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 rearrayed
- 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.



Balanced Block Design



 $A_i = i$ th specimen from class A $B_i = i$ th specimen from class B



BRCA1 v BRCA2 v S Class comparison between groups of arrays	
20 Sporadic	
1 BRCA1 This procedure finds genes differentially expressed among classes of samples. The classes are pre-defined based	
5 BRCA1 on columns of the experiment descriptor file. Each array should represent one sample, either as a single-label	
3 BRCA1 tool for class comparison between red and green samples.	
7 BRCA1	
2 BRCA1	
4 BRCA1 Experimental design: Find gene lists determined by:	
10 BRCA2 Column defining classes:	
9 BRCA2 • 0.001	
8 BRCA2	
22 BRCA2	
16 Sporadic Restriction on proportion of false discoveries:	
17 Sporadic Block by: Maximum proportion of false discoveries: 0,1	
15 Sporadic Confidence level (between 0 and 100%):	
18 Sporadic	
19 Sporadic	
21 Sporadic	
6 BRCA1 Maximum number of false discoveries: 10	
13 BRCA2 Confidence level (between 0 and 100%):	
14 BRCA2	
11 BRCA2	
12 BRCA2 Variance model:	
✓ Use random variance model for univariate tests.	
C Paired samples:	
Pair samples by:	
NOTE: This analysis is currently Select gene subsets	
the filter.	
O <u>K</u> Cancel Options Reset Help	
descriptors / Filtered log rat	

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 $E(FD) \le u$
 - 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 $p \le p^*$

• 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
 - 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



В	C		
-	BRCA1 v BRCA2 v S	Class comparison between groups of arrays	×
20	Sporadic		A
1	BRCA1	This procedure finds genes differentially expressed among classes of samples. The classes are pre-defined b	ased
5	BRCA1	on columns of the experiment descriptor file. Each array should represent one sample, either as a single-labe	el 🔤
3	BRCA1	experiment or as a dual-label experiment using a common reference. For non-reference designs, consider us tool for class comparison between red and green camples.	ing the
7	BRCA1	contor class companison between red and green samples.	
2	BRCA1	Class Comparison Options	
4	BRCA1	Experimental des	
10	BRCA2	Column defining cl Perform univariate permutation tests:	
9	BRCA2	ests: 0.00	11
8	BRCA2	Number of permutations for univariate test: 10000	
22	BRCA2		
16	Sporadic	toveries:	
17	Sporadic	Block by: ies: 0,1	_
15	Sporadic		
18	Sporadic	Number of permutations for multivariate test; 1000	
19	Sporadic		
21	Sporadic	eries;	
6	BRCA1		
13	BRCA2	Average over [Perform GO Observed vs. Expected analysis. %): 80	_
14	BRCA2		
11	BRCA2	Name to use for output files:	
12	BRCA2	ClassComparison	
		ate tests.	
		Paired samples	
		Pair samples by OK Cancel Reset Help	
		set to run on all genes passing select gene subse	ts
		the filter.	
			Lisla
doccrir	tore / Eiltored les ret		
uescrip	TOIS A FILTERED IOU FAL		

В	С		D	E	F	G 🗖		
	BRCA1 v BRCA2 v Sg	ooradic	BRCA1 V BRCA2	BRCA1 v Sporadic	BRCA2 v Sporadic	BRCA1 v not	-	
20	Sporadic	Signific	anco Analysis of M	icroarrave (SAM)				
1	BRCA1	aiginni	ance Analysis of M	icidalitays (SAM)				
5	BRCA1							ר ר
3	BRCA1	SAM of th	finds genes differentially e experiment descriptor	/ expressed among clas: file	ses of samples. The cla	isses are pre-define	d based on columns	
7	BRCA1	010	ie experiment descriptor	11101				
2	BRCA1							-
4	BRCA1	Ex	perimental design: —	P	arameters			1
10	BRCA2		olumn defining classes;					
9	BRCA2			-	Target proportion of f	alse discoveries:	0.1	
8	BRCA2				Number of Permutatio	IDS!	- 100	
22	BRCA2						100	
16	Sporadic		Unpaired samples:		Percentile:		80	
17	Sporadic	П						
15	Sporadic		Average over replica	ites of:				
18	Sporadic							
19	Sporadic							
21	Sporadic				Perform Gene Ontol	ogy Observed vs. E	xpected analysis	
6	BRCA1		Daired camples					
13	BRCA2		Faireu sampies.					í
14	BRCA2		Pair samples by:	N	ame to use for output f	iles:		1
11	BRCA2							
12	BRCA2				SAM			
				/ / /				
								1
							1	
		NOTE	This analysis is current	y set to run on all gene:	passing the filter.	Sele	ect gene subsets	
							1	-
		ОК	Cancel			Re	eset Help	
				1				
descrit	ptors / Filtered log rati	o / Ge	ene annotations 🔏 G	ene identifiers /		_ ► Ē	-	
		- A -	- X -	,			4	

Patient Array

В	С	Gene Set Expression Comparison	×
	BRCA1 v BRCA2 v		
20	Sporadic	The "Gene Ontology" option finds Gene Ontology categories that have higher than expected number of genes	
1	BRCA1	differentially expressed among classes of samples. The number of comparisons is the number of GO categories and hence	
5	BRCA1	the multiple testing problem is reduced. The "Pathway" option finds pathways that have higher than expected humber of genes differentially expressed among classes of samples. The number of comparisons is the number of nathways.	
3	BRCA1	represented in the dataset and hence the multiple testing problem is reduced. The "User gene lists" option finds Gene	
7	BRCA1	Lists that have higher than expected number of genes differentially expressed among classes of samples. All classes are	
2	BRCA1	pre-defined based on columns of the experiment descriptor file.	
4	BRCA1		
10	BRCA2	Experimental design: Gene set determined by:	
9	BRCA2	Column defining classes:	
8	BRCA2		
22	BRCA2	Find pathway lists determined by:	
16	Sporadic	Human: 🔿 BioCarta Pathways	
17	Sporadic	Unpaired samples: O KEGG Pathways	
15	Sporadic	C KEGG F GG WWDYS	
18	Sporadic	Average over replicates of: C Broad/MIT Pathways and signatures	
19	Sporadic	Mouse; C BioCarta Pathways	
21	Sporadic		
6	BRCA1		
13	BRCA2	Significance threshold of permutation tests: 0.005	
14	BRCA2	C Paired samples:	
11	BRCA2	Pair samples by:	
12	BRCA2	NOTE: This analysis is currently set	
		to run on all genes passing the filter. Select gene subsets	
		Variance model: Name to use for output files:	
		Use random variance model for	
		GOComparison	
lescrin	tors / Filtered log r	OptionsResetHelp	-
ieoenip	rears A microarlog i	•	

