection
- Levels of replication
- Allocation of samples to (cDNA) array experiments
- Number of biological samples

Sources of Variability

(cDNA Array Example)



(Geschwind, *Nature Reviews Neuroscience*, 2001)

- 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

Levels of Replication

- Technical replicates
  - RNA sample divided into multiple aliquots and re-arrayed
- Biological replicates
  - Multiple subjects
  - Re-growing the cells under the defined conditions

Technical Replicates of the Same RNA Sample

- Useful to establish that experimental technique and reagents are adequate
  - Not necessary for all samples
- Protection against bad hybridizations
- Technical replicates improve precision for comparing a given sample to another given sample. For comparing classes, however, it is more efficient to use a limited number of arrays for more independent biological samples than for technical replicates.

Levels of Replication

- For comparing classes, replication of samples should generally be at the “biological/subject” level because we want to make inference to the population of “cells/tissues/subjects”, not to the population of sub-samples of a single biological specimen.

Which Genes are Differentially Expressed In Two Conditions or Two Tissues?

- Not a clustering problem
  - Global similarity measures generally used for clustering arrays may not distinguish classes
  - Feature selection should be performed in a manner that controls the false discovery rate
- Supervised methods
- Requires multiple biological samples from each class

Myth

- That comparing tissues or experimental conditions is based on looking for red or green spots on a single array
- That comparing tissues or experimental conditions is based on using Affymetrix MAS software to compare two arrays
  - Many published statistical methods are limited to comparing rna transcript profiles from two samples

Truth

- Comparing expression in two RNA samples tells you (at most) only about those two samples and may relate more to sample handling and assay artifacts than to biology. Robust knowledge requires multiple samples that reflect biological variability.

Class Comparison: Allocation of Specimens to cDNA Array Experiments

- Reference Design
- Balanced Block Design
  - Dobbin & Simon
- Loop Design
  - Kerr & Churchill

Reference Design

RED	A_1	A_2	B_1	B_2
GREEN	R	R	R	R
	Array 1	Array 2	Array 3	Array 4

A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

R = aliquot from reference pool

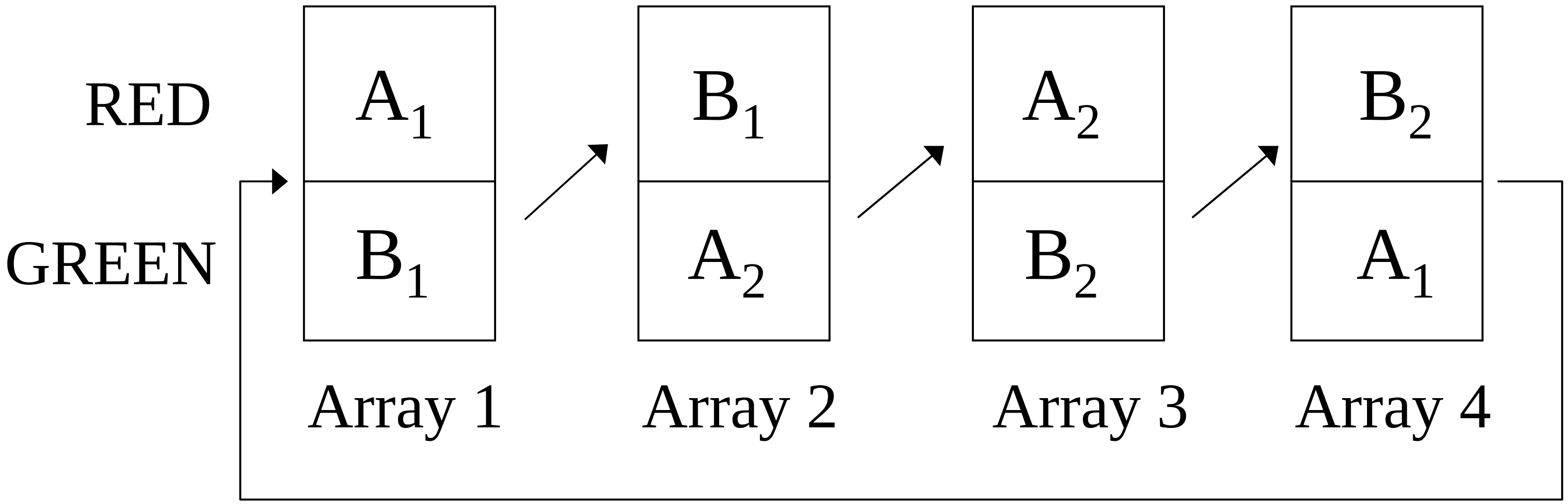
Balanced Block Design

RED	A_1	B_2	A_3	B_4
GREEN	B_1	A_2	B_3	A_4
	Array 1	Array 2	Array 3	Array 4

A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

Loop Design



	Array 1	Array 2	Array 3	Array 4
RED	A_1	B_1	A_2	B_2
GREEN	B_1	A_2	B_2	A_1

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

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

(Requires two aliquots per specimen)

- Detailed comparisons of the effectiveness of designs:
  - Dobbin K, Simon R. Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18:1462-9, 2002
  - Dobbin K, Shih J, Simon R. Statistical design of reverse dye microarrays. Bioinformatics 19:803-10, 2003
  - Dobbin K, Simon R. Questions and answers on the design of dual-label microarrays for identifying differentially expressed genes, JNCI 95:1362-1369, 2003

Myth

- Common reference designs for two-color arrays are inferior to “loop” designs.

Truth

- Common reference designs are effective for many microarray studies. They are robust, permit comparisons among separate experiments, permit unplanned types of comparisons to be performed, permit cluster analysis and class prediction analysis.

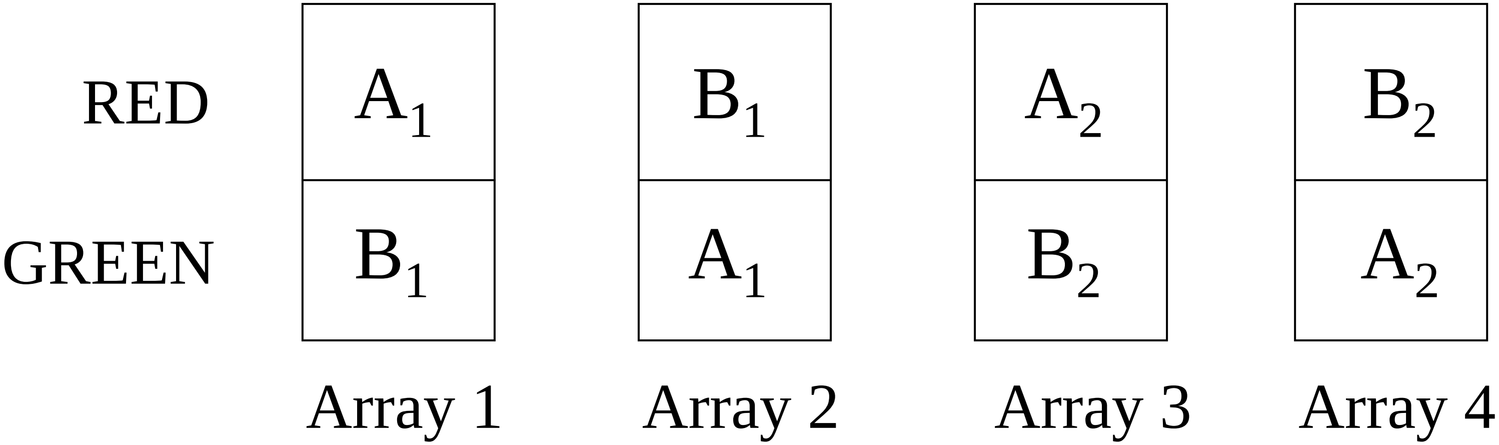
- Loop designs are non-robust, are very inefficient for class discovery (clustering) analyses, are not applicable to class prediction analyses and do not easily permit inter-experiment comparisons.

- For simple two class comparison problems, balanced block designs are the most efficient and require many fewer arrays than reference designs. They are not appropriate for class discovery or class prediction and are more difficult to apply to more complicated class comparison problems.

Myth

- For two color microarrays, each sample of interest should be labeled once with Cy3 and once with Cy5 in dye-swap pairs of arrays.

Dye Swap Design



	Array 1	Array 2	Array 3	Array 4
RED	A_1	B_1	A_2	B_2
GREEN	B_1	A_1	B_2	A_2

A_i = *i*th specimen from class A

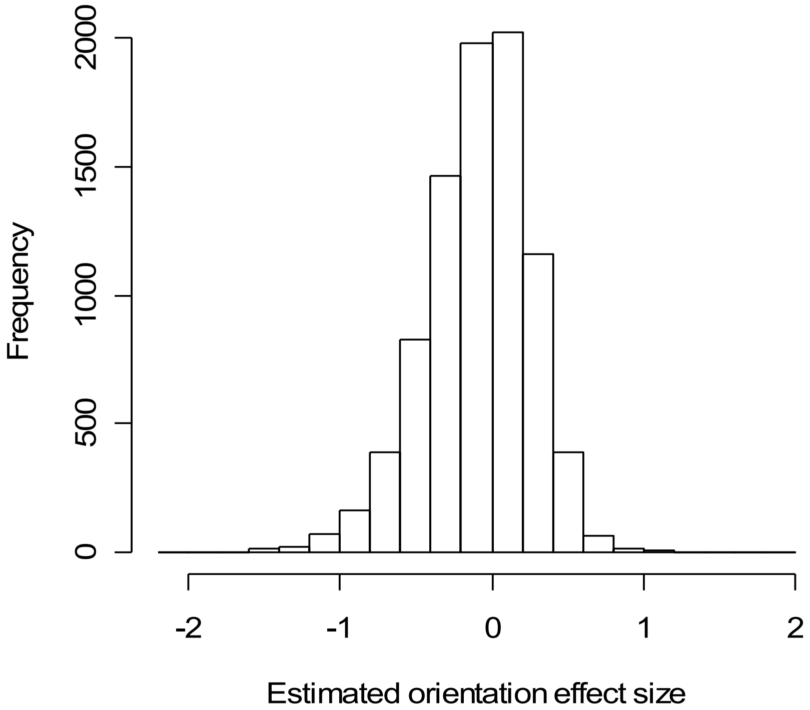
B_i = *i*th specimen from class B

Dye Bias

- Average differences among dyes in label concentration, labeling efficiency, photon emission efficiency and photon detection are corrected by normalization procedures
- Gene specific dye bias may not be corrected by normalization

- Gene-specific dye bias
  - 3681 genes with $p<0.001$ of 8604 evaluable genes

- Gene and sample specific dye bias
  - 150 genes with $p<0.001$


Frequency
0
500
1000
1500
2000
-2
-1
0
1
2
Estimated orientation effect size

- Dye swap technical replicates of the same two rna samples are rarely necessary.
- Using a common reference design, dye swap arrays are not necessary for valid comparisons of classes since specimens labeled with different dyes are never compared.
- For two-label direct comparison designs for comparing two classes, it is more efficient to balance the dye-class assignments for independent biological specimens than to do dye swap technical replicates

Balanced Block Design

RED	A_1	B_2	A_3	B_4
GREEN	B_1	A_2	B_3	A_4
	Array 1	Array 2	Array 3	Array 4

A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

Dye Swap Design

RED	A_1	B_1	A_2	B_2
GREEN	B_1	A_1	B_2	A_2
	Array 1	Array 2	Array 3	Array 4

A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

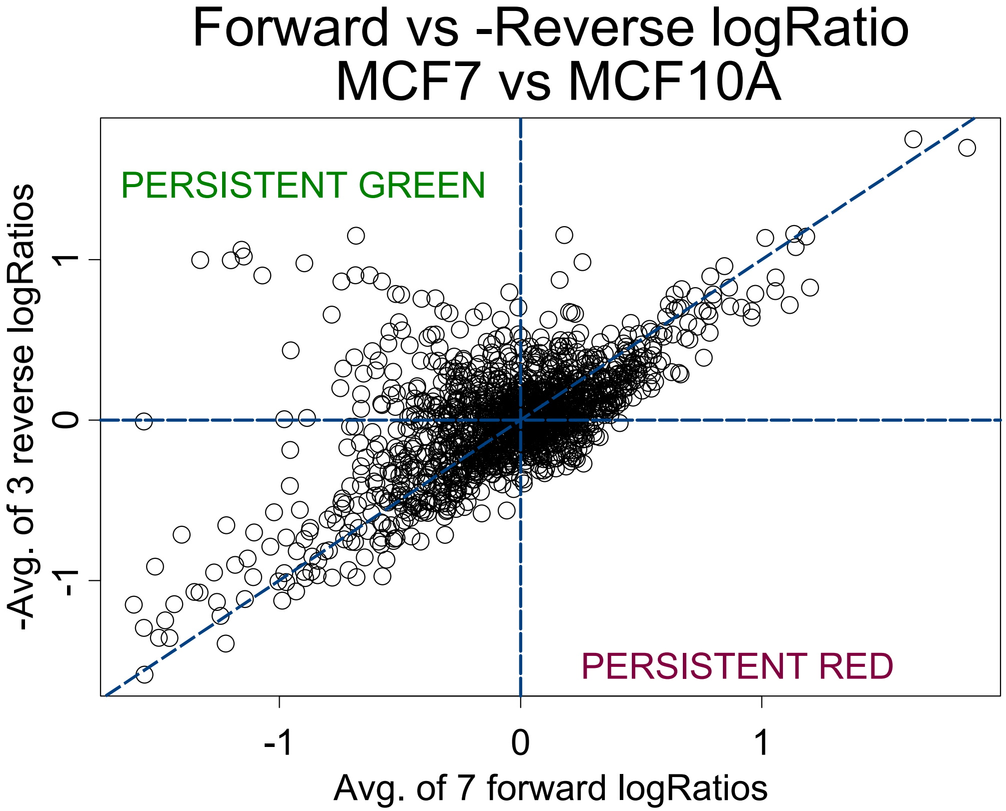
Balanced Block Designs for Two Classes

- Half the arrays have a sample from class 1 labeled with Cy5 and a sample from class 2 labeled with Cy3;
- The other half of the arrays have a sample from class 1 labeled with Cy3 and a sample from class 2 labeled with Cy5.
- Each sample appears on only one array. Dye swaps of the same rna samples are not necessary to remove dye bias and for a fixed number of arrays, dye swaps of the same rna samples are inefficient

Limitations of Balanced Block Designs

- One class comparison
- Does not support cluster analysis
- Requires ANOVA analysis of single channel log intensities

cDNA Arrays: Reverse Fluor Experiments



Reverse Labeled Arrays

- Not necessary with reference design if you are not interested in direct comparison to internal reference
  - If reference rna is consistently labeled with the same dye, dye bias effects all classes equally and does not bias comparison of classes.
  - For clustering of specimens, the reference design should be used and no reverse labeled arrays are necessary.

Reverse Labeled Arrays

- Using balanced block design to directly compare two classes, using each rna sample on only one array and balancing labels between classes is more efficient than using reverse labeled technical replicates.
  - For a fixed total number of arrays, use of reverse labeled technical replicates reduces the number of independent biological samples included

Reverse Labeled Arrays

- Necessary with reference design for some arrays if you are interested in direct comparison to internal reference
  - Gene specific dye bias not removed by normalization

Replicate Arrays of Independent Samples from Same Tissue

- Useful for establishing that clusters of samples represent different disease groups rather than just heterogeneity of individual tissues or differences in tissue handling

Sample Selection

- Experimental Samples
  - Representative of the phenotype or the population under investigation.
- Reference Sample (for 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)

Avoid Confounding Classes for Analysis With Assay Procedures

- Obtaining samples
- RNA labeling
- Hybridization
  - Print set
  - reagents

Experimental Design

- Dobbin K, Simon R. Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18:1462-9, 2002
- Dobbin K, Shih J, Simon R. Statistical design of reverse dye microarrays. Bioinformatics 19:803-10, 2003
- Dobbin K, Shih J, Simon R. Questions and answers on the design of dual-label microarrays for identifying differentially expressed genes, JNCI 95:1362-69, 2003
- Simon R, Korn E, McShane L, Radmacher M, Wright G, Zhao Y. *Design and analysis of DNA microarray investigations*, Springer Verlag (2003)
- Simon R, Dobbin K. Experimental design of DNA microarray experiments. Biotechniques 34:1-5, 2002
- Simon R, Radmacher MD, Dobbin K. Design of studies with DNA microarrays. Genetic Epidemiology 23:21-36, 2002
- Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 6:27-38, 2005.

Good Microarray Studies Have Clear Objectives

- Class Comparison
  - Find genes whose expression differs among predetermined classes
- Class Prediction
  - Prediction of predetermined class (phenotype) using information from gene expression profile
- Class Discovery
  - Discover clusters of specimens having similar expression profiles
  - Discover clusters of genes having similar expression profiles

Class Comparison and Class Prediction

- Not clustering problems
  - Global similarity measures generally used for clustering arrays may not distinguish classes
  - Don’t control multiplicity or for distinguishing data used for classifier development from data used for classifier evaluation
- Supervised methods
- Requires multiple biological samples from each class

Levels of Replication

- Technical replicates
  - RNA sample divided into multiple aliquots and re-arrayed
- Biological replicates
  - Multiple subjects
  - Replication of the tissue culture experiment

- Biological conclusions generally require independent biological replicates. The power of statistical methods for microarray data depends on the number of biological replicates.
- Technical replicates are useful insurance to ensure that at least one good quality array of each specimen will be obtained.

Microarray Platforms for Class Comparison

- Single label arrays
  - Affymetrix GeneChips

- Dual label arrays
  - Common reference design
  - Other designs

Common Reference Design

RED	A_1	A_2	B_1	B_2
GREEN	R	R	R	R
	Array 1	Array 2	Array 3	Array 4

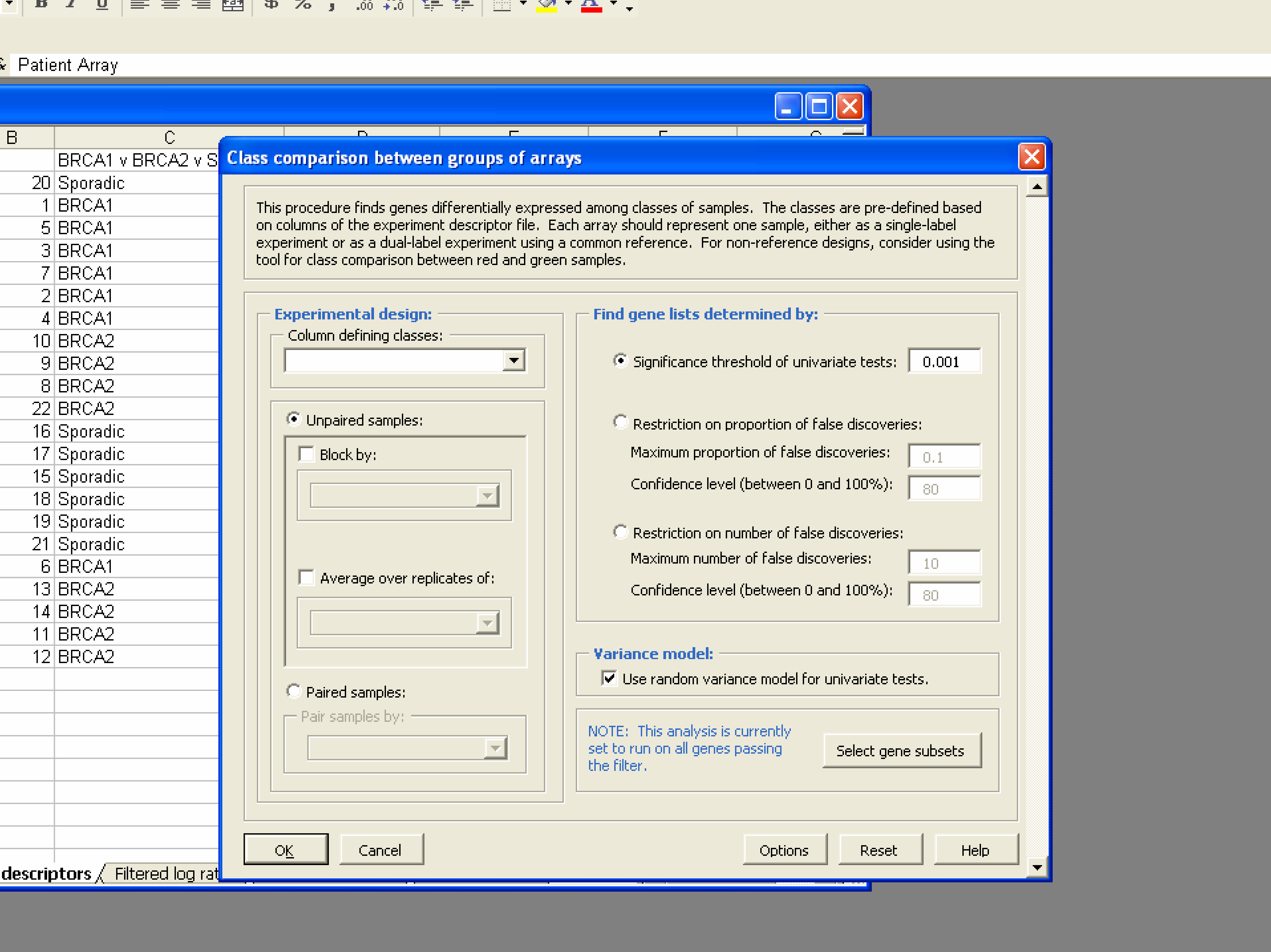
A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

R = aliquot from reference pool

Analysis Strategies for Class Comparison

- Compare classes on a gene by gene basis using statistical tests
  - Control for the large number of tests performed
  - Types of statistical significance tests
    - t-tests or F-tests
      - Hierarchical model for sharing variance information among genes
    - Univariate permutation tests
    - Analysis of variance to control for other variables
- Multivariate permutation tests



B	C
	BRCA1 v BRCA2 v S
20	Sporadic
1	BRCA1
5	BRCA1
3	BRCA1
7	BRCA1
2	BRCA1
4	BRCA1
10	BRCA2
9	BRCA2
8	BRCA2
22	BRCA2
16	Sporadic
17	Sporadic
15	Sporadic
18	Sporadic
19	Sporadic
21	Sporadic
6	BRCA1
13	BRCA2
14	BRCA2
11	BRCA2
12	BRCA2

Class comparison between groups of arrays

This procedure finds genes differentially expressed among classes of samples. The classes are pre-defined based on columns of the experiment descriptor file. Each array should represent one sample, either as a single-label experiment or as a dual-label experiment using a common reference. For non-reference designs, consider using the tool for class comparison between red and green samples.

Experimental design:

Column defining classes:

● Unpaired samples:

□ Block by:

□ Average over replicates of:

○ Paired samples:

Pair samples by:

Find gene lists determined by:

● Significance threshold of univariate tests: 0.001

○ Restriction on proportion of false discoveries:

Maximum proportion of false discoveries: 0.1

Confidence level (between 0 and 100%): 80

○ Restriction on number of false discoveries:

Maximum number of false discoveries: 10

Confidence level (between 0 and 100%): 80

Variance model:

☑ Use random variance model for univariate tests.

NOTE: This analysis is currently set to run on all genes passing the filter.

Class Comparison Blocking

- Paired data
  - Pre-treatment and post-treatment samples of same patient
  - Tumor and normal tissue from the same patient
- Blocking
  - Multiple animals in same litter
  - Any feature thought to influence gene expression
    - Sex of patient
    - Batch of arrays

Technical Replicates

- Multiple arrays on alloquots of the same RNA sample
- Select the best quality technical replicate or
- Average expression values

Controlling for Multiple Comparisons

- Bonferroni type procedures control the probability of making any false positive errors

- Overly conservative for the context of DNA microarray studies

Simple Control for Multiple Testing

- If each gene is tested for significance at level α and there are n genes, then the expected number of false discoveries is n α .
  - e.g. if n=1000 and α=0.001, then 1 false discovery
  - To control E(FD) $\leq u$
  - Conduct each of k tests at level $\alpha = u$/k

Simple Procedures

- Control E(FD) $\leq u$
  - – Conduct each of k tests at level u/k
  - – e.g. To limit of 10 false discoveries in 10,000 comparisons, conduct each test at $p<0.001$ level
- Control E(FDP) $\leq \gamma$
  - – Benjamini-Hochberg procedure

False Discovery Rate (FDR)

- FDR = *Expected* proportion of false discoveries among the tests declared significant
- Studied by Benjamini and Hochberg (1995):

	Not rejected	Rejected	Total
True null hypotheses	890	10 False discoveries	900
False null hypotheses	10	90 True discoveries	100
		100	1000

If you analyze n probe sets and select as “significant” the k genes whose p ≤ p*

- FDR ~ n p* / k

Limitations of Simple Procedures

- p values based on normal theory are not accurate in the extreme tails of the distribution
- Difficult to achieve extreme quantiles for permutation p values of individual genes
- Multiple comparisons controlled by adjustment of univariate (single gene) p values may not take advantage of correlation among genes

Gene-by-Gene Comparison of Classes

- t-test for comparing two classes
  - For dual-color arrays compare log-ratios, not ratios
  - For GeneChips compare log signals
  - t_g=(mean$_{g1}$-mean$_{g2}$)/standard-error$_g$
  - Standard-error$_g$= s_g $(1/n_1 + 1/n_2)^{1/2}$
  - s_g =within-class standard deviation
  - Computes statistical significance level as the probability of obtaining a t value as large in absolute value as actually obtained if the two classes had the same true means and the sampling variation had a Gaussian distribution
  - Gaussian distribution is symmetric “bell-shaped curve” which decreases at rate $\exp(-x^2)$

Limitations of Parametric t-test

- Expression values may not be approximately Gaussian
- t distribution approximation to the distribution of t under the null hypothesis is not accurate at the extreme tail of the distribution
- t distribution approximation is less accurate for small sample sizes
- Small sample size limits accuracy of estimation of s_g
  - Few degrees of freedom for t limits statistic power for detecting differences in mean expression levels

Gene-by-Gene Comparison of Classes

- Permutation t-test
  - Compute the t statistic comparing the two classes for a gene but don’t use the Gaussian distribution assumption to translate the t value into a p value
  - Consider all possible permutations of the labels of which arrays correspond to which class, holding fixed the number of total arrays in each class

Gene-by-Gene Comparison of Classes

- Permutation t-test (cont)
  - For each permutation of class labels re-compute the t statistic comparing the classes with regard to a specific gene
  - Determine the proportion of the permutations that gave a t value at least as large in absolute value as the one corresponding to the true data
  - That proportion is the permutation p value

Limitation of Univariate Permutation Analysis

- Statistical significance level is limited by the number of possible permutations of the class labels. For small sample sizes, statistical significance at a stringent significance level (e.g. p<0.001) either cannot be achieved or is achieved with limited statistical power

Gene-by-Gene Comparison of Classes

- All of these tests assume that the different arrays are independent. Hence replicate arrays must be either averaged, or the best quality one selected for inclusion in the analysis, or a more complex ANOVA model be used for analysis

Gene-by-Gene Comparison of Classes

- F-test
  - The generalization of the t-test when there are more than 2 classes to compare.
    - Significance indicates that the class means are more different than one expects by chance but it does not indicate which classes are different from which other classes.
    - The statistically significant genes may differ with regard to the patterns of differences among classes that they show. Clustering the set of significant genes is useful to sort the genes into sets with uniform patterns.

