Course #412
Analyzing Microarray Data using the mAdb System

April 1-2, 2008 1:00 pm - 4:00pm

madb-support@bimas.cit.nih.gov

Day 2

mAdb Analysis Tools

Use web site: http://mAdb-training.cit.nih.gov
User Name on your card
Password on the board

Esther Asaki, Yiwen He

Agenda

1. mAdb system overview
2. mAdb dataset overview
3. mAdb analysis tools for dataset
   - Class Discovery - clustering, PCA, MDS
   - Class Comparison - statistical analysis
     - t-test
     - One-Way ANOVA
     - Significance Analysis of Microarrays - SAM
   - Class Prediction - PAM

Various Hands-on exercises

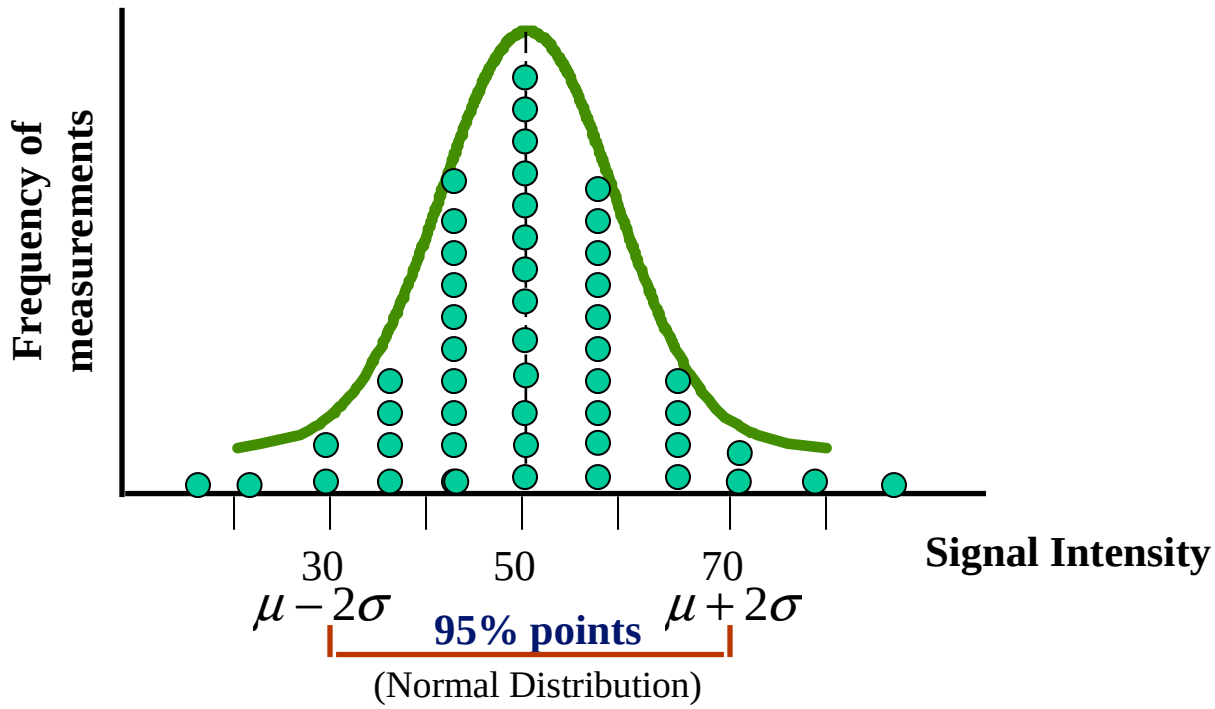
Class Comparison

- Why statistical analysis for gene expression data
- Hypothesis test and two types of errors
- mAdb statistical analysis tools for class comparison
  - t-test
  - One-way ANOVA
  - SAM

Class Comparison

- Why statistical analysis for gene expression data
- Hypothesis test and two types of errors
- mAdb statistical analysis tools for class comparison
  - t-test
  - One-way ANOVA
  - SAM

Distribution for Expression Data



Center: Mean μ

Spread: Standard deviation σ

Sources of Variation in Microarray Data

- Biological variation
  - Random
    - Stochastic mechanism of gene expression
    - Sample heterogeneity
    - Patient to patient variation
  - Due to the biological process under study

- Technical variation
  - Printed probes
  - RNA sample extraction
  - Labeling efficiency
    - Spot size
    - Sample distribution on the arrays
  - Background signals
  - Cross hybridization

Problems with Fold Change

- Genes with high fold change may exhibit high variability among cell types due to natural biological variability for these genes

- Genes with small fold changes may be highly reproducible and should be biologically essential genes

- Some systematic sources of variation are intensity-dependent. Simple, static fold-change thresholds are too stringent at high intensities and not stringent enough at low intensities.