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

_							
В			D	E	F	G 🔒	
20	BRCAT V BRCAZ V Spo	Class comp	arison betwe	en red and green :	amples		
20	Sporadic BDCA4						
- I - E		This tool is f	or finding genes	differentially expresse	among two classes fo	or dual-label arrays in wi	hich each array
3	BDCA1	contains a s	ample from one	class and a sample from	the other class. The s	samples from one class r	need not be
7	BRCA1	labeled with	i the same label : s that each biolo	on all arrays; generally Igical sample appear eith	it is best to have comp er op oply ope array (plete balance of labels a prielse always paired wit	nd class. This
2	BRCA1	sample from	the other class	. As a special case, this	tool allows to compare	samples of one class wi	ith the
	BRCA1	reference s	amples. In this c	ase, reference samples	should contian the key	y word "reference' in the	e Red-labled or
10	BRCA2	Green-label	ea sample ID col	umn of the Experiment	Descriptors worksheet	•	
.0	BRCA2						
8	BRCA2	Experim	nental design:	F	nd gene lists deter	mined by:	
22	BRCA2	Red-la	abeled sample ID	column:	G. Constitution thread	hald of contractions have	
16	Sporadic			▼	 Significance thresi 	noid of Univariate tests:	0.001
17	Sporadic				C Restriction on pro	portion of false discover	ries:
15	Sporadic	Green-	labeled sample I	D column:	Mawimum proporti	en of folgo diagonaviago	
18	Sporadic			-	Maximum proporti	on of raise discoveries:	0.1
19	Sporadic				Confidence level (between 0 and 100%):	80
21	Sporadic						
6	BRCA1	Red-la	beled sample cla	ss column:	C Restriction on nur	nber of false discoveries	
13	BRCA2			▼	Maximum number	of false discoveries:	10
14	BRCA2				Confidence level (between 0 and 100%).	
11	BRCA2	Green-	labeled sample o	lass column —	Contradence tever (Decoveen o and 100 %).	80
12	BRCA2		labelea sample (
					ariance model:		
					Use random variance	model for univariate te	sts.
		NOTE: TH	ois analysis is cur	rrently set to run on all	enes passing the filte	r Select ger	ne subsets
		NOTE: II	no anaryoio io car	ronay sector an off all	jones passing the filter		
							1
		OK	Cancel			Options Reset	Help
escrip	otors / Filtered log ratio	🔏 Gene ann	iotations 🗶 G	ene identifiers /	1		

Patient Array

В	С	D	E	F	G 🗖	
	BRCA1 v BRCA2 v Sporadic	BRCA1 V BRCA2	BRCA1 v Sporadic	BRCA2 v Sporadic	BRCA1 v noti	
20	Sporadic				notBRCA1	
1	BRCA1	BRCA1	BRCA1		BRCA1	
5	BRCA1 Ouan	titative Trait Analy	sis			
3	BRCA1	,				
7	BRCA1 TTh	is tool finds genes that a	re significantly correlate	ed with a specified quap	titative variable (trait)	
2	BRCA1	is coornings genes chara	ine significantity contract	a mara specifica qaan		,
4	BRCA1					
10	BRCA2	experimental design:	Fi	nd gene lists determ	nined by:	
9	BRCA2	- Quantitative trait colum	in:			
8	BRCA2		_	Significance threshold	old of univariate tests:	0.001
22	BRCA2	1		Course		
16	Sporadic	Nue comune comula		 Restriction on propo 	ortion of false discove	ries:
17	Sporadic	 Use Spearman Correla 	tion rest	Maximum proportion	n of false discoveries:	0.1
15	Sporadic (Use Pearson Correlatio	on Test	Confidence level (be	etween 0 and 100%):	20
18	Sporadic				,	
19	Sporadic	Average over replica	ates of:	C Restriction on numb	per of false discoveries	e,
21	Sporadic			Maximum pumber of	F Folco discoverios	"
6	BRCA1		-	Maximum number or	Taise discoveries;	10
13	BRCA2			Confidence level (be	etween 0 and 100%):	80
14	BRCA2					
11	BRCA2					
12	BRCA2					
		MOTE: This sectoris is a	weeth, est to we as all	anne annsing the filter	Select of	ane subsets
		NOTE: This analysis is co	irrendy sec to run on all	genes passing the niter		
		W Canad			Ontions Densit	
				_		
						•
					•	
descrip	otors / Filtered log ratio / Ge	ene annotations 📈 Ge	ene identifiers /	•		

Patient Array

В	С	D	E	F	G 🚽	
	BRCA1 v BRCA2 v Sporadic	BRCA1 V BRCA2	BRCA1 v Sporadic	BRCA2 v Sporadic	BRCA1 v noti	
20	Sporadic				notBRCA1	
1	BRCA1	BRCA1	BRCA1		BRCA1	
5	BRCA1 Find G	enes Correlated wi	th Survival			
3	BRCA1					
7	BRCA1 This r	procedure tests for gener	s which are significantly	associated with surviva	al.	-
2	BRCA1		o milien are organicana,			
4	BRCA1	perimental design: -	Fin	d gene lists determi	ned by:	
10	BRCA2	Status column:		_		
9	BRCA2	(0 = censored, 1 = deal	th) (Significance threshold	d of univariate tests: 0.001	
8	BRCA2		_			
22	BRCA2			C participation and annual	tion of follow discounting.	
16	Sporadic	Colump defining survival l	time: ——	Restriction on propor	tion of raise discoveries:	
17	Sporadic			Maximum proportion (of false discoveries: 0,1	
15	Sporadic		<u> </u>	Confidence level (bet	ween 0 and 100%);	
18	Sporadic /			`	/ 00	
19	Sporadic			_		
21	Sporadic	Average over replication	es of: (Restriction on numbe	r of false discoveries:	
6	BRCA1			Maximum number of f	alse discoveries: 10	
13	BRCA2			Confidence level (bet	weep 0 and 100%):	
14	BRCA2					
11	BRCA2					
12	BRCA2					
	N	OTE: This analysis is curr	rently set to run on all g	enes passing the filter.	Select gene subsets	
	O	Cancel			Options Reset He	ip 🚽
descrip	otors / Filtered log ratio / Ge	ene annotations 📈 Ge	ene identifiers /	•		

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

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

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

F = features (genes) included in model

 w_i = weight for i'th feature

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

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

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

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 $y_{j} \left(\underline{w}' \underline{x}^{(j)} + b \right) \ge 1$ 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 $S_1 \mbox{ and } S_2$
 - Compute the cross-validated classification error for distinguishing $\rm S_1$ and $\rm S_2$
- Repeat the above steps for all possible partitions in order to find the partition S₁ and S₂ for which the cross-validated classification error is minimized
- If S₁ and S₂ are not singleton sets, then repeat all of the above steps separately for the classes in S₁ and S₂ to optimally partition each of them

J	K	L	М	N	0	Р	Q	R	S	Т	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			Clas	s predictio	n								
			т	nis procedure c	omputes a cla	assifier which	can be used I	for predicting I	the class of a	new sample.			
				Column defini	ng classes: —			se rendom ver	iance model (for univeriate	tects		
			— L	J				se ranuoini vai	lance modern		lesis,		
				Average o	ver replicates	of:	Gene	e selection Idividual genes	5:				
						V	•	Significant uni Optimize over	ivariately at a the grid of a	alpha level:	0.001		
				Pair sample	paired betwe s by:	en classes.	0	With univariat rate below:	ie misclassifi	cation	0,25		
	>0	1				<u>_</u>		With fold-ratio between two	o of geometri classes exce	ic means eding:	2		
	0 >0	1		Prediction me	thods: —— covariate pre	edictor	C G	ene pairs					
	>0	1		🔽 Diagonal lir	near discrimin	ant analysis	Nur "Gr	nber of pairs s eedy pairs" me	elected by the	ne 🗌	25		
	0	1		🔽 K-nearest	neighbors (fo	or K=1 and 3))						
	-0	1		🔽 Nearest ce	ntroid		NOTE	: This analysis	s is	.	1		_
	0	1		Support ve	ctor machine	s	genes	passing the f	ilter.	Select gene	subsets		
,M	>0	1											
	U >N	1		ок	Cancel				Options	Reset	Help		
	0	1		1	Medullobia	stoma		Medullobla	Medullobla	astoma			•
ent c	lescriptors	/ Gene ann	iotations ,	/ Filtered lo	g intensity	🏑 Gene id	entifiers / 🤅	Scatterplot	/ Cluster vi	ie' 🖣			