Gene-by-Gene Comparison of Classes

- F-test
  - The standard F test computes statistical significance based on an assumption of Gaussian distribution of sampling variability.
  - The permutation F-test is a generalization of the permutation t-test and the associated p values are not based on Gaussian assumptions.

t-test Comparisons of Gene Expression for gene j

- $x_j \sim N(\mu_{j1} , \sigma_j^2)$ for class 1
- $x_j \sim N(\mu_{j2} , \sigma_j^2)$ for class 2

- H_{0j}: $\mu_{j1} = \mu_{j2}$

Estimate variances individually

Treat each σ_j^2 as a separate unknown quantity, and estimate separately for each gene.

Advantages: Allows each gene to have it’s own variance.

Disadvantages: In cases of small sample size estimate will have few degrees of freedom. Ignores the wealth of information provided by other genes

Pool Variance

Assume all genes have same residual variance so that

$$\sigma_j^2 = \sigma^2$$

Use all genes to estimate single variance value

Advantages:

Large Numbers of degrees of freedom for variance estimate

Disadvantages:

Not realistic, in observed data, some genes can be 10 times more variable than other genes

Randomized Variance Model

Assume that the variances of the genes are themselves drawn at random from an inverse Gamma distribution

$$\begin{aligned}\sigma_j^{-2} &\sim \text{Gamma}(t; a, b) \\ &= \frac{t^{a-1}\exp(-t/b)}{\Gamma(a)b^a}\end{aligned}$$

a and *b* are parameters that can be estimated from the entire set of genes.

a will indicate the shape or peakedness of the distribution of variances
b will scale the size of the variance, such that $E(1/\sigma^2) = ab$

Randomized Variance Model

Advantages:

- Allows for the variance to realistically vary between genes
- Uses information from all genes to contribute to variance estimates increasing reliability of estimate.

Disadvantages:

- Requires additional assumptions about the distribution of the variances
- Estimates of variance may still be noisy

Randomized Variance t-test

- $\Pr(\sigma^{-2}=x) = x^{a-1}\exp(-x/b)/\Gamma(a)b^{a}$

$$t = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\tilde{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\tilde{\sigma}^2 = \frac{(n-2)\hat{\sigma}^2_{pooled} + 2b^{-1}}{\left(n-2\right) + 2a}$$

Modified T-test

As an application of testing between 2 varieties, the standard T-test is usable with the following modification.

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\widehat{\sigma}_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{becomes} \quad t = \frac{\overline{Y}_1 - \overline{Y}_2}{\widetilde{\sigma}_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where,

$$\widetilde{\sigma}_{pooled} = \frac{(n-2)\widehat{\sigma}^2_{pooled} + 2b^{-1}}{(n-2) + 2a}$$

and the number of degrees of freedom increases from *n-2* to *n-2+2a.*

This is similar to a result by Baldi and Long (Bioinfomatics 2001)
Who approached this problem from a purely Bayesian standpoint.

Estimation of parameters a and b

Under our model for σ^2 we find that for each gene,

$$\frac{ab}{n-k}\left(SS_{H_1}\right) \sim F_{(n-k),2a}$$

Since we observe thousands of genes, we can arrive at quite accurate estimates of a and b by maximizing the likelihood of these observations with respect to these parameters.

Additional Procedures

- “SAM” - Significance Analysis of Microarrays
  - Tusher *et al.*, *PNAS*, 2001
  - Estimate FDR
  - Statistical properties unclear
- Multivariate permutation tests
  - Korn *et al., (Journal of Statistical Planning & Inference)*
  - Control number or proportion of false discoveries
  - Can specify confidence level of control

Multivariate Permutation Procedures

- More effective than univariate permutation tests especially with limited number of samples
  - Based on the α percentile of the distribution of the (k+1)st smallest p value under multivariate permutation distribution; not on the α/G percentile of the distribution of the univariate p value for a specific gene
- Stronger control than simple methods which control only expected number and proportion of false discoveries

Multivariate Permutation Procedures
(Simon et al. 2003, Korn et al. 2004)

Allows statements like:

- **FD Procedure**: We are 90% confident that the (actual) number of false discoveries is no greater than 5.
- **FDP Procedure**: We are 90% confident that the (actual) proportion of false discoveries does not exceed .10.

Control
$\Pr\{\text{Number of FD} > n\} \leq \alpha$

- $y = \alpha$ quantile of the distribution of the (n+1) st smallest p value under the multivariate permutation distribution.

- Include the genes corresponding to the n smallest p values in the gene list

- Include gene corresponding to $p_{(i)}$ if $p_{(i)} < y$

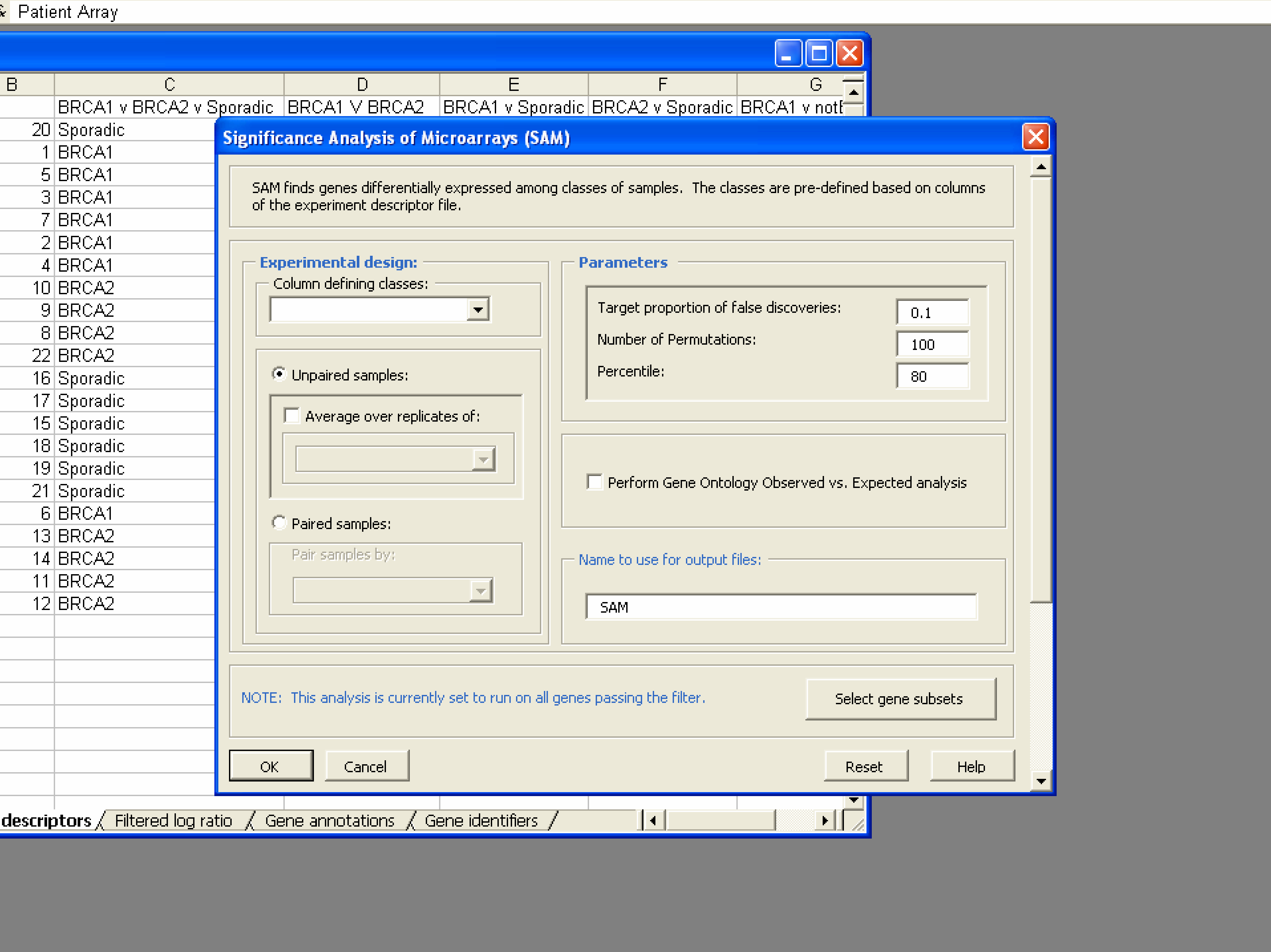
Multivariate Permutation Tests

- Distribution-free
  - even if they use t statistics
- Preserve/exploit correlation among tests by permuting each profile *as a unit*
- More effective than univariate permutation tests especially with limited number of samples

Control
$\Pr\{FDP > \gamma\} \leq \alpha$

- Determine $y(u) = \alpha$ quantile of the distribution of the (u+1)st smallest p value under the multivariate permutation distribution.
  - For $u = 1,2,3, \ldots$

- Include in the list of differentially expressed genes the gene corresponding to the i'th smallest p value as long as $p_{(i)} < y(\lfloor \gamma i \rfloor)$
  - Sequentially for $i = 1,2, \ldots$
  - $\lfloor \gamma i \rfloor$ = largest integer less than or equal to γi



Patient Array

B	C	D	E	F	G
	BRCA1 v BRCA2 v Sporadic	BRCA1 V BRCA2	BRCA1 v Sporadic	BRCA2 v Sporadic	BRCA1 v notE
20	Sporadic				
1	BRCA1				
5	BRCA1				
3	BRCA1				
7	BRCA1				
2	BRCA1				
4	BRCA1				
10	BRCA2				
9	BRCA2				
8	BRCA2				
22	BRCA2				
16	Sporadic				
17	Sporadic				
15	Sporadic				
18	Sporadic				
19	Sporadic				
21	Sporadic				
6	BRCA1				
13	BRCA2				
14	BRCA2				
11	BRCA2				
12	BRCA2				

Significance Analysis of Microarrays (SAM)

SAM finds genes differentially expressed among classes of samples. The classes are pre-defined based on columns of the experiment descriptor file.

Experimental design:

Column defining classes:

● Unpaired samples:

□ Average over replicates of:

○ Paired samples:

Pair samples by:

Parameters

Target proportion of false discoveries: 0.1

Number of Permutations: 100

Percentile: 80

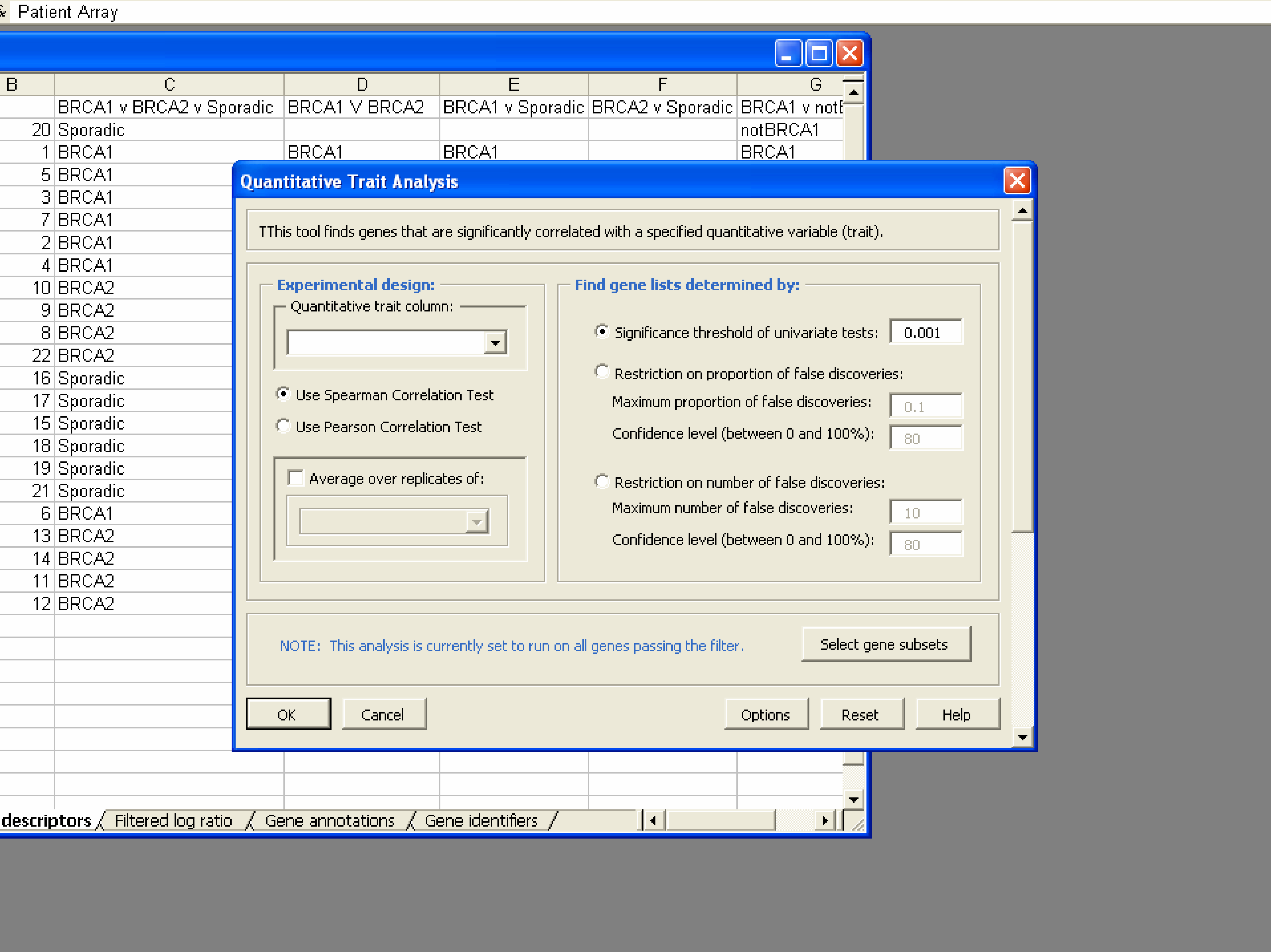
□ Perform Gene Ontology Observed vs. Expected analysis

Name to use for output files: SAM

NOTE: This analysis is currently set to run on all genes passing the filter.

Quantitative trait tool

- Selects genes which are univariately correlated with a quantitative trait such as age.
- Controls number and proportion of false discoveries in entire list: uses a multivariate permutation test which takes advantage of the correlation among genes.
- Produces a gene list which can be used for further analysis.



Patient Array

B	C	D	E	F	G
	BRCA1 v BRCA2 v Sporadic	BRCA1 V BRCA2	BRCA1 v Sporadic	BRCA2 v Sporadic	BRCA1 v notE
20	Sporadic				notBRCA1
1	BRCA1	BRCA1	BRCA1		BRCA1
5	BRCA1				
3	BRCA1				
7	BRCA1				
2	BRCA1				
4	BRCA1				
10	BRCA2				
9	BRCA2				
8	BRCA2				
22	BRCA2				
16	Sporadic				
17	Sporadic				
15	Sporadic				
18	Sporadic				
19	Sporadic				
21	Sporadic				
6	BRCA1				
13	BRCA2				
14	BRCA2				
11	BRCA2				
12	BRCA2				

Quantitative Trait Analysis

TThis tool finds genes that are significantly correlated with a specified quantitative variable (trait).

Experimental design:

Quantitative trait column:

● Use Spearman Correlation Test

○ Use Pearson Correlation Test

□ Average over replicates of:

Find gene lists determined by:

● Significance threshold of univariate tests: 0.001

○ Restriction on proportion of false discoveries:

Maximum proportion of false discoveries: 0.1

Confidence level (between 0 and 100%): 80

○ Restriction on number of false discoveries:

Maximum number of false discoveries: 10

Confidence level (between 0 and 100%): 80

NOTE: This analysis is currently set to run on all genes passing the filter.

Survival analysis tools

- Find Genes Correlated with Survival tool, selects genes which are univariately correlated with survival
- Controls number and proportion of false discoveries in entire list: uses a multivariate permutation test which takes advantage of the correlation among genes
- Produces a gene list which can be used for further analysis

Identifying Genes Correlated With Survival

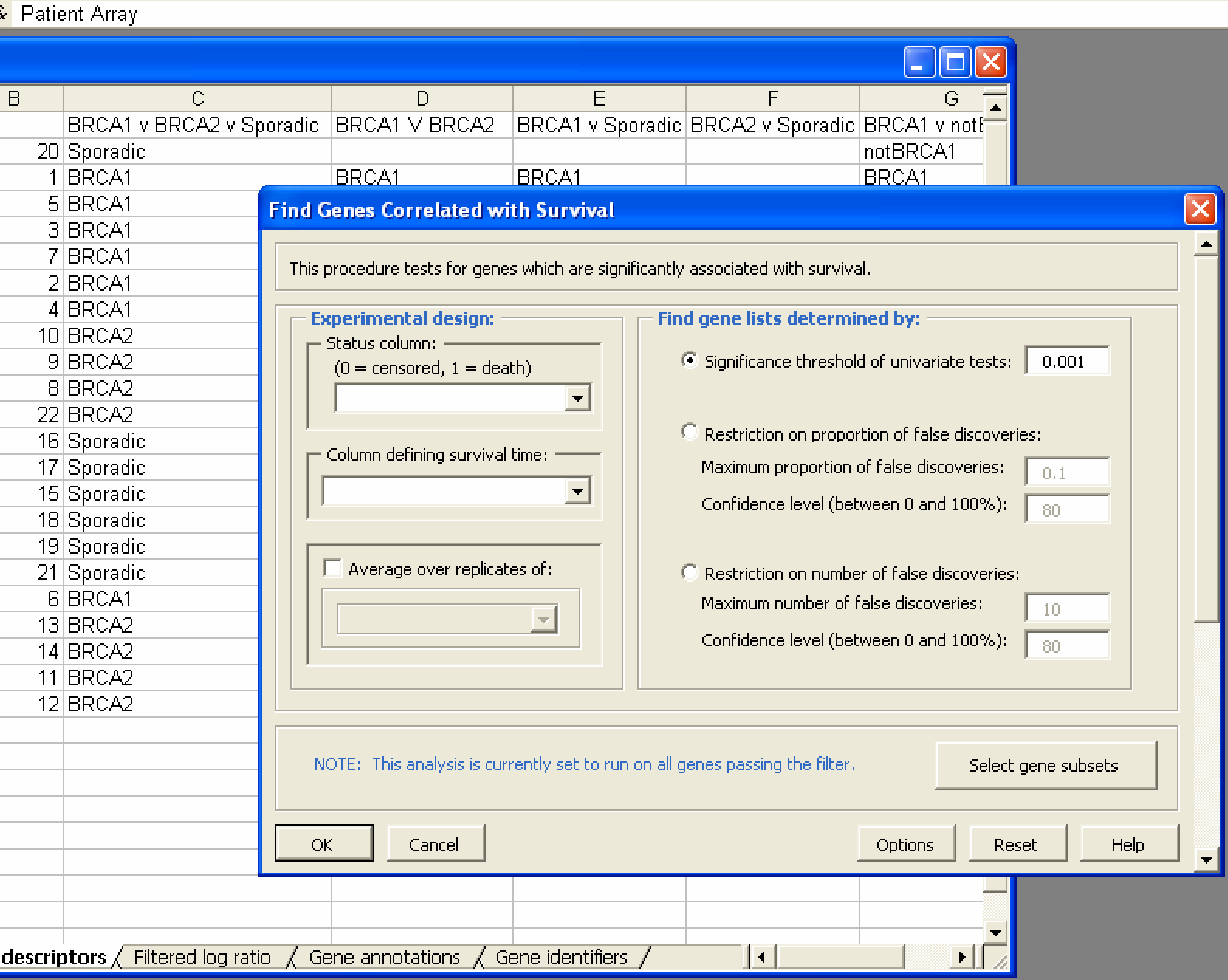
Instantaneous hazard of death at time t

$$\lambda(t) = \lambda_0(t)\exp(\beta_i x_i)$$

x_i = log ratio or log signal for gene i

Calculate p value for each gene i

Apply a multivariate permutation procedure to the p_i values, permuting survival times rather than class labels



Patient Array

B	C	D	E	F	G
	BRCA1 v BRCA2 v Sporadic	BRCA1 V BRCA2	BRCA1 v Sporadic	BRCA2 v Sporadic	BRCA1 v not
20	Sporadic				notBRCA1
1	BRCA1	BRCA1	BRCA1		BRCA1
5	BRCA1				
3	BRCA1				
7	BRCA1				
2	BRCA1				
4	BRCA1				
10	BRCA2				
9	BRCA2				
8	BRCA2				
22	BRCA2				
16	Sporadic				
17	Sporadic				
15	Sporadic				
18	Sporadic				
19	Sporadic				
21	Sporadic				
6	BRCA1				
13	BRCA2				
14	BRCA2				
11	BRCA2				
12	BRCA2				

descriptors / Filtered log ratio / Gene annotations / Gene identifiers /

Find Genes Correlated with Survival

This procedure tests for genes which are significantly associated with survival.

Experimental design:

Status column:
(0 = censored, 1 = death)

Column defining survival time:

□ Average over replicates of:

Find gene lists determined by:

● Significance threshold of univariate tests: 0.001

○ Restriction on proportion of false discoveries:
Maximum proportion of false discoveries: 0.1
Confidence level (between 0 and 100%): 80

○ Restriction on number of false discoveries:
Maximum number of false discoveries: 10
Confidence level (between 0 and 100%): 80

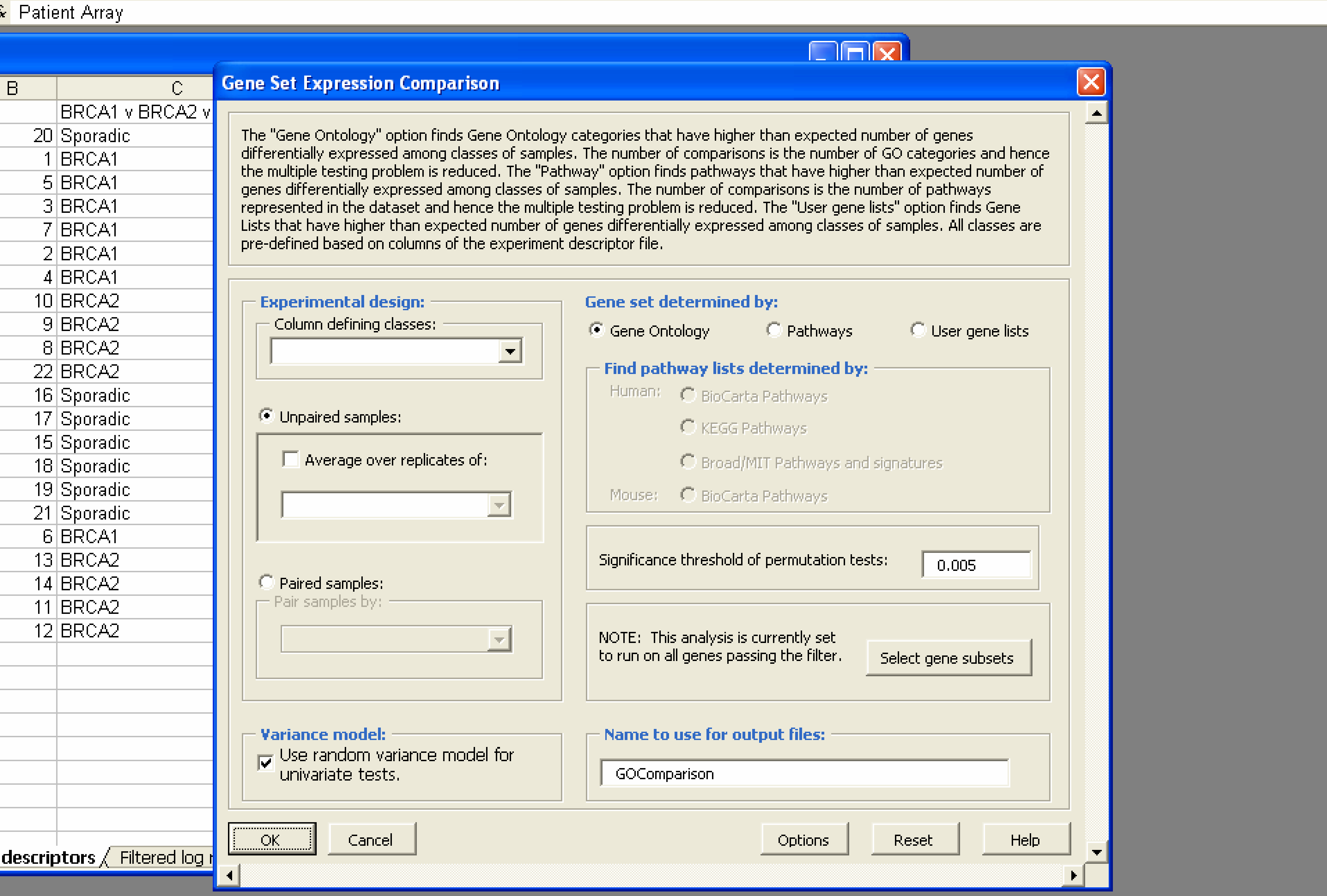
NOTE: This analysis is currently set to run on all genes passing the filter.

Gene Set Expression Comparison

- Compute p value of differential expression for each gene in a gene set (k=number of genes)
- Compute a summary (S) of these p values
- Determine whether the value of the summary test statistic S is more extreme than would be expected from a random sample of k genes (probe-sets) on that platform
- Two types of summaries provided
  - Average of log p values
  - Kolmogorov-Smirnov statistic; largest distance between the cumulative distribution of the p values and the uniform distribution expected if none of the genes were differentially expressed

Gene Set Expression Comparison

- p value for significance of summary statistic need not be as extreme as .001 usually, because the number of gene sets analyzed is usually much less than the number of individual genes analyzed
- Conclusions of significance are for gene sets in this tool, not for individual genes



Patient Array

B	C
	BRCA1 v BRCA2 v
20	Sporadic
1	BRCA1
5	BRCA1
3	BRCA1
7	BRCA1
2	BRCA1
4	BRCA1
10	BRCA2
9	BRCA2
8	BRCA2
22	BRCA2
16	Sporadic
17	Sporadic
15	Sporadic
18	Sporadic
19	Sporadic
21	Sporadic
6	BRCA1
13	BRCA2
14	BRCA2
11	BRCA2
12	BRCA2

descriptors / Filtered log r

Gene Set Expression Comparison

The "Gene Ontology" option finds Gene Ontology categories that have higher than expected number of genes differentially expressed among classes of samples. The number of comparisons is the number of GO categories and hence the multiple testing problem is reduced. The "Pathway" option finds pathways that have higher than expected number of genes differentially expressed among classes of samples. The number of comparisons is the number of pathways represented in the dataset and hence the multiple testing problem is reduced. The "User gene lists" option finds Gene Lists that have higher than expected number of genes differentially expressed among classes of samples. All classes are pre-defined based on columns of the experiment descriptor file.

Experimental design:

Column defining classes:

● Unpaired samples:

□ Average over replicates of:

○ Paired samples:

Pair samples by:

Gene set determined by:

● Gene Ontology ○ Pathways ○ User gene lists

Find pathway lists determined by:

Human: ○ BioCarta Pathways

○ KEGG Pathways

○ Broad/MIT Pathways and signatures

Mouse: ○ BioCarta Pathways

Significance threshold of permutation tests: 0.005

NOTE: This analysis is currently set to run on all genes passing the filter.

Variance model:

☑ Use random variance model for univariate tests.