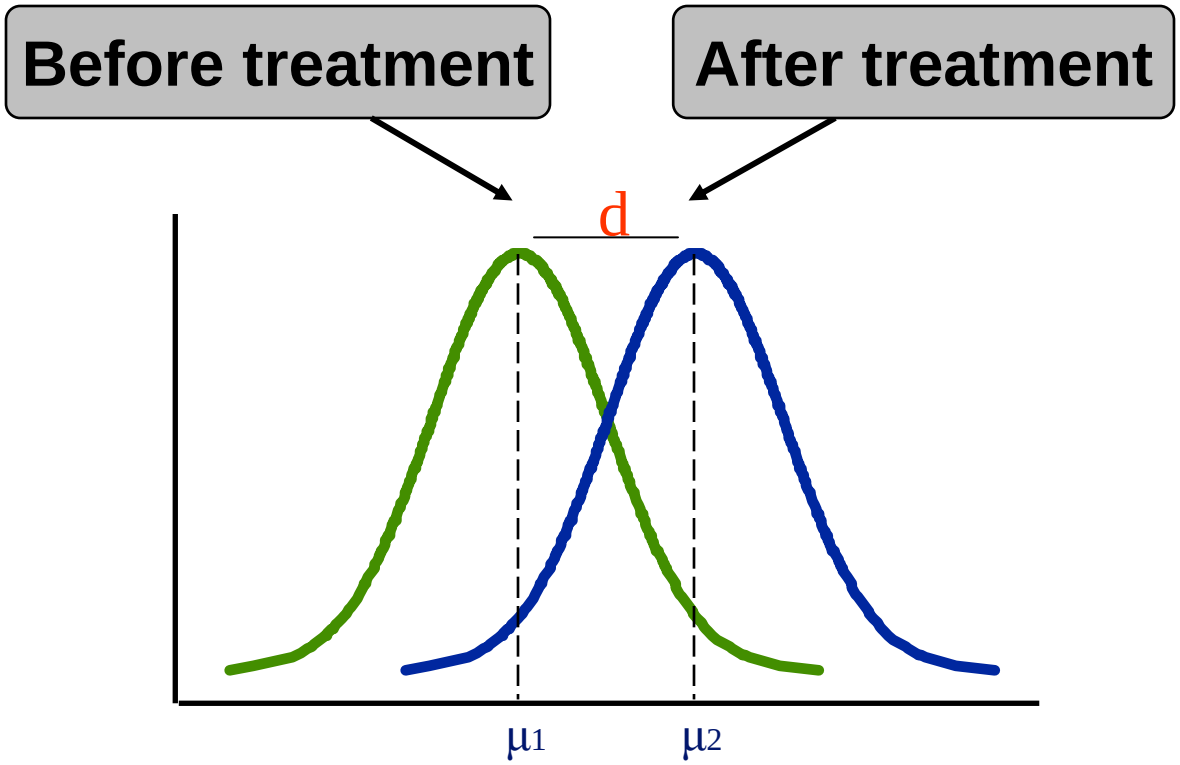
Take Home Messages

- Replicates (both biological and technical) are needed to remove random error
- Need normalization to remove systematic variability
- Need robust statistical tests
- Need additional biological validations

Class Comparison

- Why statistical analysis for gene expression data
- Hypothesis test and two types of errors
- mAdb statistical analysis tools for class comparison
  - t-test
  - One-way ANOVA
  - SAM

