BRB-ArrayTools Workshop

- Overview of gene expression analysis (2hr)

- Individual consultation as needed
  - Biometric Research Branch statisticians
  - BRB-ArrayTools Development Team

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

- http://linus.nci.nih.gov/brb
  - Powerpoint presentations and audio files
  - Reprints & Technical Reports
  - BRB-ArrayTools software
  - BRB-ArrayTools Data Archive

Assumptions

- You are somewhat familiar with BRB-ArrayTools
- You have brought your own laptop
- You have installed BRB-ArrayTools
- You have imported (collated) your data into BRB-ArrayTools

BRB-ArrayTools 3.5 alpha

- Available on cd for you to try if you'd like
- Contains
  - Data import wizzard
  - Data analysis wizzard
  - Enhanced survival risk-group prediction tool

Take Time to Clarify Your Specific Objectives

- Study Design
- Analysis Strategy

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

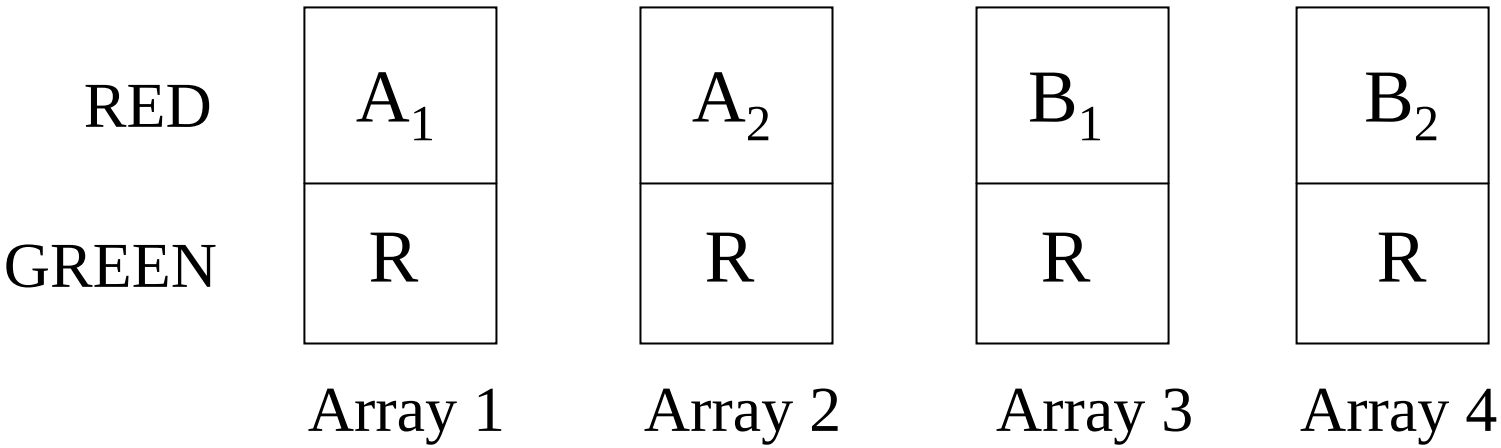
- Technical replicates
  - RNA sample divided into multiple aliquots and re-arrayed
- Biological replicates
  - Multiple subjects
  - Replication of the tissue culture experiment

- 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 Developing Predictive Classifiers

- Single label arrays
  - Affymetrix GeneChips
- Dual label arrays
  - Common reference design
  - Other designs

Common Reference Design

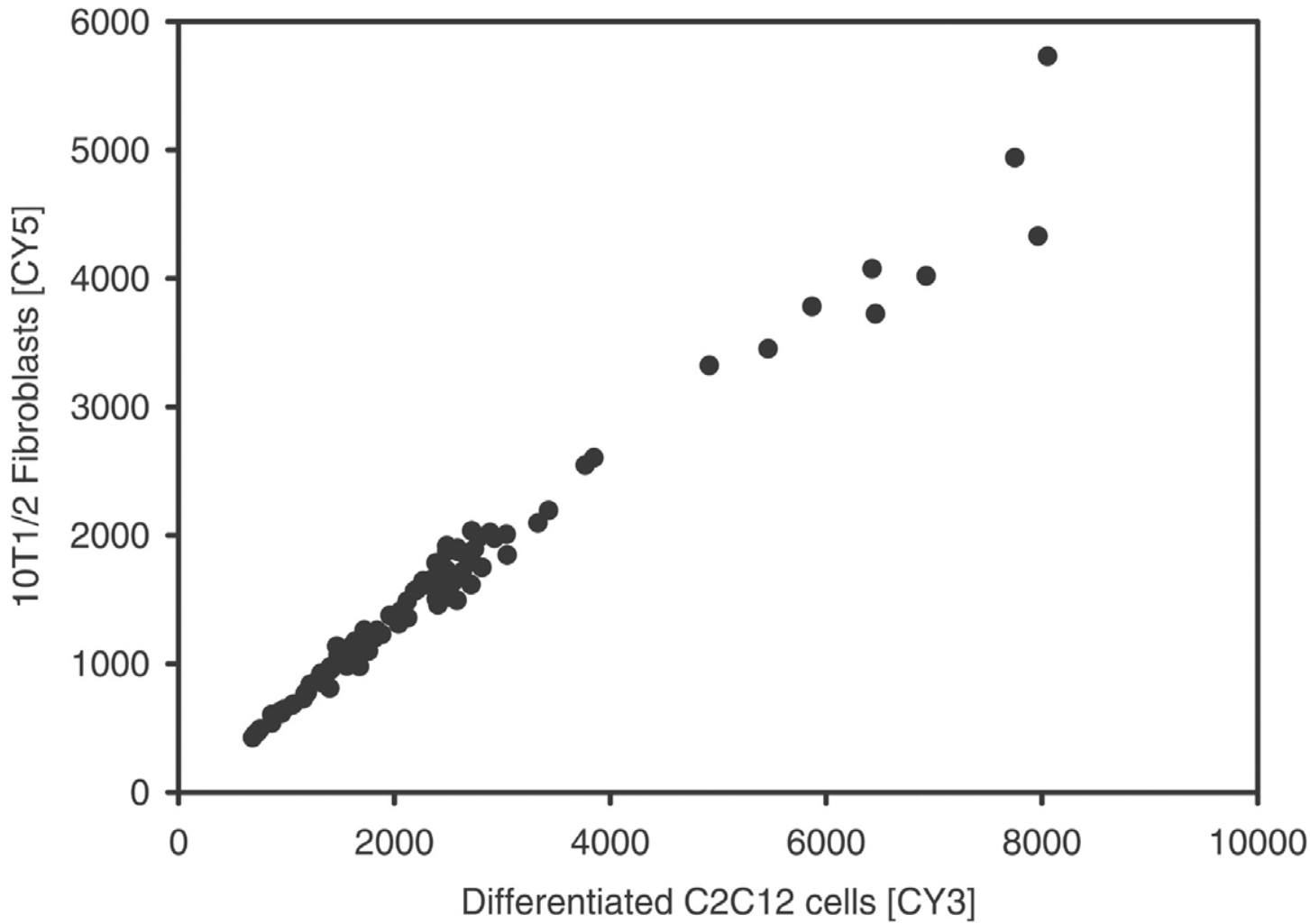


A_i = *i*th specimen from class A

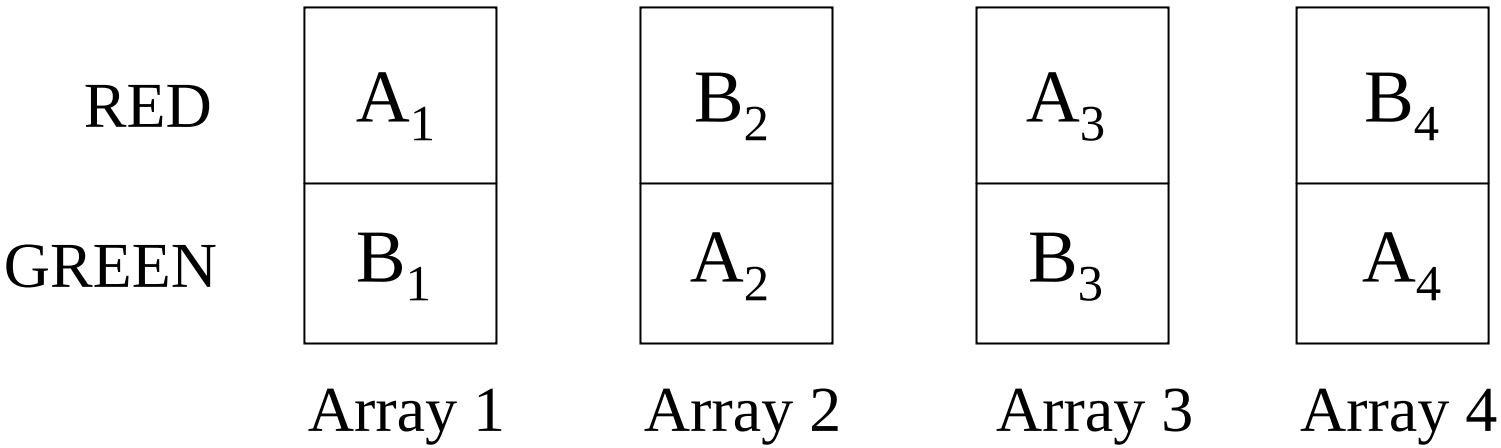
B_i = *i*th specimen from class B

R = aliquot from reference pool

- The reference generally serves to control variation in the size of corresponding spots on different arrays and variation in sample distribution over the slide.
- The reference provides a relative measure of expression for a given gene in a given sample that is less variable than an absolute measure.
- The reference is not the object of comparison.
- The relative measure of expression will be compared among biologically independent samples from different classes.

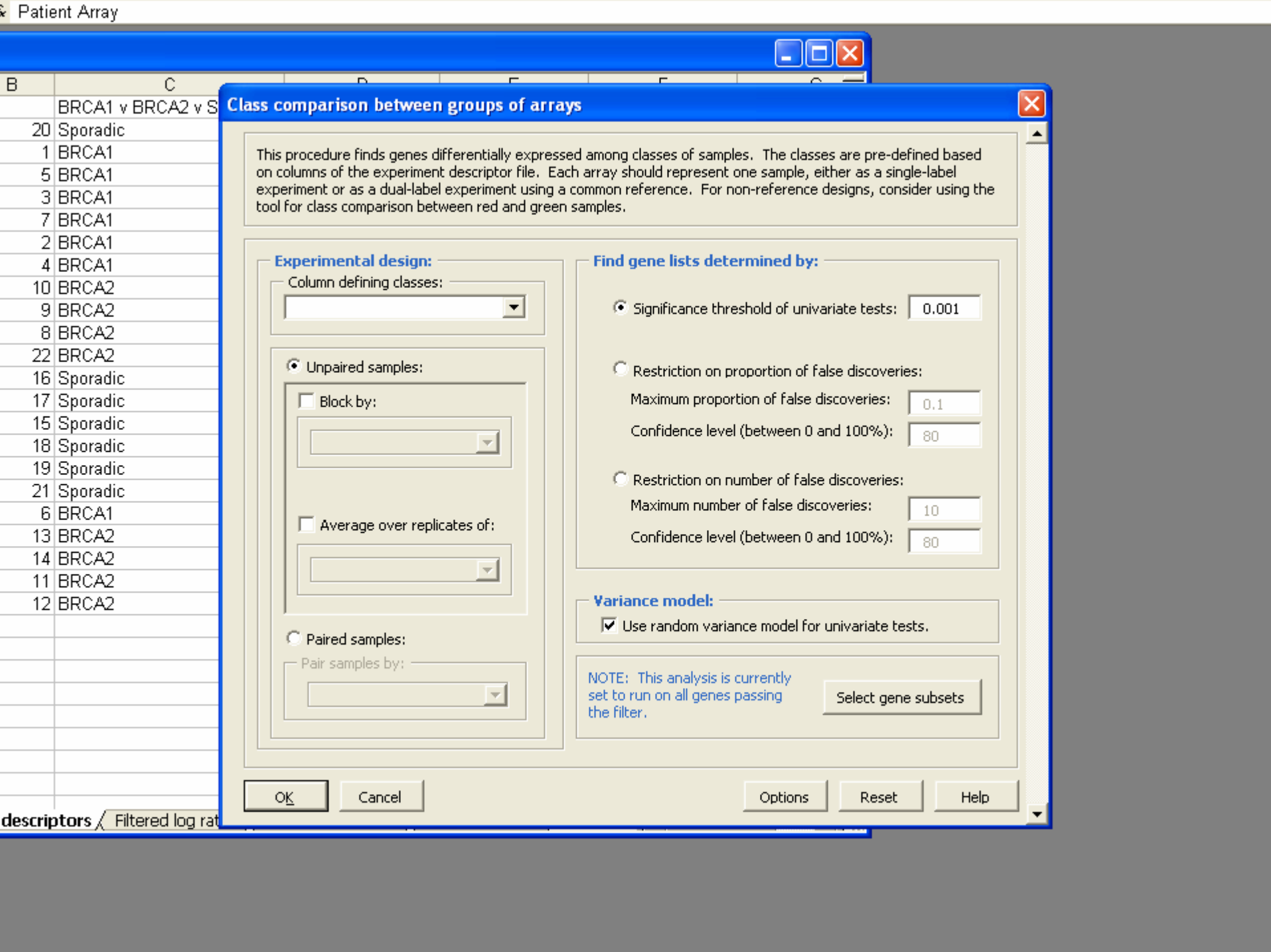

6000
5000
4000
3000
2000
1000
0
10T1/2 Fibroblasts [CY5]
0
2000
4000
6000
8000
10000
Differentiated C2C12 cells [CY3]

Balanced Block Design



A_i = *i*th specimen from class A

B_i = *i*th specimen from class B


Patient Array
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
descriptors
Filtered log rat
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.
Select gene subsets
OK
Cancel
Options
Reset
Help

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

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 $\alpha = u/k$

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 $p \leq p^*$

- 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

Additional Procedures

- "SAM" - Significance Analysis of Microarrays
  - Tusher *et al.*, *PNAS,* 2001
  - Estimate FDR
  - Statistical properties unclear
- Multivariate permutation tests
  - Korn *et al.,* 2001 (http://linus.nci.nih.gov/brb)
  - Control number or proportion of false discoveries
  - Can specify confidence level of control

Multivariate Permutation Procedures (Korn *et al.*, 2001)

Allows statements like:

FD Procedure: We are 95% confident that the (actual) number of false discoveries is no greater than 5.

FDP Procedure: We are 95% confident that the (actual) proportion of false discoveries does not exceed .10.