Name to use for output files:

GOComparison

Comparison of Gene Set Expression Comparison to O/E Analysis in Class Comparison

- Gene set expression tool is based on all genes in a set, not just on those significant at some threshold value
- O/E analysis does not provide statistical significance for gene sets

Fallacy of Clustering Classes Based on Selected Genes

- Even for arrays randomly distributed between classes, genes will be found that are “significantly” differentially expressed
- With 8000 genes measured, 400 false positives will be differentially expressed with $p < 0.05$
- Arrays in the two classes will necessarily cluster separately when using a distance measure based on genes selected to distinguish the classes

Class Prediction

- Predict membership of a specimen into pre-defined classes
  - Disease vs normal
  - Poor vs good response to treatment
  - Long vs short survival

Traditional Approach for Marker Development

- Focus on candidate protein involved in disease pathogenesis
- Develop assay
- Conduct retrospective study of whether marker is prognostic using available specimens
- Marker dies because
  - Therapeutic relevance not established
  - Inter-laboratory reproducibility not established

Genomic Approach to Diagnostic/Prognostic Marker Development

- Select therapeutically relevant population
  - Node negative well staged breast cancer patients who have not received chemotherapy and have long follow-up
  - Early stage ovarian cancer patients and normal controls
- Perform genome wide expression profiling
- Develop multi-gene/protein predictor of outcome
- Obtain unbiased estimate of prediction accuracy
- Independently confirm results

Limitations of Genomic Approach

- Difficulty relating differentially expressed genes to cause of disease
  - or to real therapeutic targets
- Availibility of tissue and clinical follow-up for therapeutically relevant questions
  - Many studies address overly simple problems or heterogeneous non-therapeutically relevant populations
  - Inclusion of advance disease patients
  - Comparing completely different types of cancer

Limitations of Genomic Approach

- Difficulty in performing adequate validation studies
- Lack of inter-laboratory reproducibility evaluations

Class Prediction Model

- Given a sample with an expression profile vector *x* of log-ratios or log signals and unknown class.
- Predict which class the sample belongs to
- The class prediction model is a function *f* which maps from the set of vectors *x* to the set of class labels {1,2} (if there are two classes).
- *f* generally utilizes only some of the components of *x* (i.e. only some of the genes)
- Specifying the model *f* involves specifying some parameters (e.g. regression coefficients) by fitting the model to the data (*learning* the data).

Components of Class Prediction

- Feature (gene) selection
  - Which genes will be included in the model
- Select model type
  - E.g. DLDA, Nearest-Neighbor, …
- Fitting parameters (regression coefficients) for model

Class Prediction Paradigm

- Select features (F) to be included in predictive model using training data in which class membership of the samples is known
- Fit predictive model containing features F using training data
  - Diagonal linear discriminant analysis
  - Neural network
- Evaluate predictive accuracy of model on completely independent data not used in any way for development of the model

Feature Selection

- Key component of supervised analysis
- Genes that are differentially expressed among the classes at a significance level α (e.g. 0.01)
  - The α level is selected to control the number of genes in the model, not to control the false discovery rate
    - Methods for class prediction are different than those for class comparison
  - The accuracy of the significance test used for feature selection is not of major importance as identifying differentially expressed genes is not the ultimate objective
  - For survival prediction, the genes with significant univariate Cox PH regression coefficients

Feature Selection

- Small subset of genes which together give most accurate predictions
  - Step-up regression
  - Combinatorial optimization algorithms
    - Genetic algorithms
- Principal components of genes
- Gene cluster averages

Linear Classifiers for Two Classes

$$l(\underline{x}) = \sum_{i \varepsilon F} w_i x_i$$

$\underline{x} =$ vector of log ratios or log signals

$F =$ features (genes) included in model

$w_i =$ weight for i'th feature

decision boundary $l(\underline{x})$ > or < d

Linear Classifiers for Two Classes

- Compound covariate predictor

$$w_i \propto \frac{\overline{x}_i^{(1)} - \overline{x}_i^{(2)}}{\hat{\sigma}_i}$$

Instead of for DLDA

$$w_i \propto \frac{\overline{x}_i^{(1)} - \overline{x}_i^{(2)}}{\hat{\sigma}_i^2}$$

Linear Classifiers for Two Classes

- Support vector machines with inner product kernel are linear classifiers with weights determined to minimize errors
- Perceptrons with principal components as input are linear classifiers with no well defined criterion for defining weights

Advantages of Simple Linear Classifiers

- Do not over-fit data
  - Incorporate influence of multiple variables without attempting to select the best small subset of variables
  - Do not attempt to model the multivariate interactions among the predictors and outcome

Evaluating a Classifier

- “Prediction is difficult, especially the future.”
  - – Neils Bohr

- Fit of a model to the same data used to develop it is no evidence of prediction accuracy for independent data.

Split-Sample Evaluation

- Training-set
  - Used to select features, select model type, determine parameters and cut-off thresholds
- Test-set
  - Withheld until a single model is fully specified using the training-set.
  - Fully specified model is applied to the expression profiles in the test-set to predict class labels.
  - Number of errors is counted

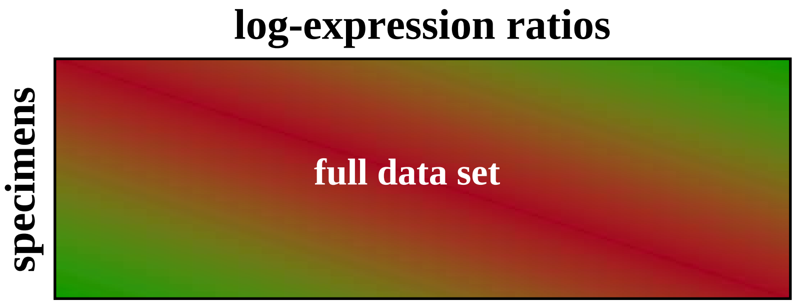
Split-Sample Evaluation

- Used for Rosenwald et al. study of prognosis in DLBL lymphoma.
  - 200 cases training-set
  - 100 cases test-set

Leave-one-out Cross Validation

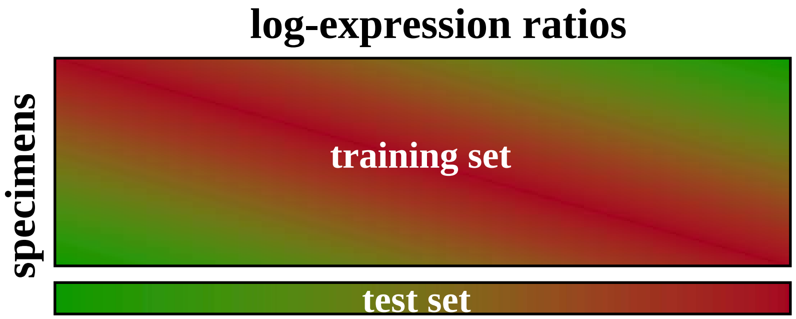
- Leave-one-out cross-validation simulates the process of separately developing a model on one set of data and predicting for a test set of data not used in developing the model

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 **from scratch** 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.

Cross-validated Misclassification Rate of Any Multivariable Classifier

- Omit sample 1
  - – Develop multivariate classifier from scratch on training set with sample 1 omitted
  - – Predict class for sample 1 and record whether prediction is correct

Cross-validated Misclassification Rate of Any Multivariate Classifier

- Repeat analysis for training sets with each single sample omitted one at a time
- e = number of misclassifications determined by cross-validation

- Cross validation is only valid if the training set is not used in any way in the development of the model. Using the complete set of samples to select genes violates this assumption and invalidates cross-validation.
- With proper cross-validation, the model must be developed from scratch for each leave-one-out training set. This means that gene selection must be repeated for each leave-one-out training set.
- The cross-validated estimate of misclassification error applies to the model building process, not to the particular model or the particular set of genes used in the model.

Prediction on Simulated Null Data

Generation of Gene Expression Profiles

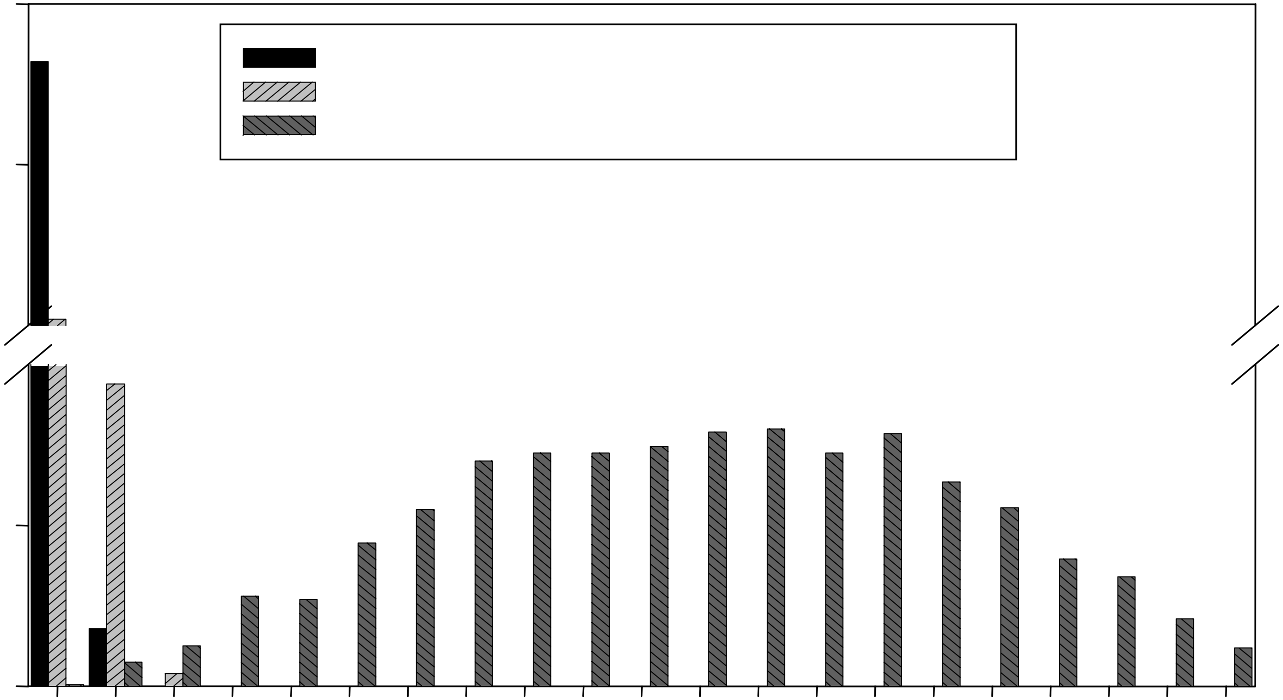
- 14 specimens (P_i is the expression profile for specimen *i*)
- Log-ratio measurements on 6000 genes
- $P_i \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}_{\mathbf{6000}})$
- Can we distinguish between the first 7 specimens (Class 1) and the last 7 (Class 2)?

Prediction Method

- Compound covariate prediction (*discussed later*)
- Compound covariate built from the log-ratios of the 10 most differentially expressed genes.

Percentage of simulated data sets with *m* or fewer misclassifications

m	Non-cross-validated class prediction	Cross-validated class prediction
0	99.85	0.60
1	100.00	2.70
2	100.00	6.20
3	100.00	11.20
4	100.00	16.90
5	100.00	24.25
6	100.00	34.00
7	100.00	42.55
8	100.00	53.85
9	100.00	63.60
10	100.00	74.55
11	100.00	83.50
12	100.00	91.15
13	100.00	96.85
14	100.00	100.00



Incomplete (incorrect) Cross-Validation

- Biologists and computer scientists are using all the data to select genes and then cross-validating only the parameter estimation (learning) component of model development
  - Highly biased
  - Many published complex methods which make strong claims based on incorrect cross-validation.
    - Frequently seen in complex feature set selection algorithms
    - Also seen in proposals for decision tree classifiers and neural networks

Compound covariate predictor

- Feature selection
  - Select genes with two-class t-statistics significant at $p<p^*$
- Form a compound covariate predictor as:

$\sum_i t_i x_i$ $\left\{ \begin{array}{l} \textbf{where } t_i \textbf{ = t-statistic, } x_i \textbf{ = log-ratio,} \\ \textbf{and sum is taken over all significant genes} \end{array} \right.$

- Determine the cutpoint of the predictor as the midpoint between its mean in one class and its mean in the other class

Advantages of Compound Covariate Classifier

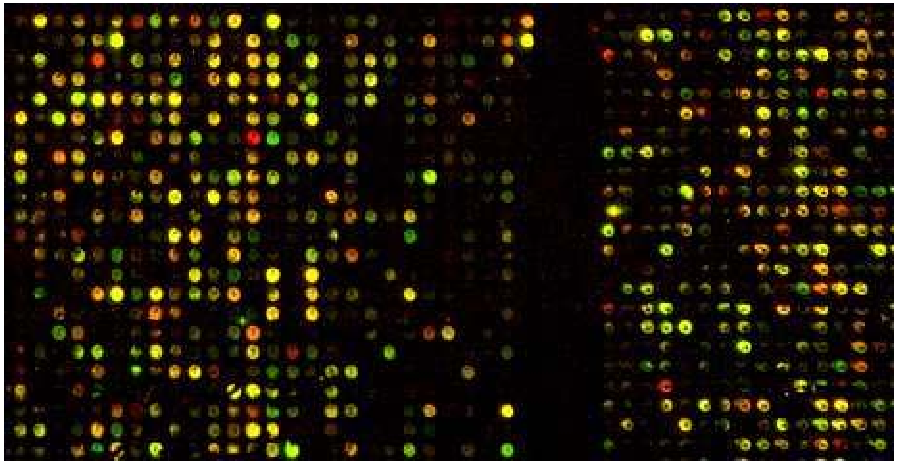
- Does not over-fit data
  - Incorporates influence of multiple variables without attempting to select the best small subset of variables
  - Does not attempt to model the multivariate interactions among the predictors and outcome
  - A one-dimensional classifier with contributions from variables correlated with outcome

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

<u>RESEARCH QUESTION</u>

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

Classification of hereditary breast cancers with the compound covariate predictor

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

BRCA1

α_g	# of significant genes	m = # of misclassified elements (misclassified samples)	% of random permutations with m or fewer misclassifications
10^{-2}	182	3 (s13714, s14510, s14321)	0.4
10^{-3}	53	2 (s14510, s14321)	1.0
10^{-4}	9	1 (s14321)	0.2

BRCA2

α_g	# of significant genes	m = # of misclassified elements (misclassified samples)	% of random permutations with m or fewer misclassifications
10^{-2}	212	4 (s11900, s14486, s14572, s14324)	0.8
10^{-3}	49	3 (s11900, s14486, s14324)	2.2
10^{-4}	11	4 (s11900, s14486, s14616, s14324)	6.6

Permutation Distribution of Cross-validated Misclassification Rate of a Multivariate Classifier

- Randomly permute class labels and repeat the entire cross-validation
- Re-do for all (or 1000) random permutations of class labels
- Permutation p value is fraction of random permutations that gave as few misclassifications as e in the real data

Exact Permutation Test

Premise: Under the null hypothesis of no systematic difference in expression profiles between the two classes, it can be assumed that assignment of class labels to expression profiles is purely coincidental.

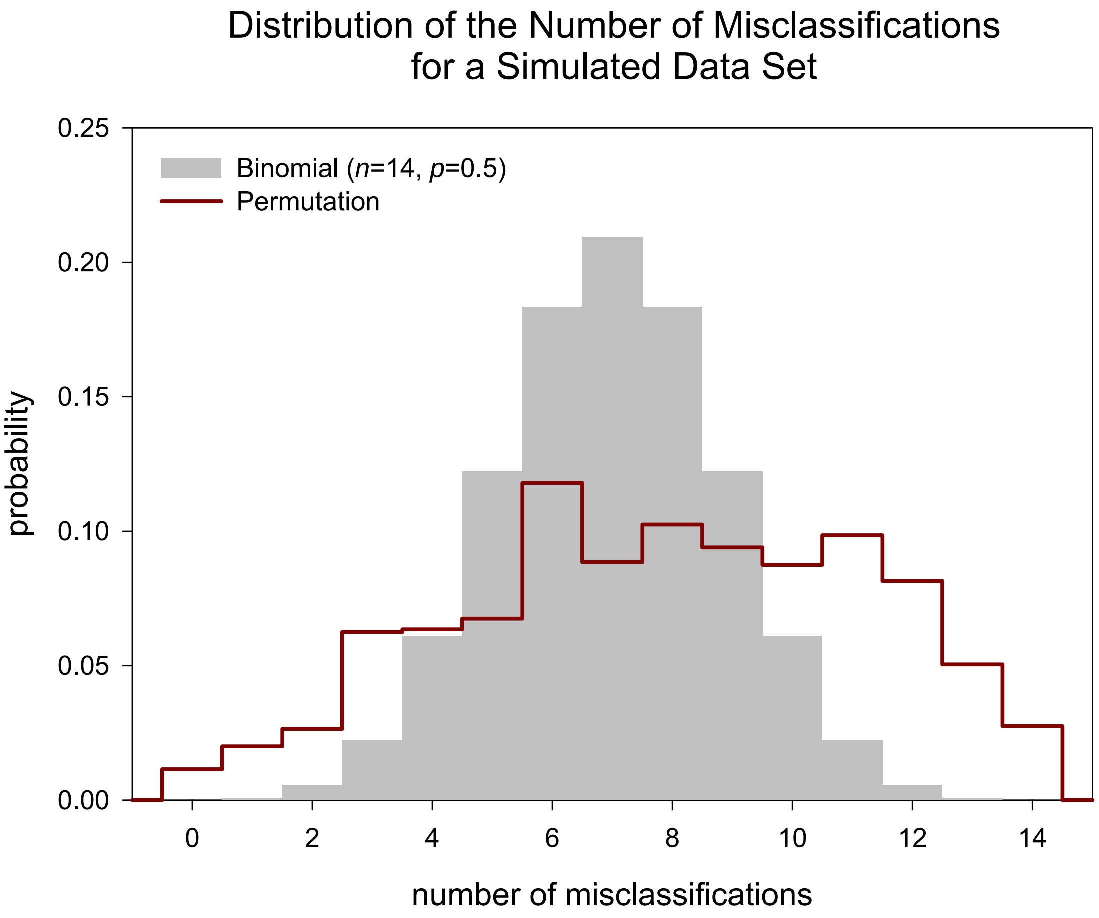
Performing the test

1. Consider every possible permutation of the class labels among the gene expression profiles.
2. Determine the proportion of the permutations that result in a misclassification error rate less than or equal to the observed error rate.
3. This proportion is the achieved significance level in a test of the null hypothesis.

Monte Carlo Permutation Test

Examining all permutations is computationally burdensome. Instead, a Monte Carlo method is used…

- n_{perm} permutations of the labels are randomly generated.
- The proportion of these permutations that have m or fewer misclassifications is an estimate of the achieved significance level in a test of the null hypothesis.
- n_{perm} is chosen such that the variability in the estimate is less than an acceptable level.
- If the true proportion of permutations with $m \leq 2$ is 0.05, $n_{\text{perm}}= 2000$ ensures the coefficient of variation of the estimate of the achieved significance level is less than 0.1.


Distribution of the Number of Misclassifications
for a Simulated Data Set
Binomial (n=14, p=0.5)
Permutation
probability
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12
14
number of misclassifications

Compound Covariate Classification of DLCL Data. (GCC vs Activated B) 42 Samples, 2317 Genes

Nominal alpha	Number of DEGs	Number of misclassifications	Permutational p value
0.01	275	3	0.00
0.001	97	5	0.00
0.0001	39	4	0.00
0.00001	16	4	0.00
0.000001	3	7	0.01

Compound Covariate Classifier
CLL Mutational Status
18 Samples

Nominal Alpha	Number of DEGs	X-validation Errors	Permutation p value	Misclassific's in 10 new samples
0.001	56	1	0.001	1
0.0001	7	5	0.107	1

Quadratic Discriminant Analysis

- Assumes that log-ratios (log intensities) have a multi-variate Gaussian distribution.
- The two classes have different mean vectors and potentially different covariance matrices.
- Using the training data, estimate the mean vector and covariance matrix for each class.

Quadratic Discriminant Analysis

- To classify a new sample, compute the probability density for the log-ratio expression profile of the new sample for each class. Compute these two values using the class-specific mean vectors and covariance matrices estimated in the training data. The computation also utilizes the Gaussian distribution assumption.
- Classify the new sample in the class with the larger value of the probability density for the expression profile of the new sample.

Quadratic Discriminant Analysis

- With G genes in the model, there are G components of the mean vector to be estimated for each class and G(G+1)/2 components of the covariance matrix for each class. Hence a total of G(G+3) parameters to be estimated.
- With N samples, one has only NG pieces of data.

Diagonal Linear Discriminant Analysis

- Full QDA performs poorly when G >N. One can help somewhat by selecting the G genes to include based on univariate discrimination power.
- The number of parameters can be dramatically reduced by assuming that the variances are the same in the two classes and that covariance among genes can be ignored. This reduces the number of parameters to 3G. This is DLDA. It has performed as well as much more complex methods in comparisons conducted by Dudoit et al.

Diagonal Linear Discriminant Analysis

- Golub’s Weighted Voting Method and Radmacher et al’s Compound Covariate Predictor are similar to DLDA.
- These methods, as well as other, are generally implemented with feature (gene) selection based on univariate classification power. In performing cross-validation to estimate mis-classification rate, the gene selection step **must** be repeated starting with the full set of genes for each leave-one-out training set.

Neural Network Classification

Kahn et al. Nature Med. 2001

- Not really a neural network (fortunately).
- A perceptron with no hidden nodes and a linear transfer function at each node.
- Inputs are first 10 principal components
  - The linear combinations of the genes that have greatest variation among samples and are orthogonal
- The method is essentially equivalent to DLDA based on the 10 PC's as predictors
- They didn't cross-validate the computation of the 10 PC's.

Linear Classifiers for Two Classes

- Support vector machines with inner product kernel are linear classifiers with weights determined to separate the classes with a hyperplain that minimizes the length of the weight vector

Support Vector Machine

minimize $\sum_{\mathrm{i}} \mathrm{w}_{\mathrm{i}}^{2}$

subject to $y_{j}\left(\underline{w}'\underline{x}^{(j)}+b\right)\geq 1$

where $\mathrm{y}_{\mathrm{j}}=\pm 1$ for class 1 or 2.

Compound Covariate Bayes Classifier

- Compound covariate $y = \Sigma t_i x_i$
  - Sum over the genes selected as differentially expressed
  - x_i the expression level of the ith selected gene for the case whose class is to be predicted
  - t_i the t statistic for testing differential expression for the i’th gene
- Proceed as for the naïve Bayes classifier but using the single compound covariate as predictive variable
  - GW Wright et al. PNAS 2005.

Other Simple Methods

- Nearest neighbor classification
- Nearest k-neighbors
- Nearest centroid classification
- Shrunken centroid classification

Nearest Neighbor Classifier

- To classify a sample in the validation set as being in outcome class 1 or outcome class 2, determine which sample in the training set it's gene expression profile is most similar to.
  - Similarity measure used is based on genes selected as being univariately differentially expressed between the classes
  - Correlation similarity or Euclidean distance generally used
- Classify the sample as being in the same class as it's nearest neighbor in the training set

K-Nearest Neighbor Classifier

- Find the k samples that are most similar to the sample to be classified
- Identify the majority class among the k nearest neighbor samples
- Classify the unknown sample as being in the majority class

Nearest Centroid Classifier

- For a training set of data, select the genes that are informative for distinguishing the classes
- Compute the average expression profile (*centroid*) of the informative genes in each class
- Classify a sample in the validation set based on which centroid in the training set it’s gene expression profile is most similar to.

Nearest Shunken Centroids

$\overline{x}_{ik} =$ mean of gene i in class k

centroid = vector of class means

$\overline{\mathrm{x}}_{\mathrm{i}} =$ overall mean of gene i

$$c = \frac{\overline{x}_{ik} - \overline{\mathrm{x}}_{\mathrm{i}}}{m_k s_i}$$

$s_i =$ standard deviation of gene i expression within each class

$$\mathrm{m}_{\mathrm{k}} = \sqrt{\tfrac{1}{n_k} + \tfrac{1}{n}}$$

$$\overline{x}'_{ik} = \overline{\mathrm{x}}_{\mathrm{i}} + m_k s_i d'_{ik}$$

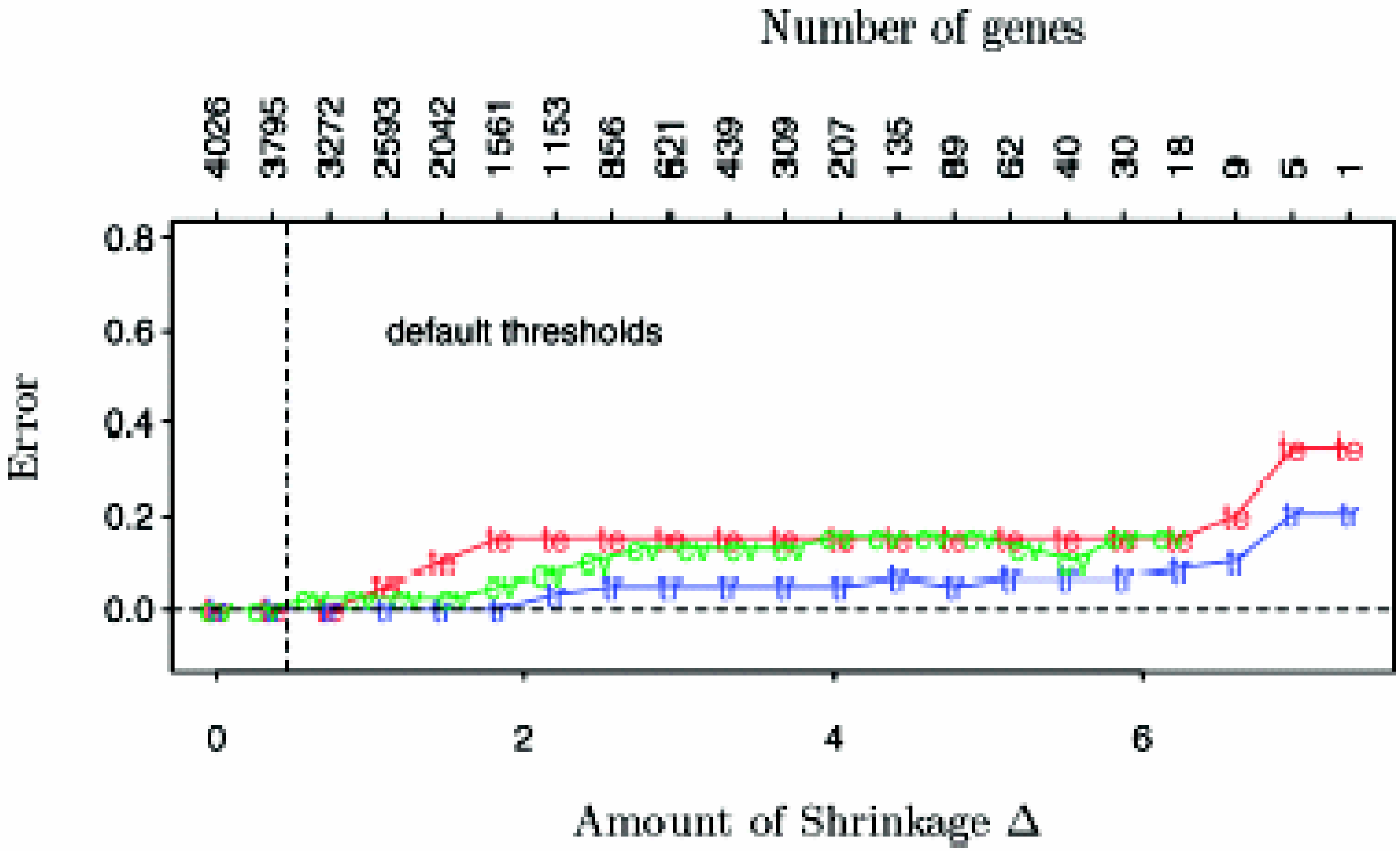
$$d'_{ik} = sign(d_{ik})\left\{|\, d_{ik}\,| - \Delta\right\}_+$$