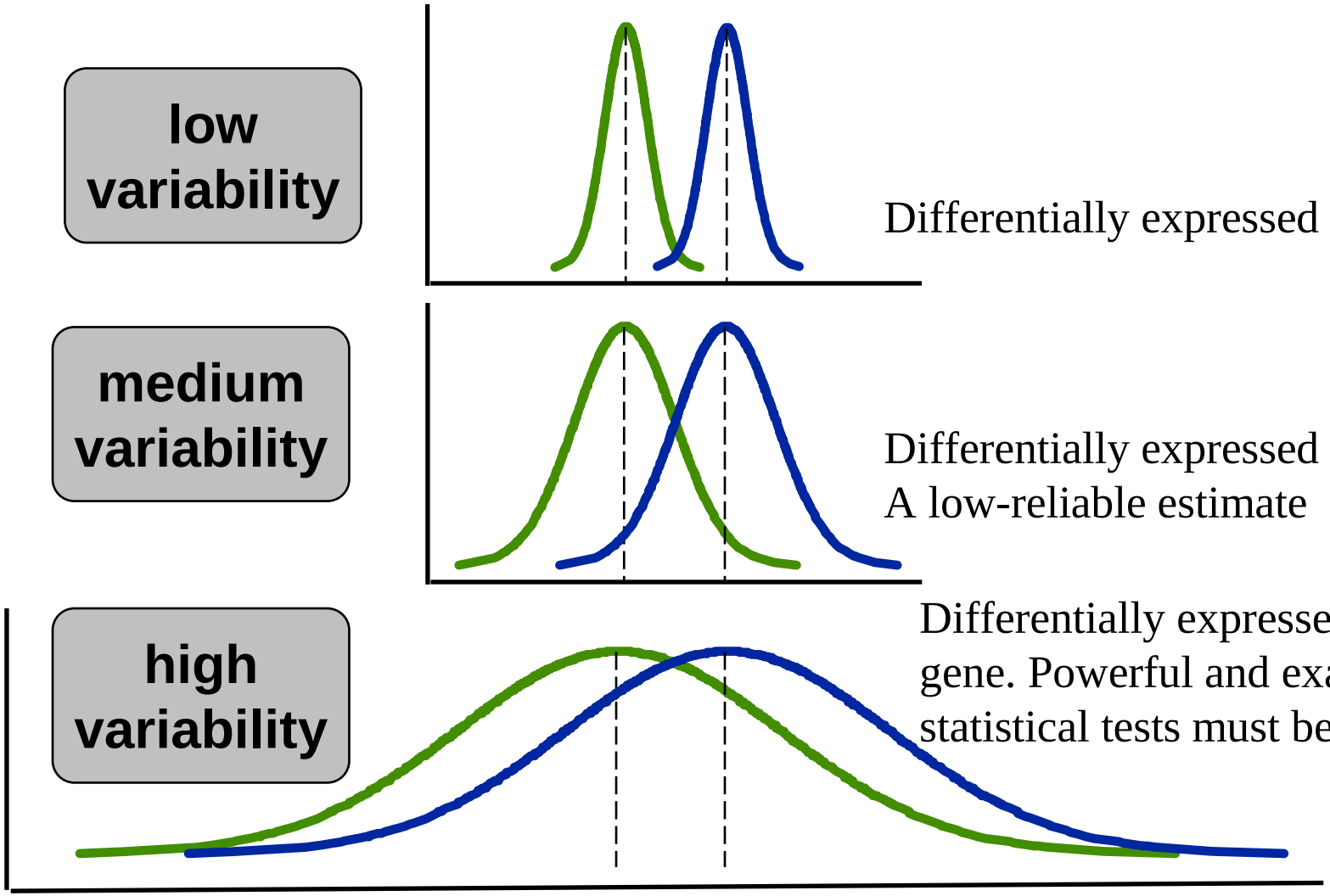
Hypothesis Test



Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypotheses $H_1 : \mu_1 \neq \mu_2$