J	K	L	M	N	0	Р	Q	R	S	Т	U	\sim	/
Γ Site	M Stage	SurvStatus	T Stage	Age Code	Medulo vs	Glio vs PN	Rhabdo vs	Medulo vs	Medulo vs	AT/RT			
			Clas	s predictio	n								
CNS			- C TUIS	is prediction									
CNS			TH	nis procedure c	omputes a cla	assifier which	can be used	for predicting	the class of a	a new sample.			
CNS													
CNS			Clas	ss Predictio	n Options							$\mathbf{\Sigma}$	
				Cross-validati Leave-one 10 Repeat 0.632 boo	on method: - -out validatio - fold valida ed 1	in tion times tion	Г N si m	– Do statistica mis-classific umber of perm gnificance tesl is-classification	al significance ation rate. hutations for t of cross-val n rate:	test of cross	-validated		
	>0 0 >0 >0 >0	1 1 1 1 1 1 1		Column con "predict", "	ate test set; taining "train 'exclude" lab;	ing",		ne to use for d	output files:				
	U \S0	1		ок	Cancel				Ontions	Reset	Help		
	~0 	1			cancer				options	Kosoc		1	
	0 0	1		Support ve	ctor machine	s	depe	n ay see to rar s passing the l	filter.	Select gene	subsets		
,M	>0	1					gono	passing criter					
	0	1						Г	0.11	Deceb	Lisle	1	
	>0	1							Options	Reset	Heip	l 🖵	
	0	1		1	Medullobla	istoma		Medullobla	Medullobia	astoma			-
ent d	lescriptors	/ Gene ann	notations ,	/ Filtered log	g intensity	/ Gene ide	entifiers 🗶	Scatterplot	🖌 Cluster v	ie			

J	K	L	M	N	0	Р	Q	R	S	Т	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			Clas	s prediction	n		I I I I	de las	11			
				his procedure c	omputes a cla	assifier which	can be used h	or predicting	the class of a	a new sample.	·	
			Clas	s Predictio	n Options							
	>0	1 1 1 1		Cross-validati Leave-one 10 Repeat 0.632 boo Use separ Column con "predict", "	on me Class -out v - fol ed [itstra ate te tainin 'exclu	Support vect Cost (tuni Weight of 1 relative denotes th come first of the class	n Options 2 cor machine pa ng parameter) misclassificati to Class 2 (wh ne class label (in an alphanu is labels):	arameters:	1		-validated	
	>0 >0 0	1				ок		Rese	t			
	>0 0	1		ок	Cancel				Options	Reset	Help	
.м	0 >0	1		Support ve	ctor machine	s	genes	passing the l	filter.	Select gene	subsets	
•	0 >0	1		ок	Cancel				Options	Reset	Help	
ent d	0 lescriptors	1 / Gene ann	otations ,	1 / Filtered log	Medullobia g intensity	istoma / Gene ide	entifiers / S	Medullobia catterplot	(Medullobia / Cluster v	astoma ie 🚺		

1	K	L	M	N	0	Р	Q	R	S	Т	U	\vee	-
Γ Site	M Stage	SurvStatus	T Stage	Age Code	Medulo vs	Glio vs PN	Rhabdo v:	s Medulo vs	Medulo vs	AT/RT			
							AT/RT		AT/RT				
CNS							AT/RT		AT/RT				
CNS			Dee	diction An	lumin of M		(DAAA)		0 7 // 7 7				_
-NO			Pre		atysis of m	icroarrays	(PAM)						_
NNG NNG			[·		- Duadiatian (: (DA)				_	_
			R	nis tooi is an ir LTibshirani, T.	Hastie, B. Na	e Prediction A arasimha and	4naiysis or M G. Chu, Shri	icroarrays (PAI unken centroid:	м) Package o s aloorithm is	eveloped by used for clas:	s		_
2140			P	redictions.	,								_
				– Colump defir	ing classes			– Name to use	for output fi	les			
							-	PAM					
				🔲 Use sepa	rate test set			NOTE: This	; analysis is c	irrently set to	n run on		
				– Column co	ntaining "trair	ning", "predici		all genes pa	assing the filte	er.			
	<u></u>	1		"exclude"	labels;		Í.						
	/U 0	1					-		Select gene s	subsets			
	0	1											_
	>0	1		= .									
	>0	1		Average	over replicate	es of:							
	0	1											
	0	1											
	>0	1											
	0	1			1				-	1			
	0	1		OK	Cancel					Reset	Help		
,M	>0	1			k.d			kata at 10 t t					
	U >0	1		U	Medullobla	istoma		Medullobla	Medullobla	astoma			
	20 0	1		1	Modullobla	istoma		Modullobla	Modullobla	istoma			_
ent r	lescrinters	/ Gene and	notations 4	/ Filtered lo	niedunobia niintensity	/ Gene ide	entifiers /	Scatterniot	/ Cluster vi	istoma		-	

														3
J	K	L	M	N	0	Р	Q	R	S	Т	U		< 1	
T Site	M Stage	SurvStatus	T Stage	Ane Code	Medulo ve	Glio ve PN	Rhahdo ve	Medulo ve	Medulo ve	AT/RT				
			Bir	iary tree cl	ass predic	tion						\mathbf{X}		
CINS			1 	his tool compu each stage (tre	utes a binary l e node), clas	tree classifier ises are divide	which can be ed into two gr	used for pred oups, Cross-v	dicting the cla validation mis-	ss of a new s classification	ample. At rate is used			
CNS			t	o characterize ode of the tre	the quality o e. Then, pro	t the division. Tedure is repe	. A division wit eated for each	h the lowest i branch with	mis-classificat two or more (ion rate is us tlasses.	ed as a			
CNS					or mony pro-			r branch wich						
CNS				Column defi	ning classes:			- Prediction	method:					
							-	Compou	und covariate	predictor				
								C K-neare	est neighbors	(for K=1)				
				🗌 Use sepa	rate test set			C K-neare	est neighbors	(for K=3)				
				Column co	ntaining "trai	ning", "predic	t", —	C Nearest	t centroid					
				"exclude"	'labels;			C Support	t vector mach	ines				
	. 0	4						C Diagona	al linear discrir	minant analys	is			
	>0	1		_				- Predictors	should only ir	nclude genes:				
	0	1		Average	over replicate	es of:		G. Carlos						
	>0	1						Significa	ant univariate	iy at level:	0.001			
	>0	1												
	0	1						With un	iivariate ification rate	below	0,25			
	0	1		NOTE: This	analysis is			misciass						
	>0	1		currently set	to run on	Soloct conce			ia-ratio of geo between two	classes	2			
	0	1		all genes pas	sing the	Select gene	subsets	exceed	ing:		-			
	U >0	1		riicer.										
,IVI	20 0	1												
	∪ >N	1		ок	Cancel			0	Dotions	Reset	Help			
	ло П	1			39/100/							-		-
ent d	escriptors	/ Gene ann	otations	Filtered lo	a intensitv	/ Gene ide	entifiers 7.9	Scatternlot	/ Cluster vi	e 📢			_ I ► I Ē	
	couptors		is callor is y	(garconord	V dono ide		a a comport	V closed a					11.