Nearest Shunken Centroids Discriminant Score

$x^* =$ expression profile of sample to be classified

$$\delta_{\mathrm{k}}(x^*) = \sum_i \frac{\left(x_i^* - \overline{x}'_{ik}\right)^2}{s_i^2} - 2\log \pi_k$$

$\pi_k =$ prior probability that sample is in class k


Number of genes
4026
3795
3272
2593
2042
1561
1153
856
621
439
309
207
135
89
62
40
30
18
9
5
1
0.8
0.6
0.4
0.2
0.0
Error
default thresholds
te
cv
tr
0
2
4
6
Amount of Shrinkage Δ

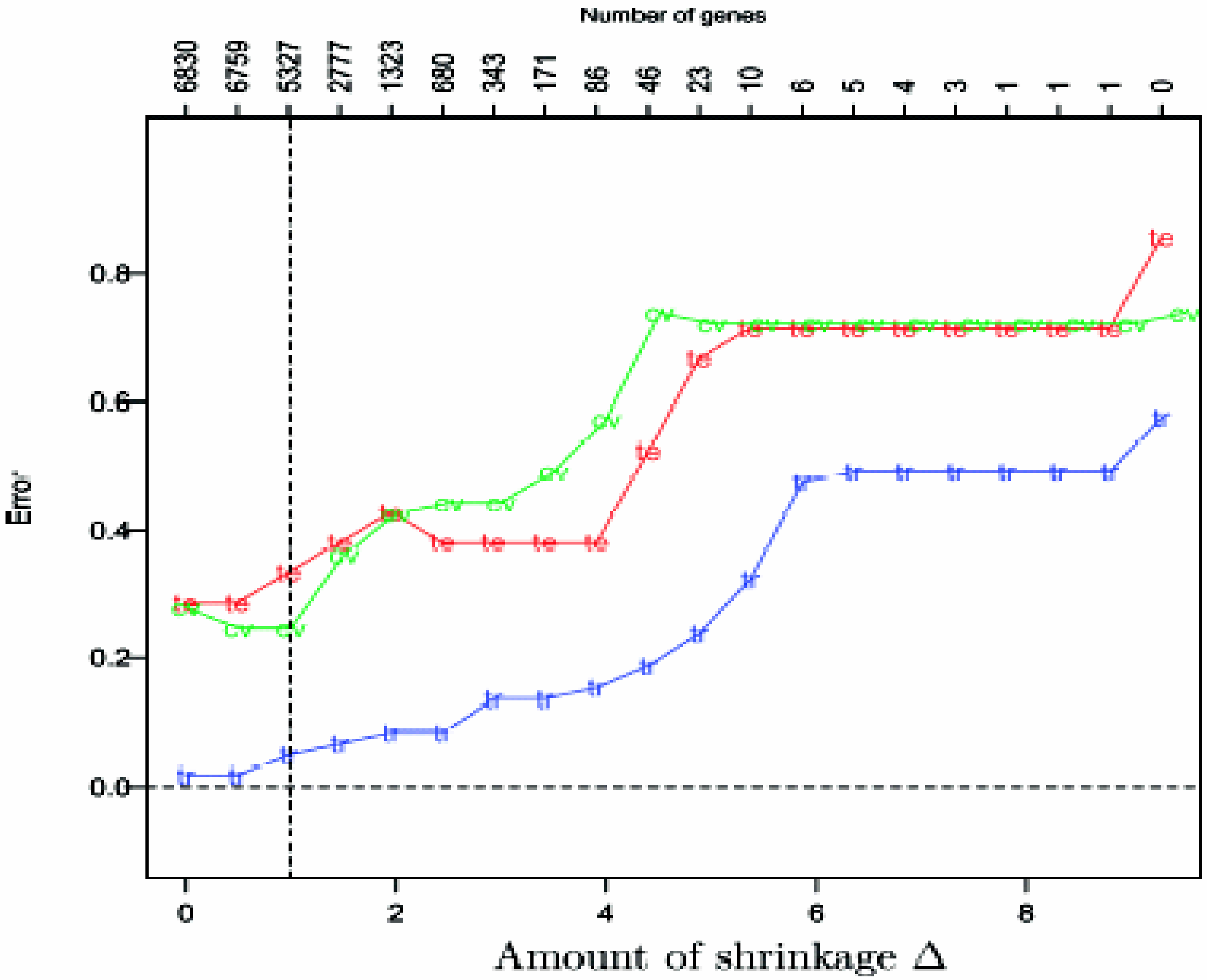


FIG. 11. *NCI cancer cell lines: training, cross-validation and test error curves.*

Other Methods

- Top-scoring pairs
  - Claim that it gives accurate prediction with few pairs because pairs of genes are selected to work well together
- Random Forrest
  - Very popular in machine learning community
  - Complex classifier

When There Are More Than 2 Classes

- Nearest neighbor type methods

- Decision tree of binary classifiers

Decision Tree of Binary Classifiers

- Partition the set of classes {1,2,…,K} into two disjoint subsets S_1 and S_2
- Develop a binary classifier for distinguishing the composite classes S_1 and S_2
  - Compute the cross-validated classification error for distinguishing S_1 and S_2
- Repeat the above steps for all possible partitions in order to find the partition $S_1$and S_2 for which the cross-validated classification error is minimized
- If $S_1$and S_2 are not singleton sets, then repeat all of the above steps separately for the classes in $S_1$and S_2 to optimally partition each of them

Myth

- That complex classification algorithms such as neural networks perform better than simpler methods for class prediction.

Truth

- Artificial intelligence sells to journal reviewers and peers who cannot distinguish hype from substance when it comes to microarray data analysis.
- Comparative studies have shown that simpler methods work as well or better for microarray problems because the number of candidate predictors exceeds the number of samples by orders of magnitude.

When p>>n
The Linear Model is Too Complex

- It is always possible to find a set of features and a weight vector for which the classification error on the training set is zero.
- Why consider more complex models?

Classification of BRCA2 Germline Mutations

Classification Method	Correct Prediction with LOO-CV
Compound Covariate Predictor	86%
Fisher LDA	64%
Diagonal LDA	86%
1-Nearest Neighbor	91%
3-Nearest Neighbor	77%
Support Vector Machine (linear kernel)	82%
Classification Tree	55%

Dataset		Linear and quadratic discriminant analysis				Classification trees				Nearest neighbors
		FLDA	DLDA	Golub	DQDA	CV	Bag	Boost	CPD	
Leukemia	Median	3	**0**	1	1	3	2	1	1	1
($K = 2$, $p = 40$, $n_{TS} = 24$)	Upper quartile	4	**1**	2	2	4	2	2	2	1
Leukemia	Median	3	1	NA	1	1	1	1	1	**1**
($K = 3$, $p = 40$, $n_{TS} = 24$)	Upper quartile	4	2	NA	2	3	2	2	2	**1**
Lymphoma	Median	6	1	NA	**0**	2	2	1	1	**0**
($K = 3$, $p = 50$, $n_{TS} = 27$)	Upper quartile	8	1	NA	**1**	3	3	3	3	**1**
NCI 60	Median	11	**7**	NA	9	12	10	9	9	8
($K = 8$, $p = 30$, $n_{TS} = 21$)	Upper quartile	11	**8**	NA	10	13	11	11	10	10

Comparison of discrimination methods

Speed et al

In this field many people are inventing new methods of classification or using quite complex ones (e.g. SVMs). Is this necessary?

We did a study comparing several methods on three publicly available tumor data sets: the Leukemia data set, the Lymphoma data set, and the NIH 60 tumor cell line data, as well as some unpublished data sets.

We compared NN, FLDA, DLDA, DQDA and CART, the last with or without aggregation (bagging or boosting).

The results were unequivocal: simplest is best!

Approaches to Intra-study Validation

- Split data into training set and validation set
  - Validation set not accessed until proposed classification system is fully specified based on training set

- Algorithmic cross-validation or bootstrap

Invalid Criticisms of Cross-Validation

- “You can always find a set of features that will provide perfect prediction for the training and test sets.”
  - For complex models, there may be many sets of features that provide zero training errors.
  - A modeling strategy that either selects among those sets or aggregates among those models, will have a generalization error which will be validly estimated by cross-validation.

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Prediction Error Estimation: A Comparison of Resampling Methods

Annette M. Molinaro[a,b,*], Richard Simon[c], Ruth M. Pfeiffer[a]

[a]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, NCI, NIH, Rockville, MD 20852, [b]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, [c]Biometric Research Branch, Division of Cancer Treatment and Diagnostics, NCI, NIH, Rockville, MD 20852

ABSTRACT

Motivation: In genomic studies, thousands of features are collected on relatively few samples. One of the goals of these studies is to build classifiers to predict the outcome of future observations. There are three inherent steps to this process: feature selection, model selection, and prediction assessment. With a focus on prediction assessment, we compare several methods for estimating the 'true' prediction error of a prediction model in the presence of feature selection.

Results: For small studies where features are selected from thousands of candidates, the resubstitution and simple split-sample estimates are seriously biased. In these small samples, leave-one-out (LOOCV), 10-fold cross-validation (CV), and the .632+ bootstrap have the smallest bias for diagonal discriminant analysis, nearest neighbor, and classification trees. LOOCV and 10-fold CV have the smallest bias for linear discriminant analysis. Additionally, LOOCV, 5- and 10-fold CV, and the .632+ bootstrap have the lowest mean square error. The .632+ bootstrap is quite biased in small sample sizes with strong signal to noise ratios. Differences in performance among resampling methods are reduced as the number of specimens available increase.

Availability: A complete compilation of results in tables and figures is available in Molinaro *et al.* (2005). R code for simulations and analyses is available from the authors.

Contact: annette.molinaro@yale.edu

1 INTRODUCTION

In genomic experiments one frequently encounters high dimensional data and small sample sizes. Microarrays simultaneously monitor expression levels for several thousands of genes. Proteomic profiling studies using SELDI-TOF (surface-enhanced laser desorption and ionization time-of-flight) measure size and charge of proteins and protein fragments by mass spectroscopy, and result in up to 15,000 intensity levels at prespecified mass values for each spectrum. Sample sizes in such experiments are typically less than 100.

In many studies observations are known to belong to predetermined classes and the task is to build predictors or classifiers for new observations whose class is unknown. Deciding which genes or proteomic measurements to include in the prediction is called *feature selection* and is a crucial step in developing a class predictor. Including too many noisy variables reduces accuracy of the prediction and may lead to over-fitting of data, resulting in promising but often non-reproducible results (Ransohoff, 2004).

Another difficulty is model selection with numerous classification models available. An important step in reporting results is assessing the chosen model's error rate, or generalizability. In the absence of independent validation data, a common approach to estimating predictive accuracy is based on some form of resampling the original data, e.g., cross-validation. These techniques divide the data into a learning set and a test set and range in complexity from the popular learning-test split to v-fold cross-validation, Monte-Carlo v-fold cross-validation, and bootstrap resampling. Few comparisons of standard resampling methods have been performed to date, and all of them exhibit limitations that make their conclusions inapplicable to most genomic settings. Early comparisons of resampling techniques in the literature are focussed on model selection as opposed to prediction error estimation (Breiman and Spector, 1992; Burman, 1989). In two recent assessments of resampling techniques for error estimation (Braga-Neto and Dougherty, 2004; Efron, 2004), feature selection was not included as part of the resampling procedures, causing the conclusions to be inappropriate for the high-dimensional setting.

We have performed an extensive comparison of resampling methods to estimate prediction error using simulated (large signal to noise ratio), microarray (intermediate signal to noise ratio) and proteomic data (low signal to noise ratio), encompassing increasing sample sizes with large numbers of features. The impact of feature selection on the performance of various cross validation methods is highlighted. The results elucidate the 'best' resampling techniques for

[*] to whom correspondence should be addressed

Table 1. *Prediction Error Estimates.* The estimate $\hat{\theta}_n$ (col 4) and standard deviation (col 5) based on learning sample of size 40. The estimate $\tilde{\theta}_n$ (rows 1-4) and standard deviation based on the remaining 260 observations. Bias (col 6) and MSE (col 7) reported for each resampling technique (col 1) and algorithm (col 3). The ten genes with largest t-statistics used in algorithms. Minimums in bold.

Estimator	p	Algorithm	Est	St.Dev	Bias	MSE
$\tilde{\theta}_n$	0.87	LDA	0.078	0.093		
		DDA	0.160	0.086		
		NN	0.042	0.084		
		CART	0.121	0.133		
v-fold CV	0.5	LDA	0.357	0.126	0.279	0.097
		DDA	0.342	0.106	0.182	0.052
		NN	0.277	0.135	0.235	0.077
		CART	0.430	0.121	0.309	0.134
	0.2	LDA	0.161	0.127	0.083	0.017
		DDA	0.208	0.086	0.048	0.012
		NN	0.108	0.102	0.066	0.011
		CART	0.284	0.117	0.163	0.055
	0.1	LDA	0.118	0.120	0.040	**0.008**
		DDA	0.177	0.087	0.017	**0.007**
		NN	0.078	0.102	0.036	**0.005**
		CART	0.189	0.104	0.068	0.024
LOOCV	0.025	LDA	0.092	0.115	**0.014**	**0.008**
		DDA	0.164	0.096	**0.004**	**0.007**
		NN	0.058	0.103	**0.016**	**0.005**
		CART	0.146	0.125	**0.025**	**0.018**
SPLIT	0.333	LDA	0.205	0.184	0.127	0.053
		DDA	0.243	0.138	0.083	0.034
		NN	0.145	0.169	0.103	0.044
		CART	0.371	0.174	0.25	0.121
	0.5	LDA	0.348	0.185	0.270	0.113
		DDA	0.344	0.139	0.184	0.062
		NN	0.265	0.177	0.223	0.086
		CART	0.438	0.155	0.317	0.147
.632+ 50 reps	≈ .368	LDA	0.274	**0.084**	0.196	0.047
		DDA	0.286	**0.074**	0.126	0.028
		NN	0.200	**0.070**	0.158	0.032
		CART	0.387	**0.080**	0.266	0.100

Simulated Data
40 cases, 10 genes selected from 5000

Method	Estimate	Std Deviation
True	.078	
Resubstitution	.007	.016
LOOCV	.092	.115
10-fold CV	.118	.120
5-fold CV	.161	.127
Split sample 1-1	.345	.185
Split sample 2-1	.205	.184
.632+ bootstrap	.274	.084

DLBCL Data

Method	Bias	Std Deviation	MSE
LOOCV	-.019	.072	.008
10-fold CV	-.007	.063	.006
5-fold CV	.004	.07	.007
Split 1-1	.037	.117	.018
Split 2-1	.001	.119	.017
.632+ bootstrap	-.006	.049	.004

Simulated Data
40 cases

Method	Estimate	Std Deviation
True	.078	
10-fold	.118	.120
Repeated 10-fold	.116	.109
5-fold	.161	.127
Repeated 5-fold	.159	.114
Split 1-1	.345	.185
Repeated split 1-1	.371	.065

Common Problems With Internal Classifier Validation

- Pre-selection of genes using entire dataset

- Failure to consider optimization of tuning parameter part of classification algorithm
  - Varma & Simon, BMC Bioinformatics 2006

- Erroneous use of predicted class in regression model

Incomplete (incorrect) Cross-Validation

- Let M(b,D) denote a classification model developed on a set of data D where the model is of a particular type that is parameterized by a scalar b.
- Use cross-validation to estimate the classification error of M(b,D) for a grid of values of b; Err(b).
- Select the value of b* that minimizes Err(b).
- Caution: Err(b*) is a biased estimate of the prediction error of M(b*,D).
- This error is made in some commonly used methods

Complete (correct) Cross-Validation

- Construct a learning set D as a subset of the full set S of cases.
- Use cross-validation restricted to D in order to estimate the classification error of M(b,D) for a grid of values of b; Err(b).
- Select the value of b* that minimizes Err(b).
- Use the mode M(b*,D) to predict for the cases in S but not in D (S-D) and compute the error rate in S-D
- Repeat this full procedure for different learning sets D_1 , D_2 and average the error rates of the models M(b_i*,D_i) over the corresponding validation sets S-D_i

Does an Expression Profile Classifier Predict More Accurately Than Standard Prognostic Variables?

- Not an issue of which variables are significant after adjusting for which others or which are *independent* predictors
  - Predictive accuracy and inference are different
- The two classifiers can be compared with regard to predictive accuracy
- The predictiveness of the expression profile classifier can be evaluated within levels of the classifier based on standard prognostic variables

Does an Expression Profile Classifier Predict More Accurately Than Standard Prognostic Variables?

- Some publications fit logistic model to standard covariates and the cross-validated predictions of expression profile classifiers

$$\log it(p) = \alpha + \beta y(\underline{x}_i \mid -i) + \gamma z_i$$

- This is valid only with split-sample analysis because the cross-validated predictions are not independent

External Validation

- Should address clinical utility, not just predictive accuracy
  - Therapeutic relevance
- Should incorporate all sources of variability likely to be seen in broad clinical application
  - Expression profile assay distributed over time and space
  - Real world tissue handling
  - Patients selected from different centers than those used for developing the classifier

Survival Risk Group Prediction

- Evaluate individual genes by fitting single variable proportional hazards regression models to log signal or log ratio for gene
- Select genes based on p-value threshold for single gene PH regressions
- Compute first k principal components of the selected genes
- Fit PH regression model with the k pc's as predictors. Let $b_1, \ldots, b_k$ denote the estimated regression coefficients
- To predict for case with expression profile vector x, compute the k supervised pc's $y_1, \ldots, y_k$ and the predictive index $\lambda = b_1 y_1 + \ldots + b_k y_k$

Survival Risk Group Prediction

- LOOCV loop:
  - Create training set by omitting i’th case
- Develop supervised pc PH model for training set
- Compute cross-validated predictive index for i’th case using PH model developed for training set
- Compute predictive risk percentile of predictive index for i’th case among predictive indices for cases in the training set

Survival Risk Group Prediction

- Plot Kaplan Meier survival curves for cases with cross-validated risk percentiles above 50% and for cases with cross-validated risk percentiles below 50%
  - Or for however many risk groups and thresholds is desired
- Compute log-rank statistic comparing the cross-validated Kaplan Meier curves

Survival Risk Group Prediction

- Repeat the entire procedure for all (or large number) of permutations of survival times and censoring indicators to generate the null distribution of the log-rank statistic
  - The usual chi-square null distribution is not valid because the cross-validated risk percentiles are correlated among cases
- Evaluate statistical significance of the association of survival and expression profiles by referring the log-rank statistic for the unpermuted data to the permutation null distribution

Survival Risk Group Prediction

- Other approaches to survival risk group prediction have been published
- The supervised pc method is implemented in BRB-ArrayTools
- BRB-ArrayTools also provides for comparing the risk group classifier based on expression profiles to one based on standard covariates and one based on a combination of both types of variables

BRB-ArrayTools
Class Prediction

- Classifiers
  - Compound covariate predictor
  - Diagonal LDA
  - K-Nearest Neighbor Classification
  - Nearest Centroid
  - Support Vector Machines
  - Random Forest Classifier
  - Shrunken Centroids (PAM)
  - Top Scoring Pairs
  - Binary Tree Classifier

BRB-ArrayTools
Class Prediction

- Validation
  - Split Sample
  - Leave one out cross validation
  - K-fold cross validation
  - Repeated K-fold cross validation
  - .632+ Bootstrap resampling

BRB-ArrayTools
Class Prediction

- Gene Selection
  - Re-done for each re-sampled training set
  - Univariate significance level less than specified threshold
    - Option for threshold for gene selection optimized by inner loop of cross-validation
  - Pairs of genes that work well together
  - Shrunken centroids

BRB-ArrayTools
Class Prediction

- Permutation test of significance of cross-validated misclassification rate
- Predictions for new patients

BRB-ArrayTools
Survival Risk Group Prediction

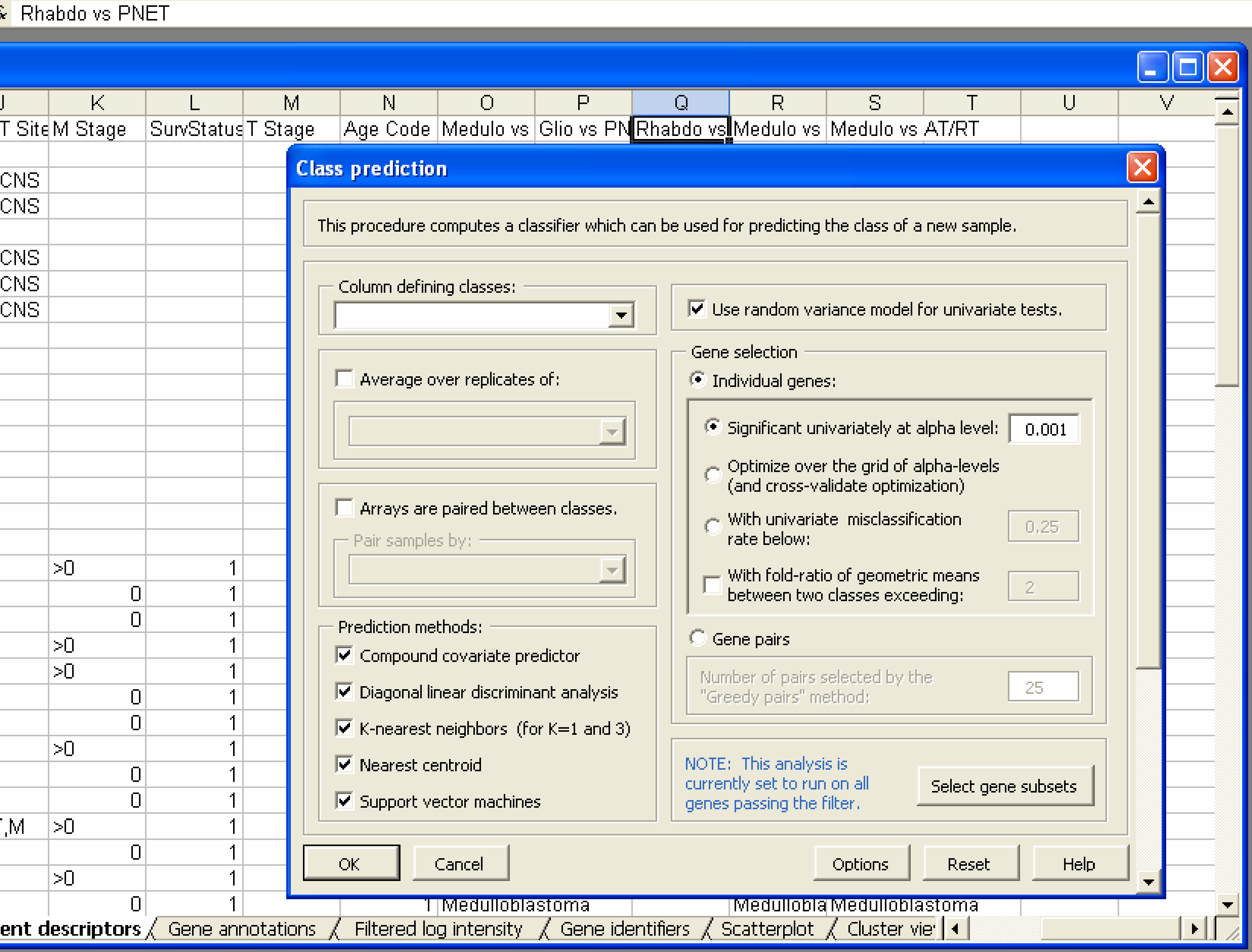
- No need to transform data to good vs bad outcome. Censored survival is directly analyzed
- Gene selection based on significance in univariate Cox Proportional Hazards regression
- Uses k principal components of selected genes
- Gene selection re-done for each resampled training set
- Develop k-variable Cox PH model for each leave-one-out training set

BRB-ArrayTools
Survival Risk Group Prediction

- Classify left out sample as above or below median risk based on model not involving that sample
- Repeat, leaving out 1 sample at a time to obtain cross-validated risk group predictions for all cases
- Compute Kaplan-Meier survival curves of the two predicted risk groups
- Permutation analysis to evaluate statistical significance of separation of K-M curves

BRB-ArrayTools
Survival Risk Group Prediction

- Compare Kaplan-Meier curves for gene expression based classifier to that for standard clinical classifier
- Develop classifier using standard clinical staging plus genes that add to standard staging



Rhabdo vs PNET
J K L M N O P Q R S T U V
T Site M Stage SurvStatus T Stage Age Code Medulo vs Glio vs PN Rhabdo vs Medulo vs Medulo vs AT/RT
Medulloblastoma Medullobla Medulloblastoma
ent descriptors / Gene annotations / Filtered log intensity / Gene identifiers / Scatterplot / Cluster vie

Class prediction

This procedure computes a classifier which can be used for predicting the class of a new sample.

Column defining classes:

□ Average over replicates of:

□ Arrays are paired between classes.

Pair samples by:

Prediction methods:

- ☑ Compound covariate predictor
- ☑ Diagonal linear discriminant analysis
- ☑ K-nearest neighbors (for K=1 and 3)
- ☑ Nearest centroid
- ☑ Support vector machines

☑ Use random variance model for univariate tests.

Gene selection

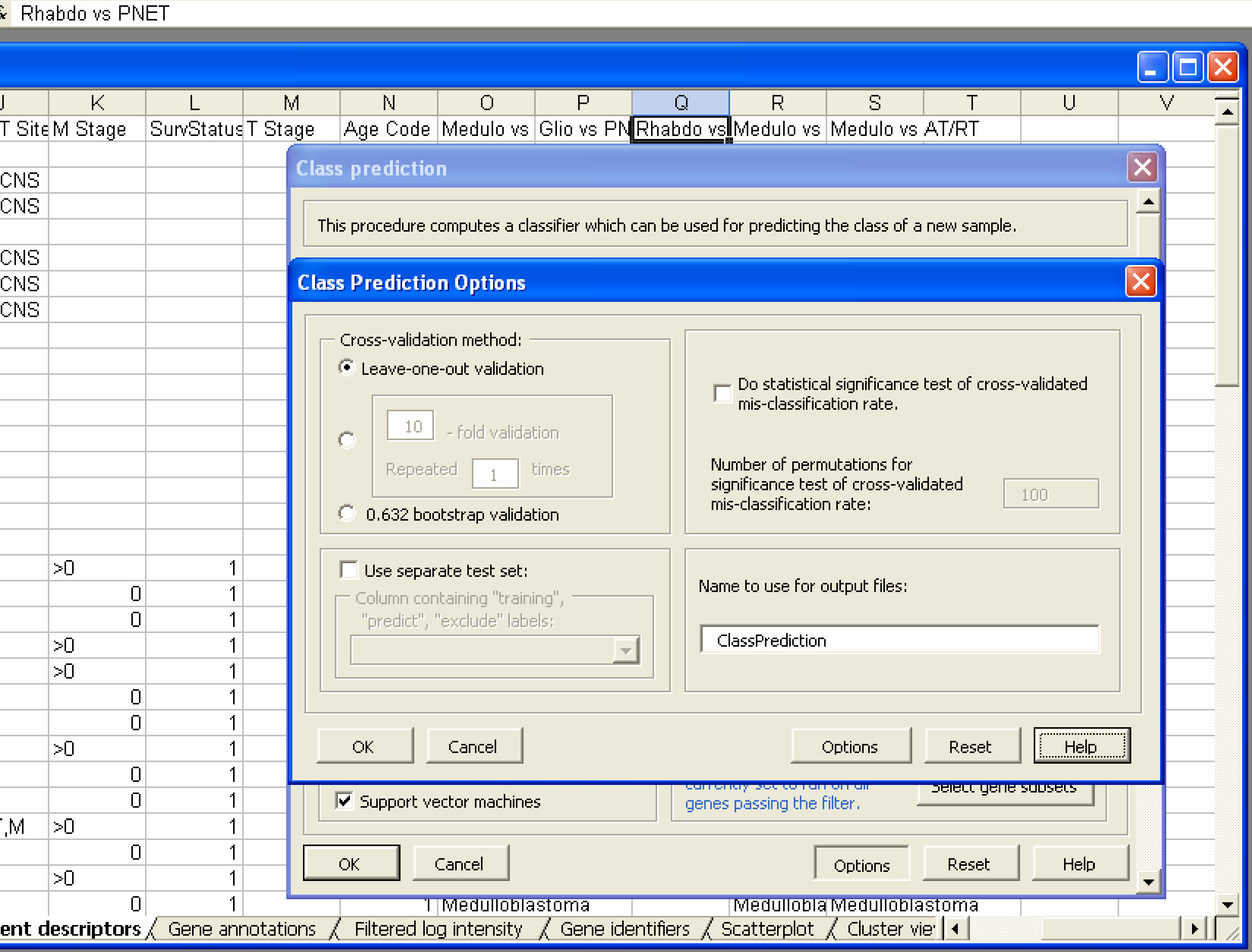
● Individual genes:

- ● Significant univariately at alpha level: 0.001
- ○ Optimize over the grid of alpha-levels (and cross-validate optimization)
- ○ With univariate misclassification rate below: 0.25
- □ With fold-ratio of geometric means between two classes exceeding: 2

○ Gene pairs

Number of pairs selected by the "Greedy pairs" method: 25

NOTE: This analysis is currently set to run on all genes passing the filter.