Spread (Variability) of Measurements



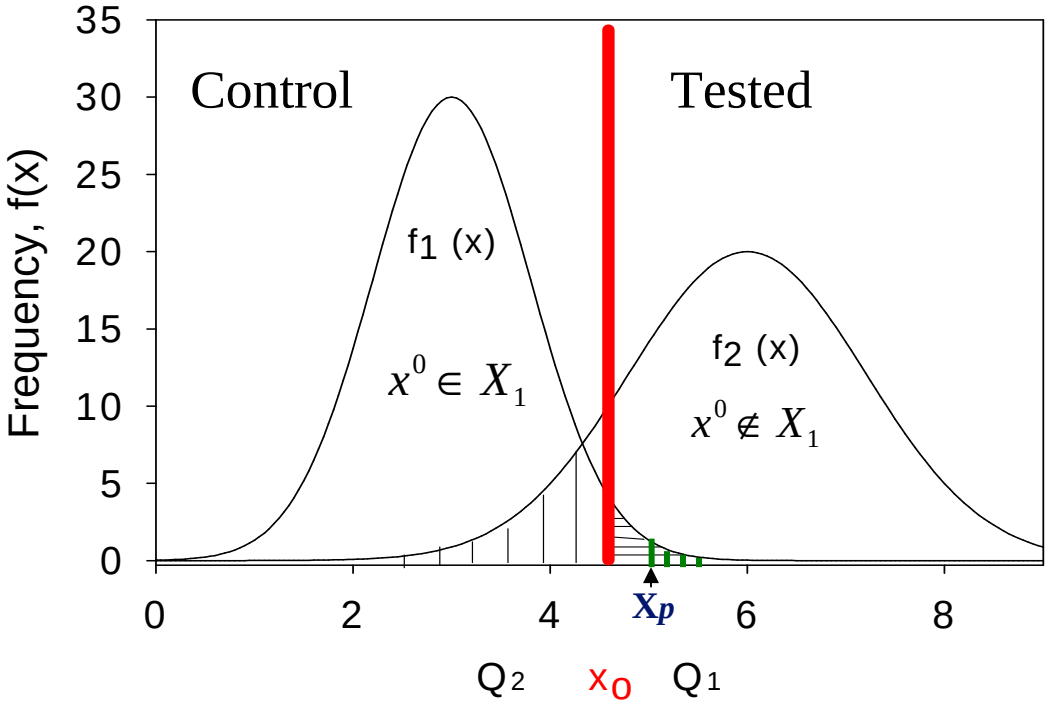
Two Types of Errors

Type I error: Rejecting the null hypothesis while it's true;

Type II error: Accepting the null hypothesis while it's not true.

	Accept *Ho*	Reject *Ho*
Ho is true	Correct decision	Type 1 error False positive
Ho is false	Type II error False negative	Correct decision

Relation of Type I & Type II Errors



$f_1(x)$: expression in control population
$f_2(x)$: expression in tested population
x^o : the observed value of x
x_0 : the critical (rejection) value of x

Q_1=The probability of a type I error (false-positive)
Q_2=The probability of a type II error (false-negative)

- Modifications of x_0 have opposite effects on Type I and type II errors.
- Increasing the sample size (number of replicates) will reduce both errors.
- ***p-value***: the probability (significance value) of observing X_p or bigger under H_0.

Class Comparison

- Why statistical analysis for gene expression data
- Hypothesis test and two types of errors
- mAdb statistical analysis tools for class comparison
  - t-test
  - One-way ANOVA
  - SAM

Statistical Analysis

Goal: To identify differentially expressed genes, i.e. a list of genes with expression levels statistically and (more important) biologically different in two or more sets of the representative transcriptomes.

- t-test (1 or 2 groups)
- One-Way ANOVA (> 2 groups)
- SAM (1, 2, and more groups)

Data for mAdb One-Group Test

- Design: Two conditions, tumor vs. normal (or treated vs. untreated), labeled with Cy3 and Cy5, respectively.
- Data: Ratio, one group
- Null hypothesis: mean is equal to 1
- Results: A list of genes with ratio significantly different from 1. i.e. Different expression level in the two conditions.
- Note: due to dye bias, it’s better to do a dye swap.

Data for mAdb Two-Group Test

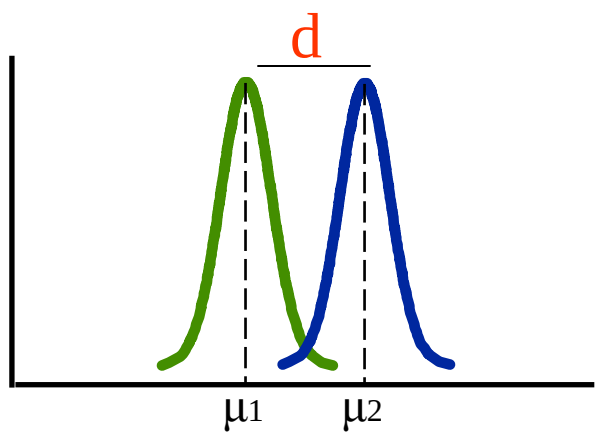
- Affymetrix
  - Normal in group 1 and tumor in group2.
  - Paired test if normal and tumor are from the same patient.
- Two-color with common reference
  - Normal as common reference with Cy3, two types of tumor (group 1 and group 2) both with Cy5.
  - Pooled as common reference, normal and tumor (group 1 and group 2) both with Cy5. Paired if normal and tumor are from the same patient.