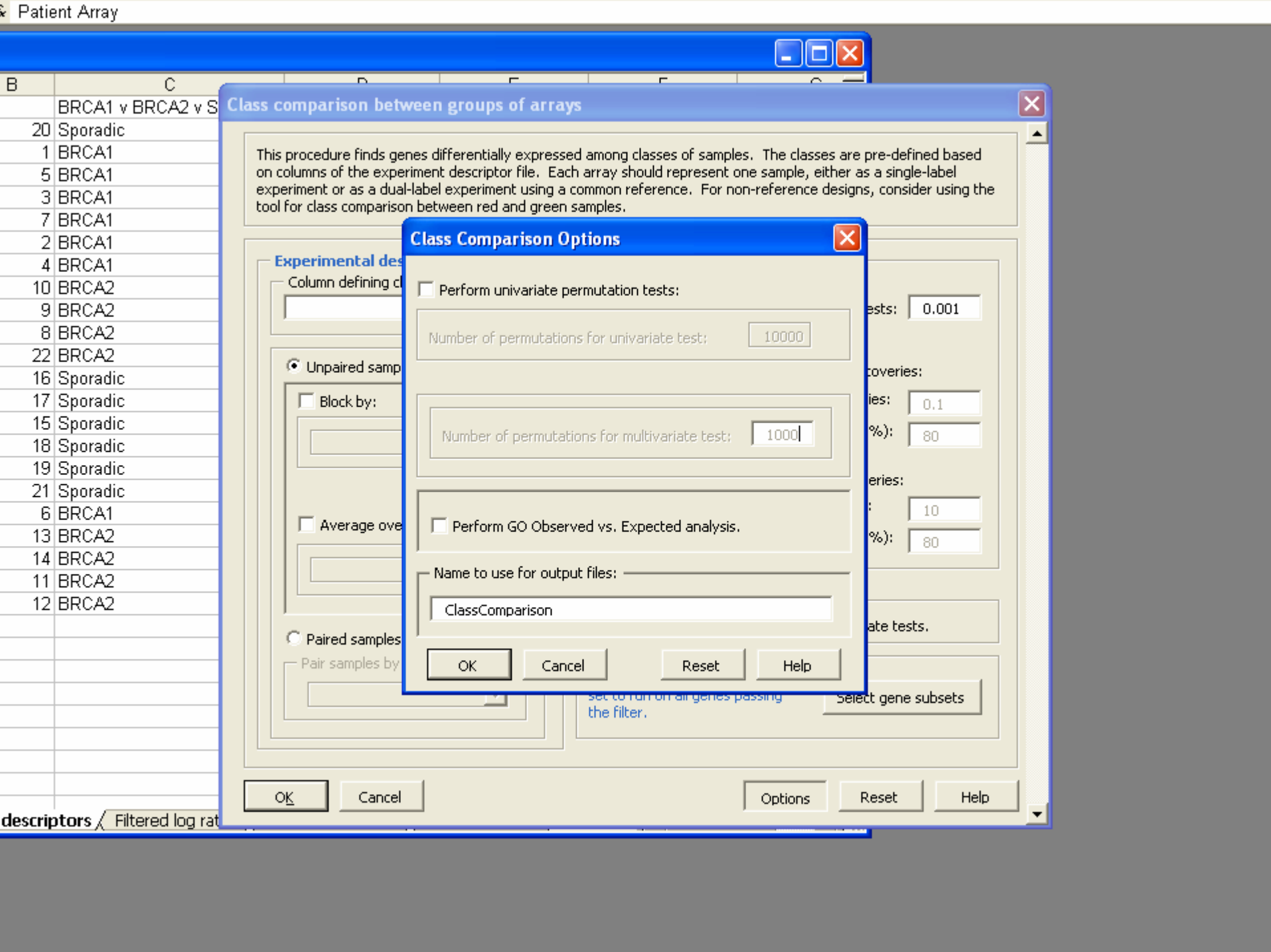
t-test Comparisons of Gene Expression for gene j

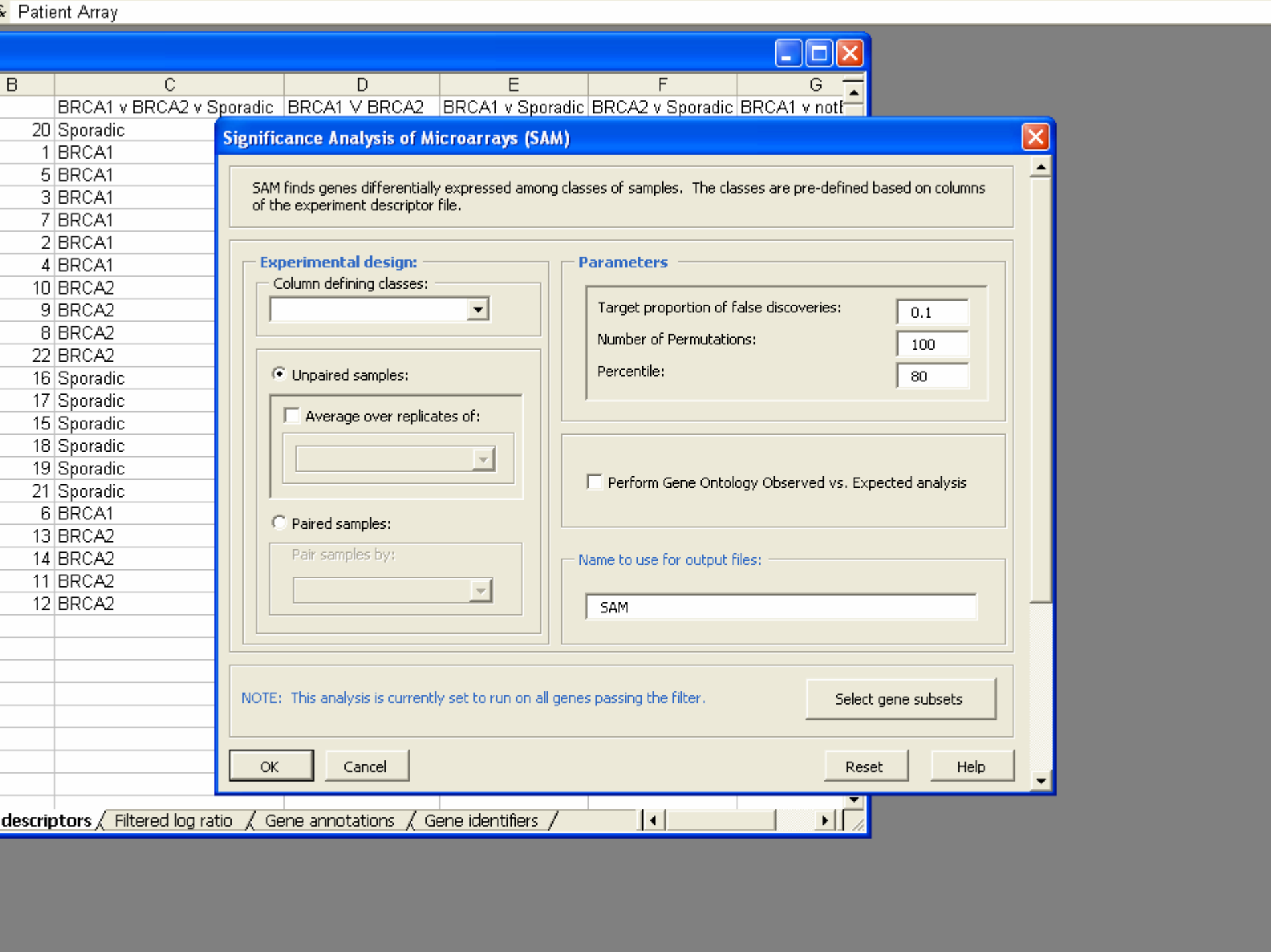
- $x_j \sim N(\mu_{j1}, \sigma_j^2)$ for class 1
- $x_j \sim N(\mu_{j2}, \sigma_j^2)$ for class 2

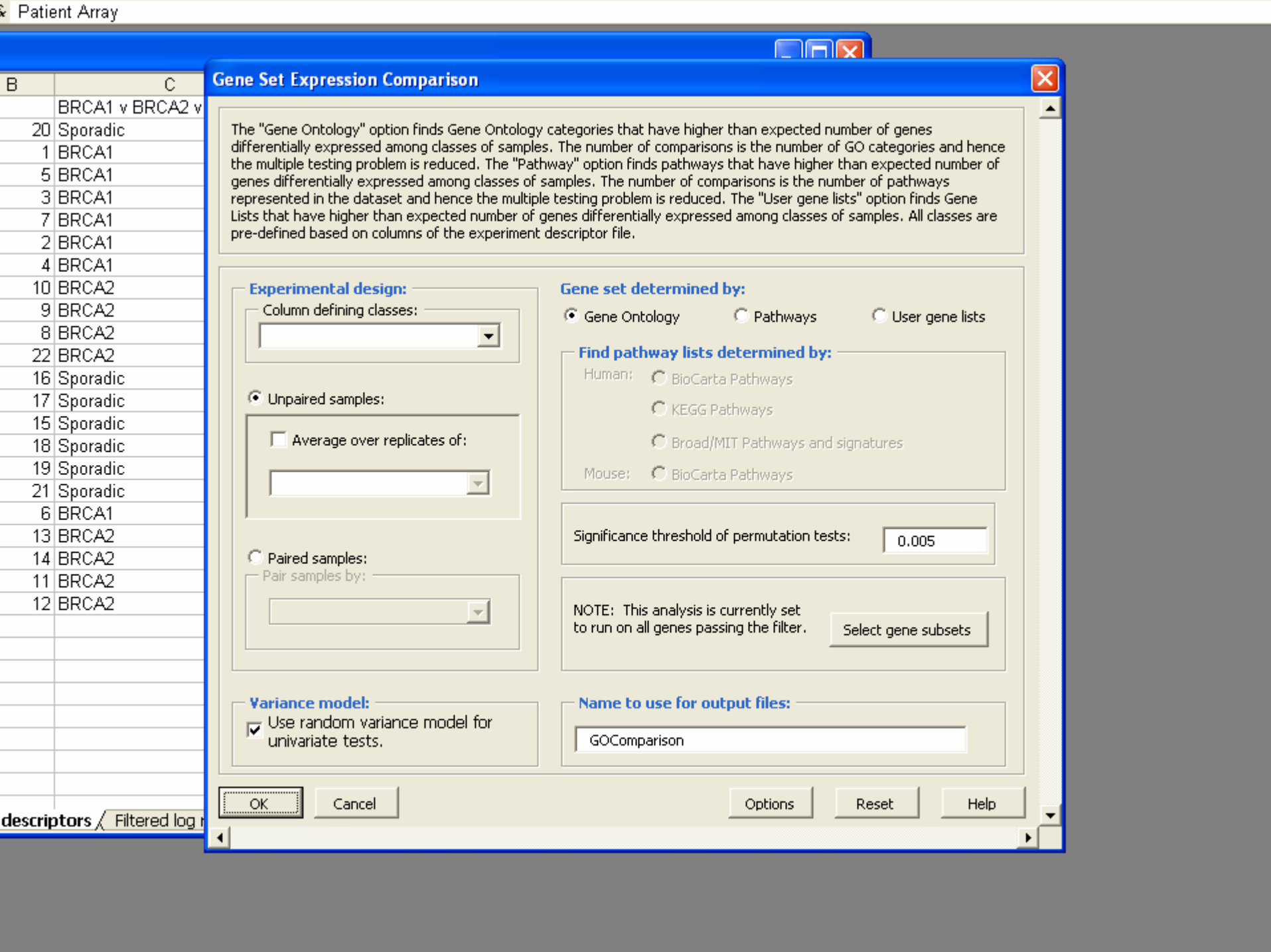
- $H_{0j}: \mu_{j1} = \mu_{j2}$

Estimation of Within-Class Variance

- Estimate separately for each gene σ_j^2
  - Limited degrees-of-freedom (precision) unless number of samples is large
  - Gene list dominated by genes with small fold changes and small variances
- Assume all genes have same variance
  - Poor assumption
- Random (hierarchical) variance model
  - Wright G.W. and Simon R. Bioinformatics19:2448-2455,2003
  - Variances are independent samples from a common distribution; Inverse gamma distribution used
  - Results in exact F (or t) distribution of test statistics with increased degrees of freedom for error variance
  - For any normal linear model


Patient Array
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
descriptors
Filtered log rat
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 des
Column defining cl
Unpaired samp
Block by:
Average ove
Paired samples
Pair samples by
esis:
0.001
coveries:
ies:
0.1
%):
80
eries:
10
%):
80
ate tests.
the filter.
Select gene subsets
OK
Cancel
Options
Reset
Help
Class Comparison Options
Perform univariate permutation tests:
Number of permutations for univariate test:
10000
Number of permutations for multivariate test:
1000
Perform GO Observed vs. Expected analysis.
Name to use for output files:
ClassComparison
OK
Cancel
Reset
Help


Patient Array
B
C
D
E
F
G
BRCA1 v BRCA2 v Sporadic
BRCA1 V BRCA2
BRCA1 v Sporadic
BRCA2 v Sporadic
BRCA1 v notB
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 ratio
Gene annotations
Gene identifiers
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.
Select gene subsets
OK
Cancel
Reset
Help


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
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.
Select gene subsets
Variance model:
Use random variance model for univariate tests.
Name to use for output files:
GOComparison
OK
Cancel
Options
Reset
Help

Gene Set Expression Comparison

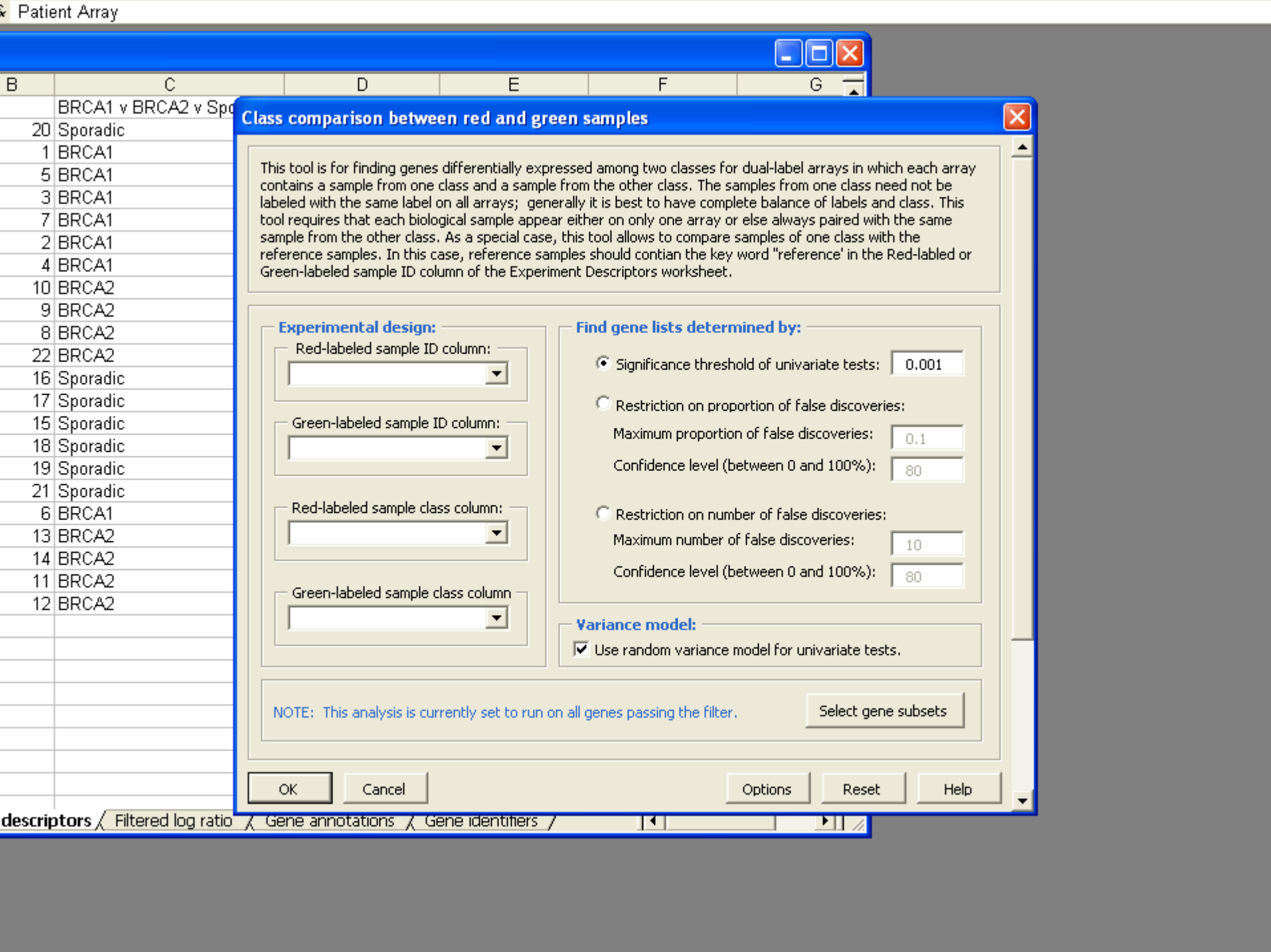
- Compute p value of differential expression for each gene in a gene set (k=number of genes)
- Compute a summary (S) of these p values
- Determine whether the value of the summary test statistic S is more extreme than would be expected from a random sample of k genes (probe-sets) on that platform
- Two 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