Rhabdo vs PNET

Class prediction

This procedure computes a classifier which can be used for predicting the class of a new sample.

Class Prediction Options

Cross-validation method:

● Leave-one-out validation

○ 10 - fold validation

Repeated 1 times

○ 0.632 bootstrap validation

□ Do statistical significance test of cross-validated mis-classification rate.

Number of permutations for significance test of cross-validated mis-classification rate: 100

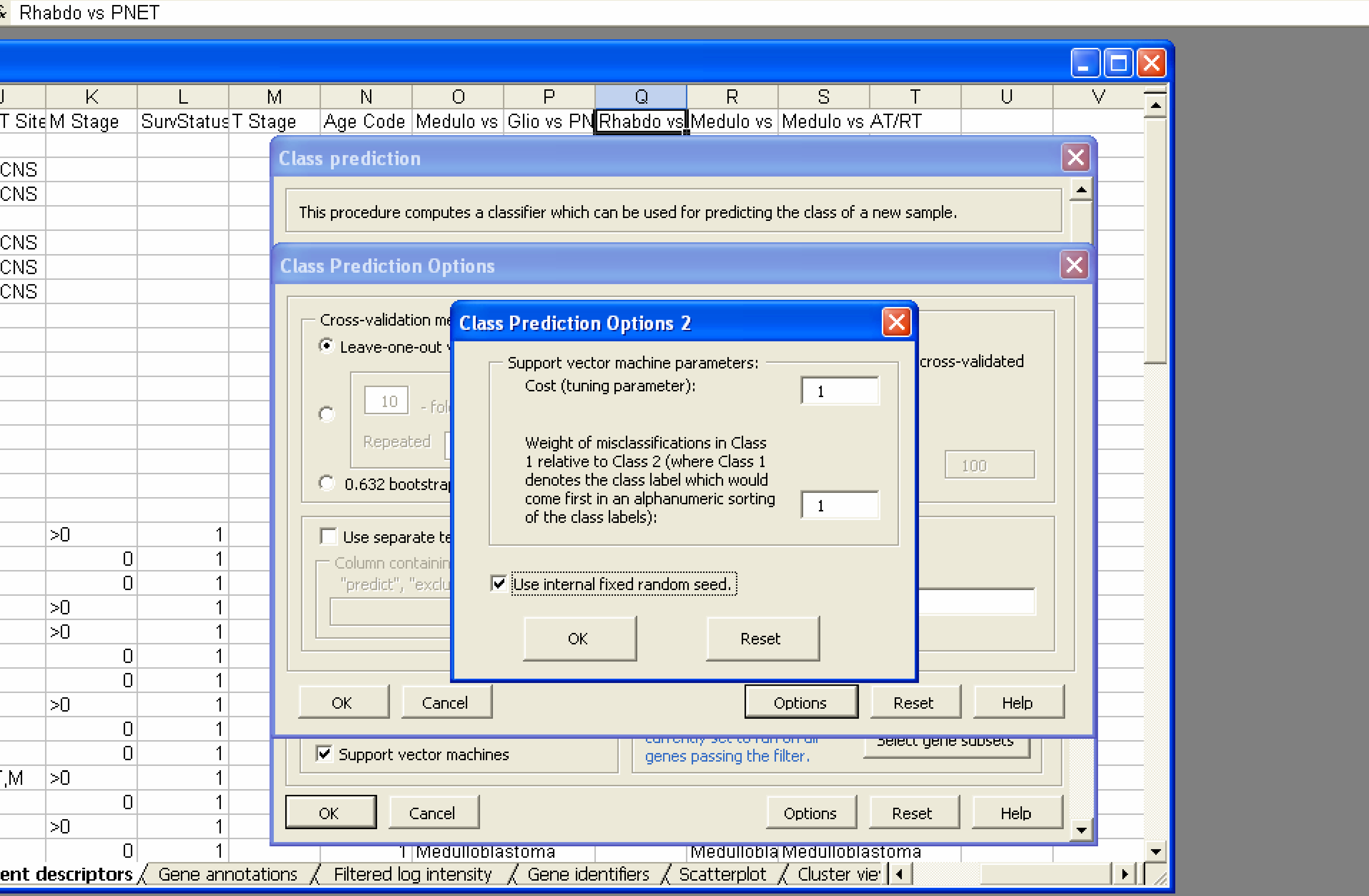
□ Use separate test set:

Column containing "training", "predict", "exclude" labels:

Name to use for output files: ClassPrediction

☑ Support vector machines

genes passing the filter.



Class prediction

This procedure computes a classifier which can be used for predicting the class of a new sample.

Class Prediction Options

Cross-validation me

● Leave-one-out

○ 10 - fol

Repeated

○ 0.632 bootstrap

cross-validated

100

□ Use separate te

Column containin "predict", "exclu

Class Prediction Options 2

Support vector machine parameters:

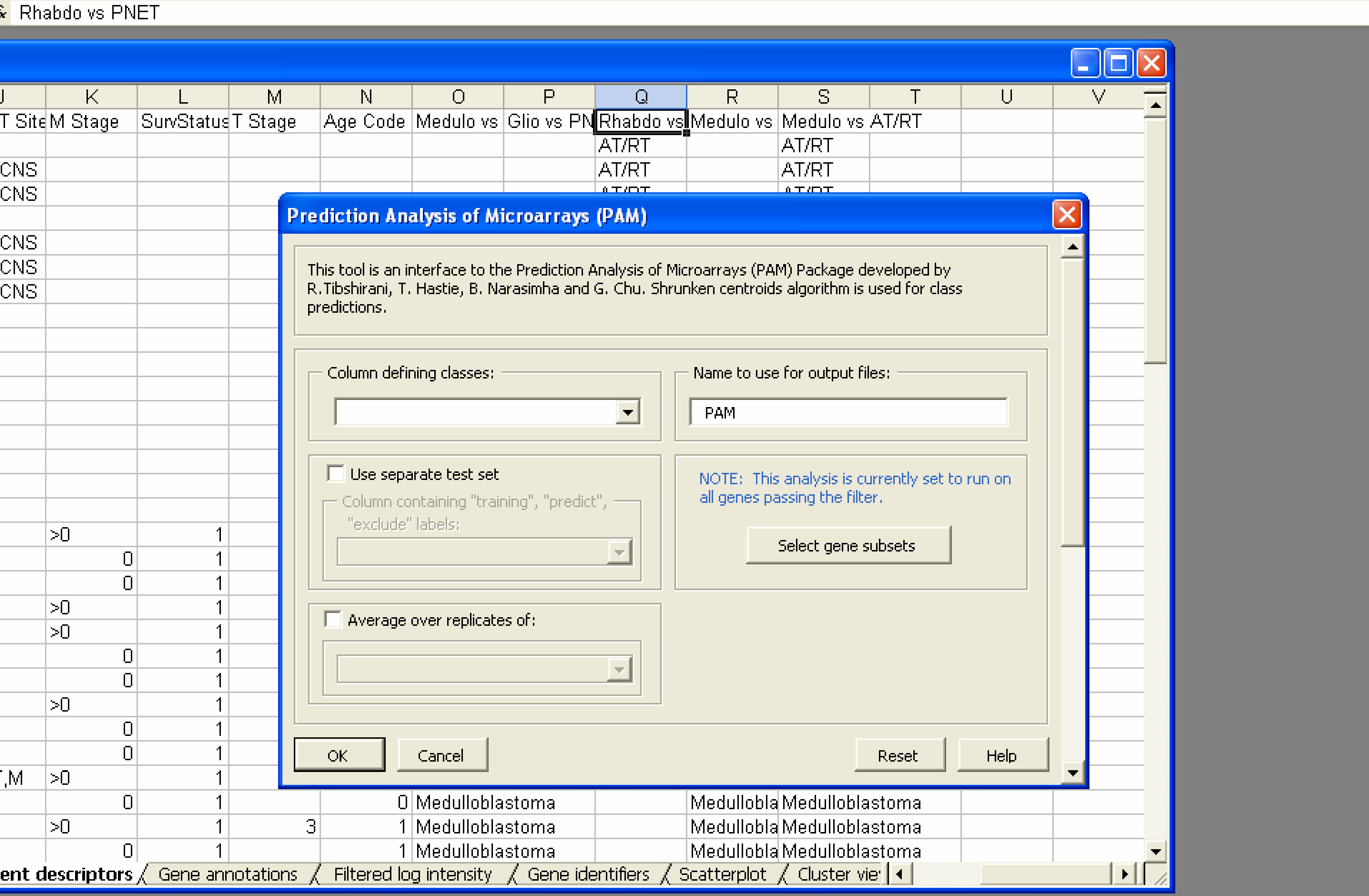
Cost (tuning parameter): 1

Weight of misclassifications in Class 1 relative to Class 2 (where Class 1 denotes the class label which would come first in an alphanumeric sorting of the class labels): 1

☑ Use internal fixed random seed.

☑ Support vector machines

currently set to run on all genes passing the filter.



Prediction Analysis of Microarrays (PAM)

This tool is an interface to the Prediction Analysis of Microarrays (PAM) Package developed by R.Tibshirani, T. Hastie, B. Narasimha and G. Chu. Shrunken centroids algorithm is used for class predictions.

Column defining classes:

Name to use for output files: PAM

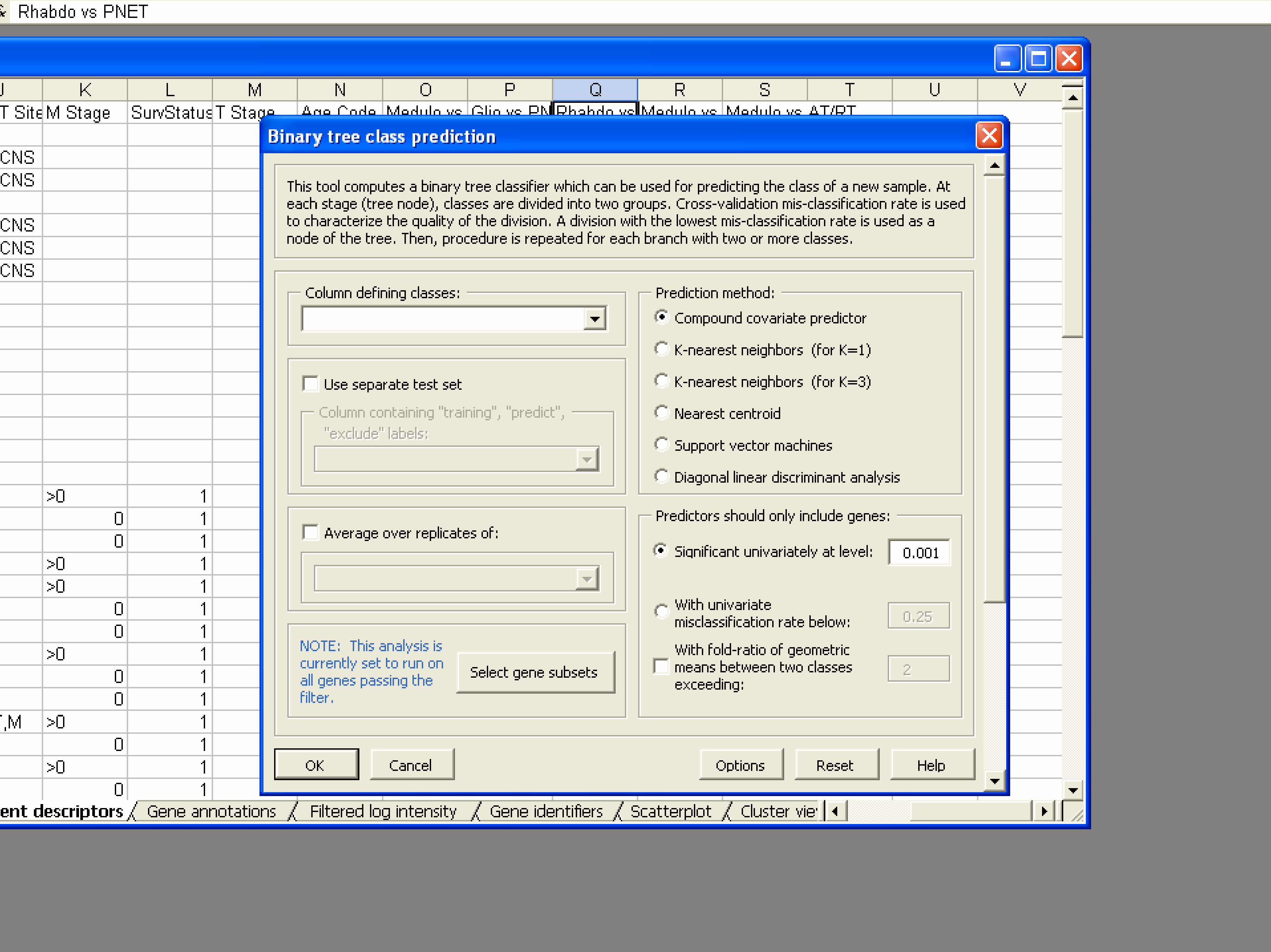
□ Use separate test set

Column containing "training", "predict", "exclude" labels:

NOTE: This analysis is currently set to run on all genes passing the filter.

Select gene subsets

□ Average over replicates of:


Rhabdo vs PNET
J K L M N O P Q R S T U V
T Site M Stage SurvStatus T Stage Age Code Medulo vs Glio vs PN Rhabdo vs Medulo vs Medulo vs AT/RT
ent descriptors
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie

Binary tree class prediction

This tool computes a binary tree classifier which can be used for predicting the class of a new sample. At each stage (tree node), classes are divided into two groups. Cross-validation mis-classification rate is used to characterize the quality of the division. A division with the lowest mis-classification rate is used as a node of the tree. Then, procedure is repeated for each branch with two or more classes.

Column defining classes:

□ Use separate test set

Column containing "training", "predict", "exclude" labels:

□ Average over replicates of:

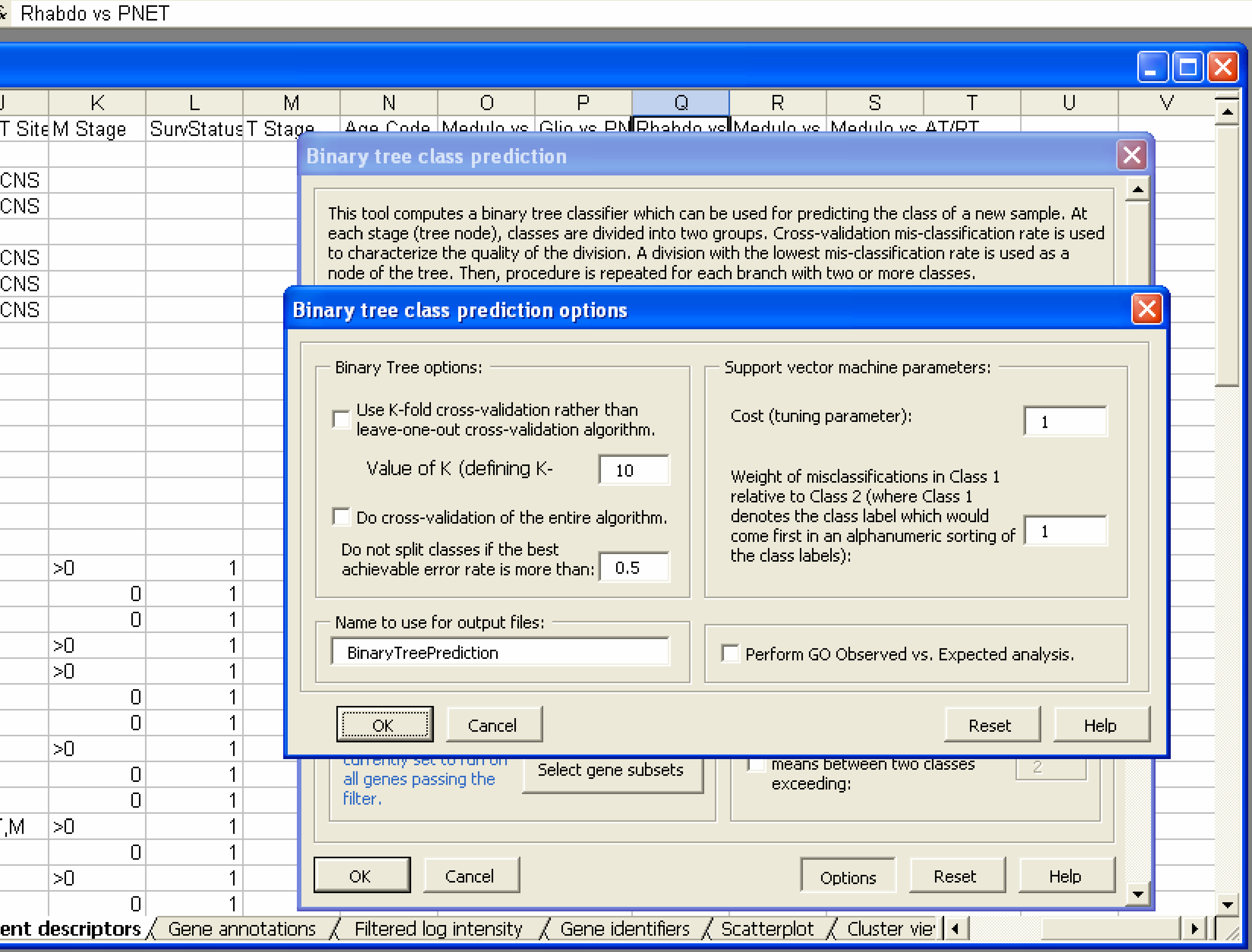
NOTE: This analysis is currently set to run on all genes passing the filter.

Prediction method:

- ● Compound covariate predictor
- ○ K-nearest neighbors (for K=1)
- ○ K-nearest neighbors (for K=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



Rhabdo vs PNET

J	K	L	M	N	O	P	Q	R	S	T	U	V
T Site	M Stage	SurvStatus	T Stage	Age Code	Medulo vs	Glio vs PN	Rhabdo vs	Medulo vs	Medulo vs	AT/RT		
CNS												
CNS												
CNS												
CNS												
CNS												
	>0	1										
	0	1										
	0	1										
	>0	1										
	>0	1										
	0	1										
	0	1										
	>0	1										
	0	1										
	0	1										
T,M	>0	1										
	0	1										
	>0	1										
	0	1										

ent descriptors / Gene annotations / Filtered log intensity / Gene identifiers / Scatterplot / Cluster vie

Binary tree class prediction

This tool computes a binary tree classifier which can be used for predicting the class of a new sample. At each stage (tree node), classes are divided into two groups. Cross-validation mis-classification rate is used to characterize the quality of the division. A division with the lowest mis-classification rate is used as a node of the tree. Then, procedure is repeated for each branch with two or more classes.

currently set to run on all genes passing the filter.

Select gene subsets

□ means between two classes exceeding: 2

OK | Cancel | Options | Reset | Help

Binary tree class prediction options

Binary Tree options:

□ Use K-fold cross-validation rather than leave-one-out cross-validation algorithm.

Value of K (defining K- 10

□ Do cross-validation of the entire algorithm.

Do not split classes if the best achievable error rate is more than: 0.5

Support vector machine parameters:

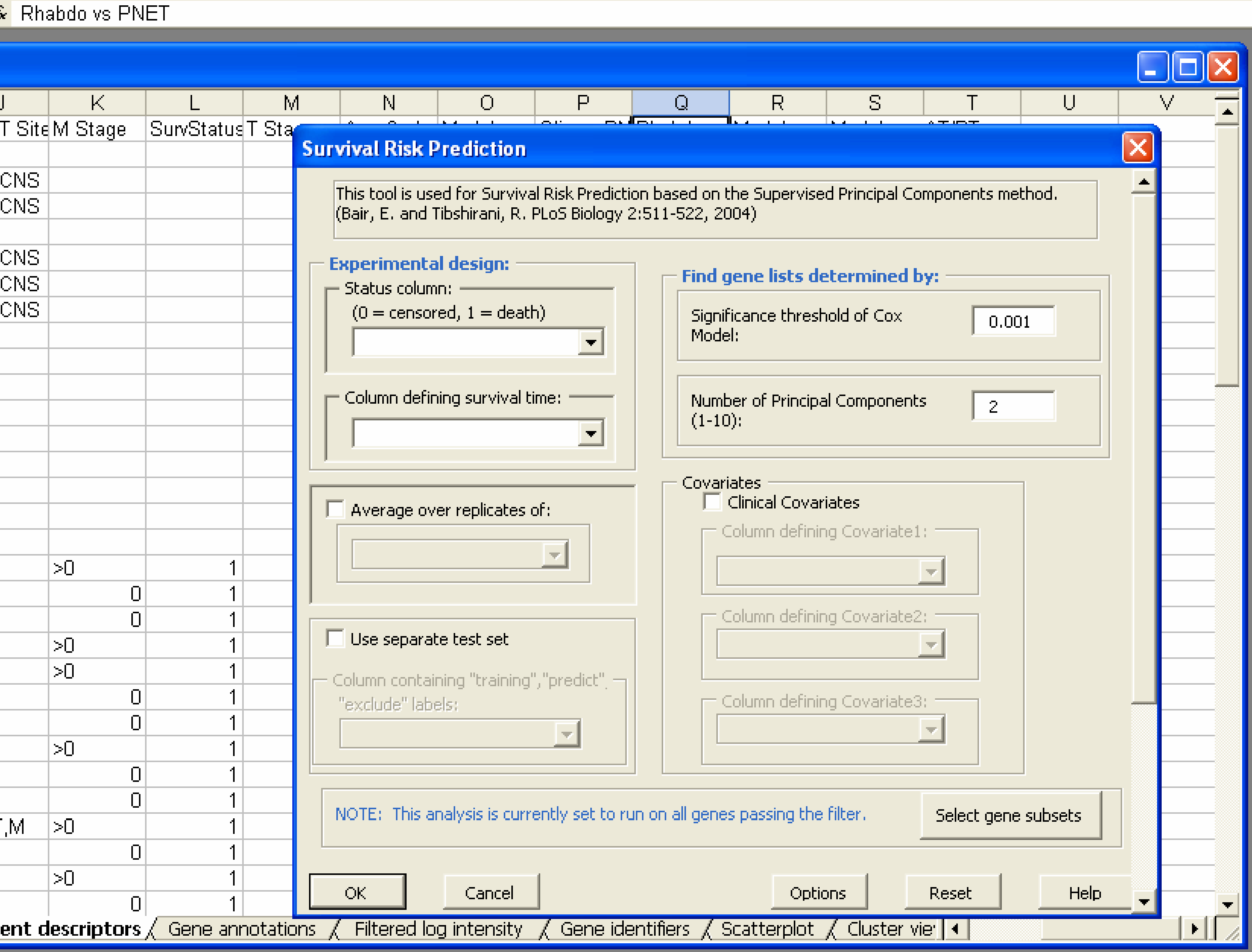
Cost (tuning parameter): 1

Weight of misclassifications in Class 1 relative to Class 2 (where Class 1 denotes the class label which would come first in an alphanumeric sorting of the class labels): 1

Name to use for output files: BinaryTreePrediction

□ Perform GO Observed vs. Expected analysis.

OK | Cancel | Reset | Help



Survival Risk Prediction

This tool is used for Survival Risk Prediction based on the Supervised Principal Components method. (Bair, E. and Tibshirani, R. PLoS Biology 2:511-522, 2004)

Experimental design:

Status column:
(0 = censored, 1 = death)

Column defining survival time:

□ Average over replicates of:

□ Use separate test set

Column containing "training","predict", "exclude" labels:

Find gene lists determined by:

Significance threshold of Cox Model: 0.001

Number of Principal Components (1-10): 2

Covariates

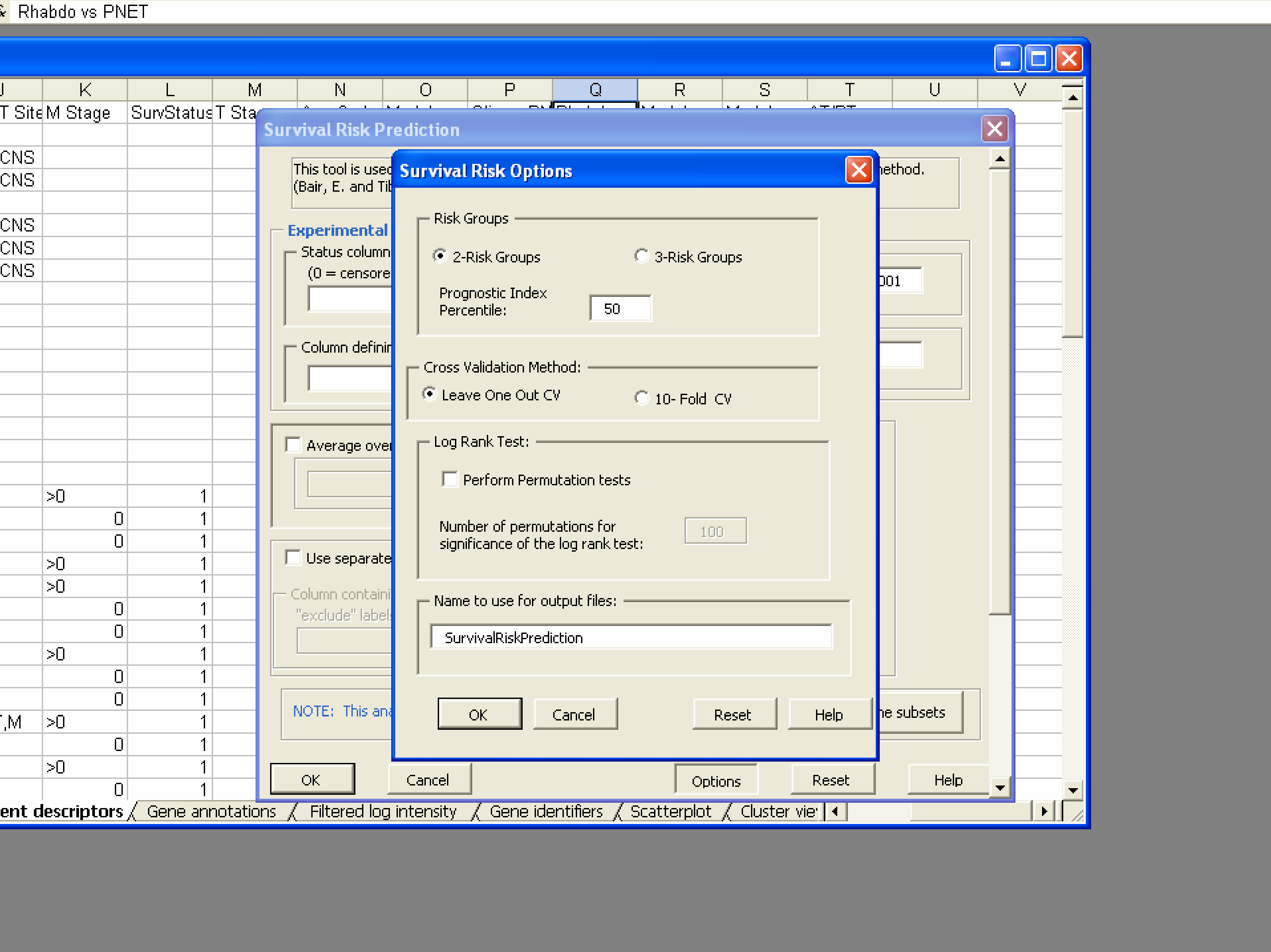
□ Clinical Covariates

Column defining Covariate1:

Column defining Covariate2:

Column defining Covariate3:

NOTE: This analysis is currently set to run on all genes passing the filter.