Two-group t-Test

The t-test assesses whether the means of two groups are statistically different

The null hypothesis:

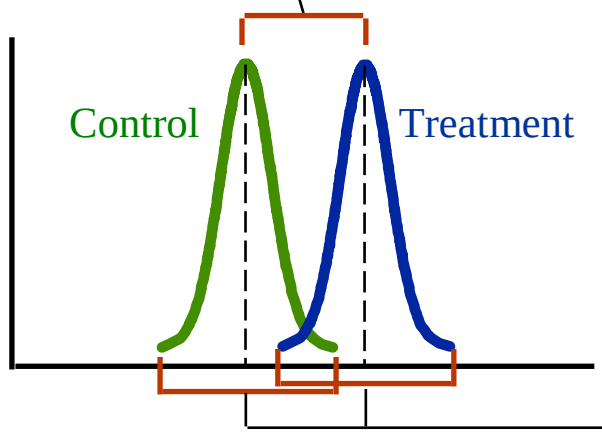
$$H_o : \mu_1 - \mu_2 = 0$$



t-Test (Cont'd)

$$\text{t-statistic} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

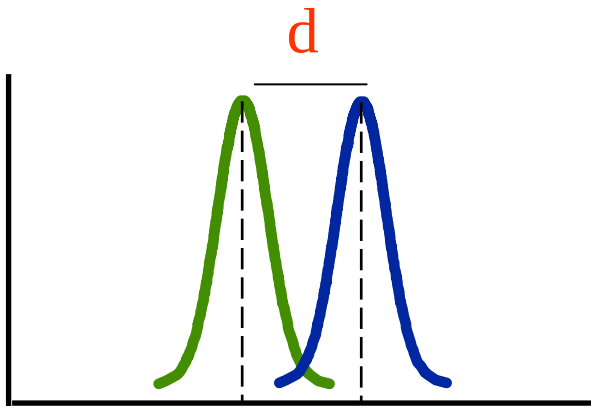


Calculating p-Value (t-Test)

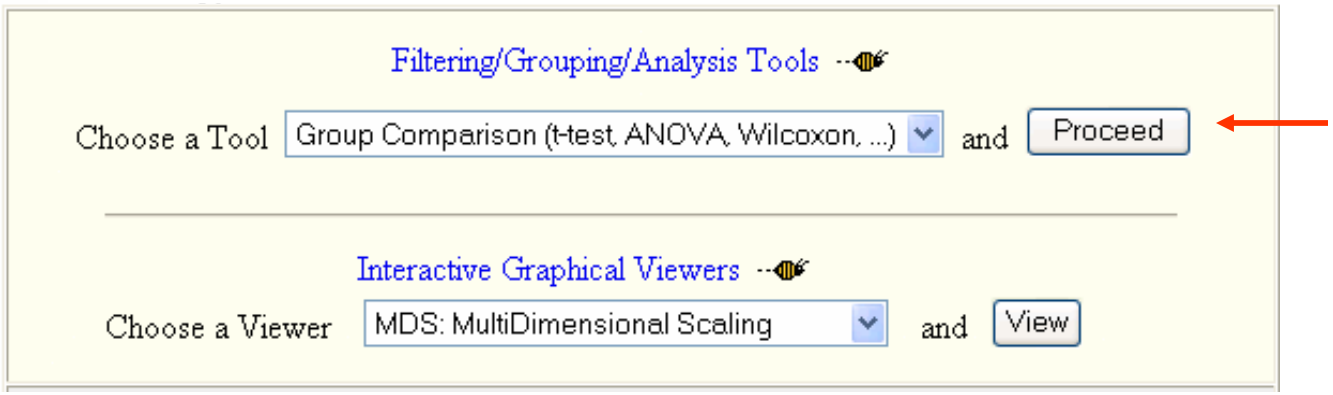
- The p-value is the probability to reject the null hypothesis ($H_o : \mu_1 - \mu_2 = 0$) when it is true (e.g. p=0.0001)

- Calculated based on t and the sample sizes n_1 and n_2.