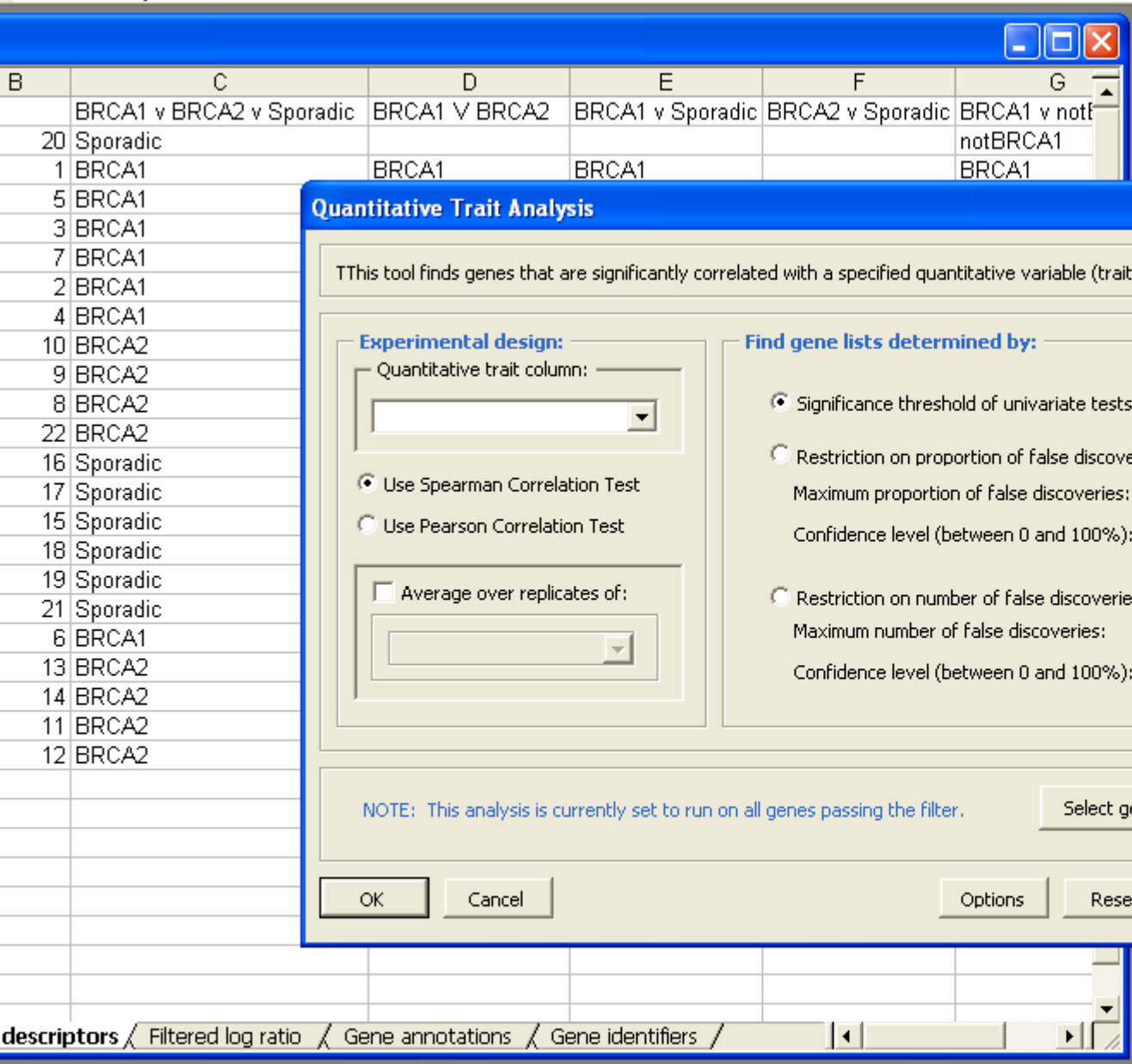
- p value for significance of summary statistic need not be as extreme as .001 usually, because the number of gene sets analyzed is usually much less than the number of individual genes analyzed
- Conclusions of significance are for gene sets in this tool, not for individual genes

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


Patient Array
B
C
D
E
F
G
BRCA1 v BRCA2 v Spo
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 ratio
Gene annotations
Gene identifiers
Class comparison between red and green samples
This tool is for finding genes differentially expressed among two classes for dual-label arrays in which each array contains a sample from one class and a sample from the other class. The samples from one class need not be labeled with the same label on all arrays; generally it is best to have complete balance of labels and class. This tool requires that each biological sample appear either on only one array or else always paired with the same sample from the other class. As a special case, this tool allows to compare samples of one class with the reference samples. In this case, reference samples should contian the key word "reference' in the Red-labled or Green-labeled sample ID column of the Experiment Descriptors worksheet.
Experimental design:
Red-labeled sample ID column:
Green-labeled sample ID column:
Red-labeled sample class column:
Green-labeled sample class column
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.
Select gene subsets
OK
Cancel
Options
Reset
Help

Patient Array



B	C	D	E	F	G
	BRCA1 v BRCA2 v Sporadic	BRCA1 V BRCA2	BRCA1 v Sporadic	BRCA2 v Sporadic	BRCA1 v notB
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 /

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.

Select gene subsets

OK Cancel Options Reset Help

Patient Array

B	C	D	E	F	G
	BRCA1 v BRCA2 v Sporadic	BRCA1 V BRCA2	BRCA1 v Sporadic	BRCA2 v Sporadic	BRCA1 v notB
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.

Select gene subsets

OK | Cancel | Options | Reset | Help

Statistical Methods Appropriate for Prediction are Different than Those Appropriate for Gene Finding

- Demonstrating statistical significance of prognostic factors is not the same as demonstrating predictive accuracy.
- Demonstrating goodness of fit of a model to the data used to develop it is not a demonstration of predictive accuracy.
- Statisticians are used to inference, not prediction
- Most statistical methods were not developed for p>>n prediction problems

Components of Class Prediction

- Feature (gene) selection
  - Which genes will be included in the model
- Select model type
  - E.g. Diagonal linear discriminant analysis, Nearest-Neighbor, …
- Fitting parameters (regression coefficients) for model
  - Selecting value of tuning parameters

Feature Selection

- Genes that are differentially expressed among the classes at a significance level α (e.g. 0.01)
  - The α level is selected only to control the number of genes in the model

Feature Selection

- Small subset of genes which together give most accurate predictions
  - Combinatorial optimization algorithms
    - Genetic algorithms
- Little evidence that complex feature selection is useful in microarray problems
  - Failure to compare to simpler methods
  - Some published complex methods for selecting combinations of features do not appear to have been properly evaluated

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

- Fisher linear discriminant analysis
  - Requires estimating correlations among all genes selected for model
- Diagonal linear discriminant analysis (DLDA) assumes gene expressions are uncorrelated
- Compound covariate predictor (Radmacher) and Golub's method are similar to DLDA in that they can be viewed as weighted voting of univariate classifiers

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 separate the classes with a hyperplain that minimizes the length of the weight vector

Support Vector Machine

minimize $\sum_{i} w_i^2$

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

where $y_j = \pm 1$ for class 1 or 2.

When p>>n

- 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?

Myth

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

- 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 they avoid overfitting the data.

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, determine it's *nearest neighbor* in the training set; i.e. which sample in the training set is its 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

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.

Other Methods

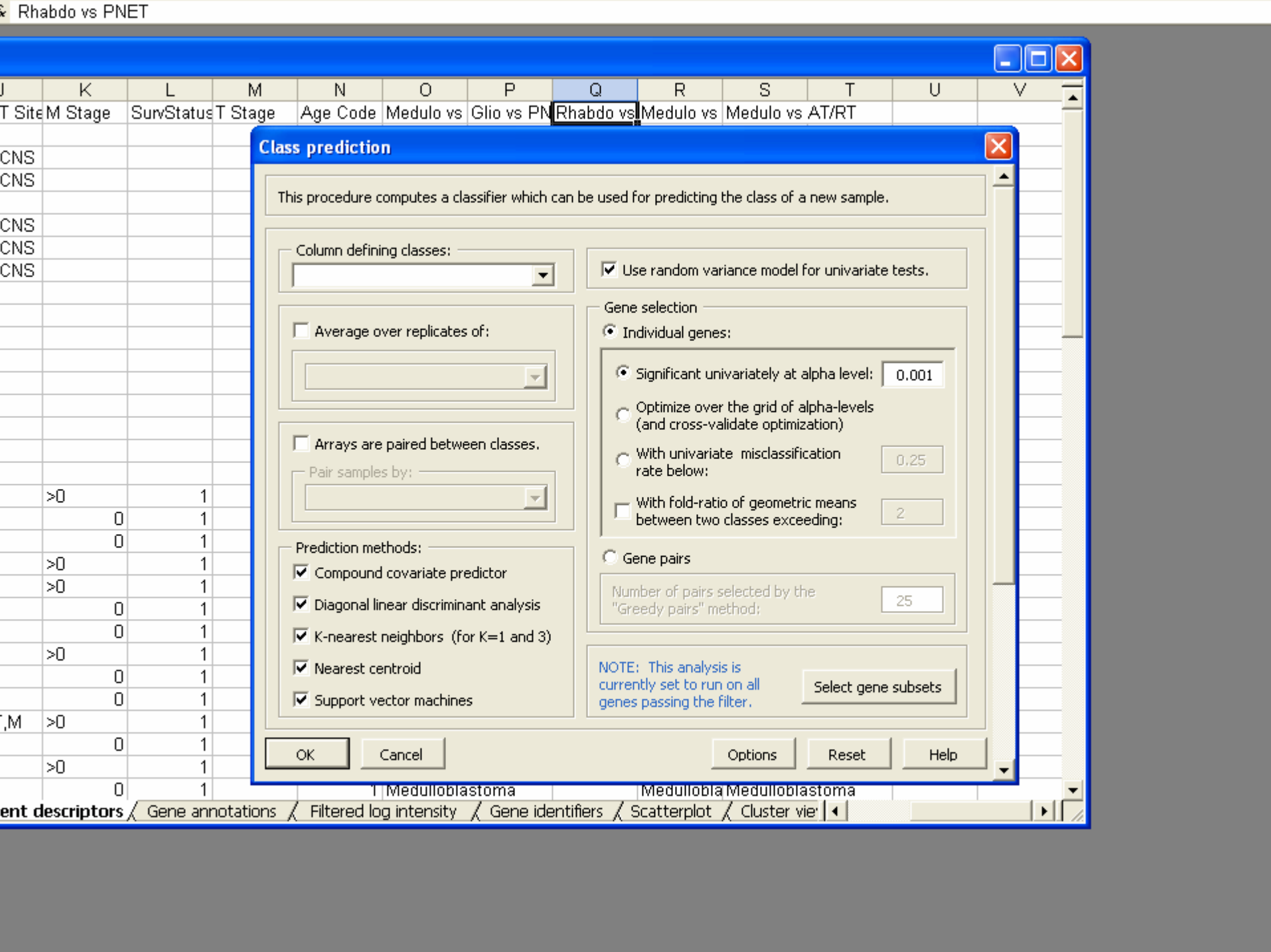
- Top-scoring pairs
  - Claim that it gives accurate prediction with few pairs because pairs of genes are selected to work well together
- Random Forrest
  - Very popular in machine learning community
  - Complex classifier