Rhabdo vs PNET

Survival Risk Prediction

This tool is used ... method.
(Bair, E. and Tib...

Experimental ...

Status column ...
(0 = censore...

Column defini...

□ Average over...

□ Use separate...

Column containi...
"exclude" label...

NOTE: This ana...

...he subsets

...001

Survival Risk Options

Risk Groups

● 2-Risk Groups ○ 3-Risk Groups

Prognostic Index Percentile: 50

Cross Validation Method:

● Leave One Out CV ○ 10- Fold CV

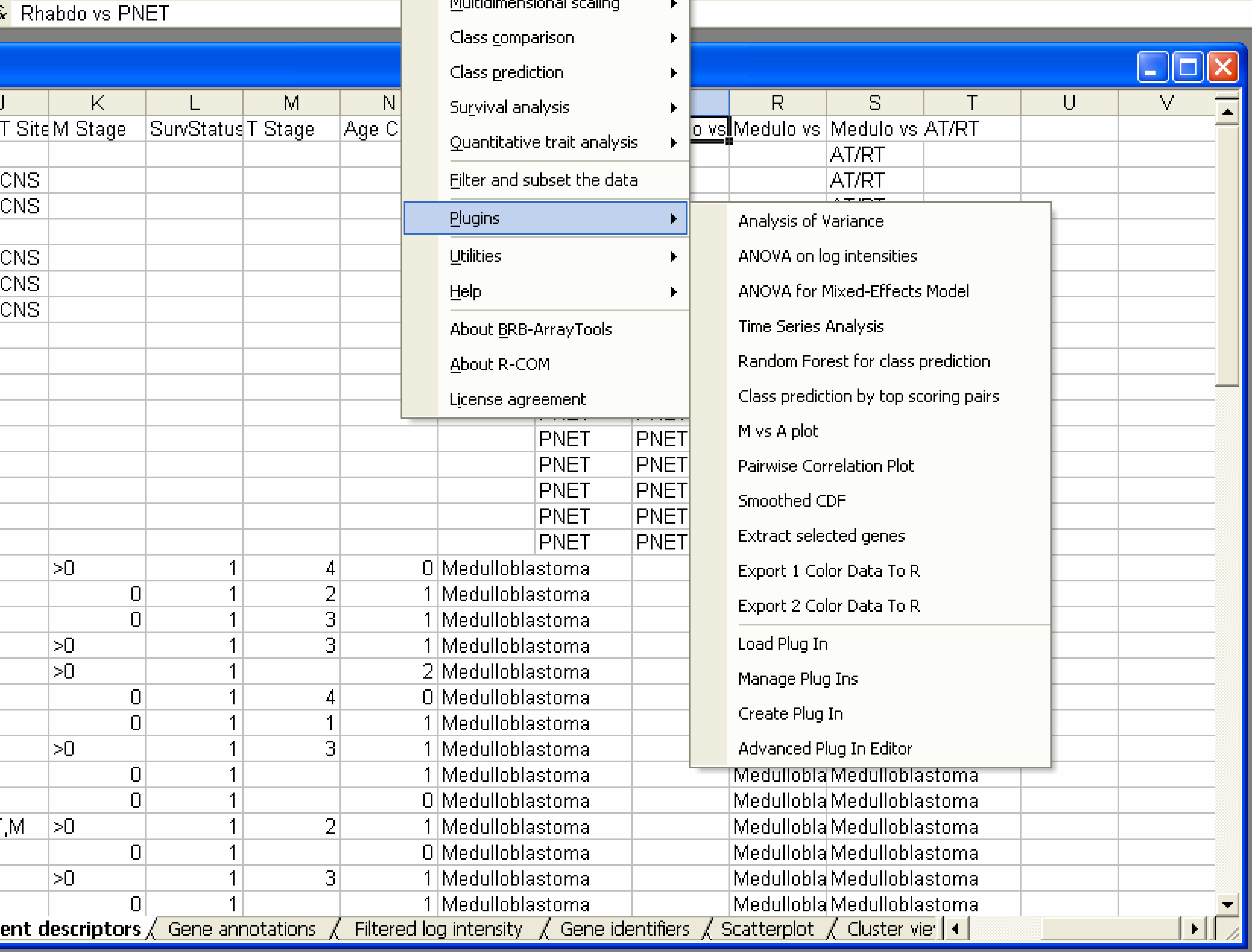
Log Rank Test:

□ Perform Permutation tests

Number of permutations for significance of the log rank test: 100

Name to use for output files:

SurvivalRiskPrediction



J	K	L	M	N				R	S	T	U	V
T Site	M Stage	SurvStatus	T Stage	Age C			o vs	Medulo vs	Medulo vs AT/RT			
									AT/RT			
CNS									AT/RT			
CNS									AT/RT			
CNS												
CNS												
CNS												
						PNET	PNET					
						PNET	PNET					
						PNET	PNET					
						PNET	PNET					
						PNET	PNET					
	>0	1	4	0	Medulloblastoma							
	0	1	2	1	Medulloblastoma							
	0	1	3	1	Medulloblastoma							
	>0	1	3	1	Medulloblastoma							
	>0	1		2	Medulloblastoma							
	0	1	4	0	Medulloblastoma							
	0	1	1	1	Medulloblastoma							
	>0	1	3	1	Medulloblastoma							
	0	1		1	Medulloblastoma			Medullobla	Medulloblastoma			
	0	1		0	Medulloblastoma			Medullobla	Medulloblastoma			
,M	>0	1	2	1	Medulloblastoma			Medullobla	Medulloblastoma			
	0	1		0	Medulloblastoma			Medullobla	Medulloblastoma			
	>0	1	3	1	Medulloblastoma			Medullobla	Medulloblastoma			
	0	1		1	Medulloblastoma			Medullobla	Medulloblastoma			

Class Prediction

- Cluster analysis is frequently used in publications for class prediction in a misleading way

Fallacy of Clustering Classes Based on Selected Genes

- Even for arrays randomly distributed between classes, genes will be found that are “significantly” differentially expressed
- With 10,000 genes measured, about 500 false positives will be differentially expressed with $p < 0.05$
- Arrays in the two classes will necessarily cluster separately when using a distance measure based on genes selected to distinguish the classes

Class Discovery

Two Types of “Classification”?

Class Discovery

- Identification of previously unknown classes of specimens
- Use of “unsupervised” methods
  - Hierarchical Clustering
  - *k*-means Clustering
  - SOMs
  - Others
- Prevalent method used in literature for analysis of gene expression data.

Class Prediction

- Assignment of specimens into known classes
- Use of “supervised” methods
  - CART
  - Discriminant Analysis
  - SVM
  - CCP
- Class prediction is more powerful than class discovery for distinguishing specimens based on *a priori* defined classes.

Class Discovery

- For determining whether a set of samples (eg tumors) is homogeneous with regard to expression profile
- To identify set of co-expressed, and perhaps co-regulated genes

Class Discovery of Samples

- Complex diseases often represent umbrella diagnoses and that heterogeneity limits the power of linkage and association studies
- Treatment selection and therapeutics development may be enhanced by biologically meaningful classification

Cluster Analysis

- Distance measure
  - Euclidean distance
  - Mahalanobis distance
  - 1- correlation
  - Mutual information
  - Bayes factor
- Feature (gene) set
  - All
  - Variably expressed
  - Selected to optimize clustering

- Algorithm
  - Hierarchical
  - K means
  - Self Organized Map
  - Generative Topographical Map
  - Autoclass
  - Bioclust
  - Gene shaving
  - Plaid
  - Splash
  - BiClustering

Clustering and Classification

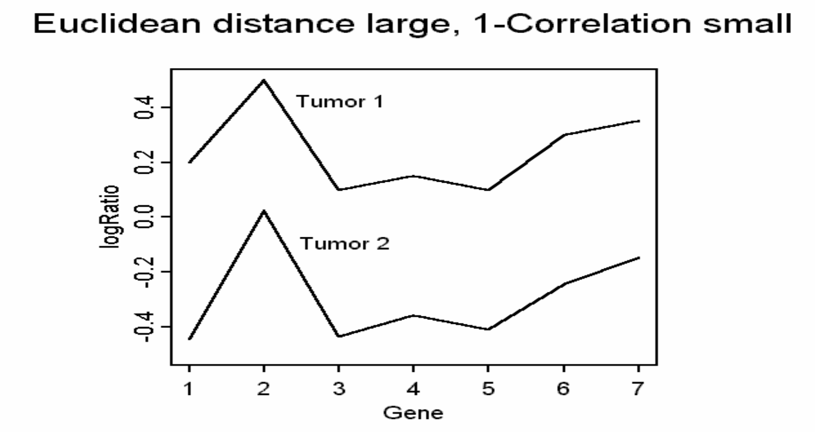
- Analysis performed on log-ratios with two channel arrays
- Analysis performed on log signal values for GeneChips

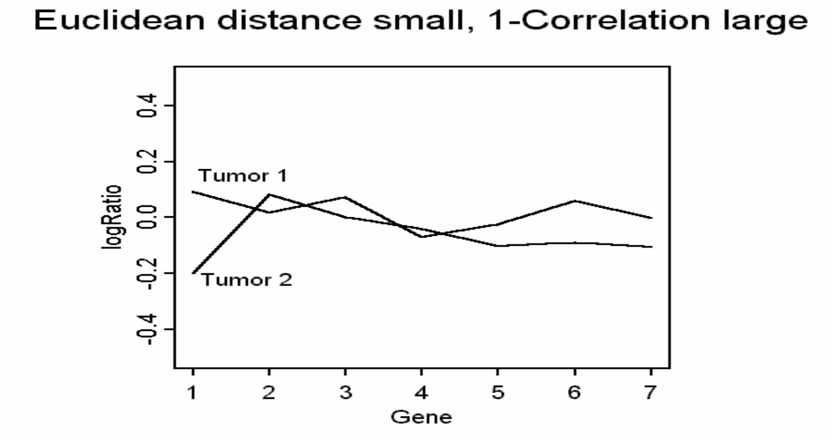
Hierarchical Agglomerative Clustering Algorithm

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

Common Distance Metrics for Hierarchical Clustering

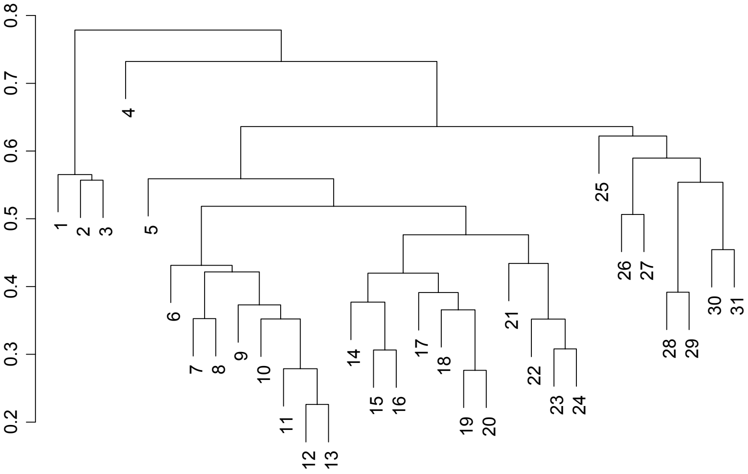
- Euclidean distance
  - Measures absolute distance (square root of sum of squared differences)
- 1-Correlation
  - Large values reflect lack of linear association (pattern dissimilarity)



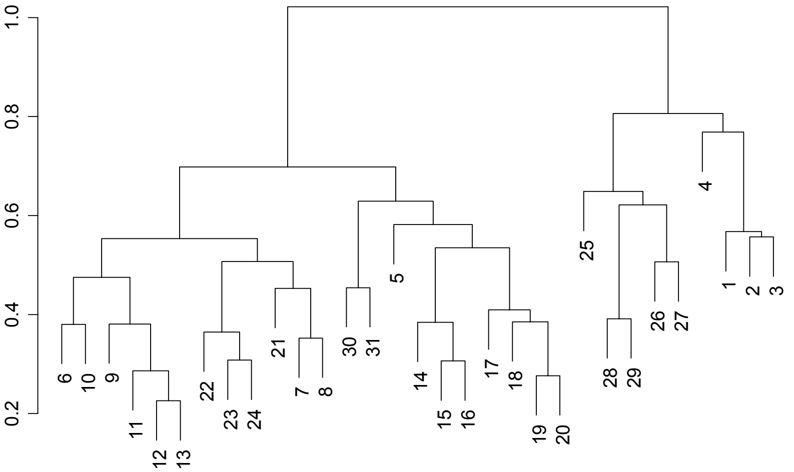


Linkage Metrics

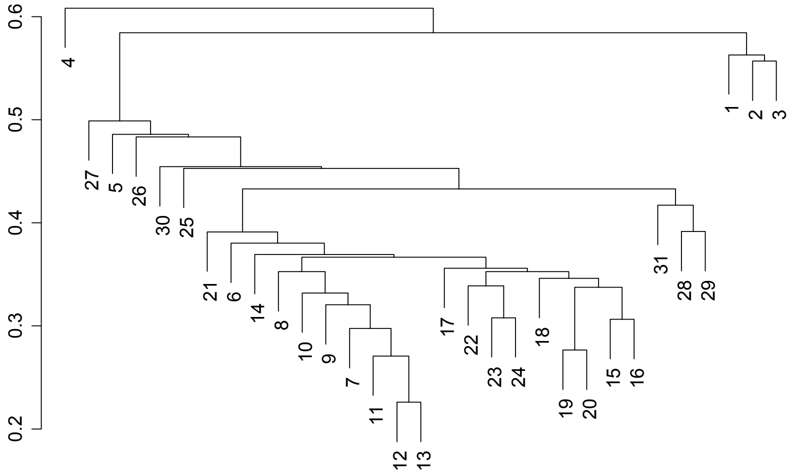
- Average linkage
  - Distance between clusters is average of all pair-wise distances between members of the two clusters
- Complete linkage
  - Distance is maximum of pair-wise distances
  - Tends to produce compact clusters
- Single linkage
  - Distance is minimum of pair-wise distances
  - Prone to “chaining” and sensitive to noise
- Centroid/Eisen linkage
  - Distance is the distance between the centroids of the two clusters



Clustering of Melanoma Tumors Using Average Linkage



Clustering of Melanoma Tumors Using Complete Linkage



Clustering of Melanoma Tumors Using Single Linkage

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

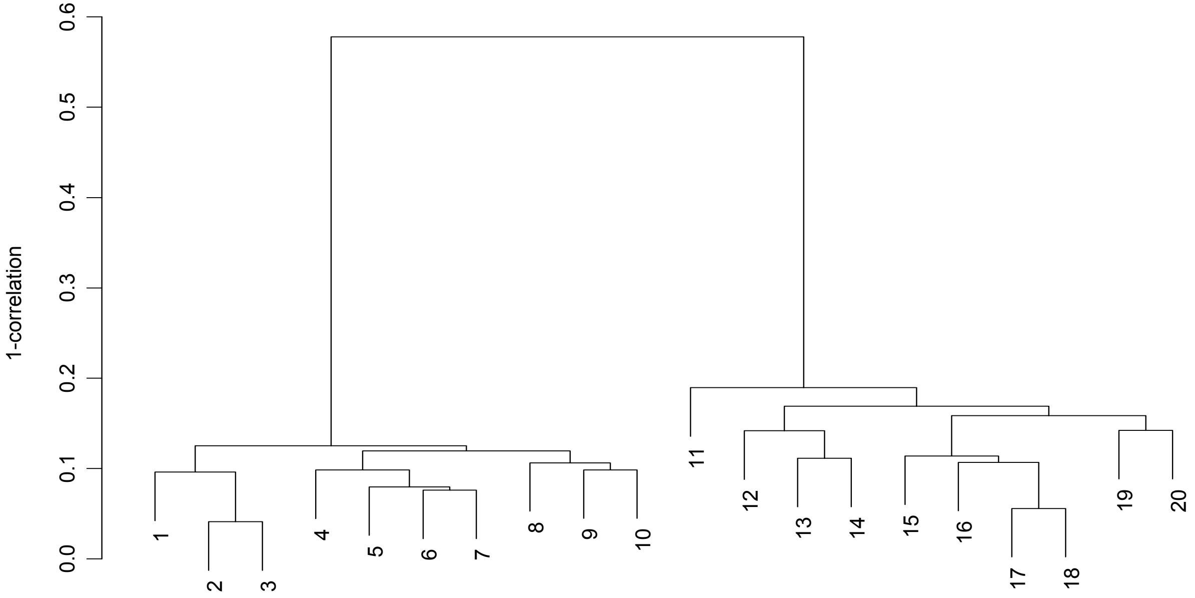
(Data from Bittner *et al.*, *Nature*, 2000)

Gene Centering

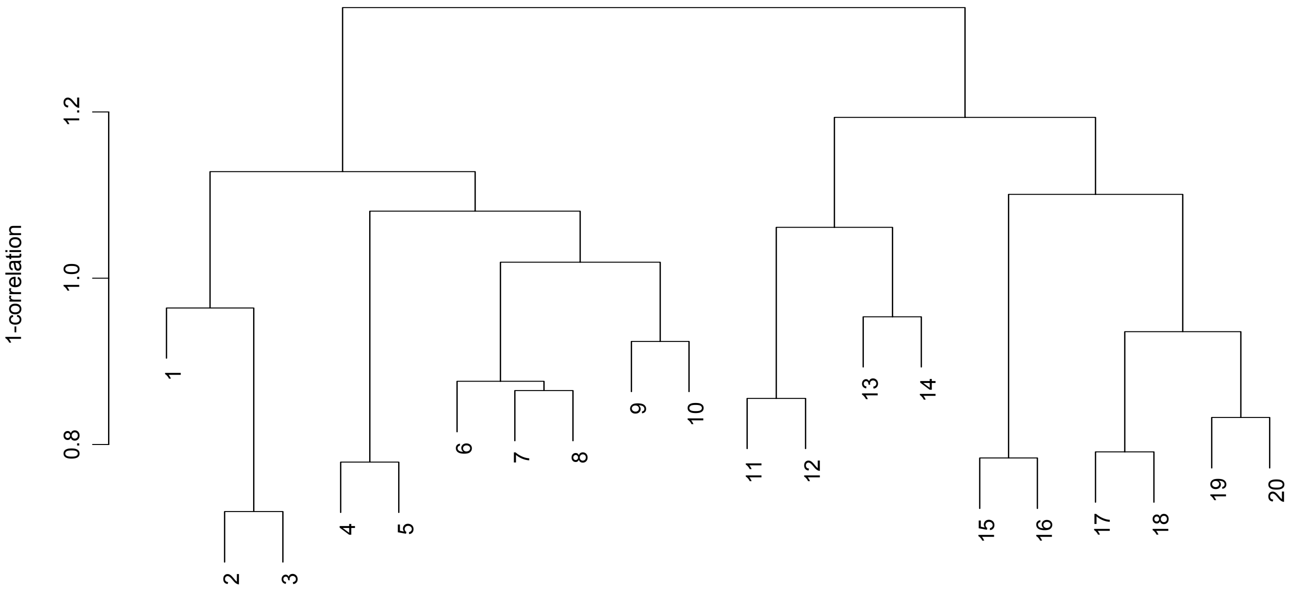
- Subtracting overall average value for each gene
- Does not influence Euclidean distance between samples
- Does influence correlation between samples
  - Reduces influence of genes whose expression in internal reference is very different than in samples
  - Beneficial for clustering samples when internal reference is arbitrary

Gene Standardization

- Dividing gene expression by standard deviation or maximum value for that gene
- Useful for single channel data to reduce the influence of high intensity genes
- Not useful for two channel data because it amplifies noise for non-informative genes



Two Clusters



Clusters?

Genes Used for Clustering

- Samples may cluster differently with regard to different gene sets
- All genes
- All “well measured” genes
  - Genes with fewer than specified percentage of values filtered because of low intensity or poor imaging
- Genes with most variation across samples

Measures of Gene Expression Variability

- Proportion of arrays in which the gene is two-fold different from it's mean or median value
- Variance of gene expression across the arrays (V_i)
  - In the upper k'th percentile of variance for all genes
  - Is statistically significantly greater than the median variance for all genes

Genes Used for Clustering Samples

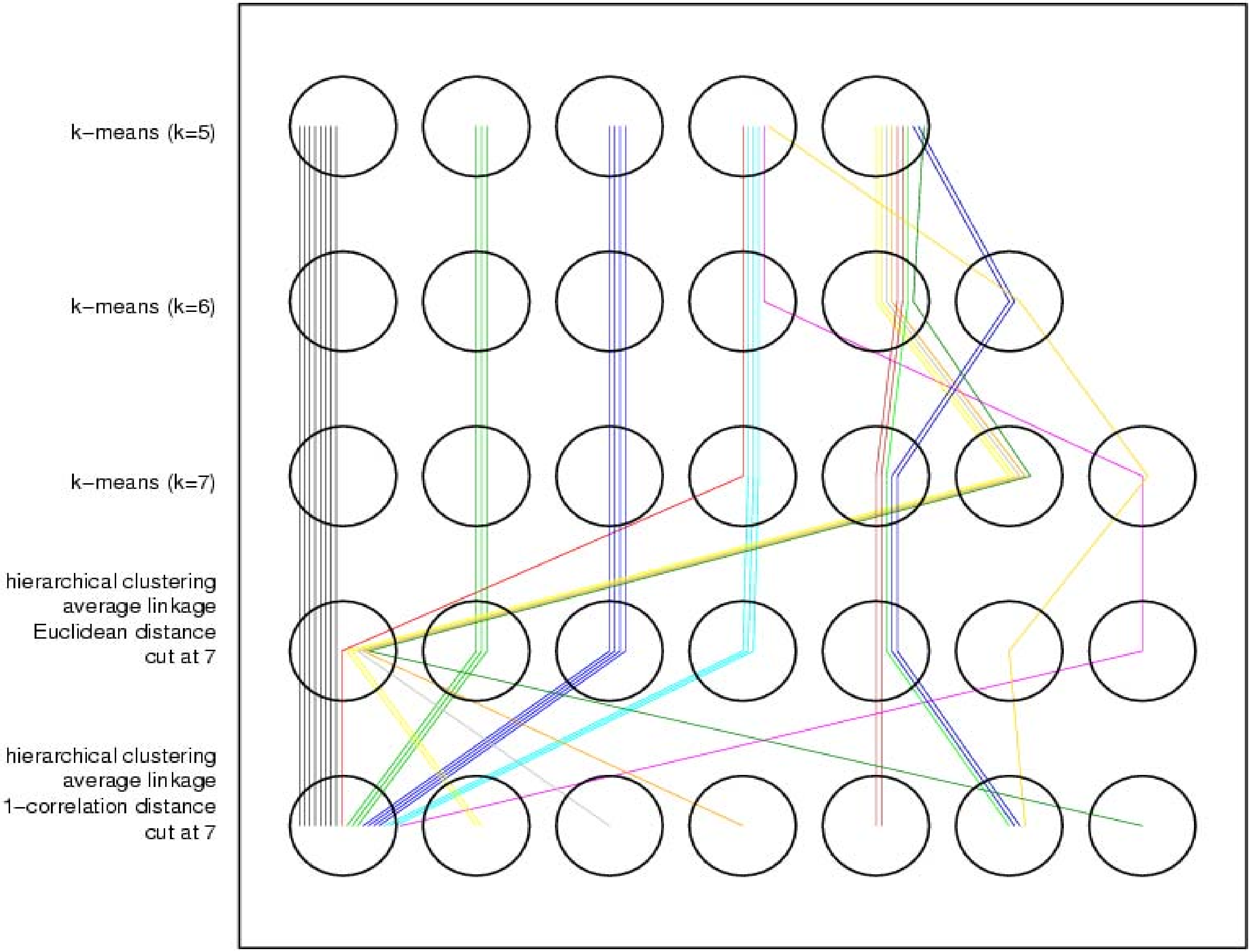
- Genes selected as being differentially expressed between pre-defined classes
  - The cluster dendrogram is a visual display that the samples are distinct with regard to these genes, but it is not independent evidence of the biological relevance of the genes

Fallacy of Clustering Classes Based on Selected Genes

- Even for arrays randomly distributed between classes, genes will be found that are “significantly” differentially expressed
- With 8000 genes measured, 400 false positives will be differentially expressed with $p < 0.05$
- Arrays in the two classes will necessarily cluster separately when using a distance measure based on genes selected to distinguish the classes

Clustering Algorithms

- K-means
  - Pre-specify K.
  - Initialize with a center for each cluster
  - Grow each cluster by adding unassigned elements (samples or genes) to the cluster center they are nearest
  - Redefine cluster centers as clusters grow and permit elements to shift clusters
  - Various implementations and variants



Bittner Melanoma Data

Self Organizing Maps

SOM’s

- Often described in the language of artificial and natural neural networks but really just a clustering algorithm
- A spatially smooth version of K-means with large K
- Cluster centers are computationally determined so that the distances among centers corresponds to the distances among points arranged in a regular 2-d or 3-d lattice. Each cluster center projects to a single lattice point.
- Clusters corresponding to adjacent lattice points are similar giving a continuous variation appearance to the clusters
- Can be useful for clustering genes using time series data