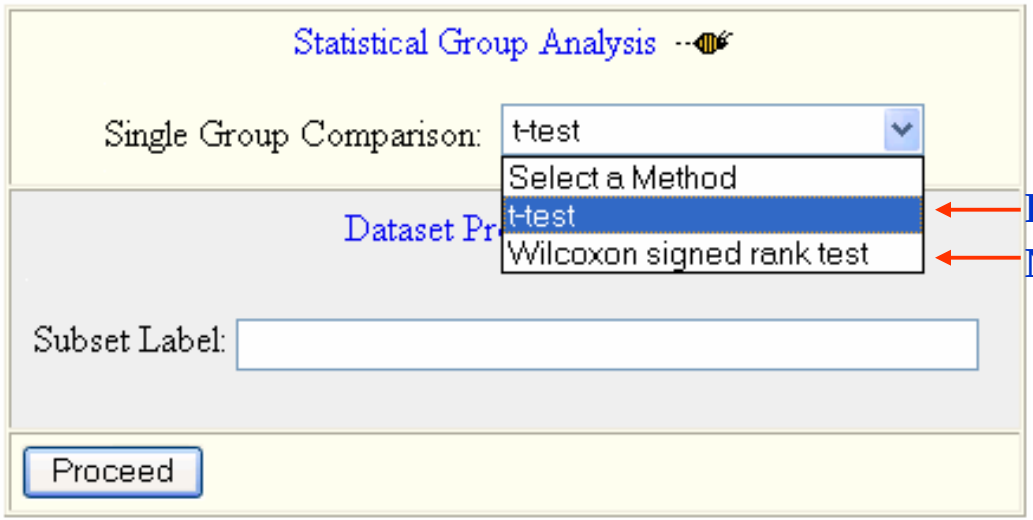
Large distance d,
low variability,
large sample sizes,
then small *p*,
i.e. more significant.



mAdb One-Group Test



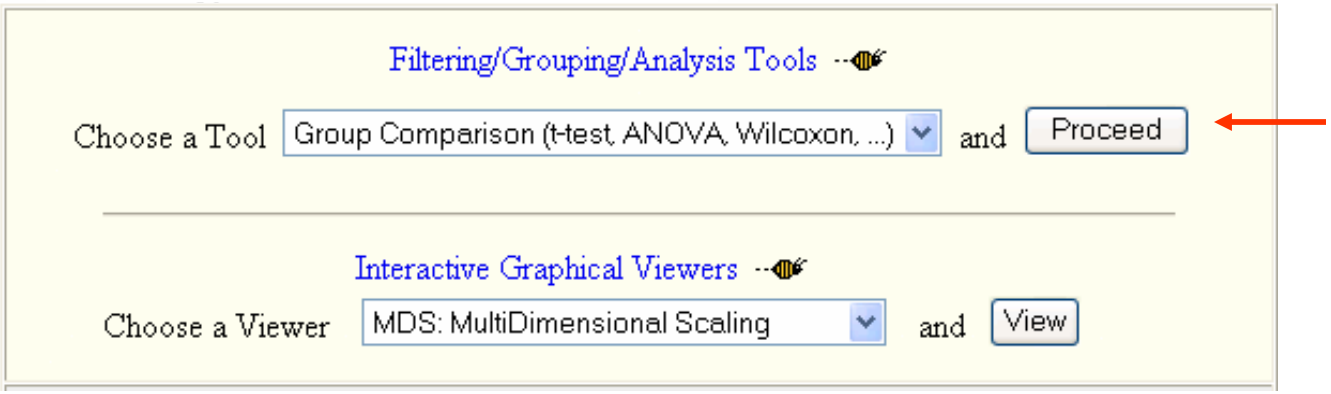
1 group statistic analysis automatically selected for a single group dataset