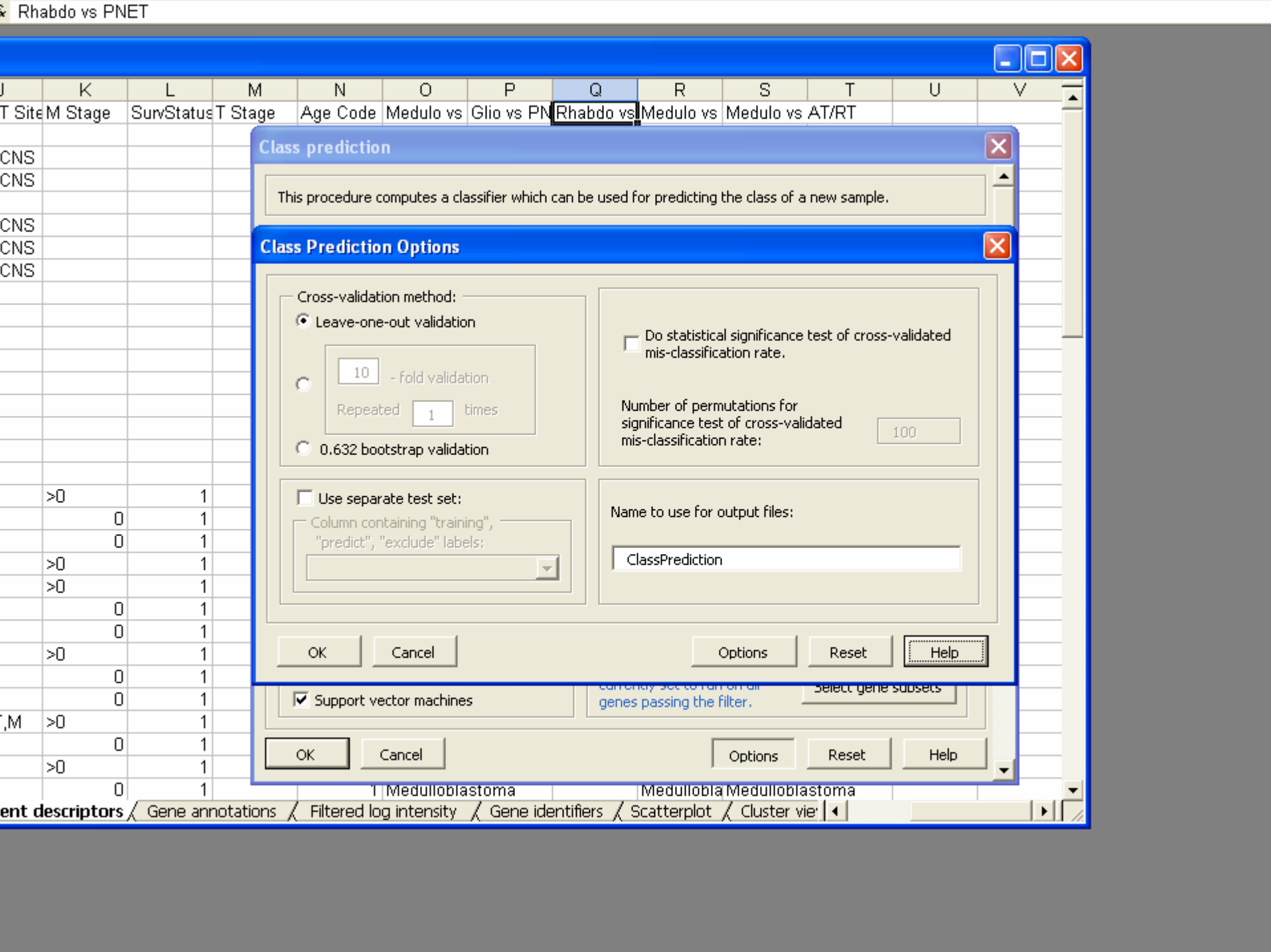
When There Are More Than 2 Classes

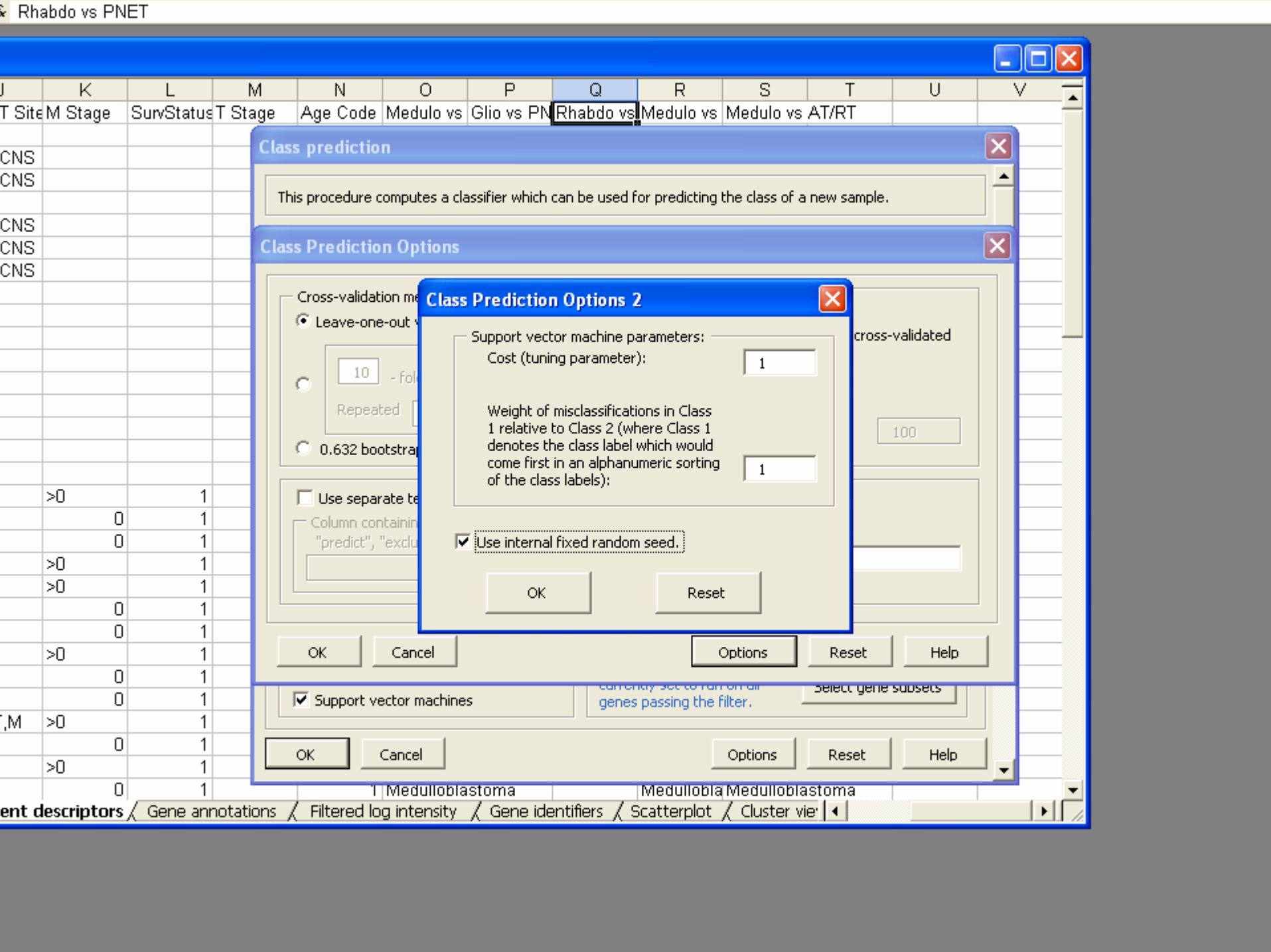
- Nearest neighbor type methods
- Decision tree of binary classifiers

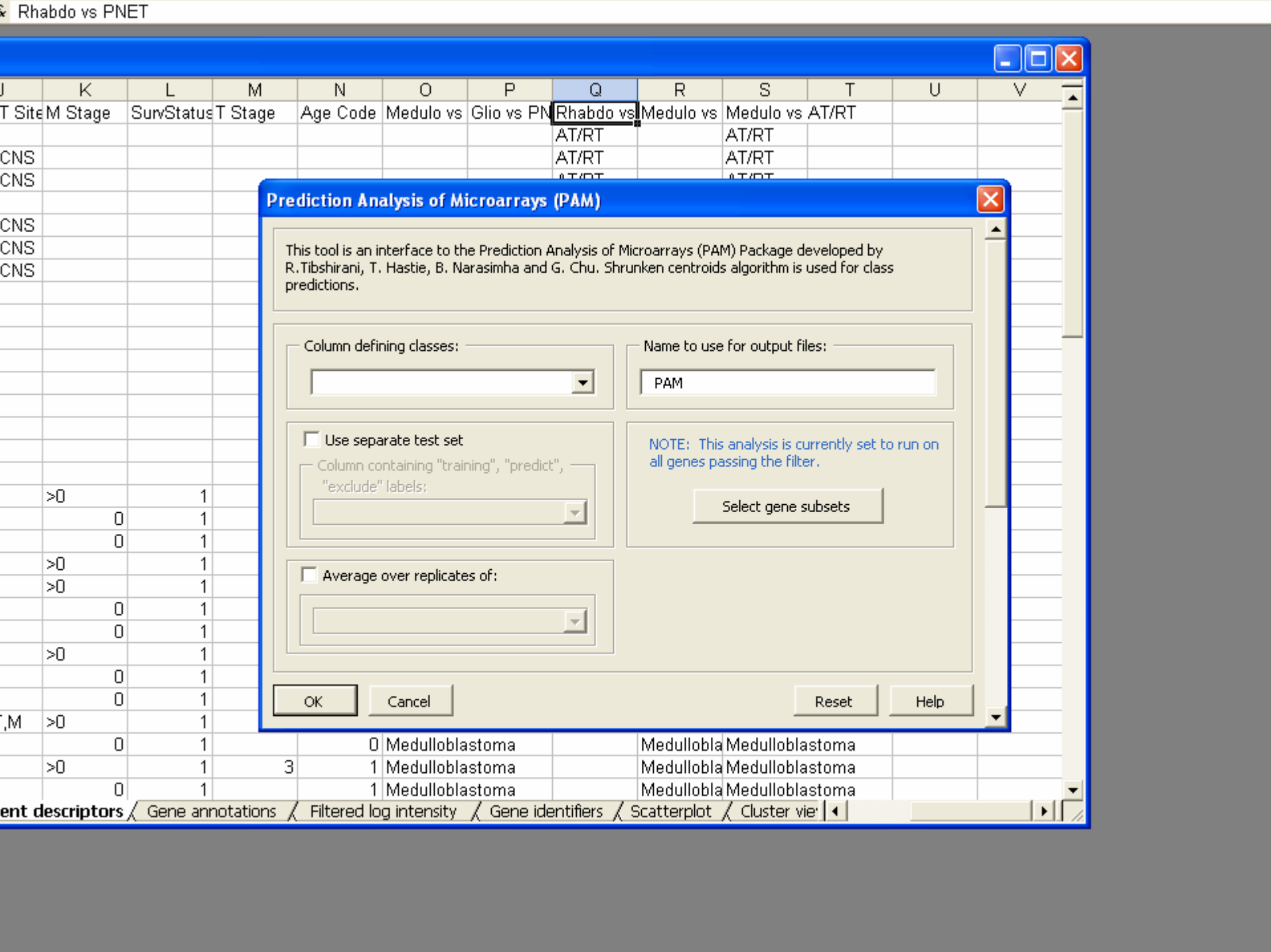
Decision Tree of Binary Classifiers

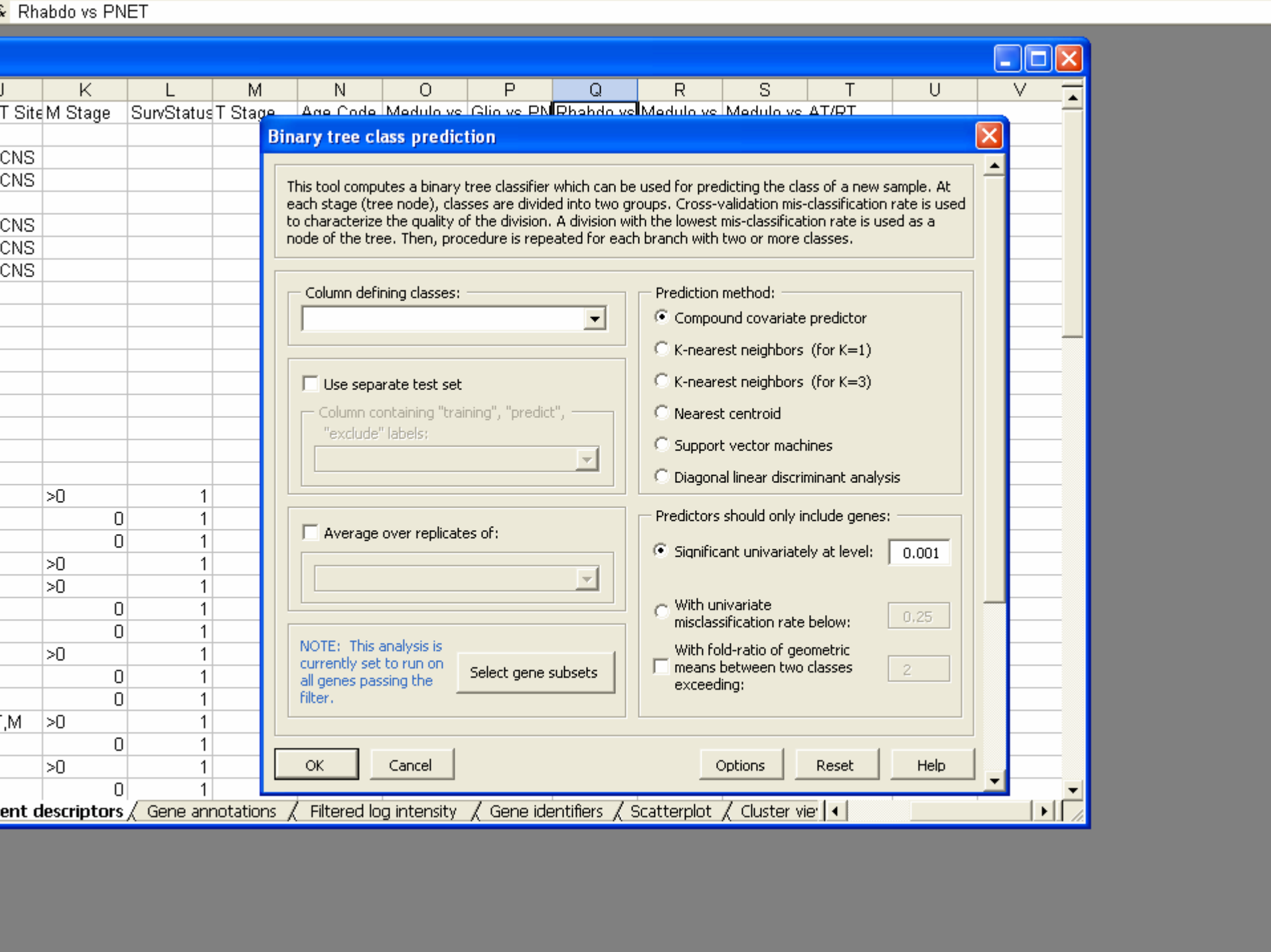
- Partition the set of classes {1,2,…,K} into two disjoint subsets S_1 and S_2
- Develop a binary classifier for distinguishing the composite classes S_1 and S_2
    - Compute the cross-validated classification error for distinguishing S_1 and S_2
- Repeat the above steps for all possible partitions 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