J	K	L	M N O	P	Q	R	S	Т	U	\sim	
T Site	M Stage	SurvStatus T St	age Ane Code Medul	n vel Glio ve PN	Rhahdo v	e Modulo ve	Medulo ve	AT/RT			
			Binary tree class pre	diction						× _	
CNS			This tool computes a bir	hary tree classifier	which can b	e used for pre	dicting the cla	ss of a new s	ample. At		
CNS			to characterize the gua	, classes are uiviu. lity of the division.	. A division w	ith the lowest	mis-classificat	ion rate is us	ed as a		
			node of the tree. Then	, procedure is repe	eated for eac	h branch with:	two or more (lasses.			
CNS			Binary tree class pred	liction options						X	
				,							
			Binary Tree options: -			Support vecto	r machine pai	rameters: —		1	
			Une K feld averaging	!:							
			leave-one-out cross-va	s-validation algorit	n hm.	Cost (tuning	parameter):		1		_
			Value of K (defi	-	_						
			Value OFK (uen	ning K- 10	,	Weight of mis	classification:	s in Class 1			
			🗌 🔲 Do cross-validation	of the entire algor	rithm.	denotes the	class Iabel wh	ich would			
			Do not split classes if	the best		come first in a	an alphanume	ric sorting of	1		
	>0	1	achievable error rate	is more than: 0.	5	une class labe	as);				
	0	1									
	0	1	Name to use for outpu	ıt files:						1	
	>0	1	BinaryTreePrediction	l		Perform GC	Observed v	s. Expected a	nalysis.		_
	<i>~</i> υ Π	1									
	0 0	1	OK Can	cel				Reset	Help		
	>0	1									
	0	1	all genes passing the	Select gene :	subsets	exceed	between two ina:	classes	2	T	
	0	1	filter.	-		UNCOOL	nigi				
,М	>0	1				L					
	0	1		1				I	1.1.1		
	>U	1					Options _	Reset	Help	-	
ont d	lescrinters	/ Gene annotat	ions / Filtered log interv	sity / Gene ide	ontifiors /	Scatternlet	/ Cluster vi				
ent u	escriptors		ions A intereallog inten:	ary A dene lue	endners X	Scatterpiot	A cluster vi				

													عاريار	
	K	L	M	N	0	Р	Q	R	S	Т	U	\ \	< _	
ΓSite	M Stage	SurvStatus	T Sta			Lor pr	aei i i	II.		A T 105T				
			2	ourvival Risk F	rediction									
CNS				This tool is us	ed for Surviv	al Risk Predict	ion based o	n the Supervi:	sed Principal C	omponents m	ethod.	-		
CNS				(Bair, E. and 1	Tibshirani, R.	PLoS Biology	2:511-522,	2004)						
ONIC														
CNO CNR				– Experimenta	al design: –		- Find	d gene lists (determined	by:				
					n: — deal	th)		:C:						
0140					icu, i — ucu	-	Mo	iniricance thre idel:	shold of Cox	0.0	001			
						<u> </u>								
				r Colump defin	ond survival	time:	Bhu Bhu	mhar of Dringi		ba 🔽				4
					iing sai vivai	cinic.	(1-	(10): 10):	pai componen	2				
						_		•						
							Cov	ariates ——			_			
				Average ov	er replicates	of:		Clinical Cov	ariates					
								- Column defir	ning Covariate	1:				
	>0	1				T								
	0	1												
	0	1		, 				- Column defir	ning Covariate	2:				
	>0	1		Use separal	te test set					-				
	>0	1		— Column contai	nina "trainina	o"."predict" –	L							
	0	1		"exclude" lab	els;			- Column defir	ning Covariate	3:		_		
	0	1				-				-				
	>U	1												
	U	1												
м	 ∖N	1		NOTE: This a	nalysis is cur	rently set to r	un on all ger	nes passing th	ne filter.	Select ger	ne subsets			
101	-0 N	1												
	>0	1				1				1				
	- 0	1		ОК	Cancel			Op	tions	Reset	Help		-	
ent d	escriptors	/ Gene anr	notation:	s 🏑 Filtered lo	g intensity	🏑 Gene ide	entifiers 🏑	Scatterplot	t 🖌 Cluster v	vie' 🖣				

J	K	L N	M N	0	P	Q	R	S	Т	U	V	
T Site	M Stage	SurvStatusT Sta	Council Dials D						AT 107			
			SULVIVAL RISK P	rediction								
			This tool is use	Survival	Risk Optic	ons				hethod.	-	
CNO			(Bair, E. and Ti	ł								
CNS			- Euporimontal	🗖 Risk (Groups ——				_			
CNS			- Status column	0.0	ni lu Curuna	,						
CNS			(0 = censore		RISK Groups	,	3-Risk Gro	ups		201		
				Prog	gnostic Index		_			101		
				Perc	entile:	50						
			🗖 🗖 Column defini	'								
				Cross '	Validation Mel	thod:			_			
				🖲 Lear	ve One Out (V (🔆 10- Fold 🔾	ΞV				
			🛛 🗌 🗖 Average ove		Rank Test: —							
					Derform Derm	utation tests						
	>0	1			renonnrenn							
	0	1		Num	ber of permu	tations for	100					
	0	1		signi	ificance of th	e log rank test						
	>0	1	Use separate									
	>U 	1	Column contain	- Name	e to use for o	utput files: —						
	U 0	1	"exclude" labe									
	 >∩	1		Sur	vivalRiskPred	liction						
	л П	1										
	0	1				1			1		1	
,M	>0	1	NOTE: This an		ОК	Cancel	R	Reset	Help	he subsets		
	0	1										
	>0	1		Capcel			Onti	0.05	Reset	Help		
	0	1							Keset			
ent d	lescriptors	/ Gene annotatio	ons / Filtered log	, intensity	_∕ Gene ide	entifiers / S	Scatterplot	/ Cluster	vie 📕			

Rł	iabdo vs PN	IET			Maidainensional scaling									
-					Class <u>c</u> omparison	•								
					Class prediction	+							×	
	K	L	M	N	Su <u>r</u> vival analysis	•		R	S	Т	U	V	-	
l Sit	eM Stage	SurvStatus	T Stage	Age C	Quantitative trait analys	is 🕨	o vs	Medulo vs	Medulo vs	AT/RT			_	
					Eilter and subset the dat	·	-		AI/RI				- 1	
							-		AURI				- 1	
					Plugins	•		Analysis of V	Variance					
CNS					<u>U</u> tilities	•		ANOVA on le	og intensities					
CNS					<u>H</u> elp	•		ANOVA for I	Mixed-Effects	Model			_	
CNS					About BRB-ArrayTools			Time Series	Analysis				- 1	
					About R-COM			Random For	est for class (prediction			- 1	
					Licopco pareomont			Class predic	tion by ton se	oring pairs				
					License agreemend	· · · _ ·				oring pairs				
					PNET	PNET		M VS A PIOC					_	
					PNEI	PNET		Pairwise Cor	relation Plot				-	
					PNET			Smoothed C	DF				-	
					PNET	PNET		Extract sele	cted genes				-	
	>0	1	4	0	Medulloblastoma			Export 1 Co	lor Data To R					
	0	1	2	1	Medulloblastoma			Export 2 Co	lor Data To R				_	
	0	1	3	1	Medulloblastoma			Lood Dive Ie					_	
	>0	1		2	Medulloblastoma				-				-	
	0	1	4	0	Medulloblastoma			Manage Plu	g Ins					
	0	1	1	1	Medulloblastoma			Create Plug	In					
	>0	1	3	1	Medulloblastoma			Advanced P	lug In Editor					
	0	1		1	Medulloblastoma			Medullobla	Medullobla	istoma			_	
	0	1		0	Medulloblastoma			Medullobla	Medullobla	istoma			_	
М	>0	1	2	1	Medulloblastoma			Medullobla	Medullobla	istoma			_	
	0	1		0	Medulloblastoma			Medullobla	Medullobla	istoma				
	>0	1	3	1	Medulloblastoma			Medullobla	Medullobla	istoma				
	0	11		1	Medulloblastoma			Medullobla	Medullobla	istoma			_	
ent e	descriptors	/ Gene ani	notations 🔏	 Filtered lo 	a intensity 🏒 Gene ida	entifiers	1 8	Scatterplot	🖌 Cluster vi	e'I∢I		•		