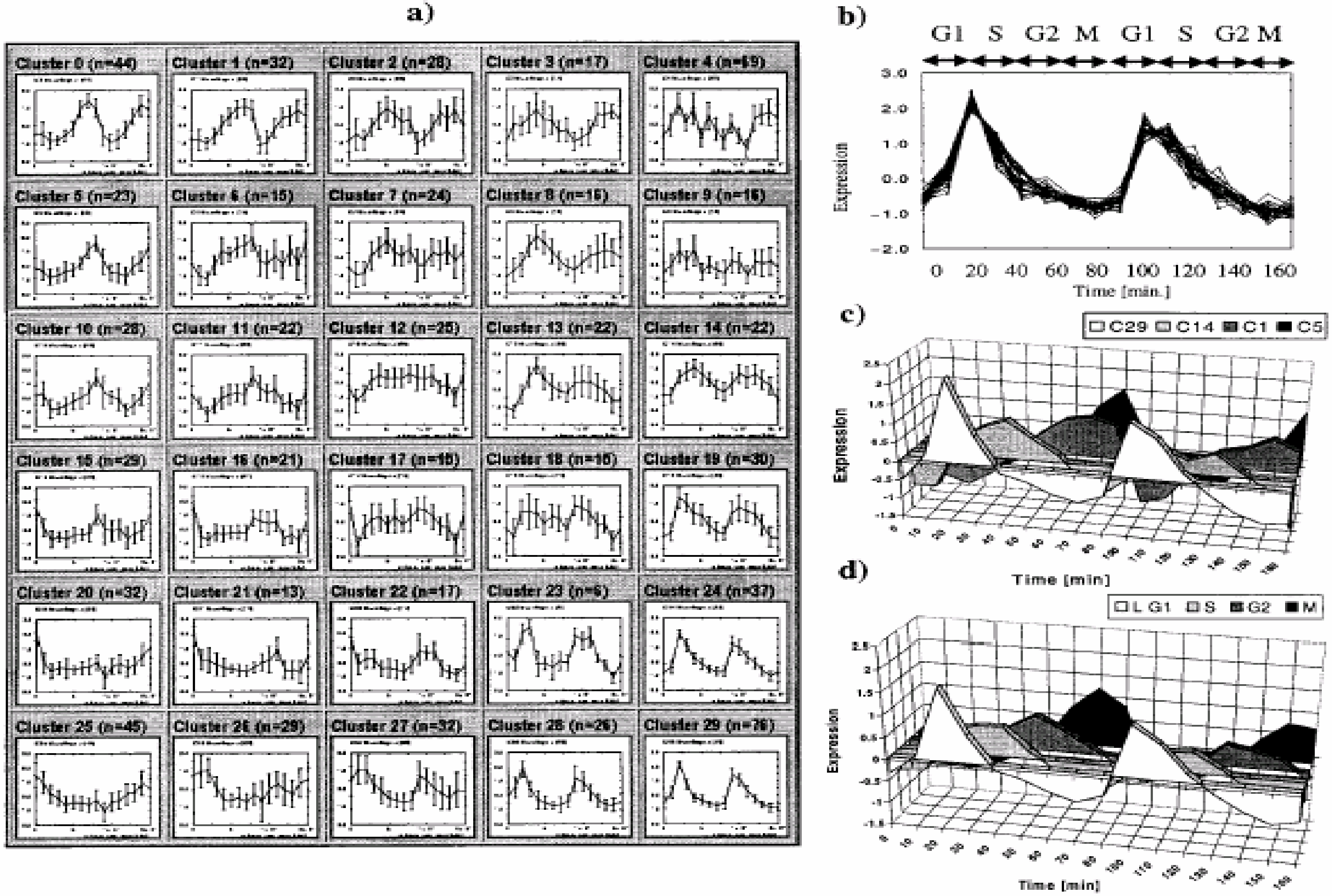
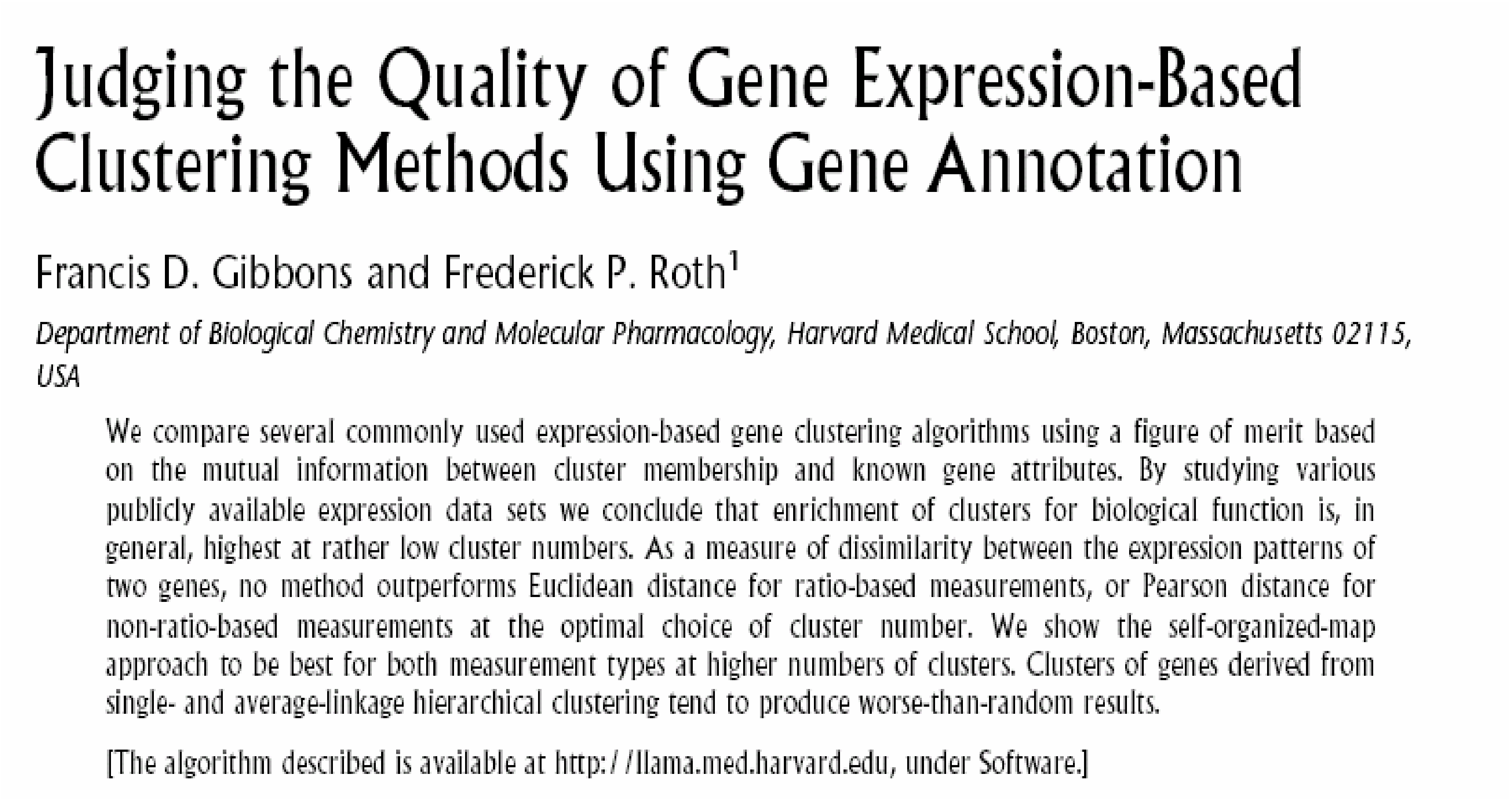


FIG. 2. Yeast Cell Cycle SOM. (*a*) 6 × 5 SOM. The 828 genes that passed the variation filter were grouped into 30 clusters. Each cluster is represented by the centroid (average pattern) for genes in the cluster. Expression level of each gene was normalized to have mean = 0 and SD = 1 across time points. Expression levels are shown on *y*-axis and time points on *x*-axis. Error bars indicate the SD of average expression. *n* indicates the number of genes within each cluster. Note that multiple clusters exhibit periodic behavior and that adjacent clusters have similar behavior. (*b*) Cluster 29 detail. Cluster 29 contains 76 genes exhibiting periodic behavior with peak expression in late G_1. Normalized expression pattern of 30 genes nearest the centroid are shown. (*c*) Centroids for SOM-derived clusters 29, 14, 1, and 5, corresponding to G_1, S, G_2 and M phases of the cell cycle, are shown. (*d*) Centroids for groups of genes identified by visual inspection by Cho *et al.* (4) as having peak expression in G_1, S, G_2, or M phase of the cell cycle are shown.



Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation

Francis D. Gibbons and Frederick P. Roth[1]

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA

We compare several commonly used expression-based gene clustering algorithms using a figure of merit based on the mutual information between cluster membership and known gene attributes. By studying various publicly available expression data sets we conclude that enrichment of clusters for biological function is, in general, highest at rather low cluster numbers. As a measure of dissimilarity between the expression patterns of two genes, no method outperforms Euclidean distance for ratio-based measurements, or Pearson distance for non-ratio-based measurements at the optimal choice of cluster number. We show the self-organized-map approach to be best for both measurement types at higher numbers of clusters. Clusters of genes derived from single- and average-linkage hierarchical clustering tend to produce worse-than-random results.

[The algorithm described is available at http://llama.med.harvard.edu, under Software.]

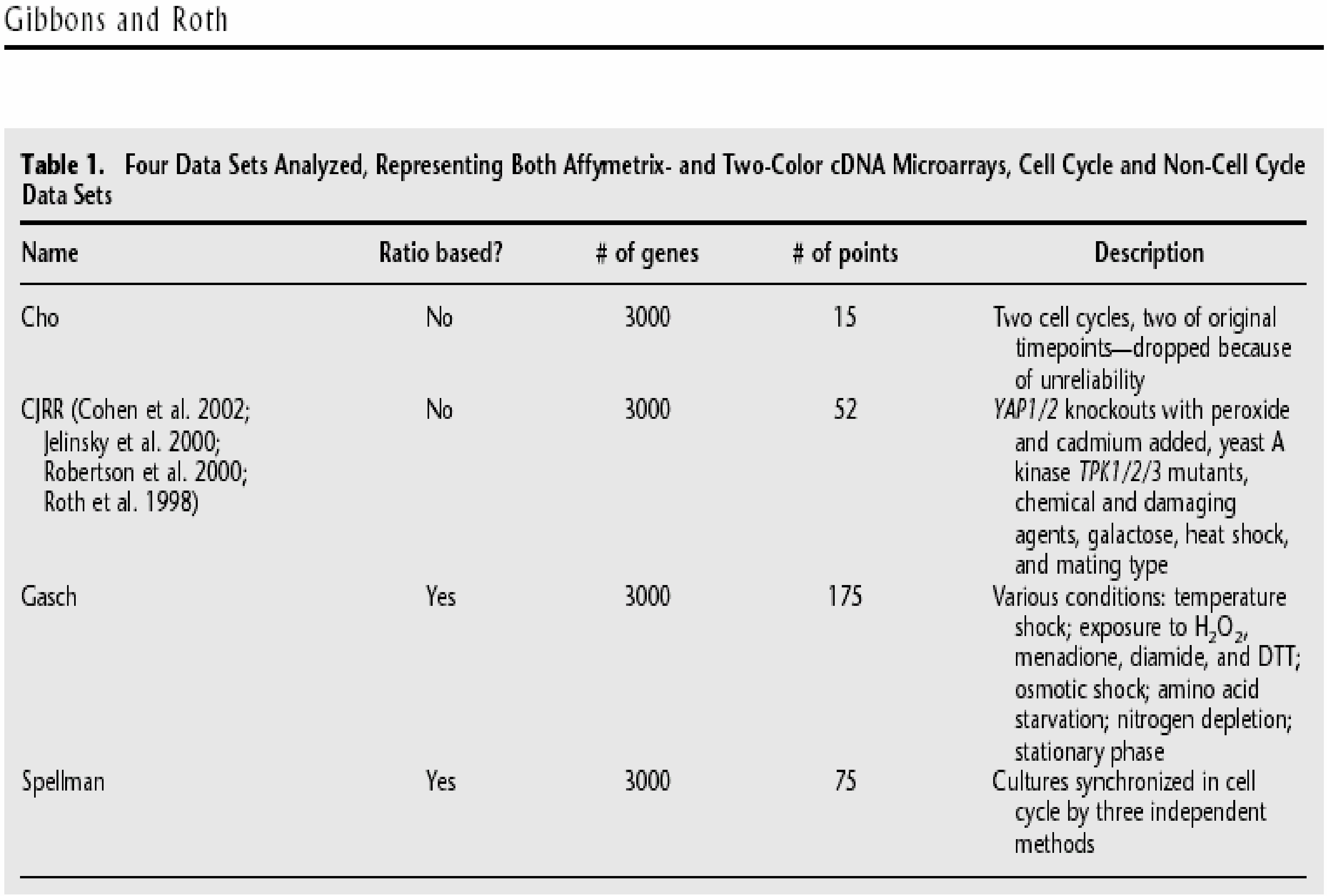


Table 1. Four Data Sets Analyzed, Representing Both Affymetrix- and Two-Color cDNA Microarrays, Cell Cycle and Non-Cell Cycle Data Sets

Name	Ratio based?	# of genes	# of points	Description
Cho	No	3000	15	Two cell cycles, two of original timepoints—dropped because of unreliability
CJRR (Cohen et al. 2002; Jelinsky et al. 2000; Robertson et al. 2000; Roth et al. 1998)	No	3000	52	*YAP1/2* knockouts with peroxide and cadmium added, yeast A kinase *TPK1/2/3* mutants, chemical and damaging agents, galactose, heat shock, and mating type
Gasch	Yes	3000	175	Various conditions: temperature shock; exposure to H_2O_2, menadione, diamide, and DTT; osmotic shock; amino acid starvation; nitrogen depletion; stationary phase
Spellman	Yes	3000	75	Cultures synchronized in cell cycle by three independent methods

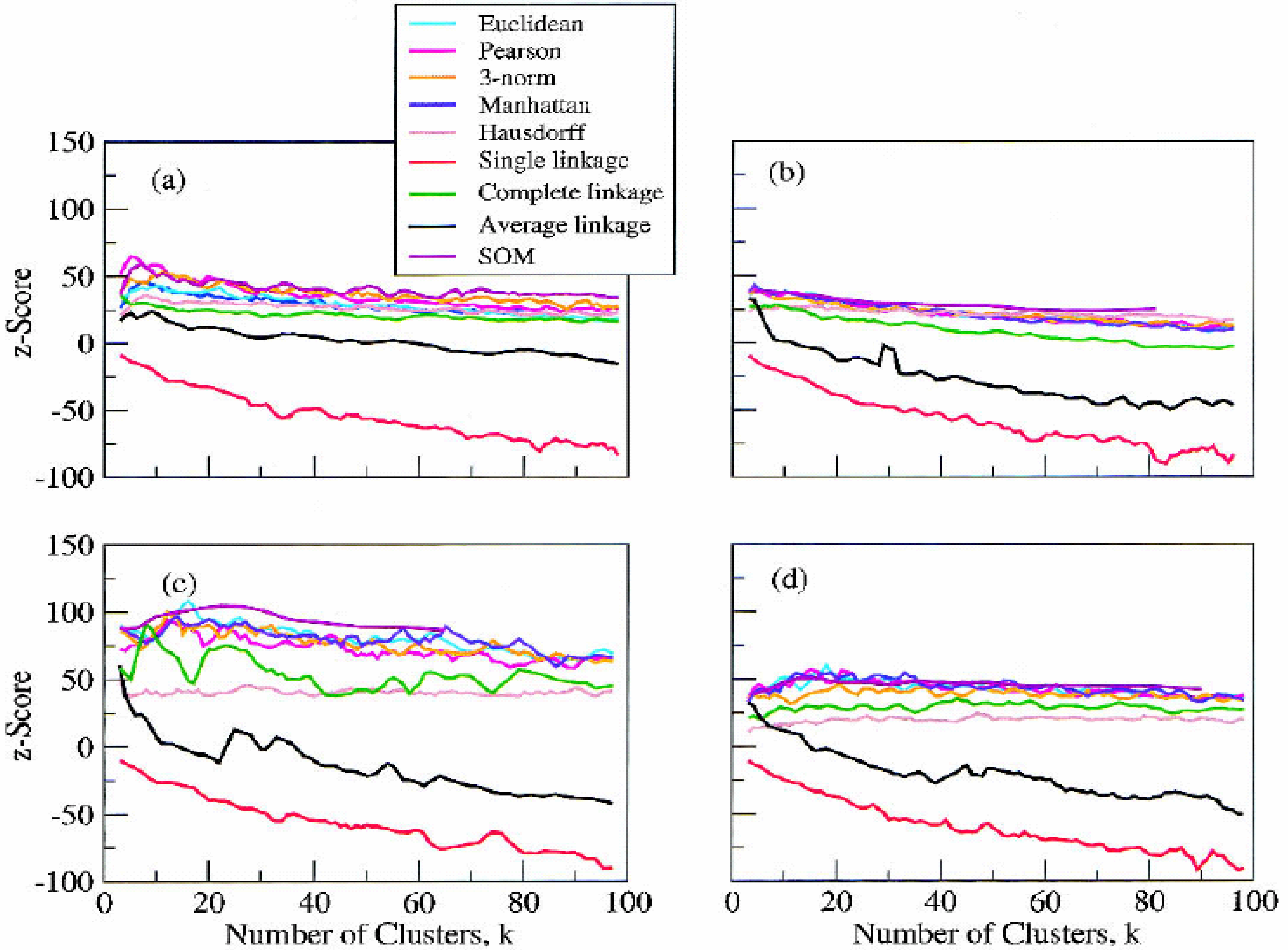


Figure 2 Four data sets clustered using *k*-means, hierarchical, and self-organized map algorithms. The horizontal axis shows the number of clusters desired, and the vertical axis shows *z*-scores. Data sets are (*a*) Cho, (*b*) CJRR, (*c*) Gasch, and (*d*) Spellman.

Validation of Clusters

- Clustering algorithms find clusters, even when they are spurious
- Clusters found may change with re-assaying tumors or selection of new tumors

Clustering Arrays

- Cluster significance
- Cluster reproducibility

Cluster Significance

McShane et al

- Transform expression data to 3-d principal component space
- Compute median of empirical distribution of distance of each sample from its nearest neighbor
- Compute distribution of above statistic for data generated from multivariate Gaussian null distribution in principal component space
- Repeat for 1000 samples from the Gaussian null distribution

Cluster Significance

- Determine the proportion of the null distribution replications that the median of nearest neighbor distances is as small as for the median of nearest neighbor distances with the actual data

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*. 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*. *D-index*.
  - Average the cluster-specific proportions over many perturbed data sets to get an *R-index* for each cluster.

Discrepancy Index

Original Data

1) Perform clustering on data.

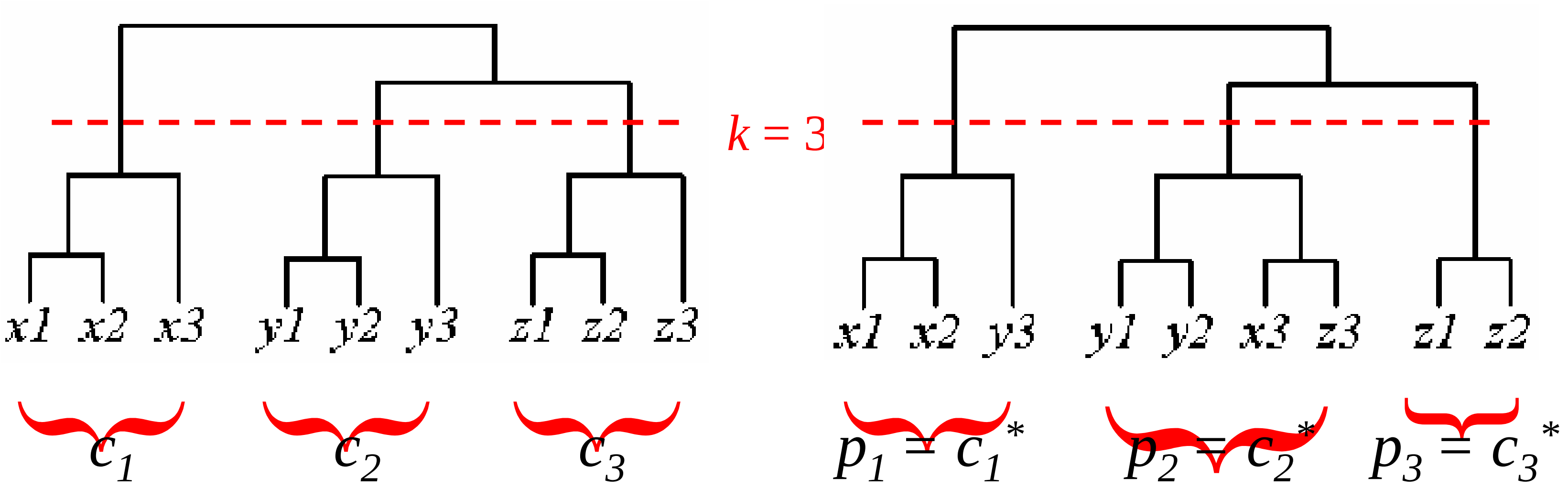
2) Cut at height that results in *k* clusters.

Perturbed Data

3) Perturb data by adding random Gaussian noise to each data point.

4) Perform clustering and cut into a similar number of clusters as original data.

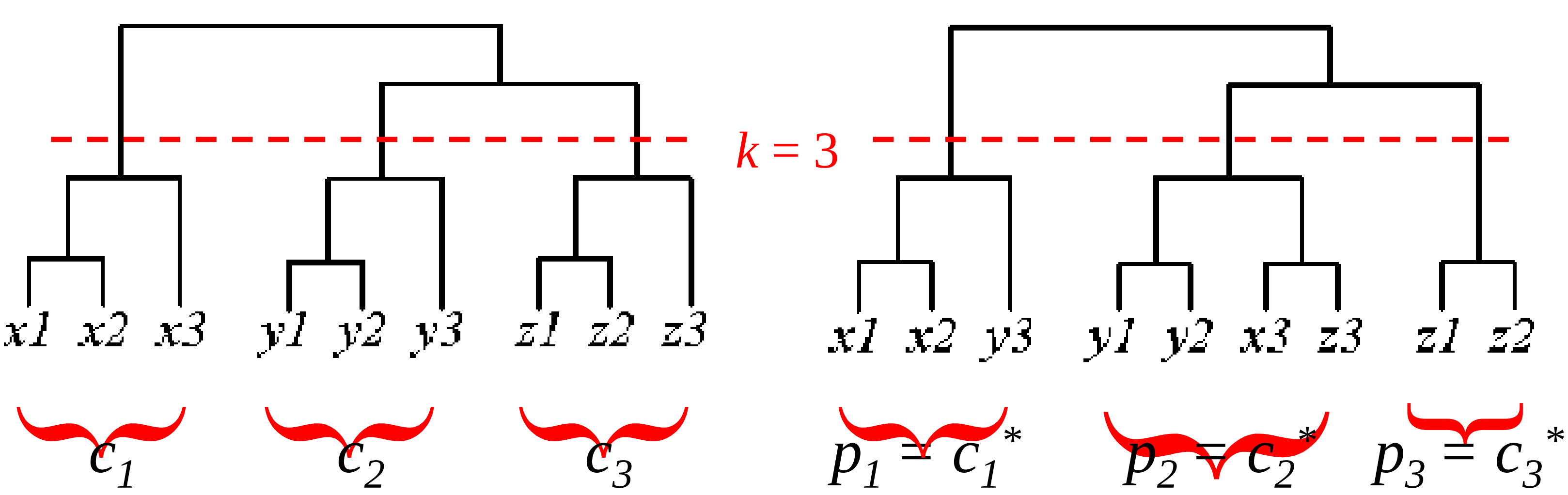
5) Map from each original cluster c_i to the perturbed cluster that minimizes the sum of missing (*m*) elements.



Discrepancy Index

Original Data

Perturbed Data



	m **(missing)**	***n*** **(contaminating)**
$c_1 \rightarrow c_1^*$:	1	1
$c_2 \rightarrow c_2^*$:	1	2
$c_3 \rightarrow c_3^*$:	1	0

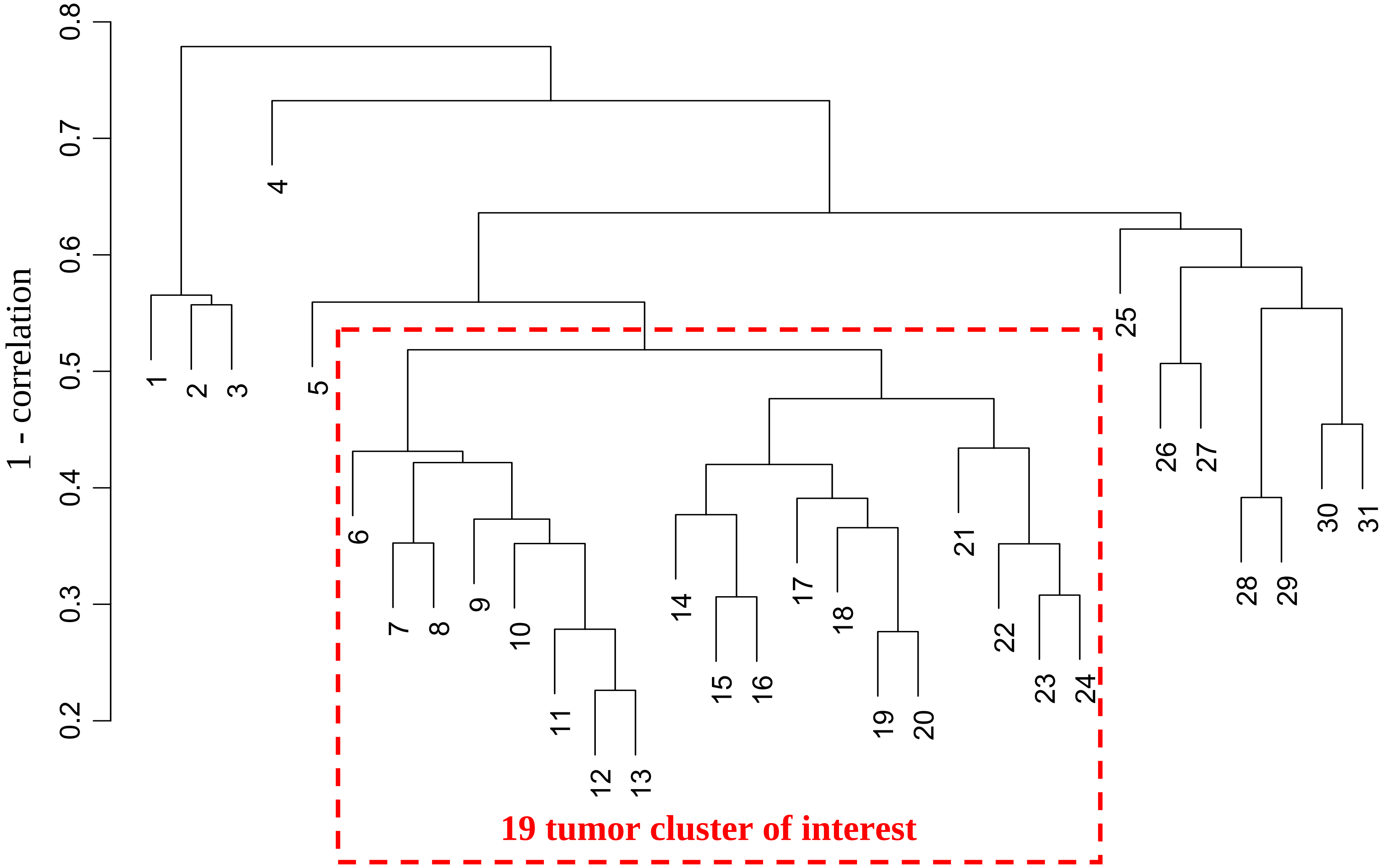
Compute minimal *mn*-error for this cut of original data:

$$\varepsilon' = \sum_{i=1}^{3}\left(m_i + n_i\right) = 6$$

ε' can be averaged over many perturbations resulting in $\overline{\varepsilon'}$.