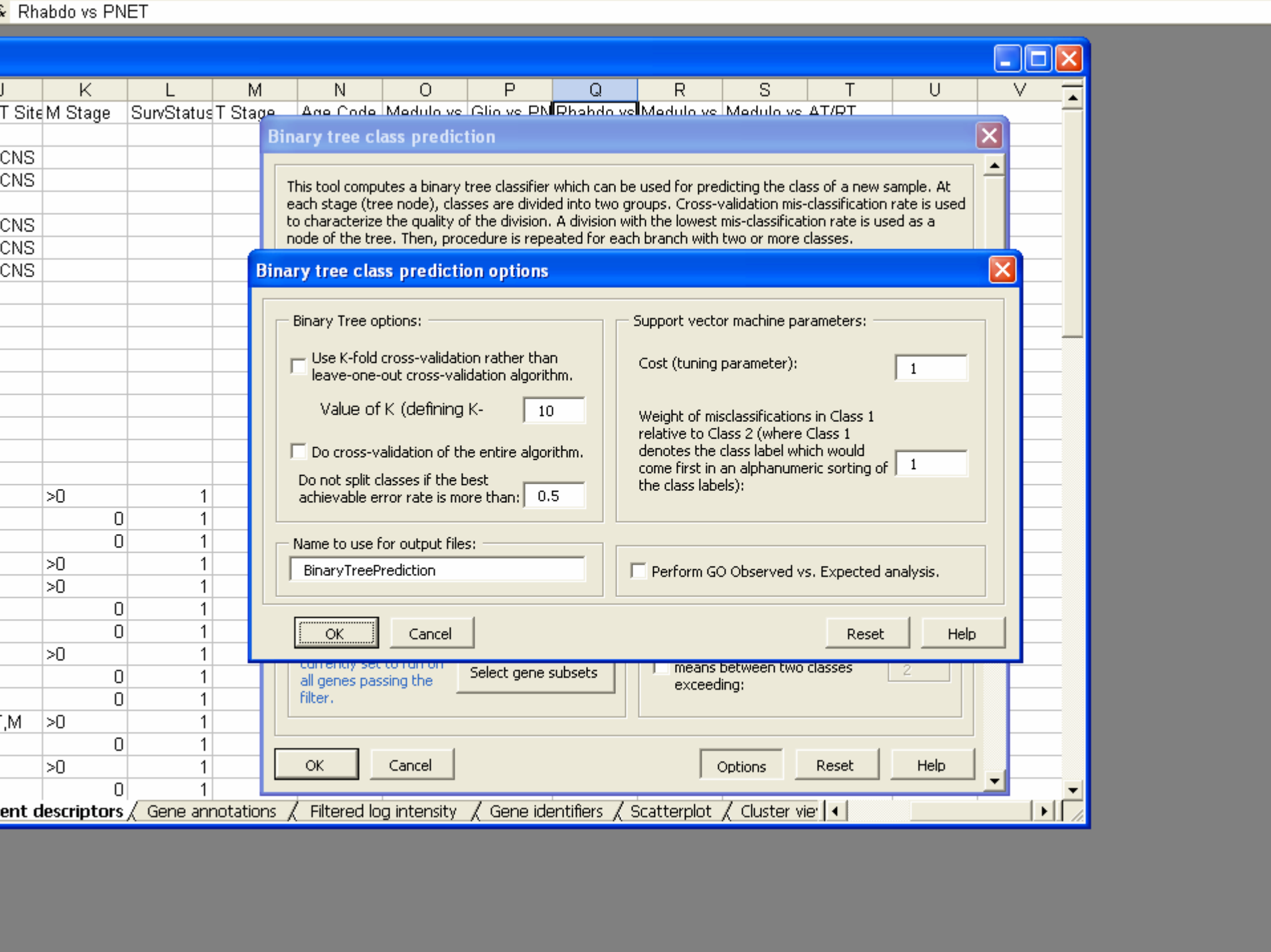

Rhabdo vs PNET
J
K
L
M
N
O
P
Q
R
S
T
U
V
T Site
M Stage
SurvStatus
T Stage
Age Code
Medulo vs
Glio vs PN
Rhabdo vs
Medulo vs
Medulo vs
AT/RT
CNS
CNS
CNS
CNS
CNS
>0
1
0
1
0
1
>0
1
>0
1
0
1
0
1
>0
1
0
1
0
1
,M
>0
1
0
1
>0
1
0
1
T
Medulloblastoma
Medullobla
Medulloblastoma
Class prediction
This procedure computes a classifier which can be used for predicting the class of a new sample.
Column defining classes:
Use random variance model for univariate tests.
Average over replicates of:
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
Arrays are paired between classes.
Pair samples by:
With fold-ratio of geometric means between two classes exceeding:
2
Prediction methods:
Compound covariate predictor
Diagonal linear discriminant analysis
K-nearest neighbors (for K=1 and 3)
Nearest centroid
Support vector machines
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.
Select gene subsets
OK
Cancel
Options
Reset
Help
ent descriptors
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie

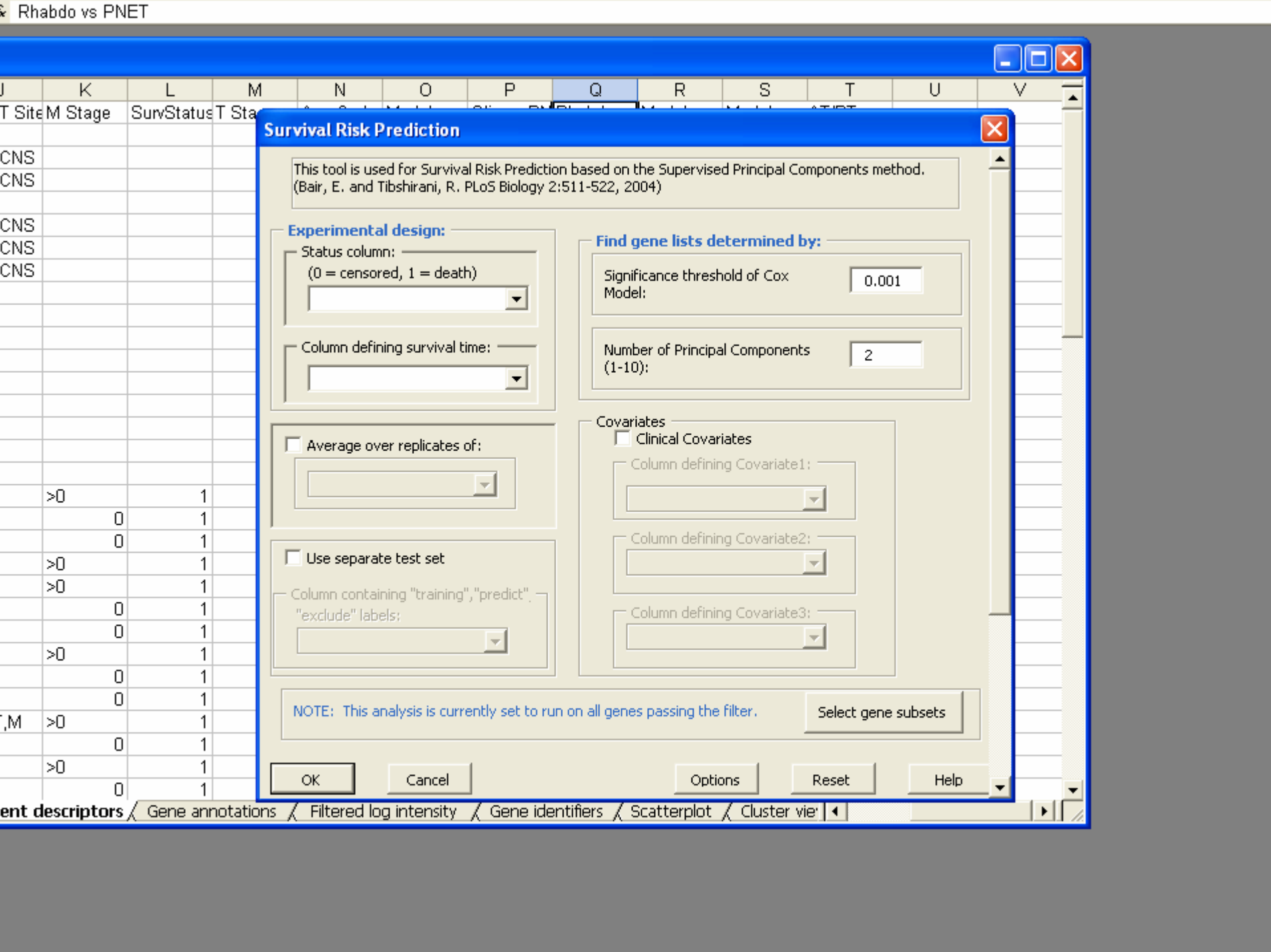

Rhabdo vs PNET
T Site
M Stage
SurvStatus
T Stage
Age Code
Medulo vs
Glio vs PN
Rhabdo vs
Medulo vs
Medulo vs
AT/RT
Class prediction
This procedure computes a classifier which can be used for predicting the class of a new sample.
Class Prediction Options
Cross-validation method:
Leave-one-out validation
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
OK
Cancel
Options
Reset
Help
Support vector machines
genes passing the filter.
Select gene subsets
OK
Cancel
Options
Reset
Help
Medulloblastoma
Medulloblastoma
ent descriptors
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie

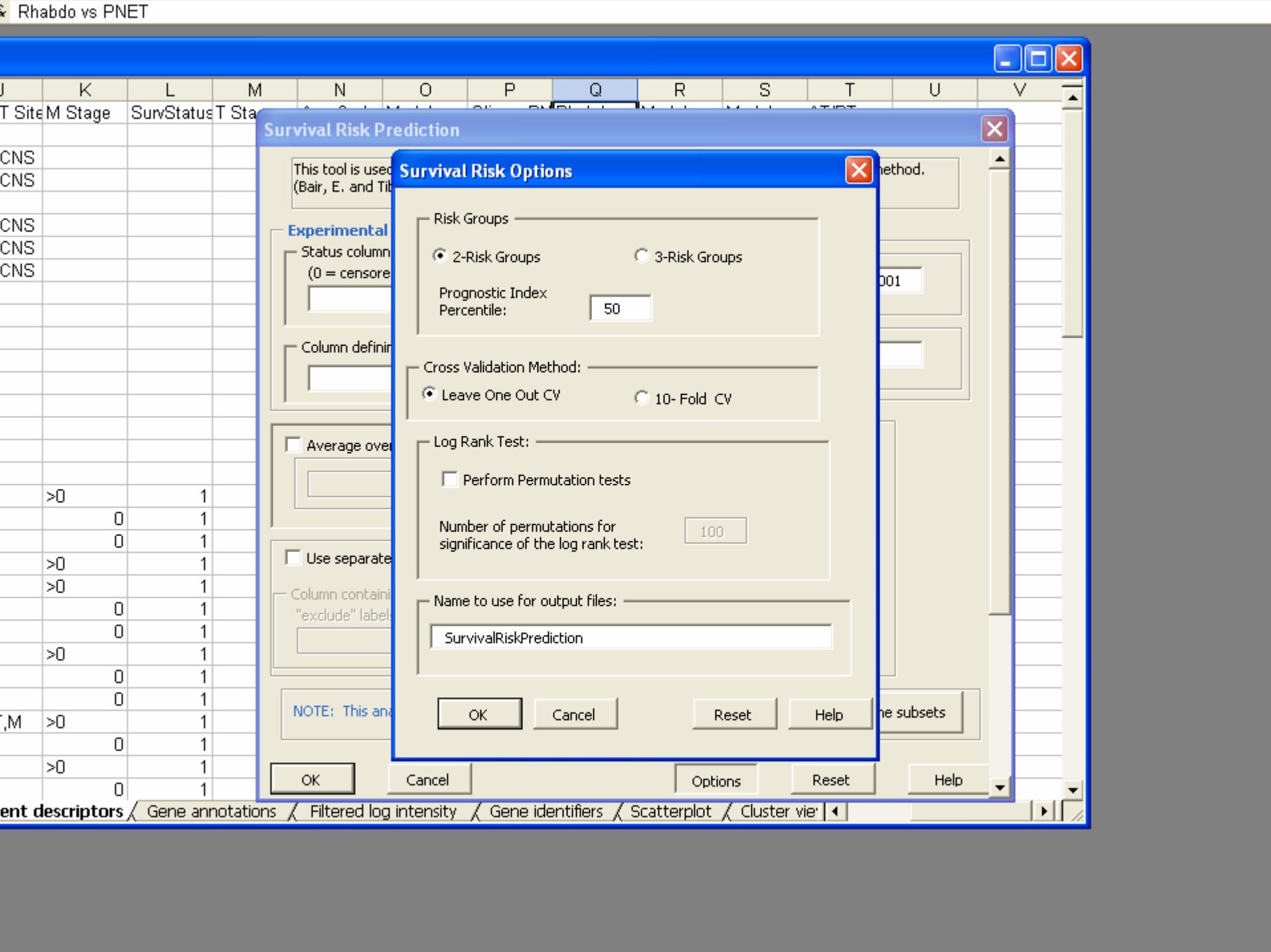

Rhabdo vs PNET
T Site
M Stage
SurvStatus
T Stage
Age Code
Medulo vs
Glio vs PN
Rhabdo vs
Medulo vs
Medulo vs
AT/RT
Class prediction
This procedure computes a classifier which can be used for predicting the class of a new sample.
Class Prediction Options
Cross-validation me
Leave-one-out
10
Repeated
0.632 bootstra
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.
OK
Reset
OK
Cancel
Options
Reset
Help
Support vector machines
genes passing the filter.
Select gene subsets
OK
Cancel
Options
Reset
Help
Medulloblastoma
Medulloblastoma
ent descriptors
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie

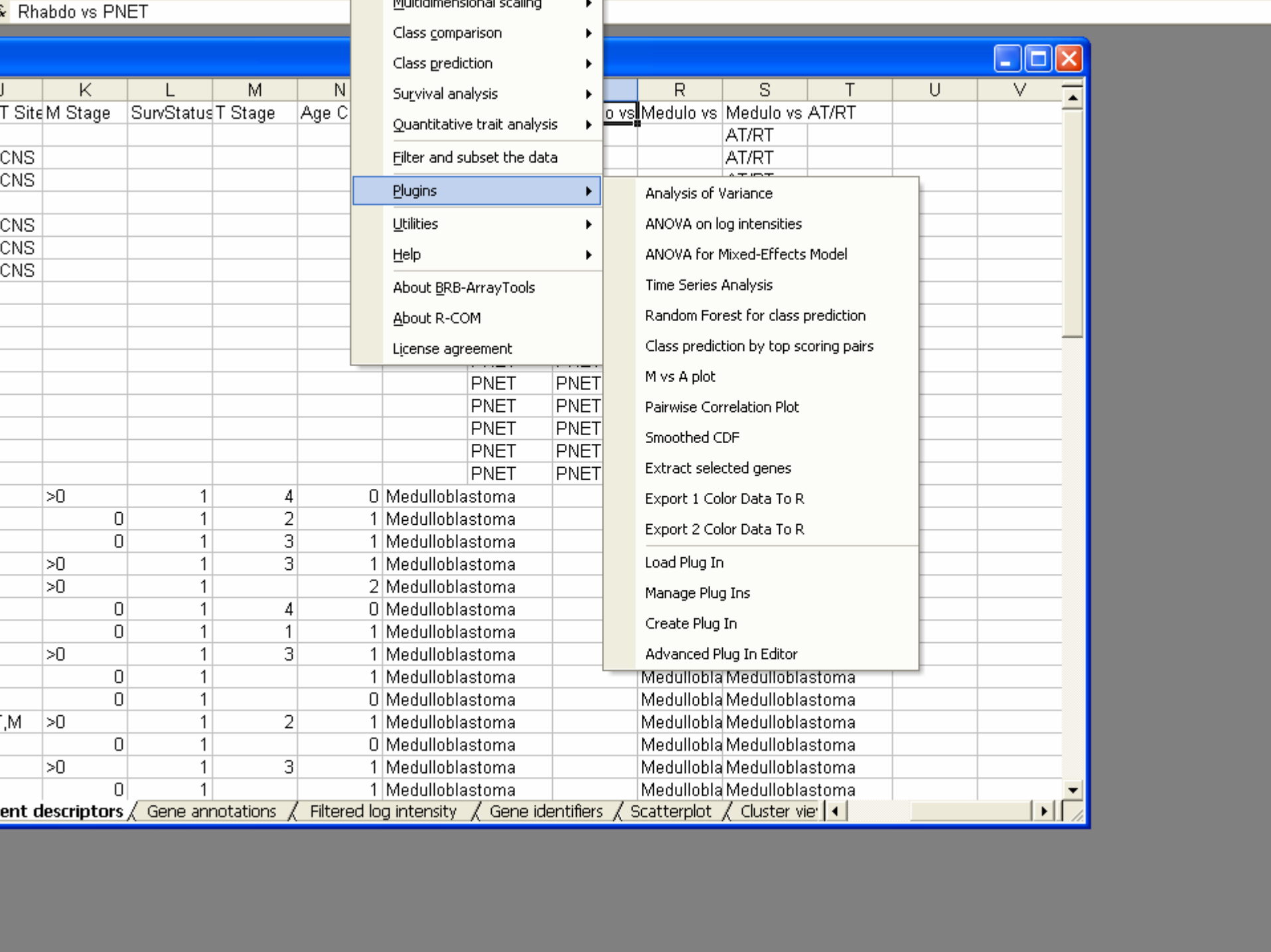

Rhabdo vs PNET
Prediction Analysis of Microarrays (PAM)
This tool is an interface to the Prediction Analysis of Microarrays (PAM) Package developed by R.Tibshirani, T. Hastie, B. Narasimha and G. Chu. Shrunken centroids algorithm is used for class predictions.
Column defining classes:
Name to use for 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:
OK
Cancel
Reset
Help
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot

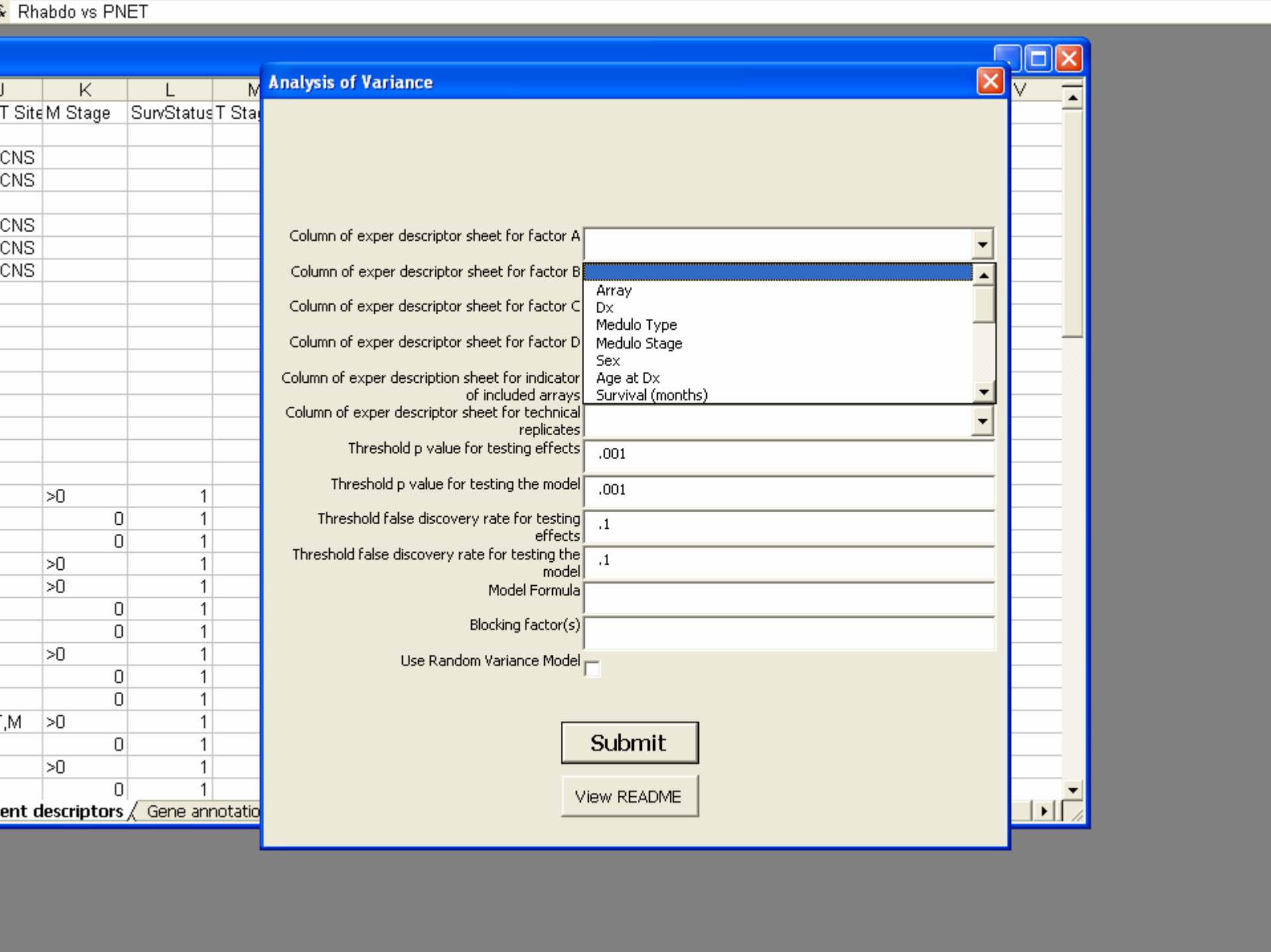

Rhabdo vs PNET
Binary tree class prediction
This tool computes a binary tree classifier which can be used for predicting the class of a new sample. At each stage (tree node), classes are divided into two groups. Cross-validation mis-classification rate is used to characterize the quality of the division. 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.
Select gene subsets
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
OK
Cancel
Options
Reset
Help
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot


Rhabdo vs PNET
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.
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
Select gene subsets
means between two classes exceeding:
2
OK
Cancel
Options
Reset
Help
T Site
M Stage
SurvStatus
T Stage
Age Code
Medulo vs
Glio vs PN
Rhabdo vs
Medulo vs
Medulo vs AT/RT
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie


Rhabdo vs PNET
T Site
M Stage
SurvStatus
T Sta
CNS
CNS
CNS
CNS
CNS
>0
0
0
>0
>0
0
0
>0
0
0
>0
0
>0
0
ent descriptors
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie
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.
Select gene subsets
OK
Cancel
Options
Reset
Help


Rhabdo vs PNET
M Stage
SurvStatus
Survival Risk Prediction
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
OK
Cancel
Reset
Help
Experimental
Status column
(0 = censore
Column definin
Average over
Use separate
Column containi
"exclude" label
NOTE: This ana
Options
ne subsets
ent descriptors
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie


Rhabdo vs PNET
Multidimensional scaling
Class comparison
Class prediction
Survival analysis
Quantitative trait analysis
Filter and subset the data
Plugins
Utilities
Help
About BRB-ArrayTools
About R-COM
License agreement
Analysis of Variance
ANOVA on log intensities
ANOVA for Mixed-Effects Model
Time Series Analysis
Random Forest for class prediction
Class prediction by top scoring pairs
M vs A plot
Pairwise Correlation Plot
Smoothed CDF
Extract selected genes
Export 1 Color Data To R
Export 2 Color Data To R
Load Plug In
Manage Plug Ins
Create Plug In
Advanced Plug In Editor
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot

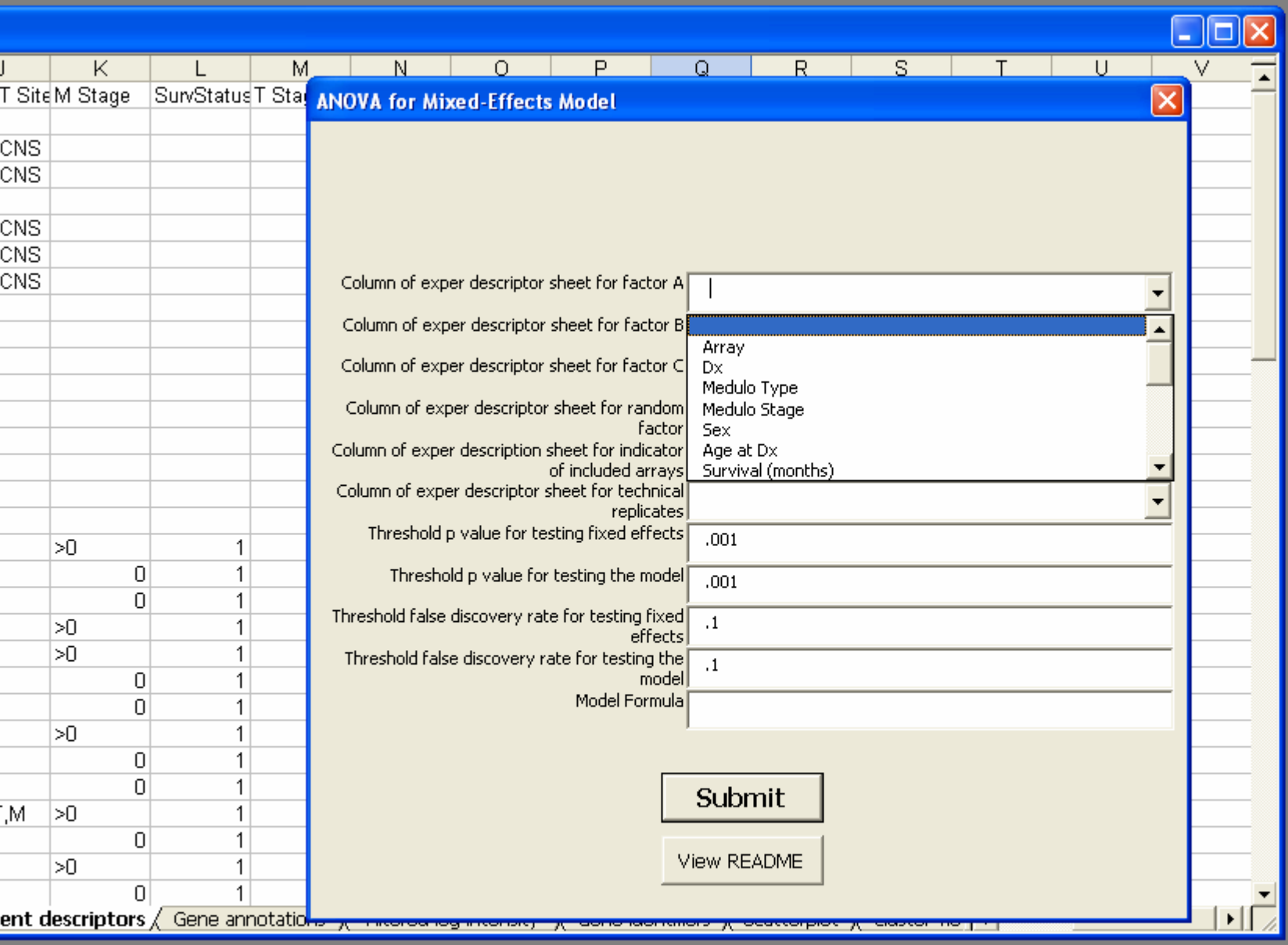

Rhabdo vs PNET
Analysis of Variance
Column of exper descriptor sheet for factor A
Column of exper descriptor sheet for factor B
Column of exper descriptor sheet for factor C
Column of exper descriptor sheet for factor D
Column of exper description sheet for indicator of included arrays
Column of exper descriptor sheet for technical replicates
Threshold p value for testing effects
.001
Threshold p value for testing the model
.001
Threshold false discovery rate for testing effects
.1
Threshold false discovery rate for testing the model
.1
Model Formula
Blocking factor(s)
Use Random Variance Model
Array
Dx
Medulo Type
Medulo Stage
Sex
Age at Dx
Survival (months)
Submit
View README
T Site
M Stage
SurvStatus
T Sta
CNS
>0
0
1
ent descriptors
Gene annotatio

Rhabdo vs DNET

anova.txt - Notepad

File Edit Format View Help

```
**********************************
**  NAME OF THIS PLUG-IN:       **
**********************************

Analysis of Variance (ANOVA) for each gene.


**********************************
**  PURPOSE OF THIS PLUG-IN:    **
**********************************

This plug-in performs an analysis of variance for relating the
log-ratio or log-signal expression to specified factors. A separate
ANOVA model is fitted for each gene. All factors are considered fixed
effects.

The F-test used for statistical significance testing is based on the likelihood
ratio test or type III sum of squares in SAS terminology. That means that the
significance of each factor is adjusted for all other factors of the model.


**********************************
**  USING THIS PLUG-IN:         **
**********************************

To run this function, the user should input the following:
. Column of exper descriptor sheet for factor A
. Column of exper descriptor sheet for factor B (can be empty)
. Column of exper descriptor sheet for factor C (can be empty)
. Column of exper descriptor sheet for factor D (can be empty)
. Column of exper descriptor sheet for indicator of included arrays
  (can be empty)
. Column of exper descriptor sheet for technical replicates
  (can be empty)
. Threshold p value for testing effects
. Threshold p value for testing the model
. Threshold false discovery rate (Benjamini & Hochberg, 1995) for testing effects
. Threshold false discovery rate for testing the model
. Model formula (e.g. A+B+A:B). See next section for details how to
  specify a model formula.
. Blocking factors (e.g. B or B,C)
. Use Random Variance Model (a checkbox)

The "Column of exper descriptor sheet for indicator of included arrays"
is used to include arrays we are only interested in. For arrays we
don't want them to be included in analyses, we should leave blank value
in this column. We can put any value other than blank in this column
for arrays we are interested in. If nothing is specified in the dialog,
all arrays with non-empty factor labels will be used.
```


J
K
L
M
N
O
P
Q
R
S
T
U
V
T Site
M Stage
SurvStatus
T Sta
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
,M
>0
1
0
1
>0
1
0
1
ent descriptors
Gene annotation
ANOVA for Mixed-Effects Model
Column of exper descriptor sheet for factor A
Column of exper descriptor sheet for factor B
Column of exper descriptor sheet for factor C
Column of exper descriptor sheet for random factor
Column of exper description sheet for indicator of included arrays
Column of exper descriptor sheet for technical replicates
Threshold p value for testing fixed effects
.001
Threshold p value for testing the model
.001
Threshold false discovery rate for testing fixed effects
.1
Threshold false discovery rate for testing the model
.1
Model Formula
Array
Dx
Medulo Type
Medulo Stage
Sex
Age at Dx
Survival (months)
Submit
View README

Rhabdo vs PNET

J	K	L	M
T Site	M Stage	SurvStatus	T Stag
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 /