J K L Malysis of Variance T Site M Stage SurvStatue T Sta CNS							
T Site M Stage SuvStatu T Stage CNS	J	K	L	M	Analysis of Variance		V -
CNS Column of exper descriptor sheet for factor A CNS Column of exper descriptor sheet for factor A CNS Column of exper descriptor sheet for factor A CNS Column of exper descriptor sheet for factor A Column of exper descriptor sheet for indicator of indicated arrays Column of exper descriptor sheet for testing effects Column of exper descriptor sheet for testing effects Column of exper descriptor sheet for testing effects .001 Threshold p value for testing the model .001 Threshold false discovery rate for testing the model .001 Threshold false discovery rate for testing the model .001 Do 1 .001 Model Formula .001 .001 .001 .001 .001 .001 .001 <td>r T Site</td> <td>M Stage</td> <td>_ SurvStatusT</td> <td>Star</td> <td></td> <td></td> <td><u> </u></td>	r T Site	M Stage	_ SurvStatusT	Star			<u> </u>
CNS Column of exper descriptor sheet for factor A Array CNS Column of exper descriptor sheet for factor B Array CNS Column of exper descriptor sheet for factor B Array CNS Column of exper descriptor sheet for factor B Array Column of exper descriptor sheet for factor C Array Column of exper descriptor sheet for factor C Array Column of exper descriptor sheet for factor C Array Column of exper descriptor sheet for factor C Array Column of exper descriptor sheet for factor C Array Column of exper descriptor sheet for factor C Array Column of exper descriptor sheet for factor C Array Column of exper descriptor sheet for factor C Array Column of exper descriptor sheet for factor C Array Column of exper descriptor sheet for testing effects .001 Threshold p value for testing effects .001 Threshold false discovery rate for testing fifters .001 D 1 .001 D 1 .001 Model Formula .001 D 1							
CNS	CNS						
CNS Column of exper descriptor sheet for factor A CNS Column of exper descriptor sheet for factor A CNS Column of exper descriptor sheet for factor B Column of exper descriptor sheet for factor C Column of exper descriptor sheet for testing fleator S Column	CNS						
CNS Column of exper descriptor sheet for factor A Array CNS Column of exper descriptor sheet for factor A Array Column of exper descriptor sheet for factor D Array Column of exper descriptor sheet for factor D Medulo Stage Column of exper descriptor sheet for indicator Medulo Stage Column of exper descriptor sheet for indicator Medulo Stage Column of exper descriptor sheet for indicator Sex Column of exper descriptor sheet for testing effects .001 Column of exper descriptor sheet for testing effects .001 Column of exper descriptor sheet for testing effects .001 Column of exper descriptor sheet for testing fields .001 Column of exper descriptor sheet for testing fields .001 Column of exper descriptor sheet for testing fields .001 Column of exper descriptor sheet for testing fields .001 Column of exper descriptor sheet for testing fields .001 Column of exper descriptor sheet for testing fields .001 Column of exper descriptor sheet for testing fields .001 Column of exper descriptor sheet for testing fields .001 Column of exper descriptor sheet for testing fields .001				_			
CNS Column of exper descriptor sheet for factor A CNS Column of exper descriptor sheet for factor B CNS Column of exper descriptor sheet for factor B Column of exper descriptor sheet for factor B Array Column of exper descriptor sheet for factor B Array Column of exper descriptor sheet for factor C Dx Column of exper descriptor sheet for factor C Medulo Stage Sex Sex Column of exper descriptor sheet for technical replaceter Column of exper descriptor sheet for technical Survival (months) Column of exper descriptor sheet for testing effects .001 Column of all Threshold p value for testing the model .001 Column of all Threshold false discovery rate for testing the model .001 Column of all Model Formula .001 Column of all Blocking factor(s) .001 Column of all Use Random Variance Model .001 Model Formula	CNS						
CNS Column of exper descriptor sheet for factor B Array Column of exper descriptor sheet for factor C Ox Column of exper descriptor sheet for factor C Ox Column of exper descriptor sheet for factor C Ox Column of exper descriptor sheet for factor C Ox Column of exper descriptor sheet for factor C Ox Column of exper descriptor sheet for factor C Ox Column of exper descriptor sheet for indicator replicates Sex Column of exper descriptor sheet for technica replicates Column of exper descriptor sheet for testing fifetts .001 Column of exper descriptor sheet for testing fifetts .001 Column of exper descriptor sheet for testing fifetts .001 Column of exper descriptor sheet for testing fifetts .001 Column of alse discovery rate for testing fifetts .001 Column of alse discovery rate for testing fifetts .001 Column of alse .001 Submit .001 Column of alse .001 Submit .001 Column of alse discovery rate for testing fifetts .001 Column of all Model Formula Column of all	CNS				Column of exper descriptor sheet for factor A	▼	
Column of exper descriptor sheet for factor D Column of exper descriptor sheet for factor D Column of exper descriptor sheet for factor D Column of exper descriptor sheet for indicator of exper descriptor sheet for indicator of induded arrays Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model 0	CNS				Column of exper descriptor sheet for factor B		
Column of exper descriptor sheet for factor C Column of exper descriptor sheet for indicator Column of exper descriptor sheet for testing effects Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Image: Column of exper descriptor sheet for testing the model Column of exper descriptor sheet for testing the model Column of exper des						Array	
Image: Second					Column of exper descriptor sheet for factor C	Dx	
Column of exper descriptor sheet for ratio b Medul Stage Sex Age at Dx Age at Dx Survival (months) Column of exper descriptor sheet for technical replicates Imedul Stage Column of exper descriptor sheet for indicator of included arrays Survival (months) Column of exper descriptor sheet for testing effects .001 Column of exper descriptor sheet for testing effects .001 Column of exper descriptor sheet for testing effects .001 Column of exper descriptor sheet for testing effects .001 Column of exper descriptor sheet for testing the model .001 Column of alse discovery rate for testing the model .001 Column of alse discovery rate for testing the model .001 Submit .001 Column of alse discovery rate for testing the model .1 Model Formula .1 Submit .1 Use Random Variance Model .1 Submit View README					Colump of experied descriptor sheet for faster D		
Column of exper description sheet for indicator Age at Dx Survival (months) Survival (months) <td< td=""><td></td><td></td><td></td><td></td><td>column of experidescriptor sheet for factor D</td><td>Sex</td><td></td></td<>					column of experidescriptor sheet for factor D	Sex	
of include tarrays Survival (months) Image: Column of exper descriptor sheet for testing effects .001 Image: Column of exper descriptor sheet for testing effects .001 Image: Column of exper descriptor sheet for testing effects .001 Image: Column of exper descriptor sheet for testing effects .001 Image: Column of exper descriptor sheet for testing effects .001 Image: Column of exper descriptor sheet for testing effects .001 Image: Column of exper descriptor sheet for testing effects .001 Image: Column of exper descriptor sheet for testing effects .001 Image: Column of exper descriptor sheet for testing effects .001 Image: Column of exper descriptor sheet for testing the model .001 Image: Column of exper descriptor sheet for testing the model .001 Image: Column of exper descriptor sheet for testing the model .001 Image: Column of exper descriptor sheet for testing the model .001 Image: Column of exper descriptor sheet for testing the model .001 .001 Image: Column of exper descriptor sheet for testing the model .001 Image: Column of exper descriptor sheet for testing the model .001 .001 Image: Column of exper descriptor sheet for testing the model .001 Image: Column of exper descriptor shee					Column of exper description sheet for indicator	Age at Dx	
Column of exper descriptor sheet for technical replicates Column of exper descriptor sheet for testing effects Column of exper descriptor sheet for testing Column of exper descripto					of included arrays	Survival (months)	
Implementation Impleme					Column of exper descriptor sheet for technical replicates		
>0 1 Interstold p value for testing the model .001 0 1 Interstold false discovery rate for testing the model .001 >0 1 Interstold false discovery rate for testing the model .001 >0 1 Interstold false discovery rate for testing the model .001 >0 1 Interstold false discovery rate for testing the model .001 >0 1 Interstold false discovery rate for testing the model .001 >0 1 Interstold false discovery rate for testing the model .001 >0 1 Interstold false discovery rate for testing the model .001 0 1 Interstold false discovery rate for testing the model .001 0 1 Interstold false discovery rate for testing the model .001 0 1 Interstold false discovery rate for testing the model .001 0 1 Interstold false discovery rate for testing the model .001 0 1 Use Random Variance Model .001 .0 1 .001 .001 .001 .0 1 .001 .001 .001					Threshold p value for testing effects	001	
>0 1 .001 .001 0 1 .001 .001 0 1 .1 .1 >0 1 .1 .1 >0 1 .1 .1 >0 1 .1 .1 >0 1 .1 .1 >0 1 .1 .1 0 1 .1 .1 0 1 .1 .1 0 1 .1 .1 0 1 .1 .1 0 1 .1 .1 .0 1 .1 .1 .0 1 .1 .1 .0 1 .1 .1 .0 1 .1 .1 .0 1 .1 .1 .0 1 .1 .1 .0 1 .1 .1 .0 1 .1 .1 .0 1 .1 .1 .1 .0 <t< td=""><td></td><td></td><td></td><td></td><td></td><td>.501</td><td></td></t<>						.501	
No 1 No <td< td=""><td></td><td>>0</td><td>1</td><td></td><td>Threshold p value for testing the model</td><td>.001</td><td></td></td<>		>0	1		Threshold p value for testing the model	.001	
0 1 >0 1 >0 1 >0 1 >0 1 ent descriptors Gene annotatio		0	1		Threshold false discovery rate for testing		
>0 1 >0 1 >0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 View README Image: Comparison of the string the model		0	1		effects	.1	
>0 1 model >0 1 0 1 0 1 >0 1 >0 1 >0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 >0 1 0 1 >0 1 0 1 View README Image: Constant		>0	1	_	Threshold false discovery rate for testing the	,1	
Image:		>0	1		model Model Formula		
Image: Solution of the striptors / Gene annotatio Image: Solution of the striptor of the stripto		n	1		Model Fornidia		
>0 1 0 1 0 1 0 1 0 1 >0 1 >0 1 0 1 View README		n	1		Blocking factor(s)		
O 1 0 1 0 1 7,M >0 0 1 >0 1 >0 1 View README View README		>0	1		Use Dan dan Usetaan we dat		
0 1 M >0 1 0 1 >0 1 0 1 0 1 View README		п П	. 1		Use Random variance Model		
Image: Second Secon			1				
Image: Second state	M	>0	1				
>0 1 0 1 ent descriptors / Gene annotatio		. <u>с</u> П	1			Submit	
ent descriptors / Gene annotatio		>0	1				
ent descriptors / Gene annotatio		. <u>с</u>	1				-
	ent o	escriptors	/ Gene anno	otatio	V		I II
			<u></u>				