Melanoma Gene Expression Data

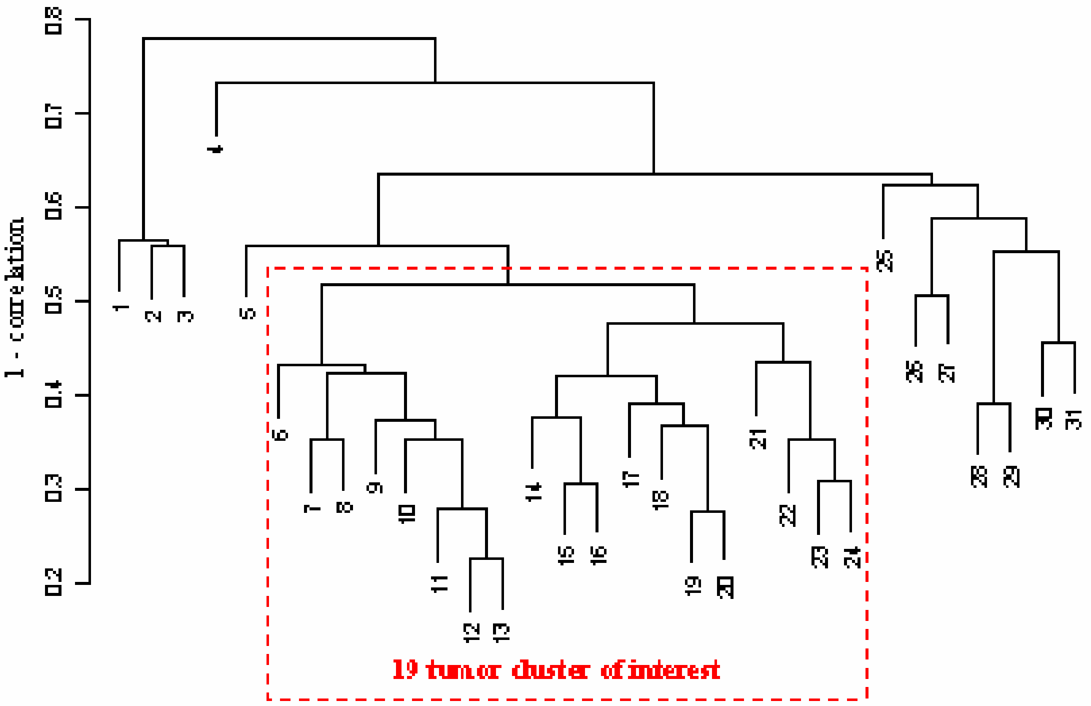
Q: Can gene expression profiles of melanoma be used to distinguish sub-classes of disease? (M. Bittner et al.)



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.



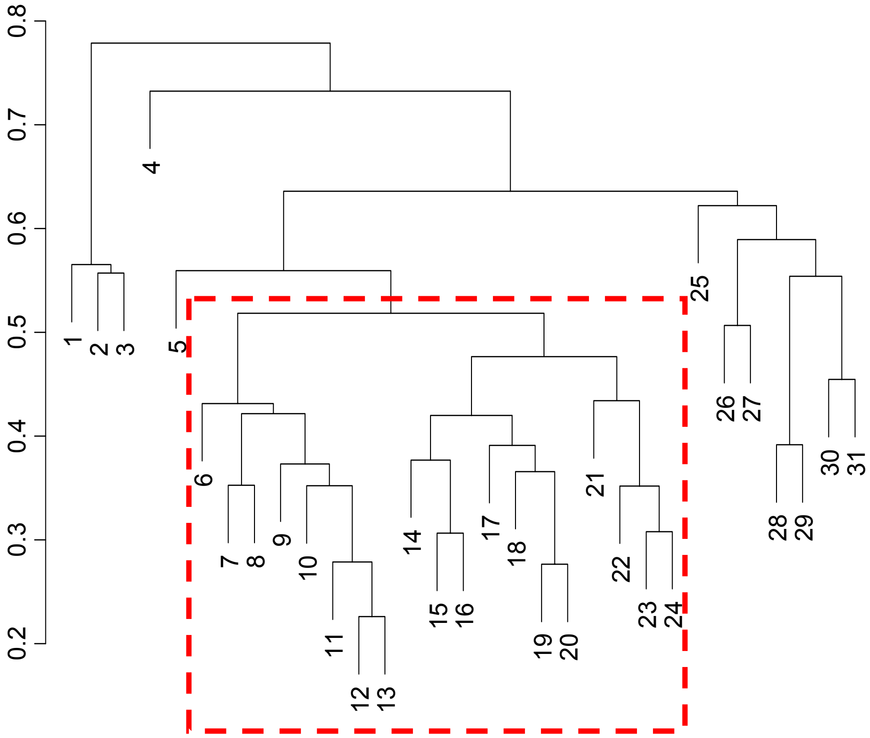
For hierarchical clustering, the cluster of interest had an *R-index* = 1.0.

⇒ highly reproducible

Melanomas in the 19 element cluster tended to have:

- reduced invasiveness
- reduced motility

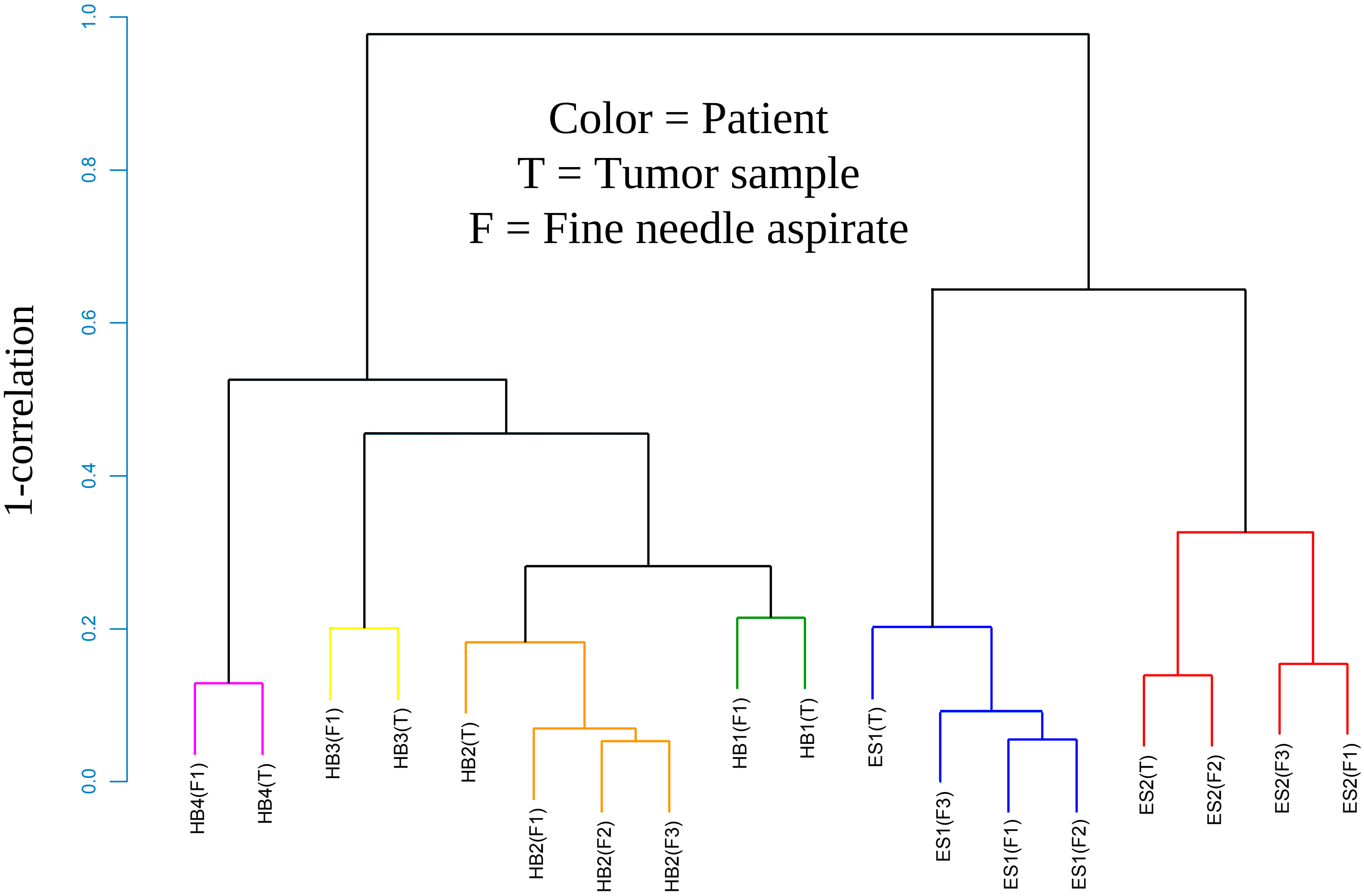
Melanoma Data:
Discrepancy Index - Individual Clusters



k	*Cluster Membership*	$\overline{m}'$	$\overline{n}'$
7	5-24	0.00	0.00
8	5	0.00	5.13
8	6-24	0.00	0.27

Visualization Tools

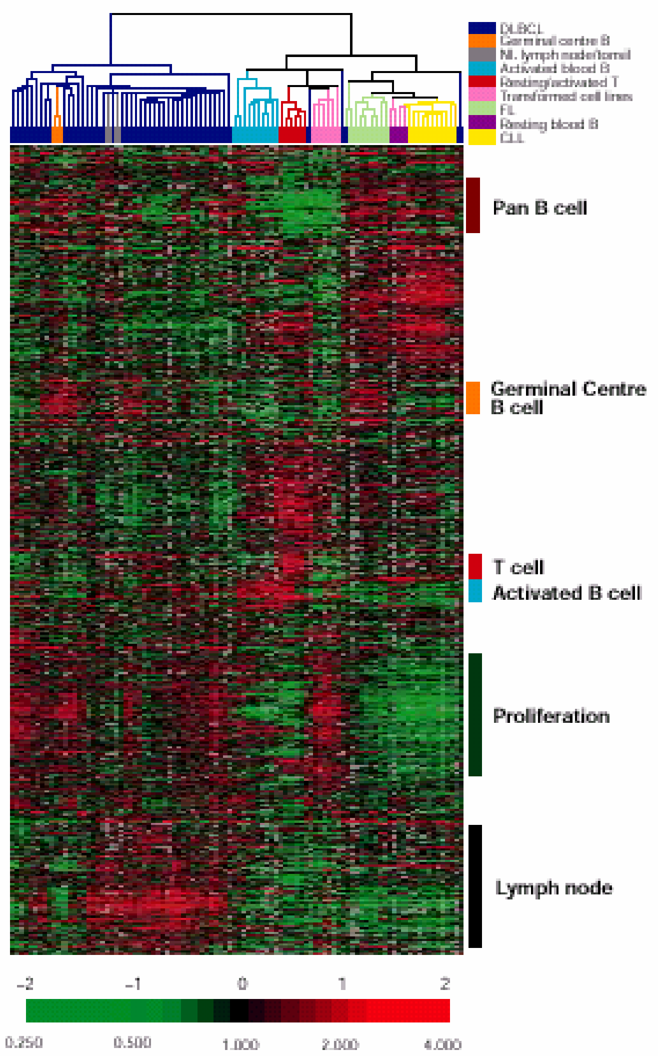
Color-Coded Hierarchical Clustering Dendrogram for Breast Tumor and FNA Samples



Color = Patient
T = Tumor sample
F = Fine needle aspirate

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

Heat Map



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

Multidimensional Scaling

- How to most accurately represent the pairwise distances between expression profiles (vector of log-ratios or vector of log intensities) in 5000-d space in a 3-dimensional plot
  - Pair-wise distance relationships cannot be exactly represented in low dimension
  - MDS gives a best approximation
  - The first three principal components are approximately optimal for 3-d representation

Principal Components for Representing Samples

- The first principal component is the linear combination of gene expression levels that has the greatest variation among all linear combinations
  - – x_i denotes the log-ratio for the i'th gene
  - – Linear combination = $a_1x_1+\ldots+a_nx_n$
  - – Where $a_1^2 + \ldots + a_n^2 = 1$

Principal Components for Representing Samples

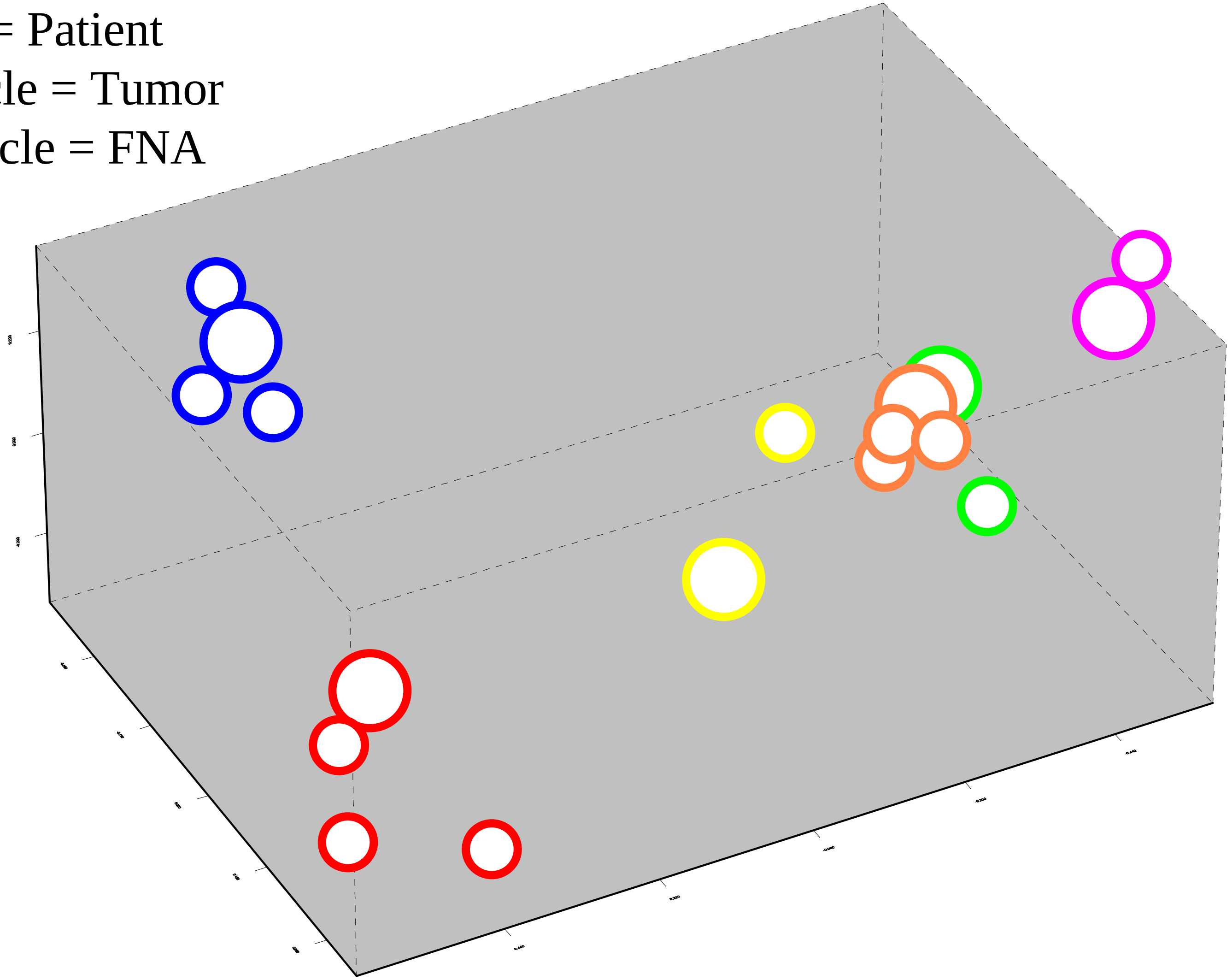
- The second principal component is the linear combination of gene expression levels that shows the largest variation among all linear combinations orthogonal to the first principal component
  - Linear combination b_1x_1+…+ b_nx_n is orthogonal if a_1b_1+…+ a_nb_n =0

Principal Components for Representing Samples

- Principal components being linear combinations of genes are not easily interpretable in terms of which genes are highly represented
- Principal components can be very useful for visualizing distance relationships between samples where they capture much of the variation eventhough they may not be identified with specific genes

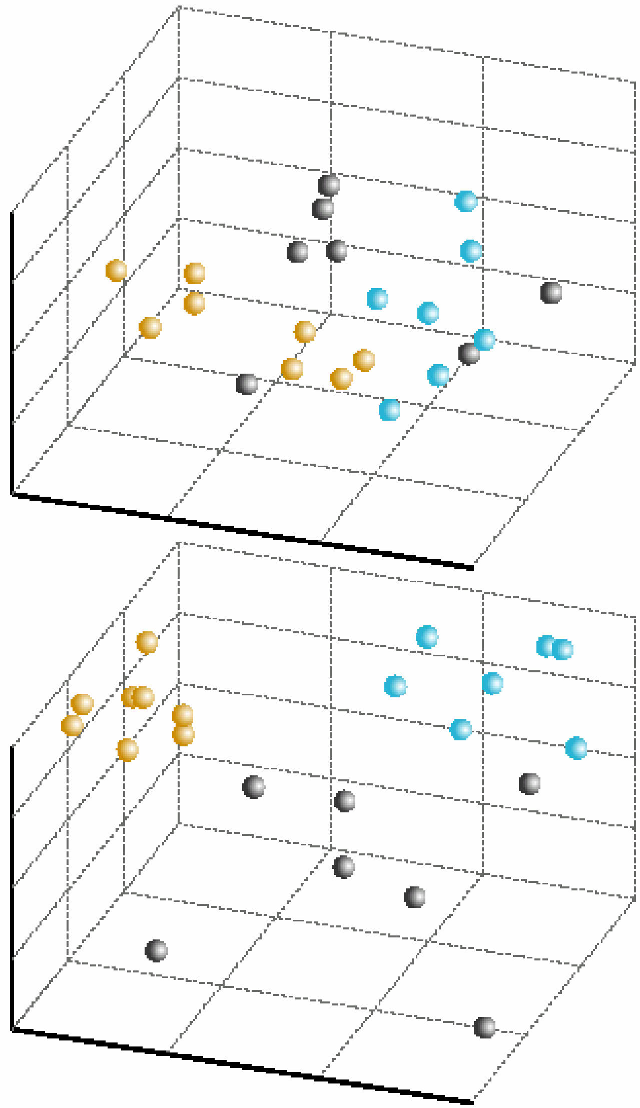
MDS: Breast Tumor and FNA Samples

Color = Patient
Large circle = Tumor
Small circle = FNA



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

MDS Plots



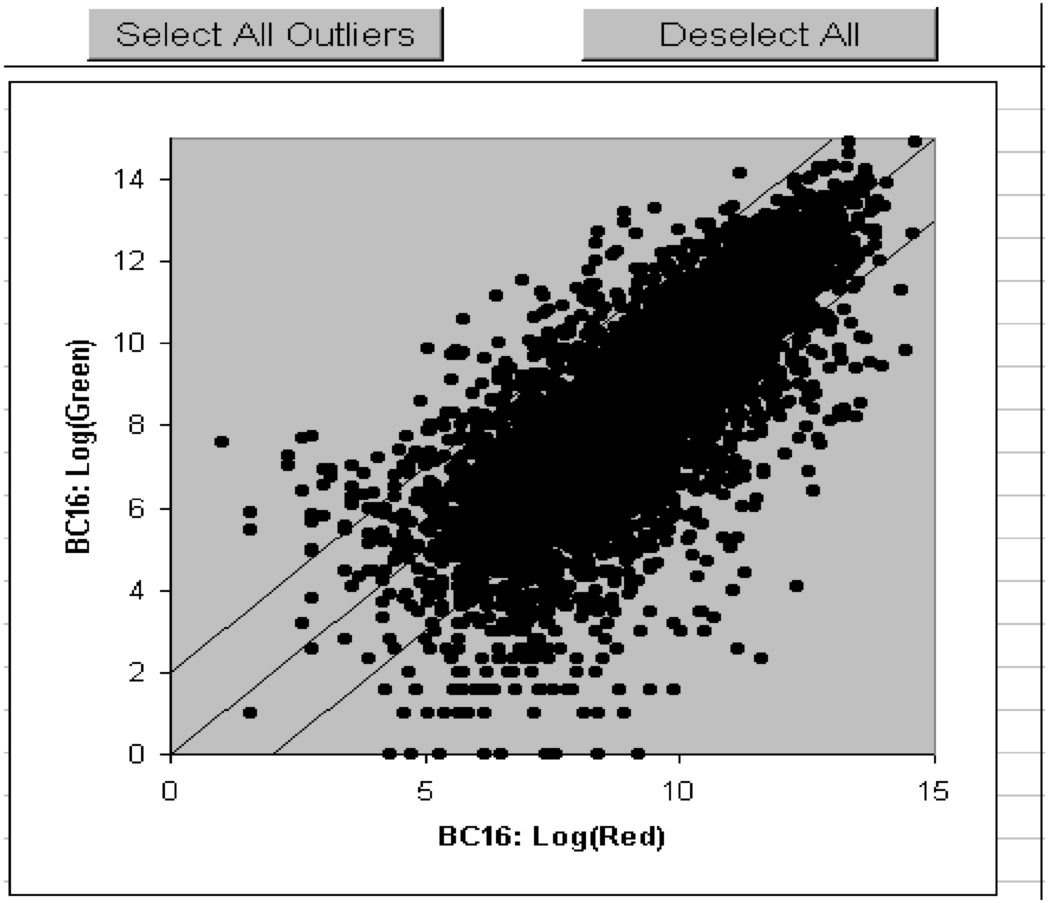
Blue Spheres: BRCA1

Gold Spheres: BRCA2

Gray Spheres: Sporadic

Using all genes

Using differentially expressed genes

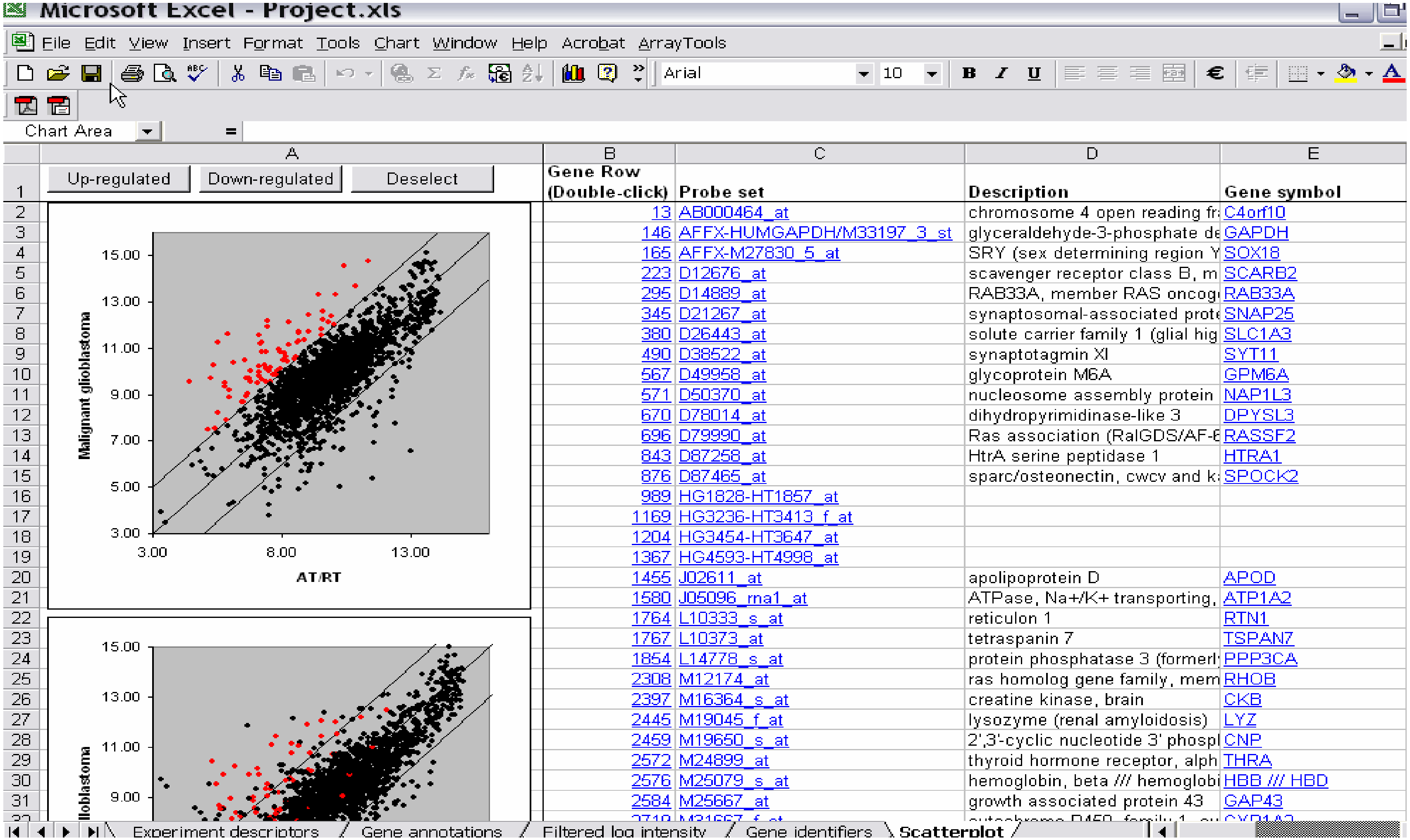

Select All Outliers
Deselect All
BC16: Log(Green)
0
2
4
6
8
10
12
14
0
5
10
15
BC16: Log(Red)

Scatterplots

- Array vs array
  - Log-ratio vs log-ratio or log signal vs log signal
- MA for single array
  - Mean of log red and log green intensities vs log-ratio
- Scatterplot of phenotype averages
  - Plots average log-ratio or log signal, averaged over classes of arrays
- Identification of outliers
- Click on plots to hyperlink to clone reports
- Double-click to view gene annotations (if available)

Scatterplot of phenotype averages

Pomeroy dataset



	B	C	D	E
1	Gene Row (Double-click)	Probe set	Description	Gene symbol
2	13	AB000464_at	chromosome 4 open reading fr	C4orf10
3	146	AFFX-HUMGAPDH/M33197_3_st	glyceraldehyde-3-phosphate de	GAPDH
4	165	AFFX-M27830_5_at	SRY (sex determining region Y	SOX18
5	223	D12676_at	scavenger receptor class B, m	SCARB2
6	295	D14889_at	RAB33A, member RAS oncog	RAB33A
7	345	D21267_at	synaptosomal-associated prote	SNAP25
8	380	D26443_at	solute carrier family 1 (glial hig	SLC1A3
9	490	D38522_at	synaptotagmin XI	SYT11
10	567	D49958_at	glycoprotein M6A	GPM6A
11	571	D50370_at	nucleosome assembly protein	NAP1L3
12	670	D78014_at	dihydropyrimidinase-like 3	DPYSL3
13	696	D79990_at	Ras association (RalGDS/AF-6	RASSF2
14	843	D87258_at	HtrA serine peptidase 1	HTRA1
15	876	D87465_at	sparc/osteonectin, cwcv and k	SPOCK2
16	989	HG1828-HT1857_at		
17	1169	HG3236-HT3413_f_at		
18	1204	HG3454-HT3647_at		
19	1367	HG4593-HT4998_at		
20	1455	J02611_at	apolipoprotein D	APOD
21	1580	J05096_rna1_at	ATPase, Na+/K+ transporting,	ATP1A2
22	1764	L10333_s_at	reticulon 1	RTN1
23	1767	L10373_at	tetraspanin 7	TSPAN7
24	1854	L14778_s_at	protein phosphatase 3 (formerl	PPP3CA
25	2308	M12174_at	ras homolog gene family, mem	RHOB
26	2397	M16364_s_at	creatine kinase, brain	CKB
27	2445	M19045_f_at	lysozyme (renal amyloidosis)	LYZ
28	2459	M19650_s_at	2',3'-cyclic nucleotide 3' phospl	CNP
29	2572	M24899_at	thyroid hormone receptor, alph	THRA
30	2576	M25079_s_at	hemoglobin, beta /// hemoglobi	HBB /// HBD
31	2584	M25667_at	growth associated protein 43	GAP43
32	2719	M31667_f_at	cytochrome P450, family 1, su	CYP1A2