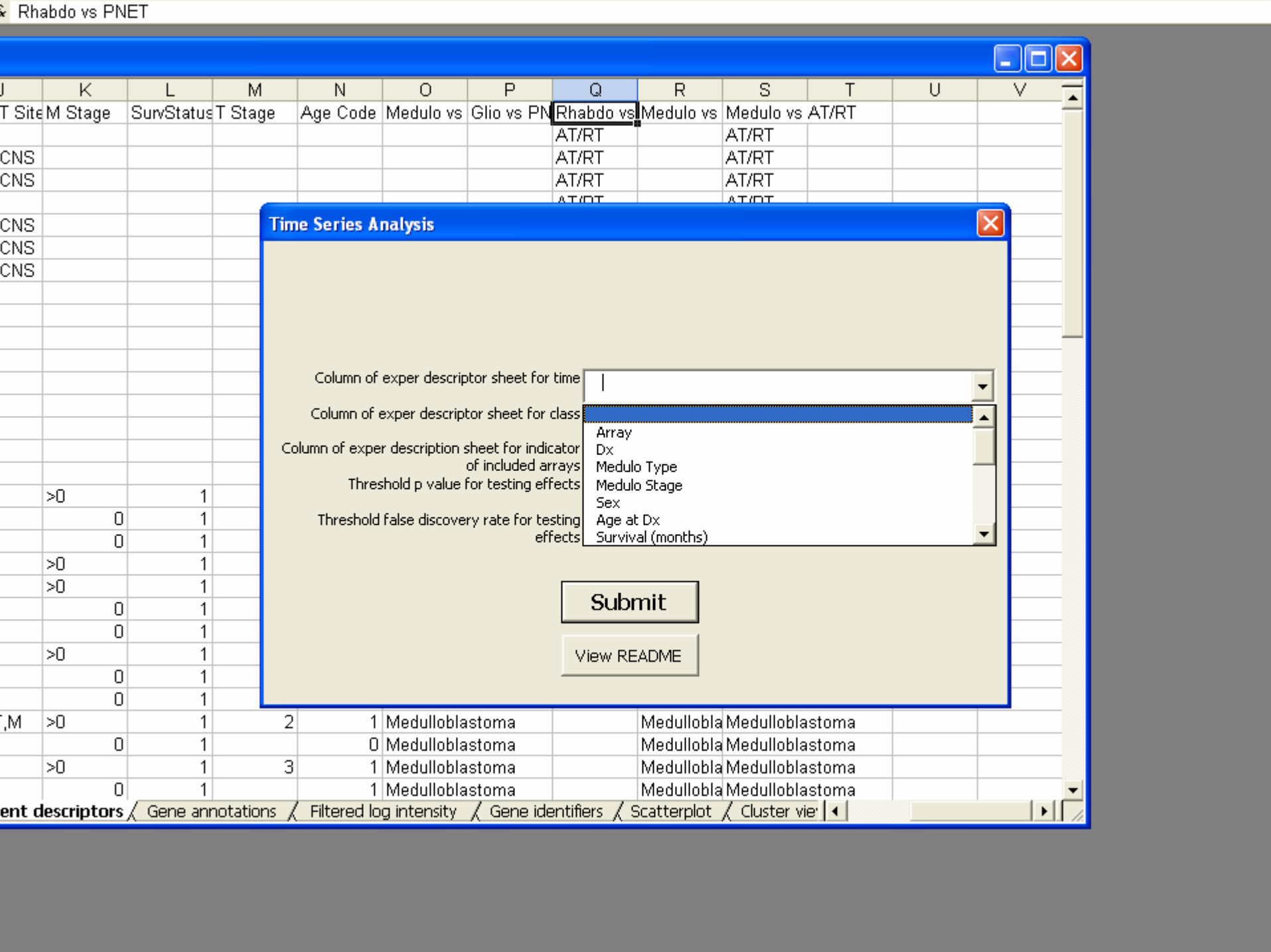
ANO

C
C
C
C
Col
Co
Thr
T

anovamix.txt - Notepad

File Edit Format View Help

```
************************************
**  NAME OF THIS PLUG-IN:     **
************************************

Linear mixed-effects model for each gene.


************************************
**  PURPOSE OF THIS PLUG-IN:  **
************************************

This plug-in fits data to a linear mixed-effects model for each gene
and computes ANOVA for a user's specified model. Only one random factor
can be specified in the mixed-effects model. This model is useful for
time series data with specimens collected from each subject at multiple
time points. The subjects may have different characteristics of interest,
such as diseased or normal, treated or non-treated, male or female, but
the genes differentially expressed among the subjects not related to those
specified factors may not be of interest. Although a fixed effects model
such as in the basic ANOVA provided in another plug-in could be used, if there
are many subjects, then considering the subjects as a random factor can provide
more degrees of freedom for error estimation and potentially greater statistical
power for testing the effects of interest. This mixed-effects model is, however,
much more computationally intensive than the standard fixed effects model provided
in the other plug-in.

The F-test is based on the likelihood ratio test or type III sum of
squares in SAS's terminology. That means that the
significance of each factor is adjusted for all other factors of the model.


************************************
**  USING THIS PLUG-IN:       **
************************************

To run this function, the user should input the following:
. Column of exper descriptor sheet for factor A
. Column of exper descriptor sheet for factor B (can be empty)
. Column of exper descriptor sheet for factor C (can be empty)
. Column of exper descriptor sheet for random factor
. Column of exper descriptor sheet for indicator of included arrays
  (can be empty)
. Column of exper descriptor sheet for technical replicates
  (can be empty)
. Threshold p value for testing fixed effects
. Threshold p value for testing the model
. Threshold false discovery rate (Benjamini & Hochberg, 1995) for testing fixed effects
. Threshold false discovery rate for testing the model
. Model formula (e.g. A+B+A:B).
```


Rhabdo vs PNET
Time Series Analysis
Column of exper descriptor sheet for time
Column of exper descriptor sheet for class
Column of exper description sheet for indicator of included arrays
Threshold p value for testing effects
Threshold false discovery rate for testing effects
Array
Dx
Medulo Type
Medulo Stage
Sex
Age at Dx
Survival (months)
Submit
View README
Rhabdo vs
Medulo vs
Medulo vs
AT/RT
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot

Rhabdo vs PNET

J	K	L	M	
T Site	M Stage	SurvStatus	T Stage	Age
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		
,M	>0	1	2	
	0	1		
	>0	1	3	
	0	1		

ent descriptors / Gene annotations / Filte

Time Ser

Colu

Colu

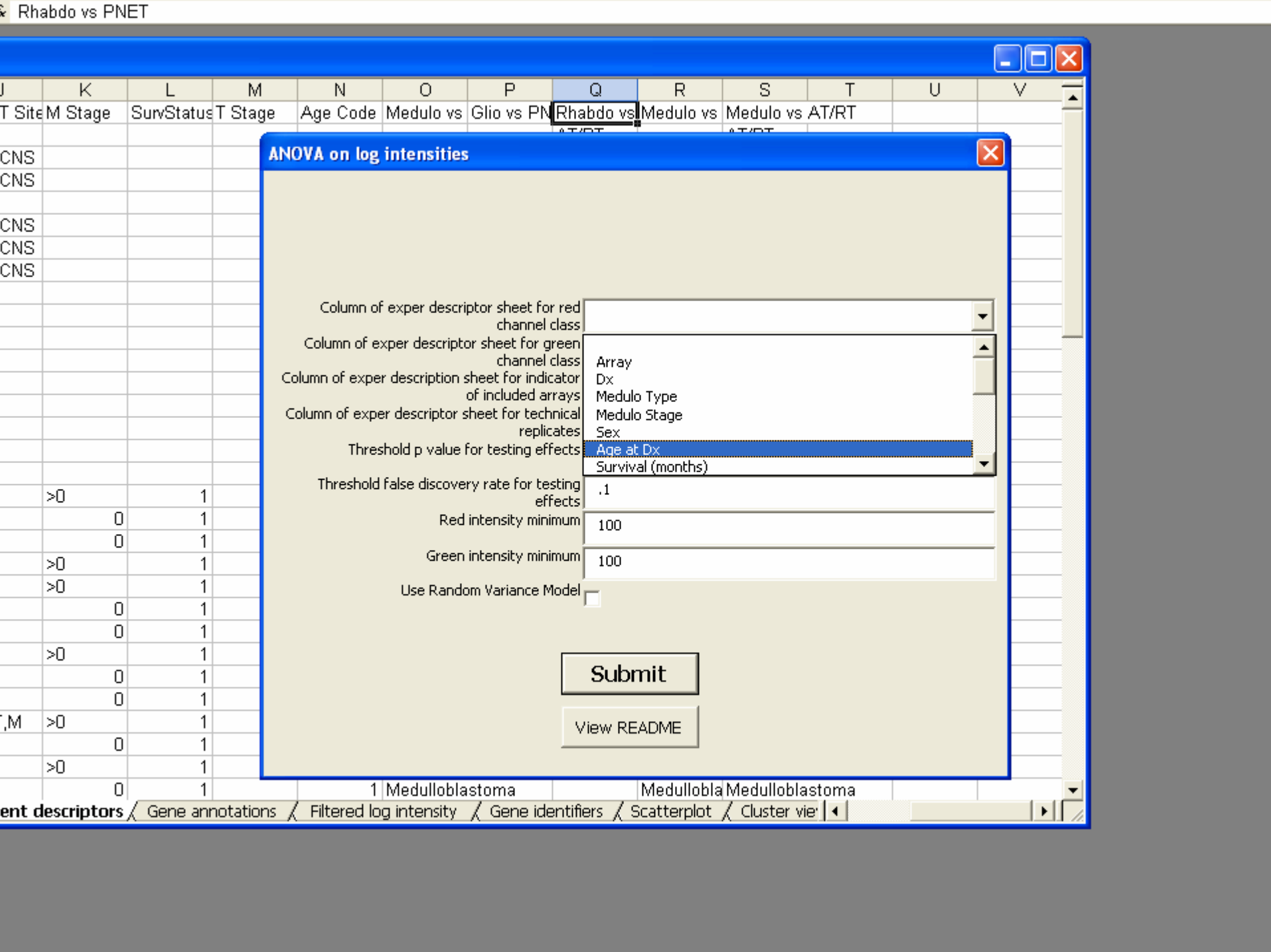
Column c

Thr

timeseries.txt - Notepad

File Edit Format View Help

```
************************************
**  NAME OF THIS PLUG-IN:     **
************************************

Time Series Analysis.


************************************
**  PURPOSE OF THIS PLUG-IN:  **
************************************

This plug-in can be used for regression analysis of time series
expression data. In its simplest form (model A), the genes whose
expression are varying over time are identified. A quadratic function
is fit to the expression data of each gene and the hypothesis is that the
linear and quadratic coefficients are simultaneously zero. The genes for
which this hypothesis is rejected are identified. The tests are performed
at a significance level specified by the user and also at a false discovery
rate (FDR) specified by the user. Two lists of significant genes are
produced, one for the specified significance level threshold and one for
the FDR threshold. To fit this model, the user must provide a column in
the experiment descriptor worksheet specifying the time point for each
array. This column should be strictly numeric and should not contain
alphabetic characters. The entry in the column should be blank if the
array is to be excluded from the analysis. The arrays at the same time
points can represent either technical or biological replicates, but
the two kinds of replicates should not be combined in the same analysis.
This plug-in is not appropriate for nested data where the same subject is
sampled at different time points.

Model B is for identifying genes that are changing over time, but
where there is a class variable to adjust for. For example, there could
be two strains of mice included in the experiment or arrays were from
two different print set batches. For model B it is assumed that the
variation in gene expression over time is the same for each class. The
output also indicates which genes are differentially expressed among
the classes uniformly over time.

Model C is similar to model B but the variation in gene expression over
time is permitted to differ among the classes. The output of model C
identifies these genes for which the variation over time is different
for different levels of the class variable. These genes are identified
based on the user specified significance level and based on the user
specified FDR. For genes whose variation over time does not significantly
vary among classes, model B is fit to determine whether the gene is varying
over time uniformly for each classes. Model C is useful for experiments
where the class variable represents a treatment indicator.

For data without a class variable, the ANOVA model takes the form:
```


Rhabdo vs PNET
ANOVA on log intensities
Column of exper descriptor sheet for red channel class
Column of exper descriptor sheet for green channel class
Column of exper description sheet for indicator of included arrays
Column of exper descriptor sheet for technical replicates
Threshold p value for testing effects
Threshold false discovery rate for testing effects
Red intensity minimum
Green intensity minimum
Use Random Variance Model
Array
Dx
Medulo Type
Medulo Stage
Sex
Age at Dx
Survival (months)
.1
100
100
Submit
View README
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie

Rhabdo vs PNET

J	K	L	M	
T Site	M Stage	SurvStatus	T Stage	Age
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 / Filte