🖌 Rhabdo	
-	🗭 anova.txt - Notepad
	File Edit Format View Help
J I TSit∈MSt	**************************************
CNS CNS	Analysis of Variance (ANOVA) for each gene.
CNS CNS	WWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW
	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.
>0	WWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW
>0	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 factor D (can be empty)
>0	(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
,M >0	. 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.
ent descrij	. 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	0	Р	Q		R	S	Т	U	\vee	
T Site	M Stage	SurvStatus T	Sta	ANOVA for Mi	xed-Effect	Model						1	\times	
CNS														
CNS														
														_
CNS														_
			_											_
CNS			_	Column of exp	per descriptor	sheet for fact	or A						·	_
			_	Column of exi	per descriptor	sheet for fact	or B				 	 1	<u>i </u>	_
			_				Arr	ay			 			
			_	Column of exp	per descriptor	sheet for fact	or C DX	dula Tura	_					
			_	Column of ex	per descriptor	sheet for ran	dom Me	dulo Type dulo Stac	= 1e					-
			_			fa	ctor Se:	<	,-					-
			_	Column of expe	er description	sheet for indic of included ar	ator Agi	e at Dx wigel (mo	othe)					-
				Column of exp	er descriptor :	sheet for tech	nical	vival (inc	naisy					-
						replica	ates							-
	>0	1		Threshold	p value for te	sting fixed effi	ects .00)1						
	0	1		Thresh	old p value for	testing the m	odel or	11					-	
	0	1					,						_	
	>0	1		I hreshold faise	e discovery rat	e for testing fi effi	ects .1							
	>0	1		Threshold fal:	se discovery r	ate for testing	the 1						-	
	0	1				m	odel						_	
	0	1				Model Forr	nula							
	>0	1					,							_
	0	1							_					
	0	1					Su	bmit						_
,M	>0	1												_
	0	1					View							_
	>U	1						READIVI	-					
	0	1	+ - +										- 1-	
ent d	iescriptors	<u>A</u> Gene anno	tatio,	o <u>A ricoroario</u>	og inteoriorej	X CONDIGO		(acare	or prove-	V classes				

					🔂 anovamix.txt - Notepad
J	K	L	M		File Edit Format View Help
T Site	M Stage	SurvStatus	T Sta	INO	************************
	Ū				** NAME OF THIS PLUG-IN: **
CNS					~~~~~~
CNS					Linear mixed-effects model for each gene.
CNS					
CNS					** PURPOSE OF THIS PLUG-IN: **
CNS				- d	
				C	This plug-in fits data to a linear mixed-effects model for each gene
				~	can be specified in the mixed-effects model. This model is useful for
				୍ୟ ସ	time series data with specimens collected from each subject at multiple
				d	time points. The subjects may have different characteristics of interest,
				1	such as diseased or normal, treated or non-treated, male or female, but
				Col	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
	>0	1			power for testing the effects of interest. This mixed-effects model is, nowever, much more computationally intensive than the standard fixed effects model provided
	0	1			in the other plug-in.
	0	1		ты	
	>0	1			The F-test is based on the likelihood ratio test or type III sum of
	>0	1		Т	significance of each factor is adjusted for all other factors of the model.
	0	1			- · · · · · · · · · · · · · · · · · · ·
	0	1			******
	>0	1			** USING THIS PLUG-IN: **
	0	1			**********************
	0	1			To sum this function, the user chauld include the following.
,Μ	>0	1			To run this function, the user should input the following:
	0	1			. Column of exper descriptor sheet for factor B (can be empty)
	>0	1			. Column of exper descriptor sheet for factor C (can be empty)
	0	1			. Column of exper descriptor sheet for random factor Column of exper descriptor sheet for indicator of included arrays
ent c	lescriptors	/ Gene ann	otatio <mark>, "</mark>	~~	(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).