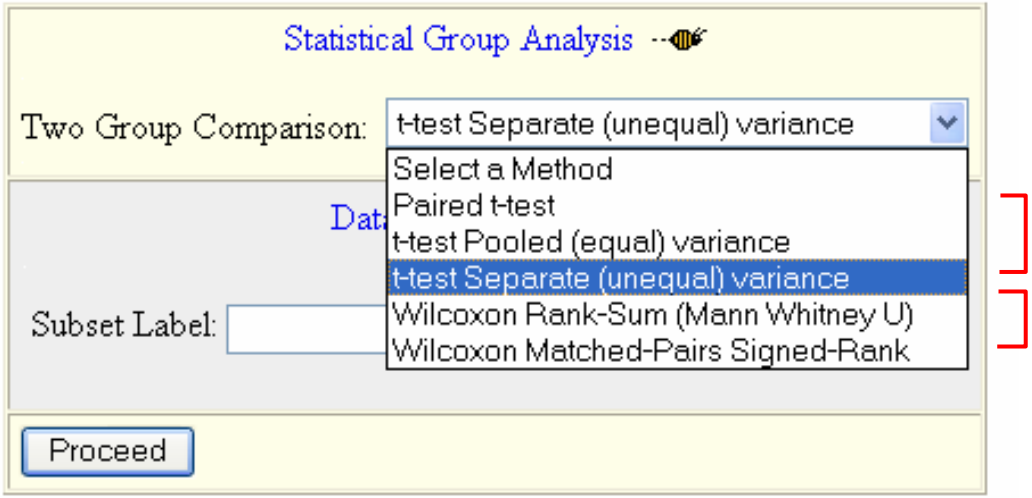
Parametric (normal distribution)

Non-Parametric (distribution free)

mAdb Two-Group Test



2 group statistic analysis automatically selected for a 2 group dataset



Parametric (normal distribution)

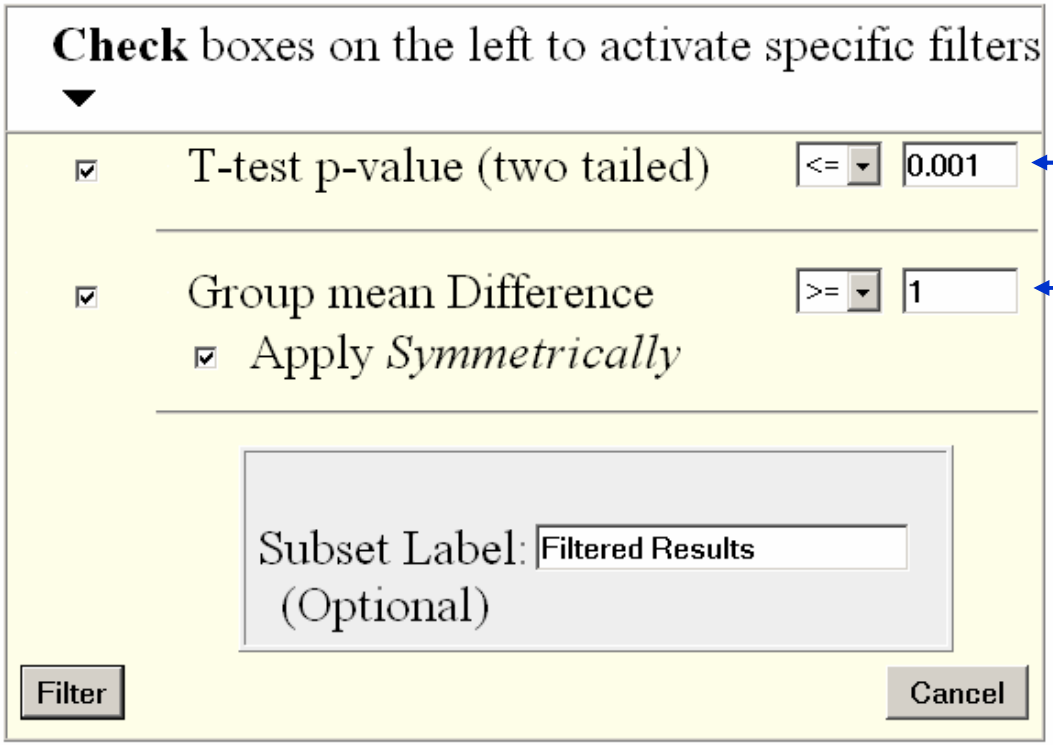
Non-Parametric (distribution free)