ANOVA c

Co

Colur

Column c

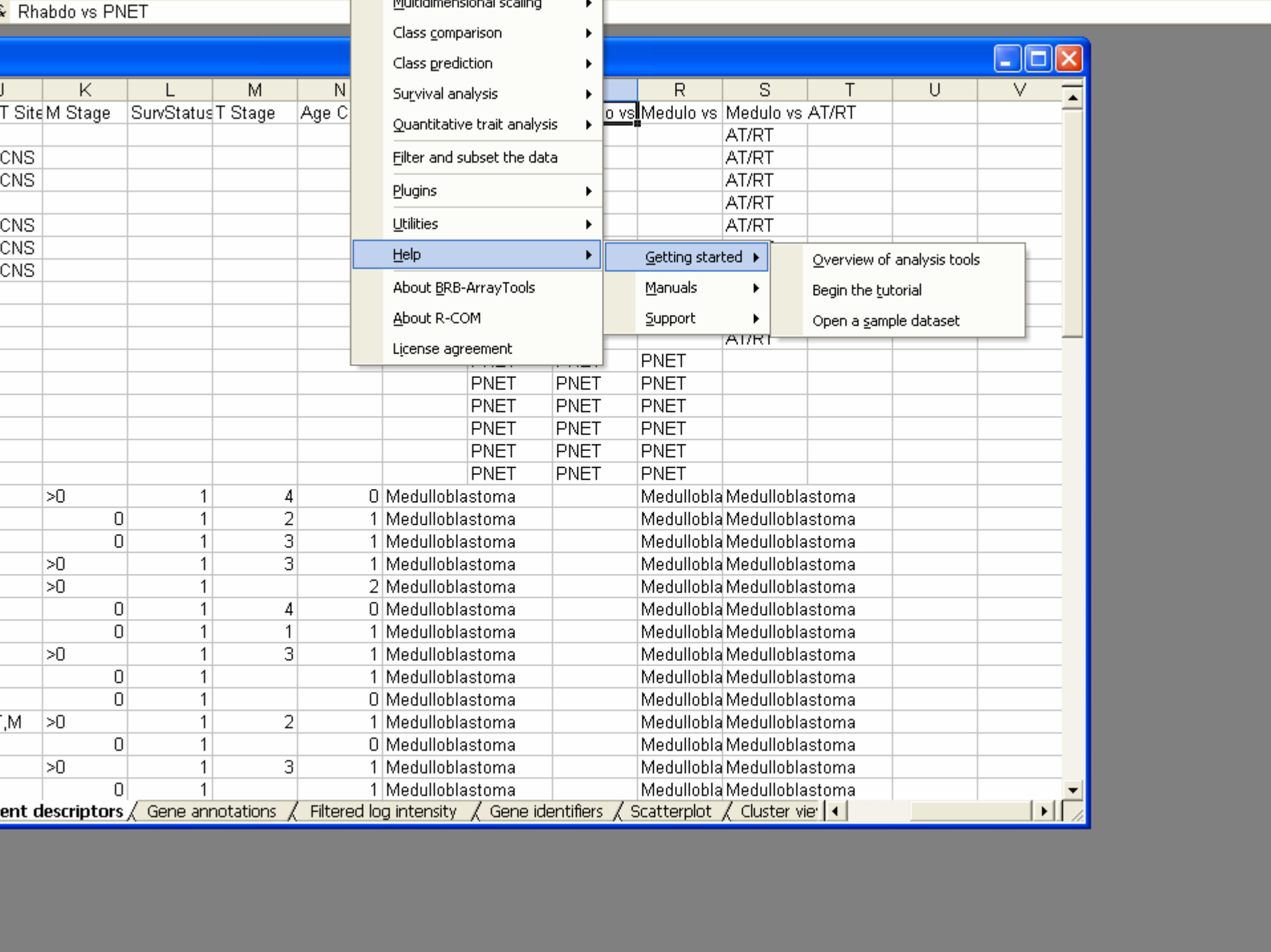
Column

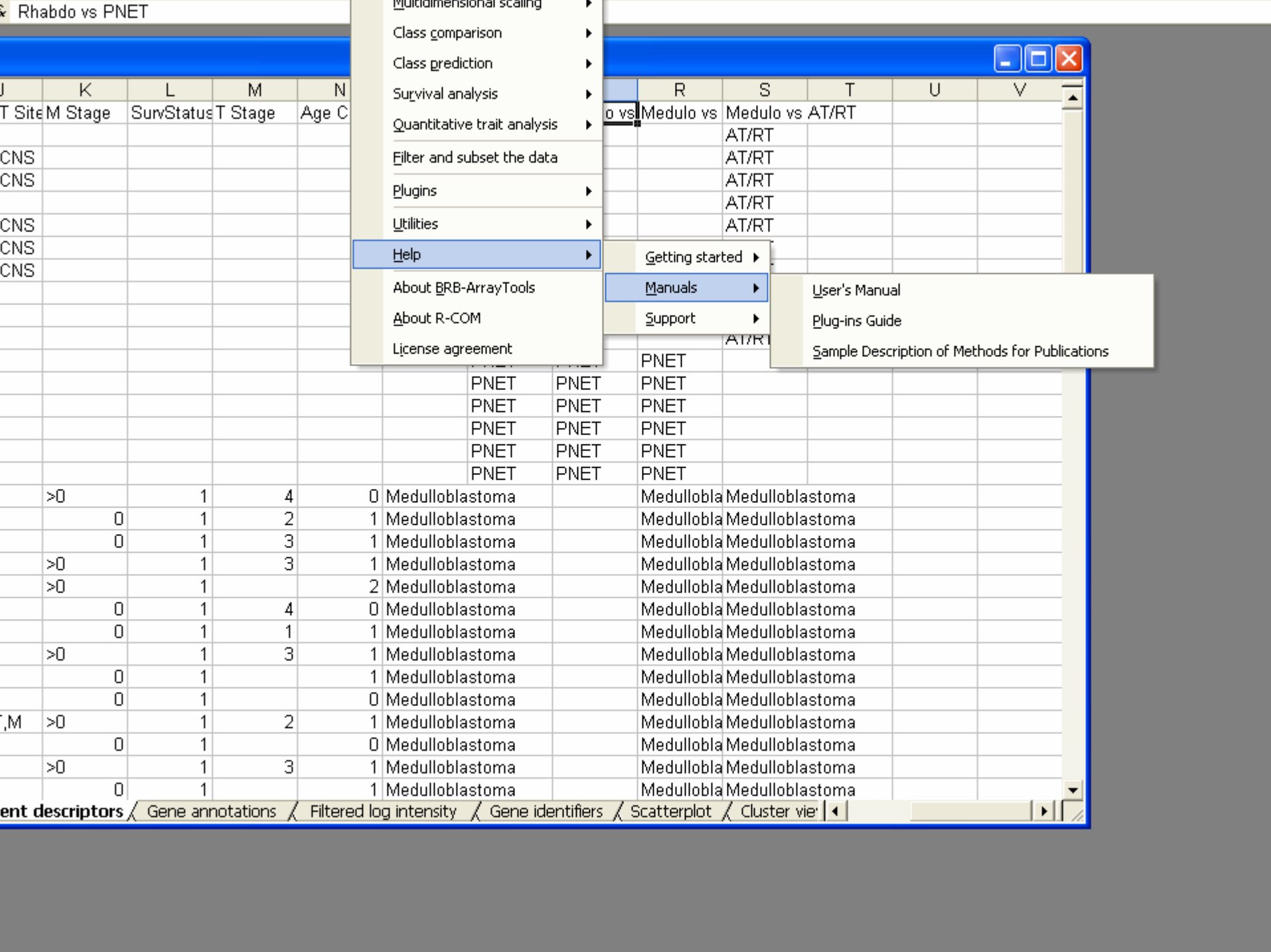
Thr

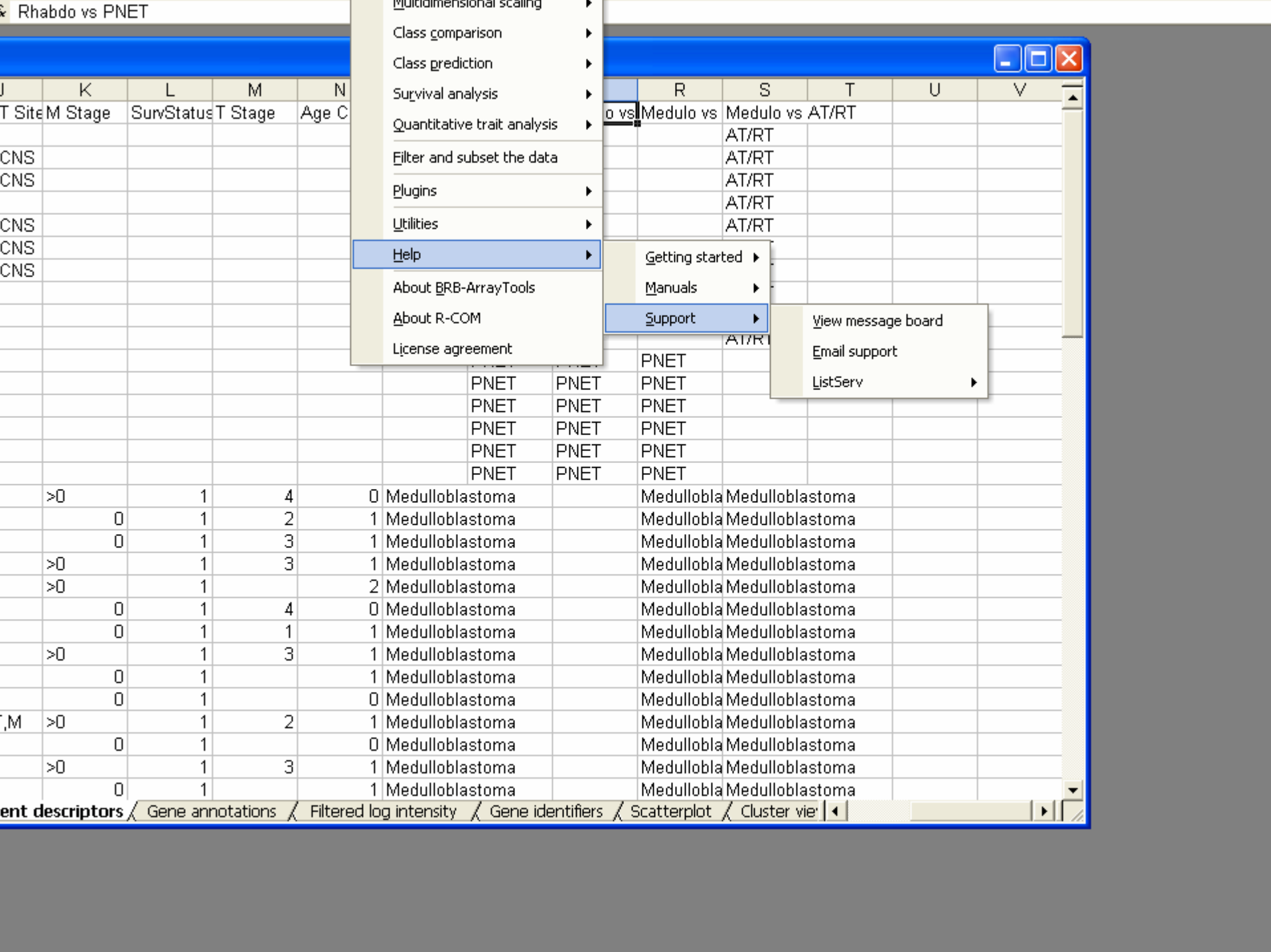
anovachnnl.txt - Notepad

File Edit Format View Help

```
************************************
**  NAME OF THIS PLUG-IN:     **
************************************

ANOVA on log intensities for each gene.


************************************
**  PURPOSE OF THIS PLUG-IN:  **
************************************

There are two steps for running ANOVA in this plugin.
The first step is to normalize log intensity for each channel
and the second step is to do ANOVA on normalized log intensity.

Step1: Normalize each log intensity by the following
       normalization model (underline denotes subscript):

         y_{adcg} = mu + A_a + AD_{ad} + e1_{adcg} ---- (1)

       where
           y_{adcg} is the log intensity,
           mu is the overall log intensity mean value,
           A_a is the effect of the array a,
           AD_{ad} is the interaction of array a and dye d,
           e1_[adcg} is the random noise,
           c is the index of class(variety).

       We assume each effects are fixed. So {A_a} and {AD_{ad}}
       satisfy some identification conditions.

       After fitting the normlization model, the residuals
       (normalized log intensity), r_{adcg} are obtained,

         r_{adcg} = y_{adcg} - \hat{mu} - \hat{A}_a - \hat{AD}_{ad} ---- (2)

Step2: Fit the normalized log intensity by the following ANOVA model

         r_{adcg} = mu_g + alpha_{ag} + beta_{dg} + class_{cg} + e2_{adcg}  ---- (3)

       where
           mu_g is the gene-specific average log intensity,
           alpha_{ag} is the gene-specific array effect (spot effect),
           beta_{dg} is the gene-specific dye effect,
           class_{cg} is the gene-specific class (variety) effect,
           e2_[adcg} is the random noise.

       Again, we assume each effects in model (3) are fixed.
```


Rhabdo vs PNET
Multidimensional scaling
Class comparison
Class prediction
Survival analysis
Quantitative trait analysis
Filter and subset the data
Plugins
Utilities
Help
About BRB-ArrayTools
About R-COM
License agreement
Getting started
Manuals
Support
Overview of analysis tools
Begin the tutorial
Open a sample dataset
Medulo vs
Medulo vs AT/RT
AT/RT
PNET
Medulloblastoma
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot


Rhabdo vs PNET
Multidimensional scaling
Class comparison
Class prediction
Survival analysis
Quantitative trait analysis
Filter and subset the data
Plugins
Utilities
Help
About BRB-ArrayTools
About R-COM
License agreement
Getting started
Manuals
Support
User's Manual
Plug-ins Guide
Sample Description of Methods for Publications
T Site
M Stage
SurvStatus
T Stage
Age C
Medulo vs
AT/RT
PNET
Medulloblastoma
ent descriptors
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie


Rhabdo vs PNET
Multidimensional scaling
Class comparison
Class prediction
Survival analysis
Quantitative trait analysis
Filter and subset the data
Plugins
Utilities
Help
About BRB-ArrayTools
About R-COM
License agreement
Getting started
Manuals
Support
View message board
Email support
ListServ
Rhabdo vs
Medulo vs
Medulo vs
AT/RT
PNET
Medulloblastoma
ent descriptors
Gene annotations
Filtered log intensity
Gene identifiers
Scatterplot
Cluster vie

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.

Evaluating a Classifier

- Fit of a model to the same data used to develop it is no evidence of prediction accuracy for independent data
  - Goodness of fit vs prediction accuracy
- Demonstrating statistical significance of prognostic factors is not the same as demonstrating predictive accuracy
- Demonstrating stability of identification of gene predictors is not necessary for demonstrating predictive accuracy

Evaluating a Classifier

- The classification algorithm includes the following parts:
  - Determining what type of classifier to use
  - Gene selection
  - Fitting parameters
  - Optimizing with regard to tuning parameters
- If a re-sampling method such as cross-validation is to be used to estimate predictive error of a classifier, **all** aspects of the classification algorithm must be repeated for each training set and the accuracy of the resulting classifier scored on the corresponding validation set

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
  - Ideally test set data is from different centers than the training data and assayed at a different time

Leave-one-out Cross Validation

- 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

Leave-one-out Cross Validation

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

- Cross validation is only valid if the test 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 feature selection must be repeated for each leave-one-out training set.
- The cross-validated estimate of misclassification error is an estimate of the prediction error for model fit using specified algorithm to full dataset
- If you use cross-validation estimates of prediction error for a set of algorithms indexed by a tuning parameter and select the algorithm with the smallest cv error estimate, you do not have a valid estimate of the prediction error for the selected model

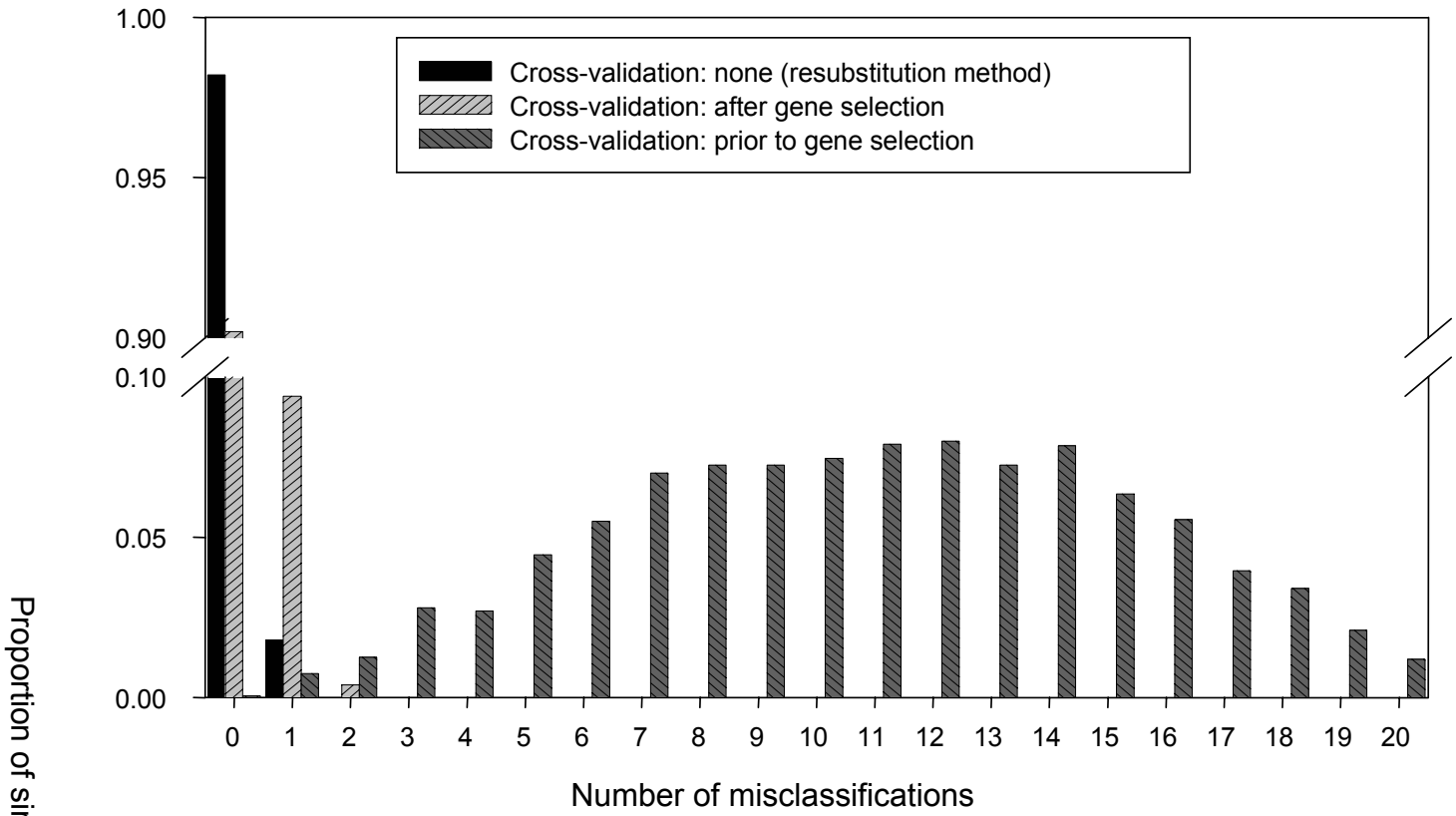
Prediction on Simulated Null Data

Generation of Gene Expression Profiles

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


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

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Prediction Error Estimation: A Comparison of Resampling Methods

Annette M. Molinaro [ab]*, 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

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

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

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

- Publications are using all the data to select genes and then cross-validating only the parameter estimation 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
    - Some software encourages inappropriate cross-validation

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 model 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

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 , …, b_k denote the estimated regression coefficients
- To predict for case with expression profile vector x, compute the k supervised pc's y_1 , …, 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

Sample Size Planning References

- K Dobbin, R Simon. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 6:27-38, 2005
- K Dobbin, R Simon. Sample size planning for developing classifiers using high dimensional DNA microarray data. Biostatistics (In Press)