													١×
J	K	L	М	N	0	Р	Q	R	S	Т	U	V	
T Site	M Stage	SurvStatus	T Stage	Age Code	Medulo vs	Glio vs PN	Rhabdo vs	Medulo vs	Medulo vs	AT/RT			_
							AT/RT		AT/RT				
CNS							AT/RT		AT/RT				
CNS							AT/RT		AT/RT				
CNS			Tin	ne Series A	nalysis							× _	_
CNS													_
CNS													_
													_
													_
				Column of	over docrie		hime I					_	
				CONTINU	exper descrip	lor sheet for						-	_
				Column of e	exper descript	tor sheet for (class						
							Array				Ĩ		_
				olumn or expe	r description :	sneet for indic of included ar	ator Dx ravsi Medul	o Type			_		
	_			Thres	hold p value i	for testing eff	ects Medul	o Stage					
	>0	1					Sex	-					_
	U	1		Threshold	talse discove	ry rate for te	sting Age a	t Dx al (months)					_
	U	1				CII		ar (monuns)					
	>0	1											_
	>0	1					Sub	mit					
	U	1											_
	U 150	1											
	20	1					View RE	:ADME					_
		1											_
. NA		1		4	Modullable	otomo		Modullable	Modullable	otomo			
'M	20	1	2		Medullobla	stoma		Modullobla	Modullobla	astoma			
	U 150	1	2	1	Medulloble	istoma		Medullobla	Medullobla	astoma			_
	20	1	3	1	Medullobla	stoma		Medullobla	Modullobla	astoma			_
ont ·	U loccrinters	Cono por	otations /	L Eiltorod Io	nviedunobla a intoncity	/ Cono.ida	potifiore / (Fiviedunobia Scottorplot	(Cluster vi	iorla			, r
ли с	rescriptors	A Gene ann	iotations A	, ritereu iu	y intensity –	A Generus	anumers 🔬 🤇	scatterpiot	A Cluster V				r

F		М	L	K	J
4	Age	T Stage	SurvStatus	M Stage	T Site
14 					
۲					CNS
Т					CNS
	- 6				~~~~
¥	e sei				CNS CNC
14 					
1					CNO
Т					
e					
i					
٦	Col				
W a	Colu				
r	00				
Ę	ilumn c	Co			
t t					
a			1	>0	
a	Thr		1	0	
p			1	0	
t			1	>0	
I S			1	>0	
1			1	U 0	
M			1	U \()	
ľ			1	-0 Π	
t			1	0	
l C		2	1	>0	M
t		£	1	0	
M		3	1	>0	
t			1	0	
j	Filte	notations 📈	(Gene anr	lescriptors	ent d
T					_

-	intes	er les. LA	1 - 110	tepau	
File	Edit	Format	View	Help	
***	***	*****	****	********	*****
**	NAM	IE OF T	HIS	PLUG-IN:	**
***	****	******	****	*********	*****

Time Series Analysis.

VERNERAL CONTRACTOR CONTRACT

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 inear 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 ate (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 he 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. his plug-in is not appropriate for nested data where the same subject is ampled 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:

J	K	L	M	N	0	Р	Q	R	S	Т	U	✓
T Site	M Stage	SurvStatus	T Stage	Age Code	Medulo vs	Glio vs PN	Rhabdo vs	Medulo vs	Medulo vs	AT/RT		
				014 1			0 T /DT		0 7 /07			<u> </u>
				UVA ON log	Intensities						<u> </u>	4 /
CNS												
ONG												
0110												
				Column of	exper descri	ptor sheet fo	r red				-	
				Column of ex	oer descripto	or sheet for a	reen					
						channel	class Array				-	
			C	olumn of expei	r description :	sheet for india of included as	tator Dx rrays Meduk	Tuna				
			(Column of expe	r descriptor s	heet for tech	nical Medulo) Stage				
				Thurs		replic	ates Sex	- -				
				Inres	nola pivalue l	ror testing eri	Surviv	ux al (months)			-	
	<u>\</u> 0	1		Threshold	false discove	ry rate for te	sting 1	_, _, _,				
	-0 Π	1			Red	intensity mini						
	0	1			1.00		100					
	>0	1			Green	intensity mini	imum 100					
	>0	1			Use Rando	om Variance M	1odel 👝					
	0	1										
	0	1										
	>0	1					Cubr	nit				
	U 0	1					Subi	IIIL				
M	U \\N	1										
141	-0 N	1					View RE	ADME				
	>0	1										
	0	1		1	Medullobla	stoma		Medullobla	Medullobla	astoma		-
ent d	lescriptors	/ Gene anr	notations ,	/ Filtered lo	g intensity	/ Gene ide	entifiers / S	catterplot	/ Cluster v	ie' 🔺		

```
🛑 anovachnnl.txt - Notepad
                    ĸ
                                                                  M
                                                                                             File Edit Format View Help
T Site M Stage
                                SurvStatusT Stage
                                                                               Ade
                                                                                            *****************************
                                                                                             ** NAME OF THIS PLUG-IN:
                                                                                             ************************
CNS.
CNS
                                                                                            ANOVA on log intensities for each gene.
CNS
                                                                                             *************************
CNS.
                                                                                             ** PURPOSE OF THIS PLUG-IN:
CNS
                                                                                             ***********************
                                                                                     Co There are two steps for running ANOVA in this plugin.
                                                                                            The first step is to normalize log intensity for each channel
                                                                                 Colur and the second step is to do ANOVA on normalized log intensity.
                                                                           Column d Step1: Normalize each log intensity by the following
                                                                                                                 normalization model (underline denotes subscript):
                                                                            Column
                                                                                                                      y_{adcq} = mu + A_a + AD_{ad} + e1_{adcq} ---- (1)
                                                                                                                 where
                                                                                                                            y_{adcq} is the log intensity,
                                                                                     The
           >0
                                                                                                                            mu is the overall log intensity mean value,
                                                                                                                            A_a is the effect of the array a,
                             0
                                                     1
                                                                                                                            AD_{ad} is the interaction of array a and dye d,
                             Π
                                                                                                                            e1_[adcg} is the random noise,
           >0
                                                                                                                            c is the index of class(variety).
           |>Π
                                                                                                                We assume each effects are fixed. So {A_a} and {AD_{ad}}
                             0
                                                                                                                 satisfy some identification conditions.
                             Π
                                                     1
                                                                                                                 After fitting the normlization model, the residuals
           >0
                                                                                                                 (normalized log intensity), r_{adcg} are obtained,
                             0
                             Π
                                                                                                                      r_{adcq} = v_{adcq} - \lambda_{adcq} 
ΙM.
          1>0
                                                                                            Step2: Fit the normalized log intensity by the following ANOVA model
                             0
           >0
                                                                                                                      r_{adcq} = mu_q + alpha_{aq} + beta_{dq} + class_{cq} + e2_{adcq} ---- (3)
                             Π
ent descriptors / Gene annotations / Filte
                                                                                                                 where
                                                                                                                            mu_q is the gene-specific average log intensity,
                                                                                                                            alpha_{aq} is the gene-specific array effect (spot effect),
                                                                                                                            beta_{dq} is the gene-specific dye effect,
                                                                                                                            class_{cq} is the gene-specific class (variety) effect,
                                                                                                                            e2_[adcq{is the random noise.
                                                                                                                 Again, we assume each effects in model (3) are fixed.
```

Rh	abdo vs PN	IET			Multidimens	sional scaling									
					Class <u>c</u> omp	arison	•								
					Class gredi	ction	•								
	K	L	M	N	Survival an	alysis	•		R	S		Т	U	\sim	
T Site	M Stage	SurvStatus	T Stage	Age C	Ouantitativ	e trait analys	ic b	o vs	Medulo vs	Medul	o vs	AT/RT			
						· · · ·		_		AT/RT					_
					Filter and s	Filter and subset the data				AT/RT					_
CNS					<u>P</u> lugins	Plugins •				AI/RI					- 11
CNS					Utilities •					AT/RT	•				- 11
CNS					Help	Help +			Getting star	tting started		Overview of	analysis tools		
CNS					About BRB-	-ArrayTools			Manuals	•		Begin the tut	orial		- 11
									Support			- Degin the <u>C</u> ut		- H	_
					ADOUL R-CU	лм		_	Sabbour			Open a <u>s</u> amp	ole dataset		
					License agr	eement			PNET						
						PNET	PNET		PNET						
						PNET	PNET		PNET						
						PNET	PNET		PNET						
						PNET	PNET		PNET						
						PNET	PNET		PNET						
	>0	1	4	0	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	0	1	2	1	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	0	1	3	1	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	>0	1	3	1	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	>0	1		2	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	0	1	4	0	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	0	1	1	1	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	>0	1	3	1	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	0	1		1	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	0	1		0	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
,M	>0	1	2	1	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	0	1		0	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	>0	1	3	1	Medullobla	astoma			Medullobla	Medul	lobla	stoma			
	0	1		1	Medullobla	astoma			Medullobla	Medul	lobla	stoma			-
ent d	descriptors	/ Gene ani	notations 📈	Filtered lo	g intensity	/ Gene id	entifiers	1	Scatterplot	/ Clust	ter vi	e' 🖣			▶ /

Rh 🕯	abdo vs PN	IET															
						Class <u>c</u> ompa	arison	•									
						Class predic	tion	•								\mathbf{X}	
J	K	L	M	N		Su <u>r</u> vival ana	alysis	•		R	S		Т	U	V		
T Site	M Stage	SurvStatus	T Stage	Age C		Quantitativ	e trait analy:	sis 🕨	o vs	Medulo vs	Medul	o vs	AT/RT			-	
CNS						<u>F</u> ilter and su	ubset the da	ita			AT/RT	•					
CNS						<u>P</u> lugins		•	-		AT/RT					- 1	
CNS						<u>U</u> tilities					AT/RT	•					
				[Help •				<u>G</u> etting star	ted 🕨	-				- 1	
						About <u>B</u> RB-ArrayTools			<u>M</u> anuals	•		<u>U</u> ser's Manua	al				
					About R-COM					<u>S</u> upport	•		<u>P</u> lug-ins Guid	e			
						License agr	eement	···· <u>-</u> ·	F	PNET	AI/RI		<u>S</u> ample Desci	ription of Me	thods for Pu	ublicatio	ins
							PNET	PNET		PNET							
							PNET	PNET		PNET						_	
							PNET	PNET		PNET						- 1	
							PNET	PNET		PNET						- 1	
	>0	1	4		0	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	0	1	2		1	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	0	1	3		1	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	>0	1	3		1	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	>0	1			2	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	0	1	4		0	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	0	1	1		1	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	>0	1	3		1	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	0	1			1	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	0	1			0	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
,M	>0	1	2		1	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	0	1			0	Medullobla	istoma			Medullobla	Medul	lobla	astoma				
	>0	1	3		1	Medullobla	istoma		Medullobla Medulloblastoma								
	0	1			1 Medulloblastoma					Medullobla	Medul	lobla	astoma			_	
ent d	lescriptors	/ Gene ann	notations 🖌	Filtere	d lo	g intensity	/ Gene id	lentifiers	18	Scatterplot	/ Clust	ter v	ie' 🖣		•		
Rhabdo vs PNET																	
----------------	------------------	------------	-------------	-------------	-----------------------------	----------------------	------------------	----------------------	-----------------	--------	-----------------	----------------------	---	---	--------------		
					Class <u>c</u> ompa	arison	•										
					Class predic	tion.	•								\mathbf{X}		
J	K	L	M	N	Survival analysis				R	S		Т	U	V			
T Site	M Stage	SurvStatus	T Stage	Age C	Oupphitative trait applusis		o vs	s Medulo vs	Medulo vs AT/RT		AT/RT						
									AT/R1	-							
CNS					Filter and subset the data				AT/R1	-				_			
CNS					<u>Plugins</u>			L		AT/RT	-				_		
-NO					Litilities N			-		AI/RI	-				_		
					<u>U</u> undes		-		AI/RI					-			
					Help •			<u>G</u> etting star	ted 🕨	-				-			
0110					About BRB-ArrayTools			<u>M</u> anuals	•	-				-			
					About R-COM			Support	•		View messag	e board					
					– Licence agreement		<i></i>		AI/RI		Empil suppor	: - L					
					License agr	· · · - ·	· · · <u>-</u> ·		PNET			<u>c</u> mail suppor	L		_		
						PNET	PNET		PNET			ListServ	•		_		
						PNET	PNET		PNET		_				_		
						PNET	PNET		PNET						_		
							PNET								_		
	>0	1	1	Π	Medullohia	rn∟i etoma			Medullohis	Modul	llohls	etomo			-		
	νo Π	1	2	1	Medulloblastoma				Medullohla	Medu	llohla	astoma			-		
	0	1	3	1	Medullohlastoma				Medullobla	Medu	Medulloblastoma				-		
	>0	1	3	1	Medulloblastoma				Medullobla	Medu	loblastoma				-		
	>0	1		2	Medullobla	istoma			Medullobla	Medu	llobla	astoma					
	0	1	4	0	Medullobla	istoma			Medullobla	Medu	llobla	astoma					
	0	1	1	1	Medullobla	istoma			Medullobla	Medu	llobla	astoma					
	>0	1	3	1	Medullobla	istoma			Medullobla	Medu	llobla	astoma			_		
	0	1		1	Medullobla	istoma			Medullobla	Medu	llobla	astoma			_		
	0	1		0	Medullobla	istoma			Medullobla	Medu	llobla	astoma			_		
,M	>U	1	2	1	Medullobla	istoma			Medullobla	Medu	llobla	astoma			_		
	U _0	1	2	U	Medullobla	istoma			Medullobla	Medu	llobla	astoma			_		
	>U 0	1	3	1	Medullobla	istoma			Modullobla	Modu	IIODI8	astoma					
ent r	U lescrintore	/ Gene and	notations /	Filtered la	nintensity	/ Gene id	entifiera	1	Scatternlot	/ Clus	tor v	io 4			1		

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.



Proportion of simulated data sets

Prediction Error Estimation: A Comparison of Resampling Methods

Annette M. Molinaro ab* Richard Simon , Ruth M. Pfeiffer*

^aBiostatistics Branch, Division of Cancer Epidemiology and Genetics, NCI, NIH, Rockville, MD 20852, ^bDepartment of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, ^cBiometric 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 splitsample 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 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 compliation 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@vale.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-offlight) 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.

"to whom correspondence should be addressed

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., crossvalidation. 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 vfold 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. Farly 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 clucidate the 'best' resampling techniques for

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 Crossvalidated 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 D₁, D₂ 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

- 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 $\lambda = b_1 y_1 + ... + b_k y_k$

- 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

- Plot Kaplan Meier survival curves for cases with cross-validated risk percentiles above 50% and for cases with crossvalidated 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

- 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

- 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)