Two-Group t-Test Results

$\overline{\log2(A)} - \overline{\log2(B)}$

A	A	A	B	B	B		
JIM3_A	JJN3_A	U266_A	HDLM2_A	L428_A	L540_A	p-Value	Difference
52.4309	54.9520	45.0046	0.7800	0.6485	0.8532	1.9737e-06	6.07
35.1142	52.4541	42.8235	0.7800	0.6485	0.8532	8.9006e-06	5.83
53.3166	74.5535	46.5118	0.7800	0.6485	0.8532	1.1662e-05	6.24
5.9693	5.9444	5.7954	9.4782	9.6511	10.0555	1.4619e-05	-0.72
12.2739	13.0063	9.6026	0.7800	0.6485	0.8532	2.4704e-05	3.93
0.6680	0.6954	0.6536	9.0445	8.4780	13.0657	3.7853e-05	-3.9
3.7943	3.4277	3.3739	7.3190	7.6012	7.2551	4.7738e-05	-1.07
0.6680	0.6954	0.6536	2.3401	2.0402	2.5358	4.9127e-05	-1.77
0.6680	0.6954	0.6536	7.6466	6.0506	9.6493	5.7477e-05	-3.51
0.6680	0.6954	0.9490	8.0788	8.5636	6.8106	5.8369e-05	-3.35
0.6680	0.6954	0.7869	68.9017	34.0804	72.9403	6.3509e-05	-6.28
34.7315	29.5014	60.8882	0.7800	0.6485	0.8532	7.1258e-05	5.71
0.6680	0.6954	0.6706	0.8424	0.8593	0.8532	8.4299e-05	-0.329
0.6680	0.6954	0.6536	39.1841	17.6407	27.2176	9.1539e-05	-5.31
3.7288	2.9875	3.1098	0.9774	0.8392	0.8532	9.9425e-05	1.88
0.6680	1.3275	0.6536	26.2949	22.3119	26.9078	0.00014347	-4.91
1.7328	1.8435	2.0412	0.8557	0.9196	0.8532	0.00014599	1.09

Statistic Results Filtering



statistical significance, i.e. p-value

$\overline{\log_2(A)} - \overline{\log_2(B)}$

Multiple Group Comparison

	Group 1	Group 2	…	Group k
Gene 1	$\mu_{1.1}$	$\mu_{1.2}$	…	$\mu_{1.k}$
Gene 2	$\mu_{2.1}$	$\mu_{2.2}$	…	$\mu_{2.k}$
…	…	…	…	…
Gene n	$\mu_{n.1}$	$\mu_{n.2}$	…	$\mu_{n.k}$

n: Number of genes/probes
k: number of groups, $k > 2$

Data for mAdb Multiple-Group Test

- Time course/Dose response
- Normal vs. multiple types of tumor
- For two-color arrays, must have common reference.
  - More than two types of tumor/treatments, with normal/untreated as common reference
  - Normal, tumor type I, tumor type II, etc. with some common reference.

Analysis of Variances (ANOVA)

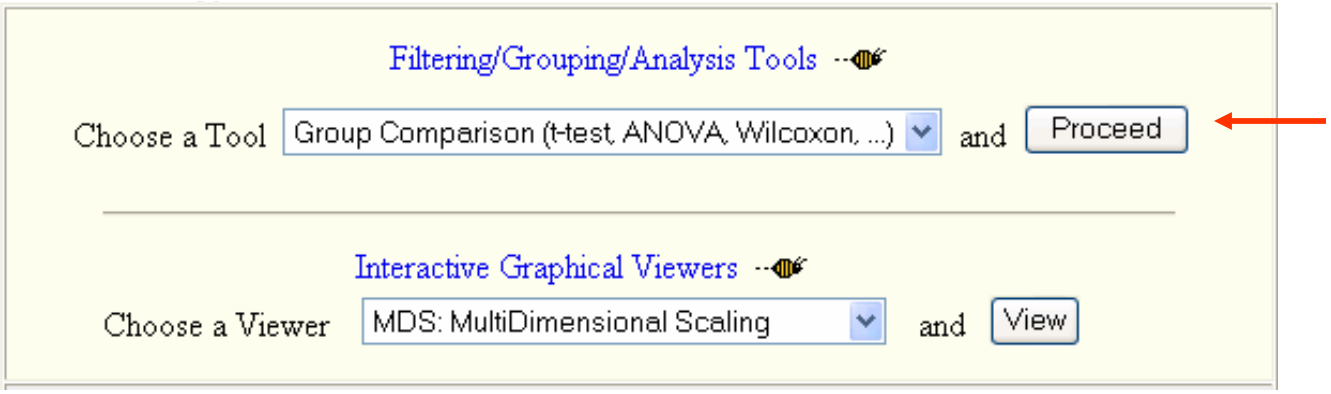
To compare several population means:

$$H_o : \mu_1 = \mu_2 = \ldots = \mu_k \quad (k > 2)$$

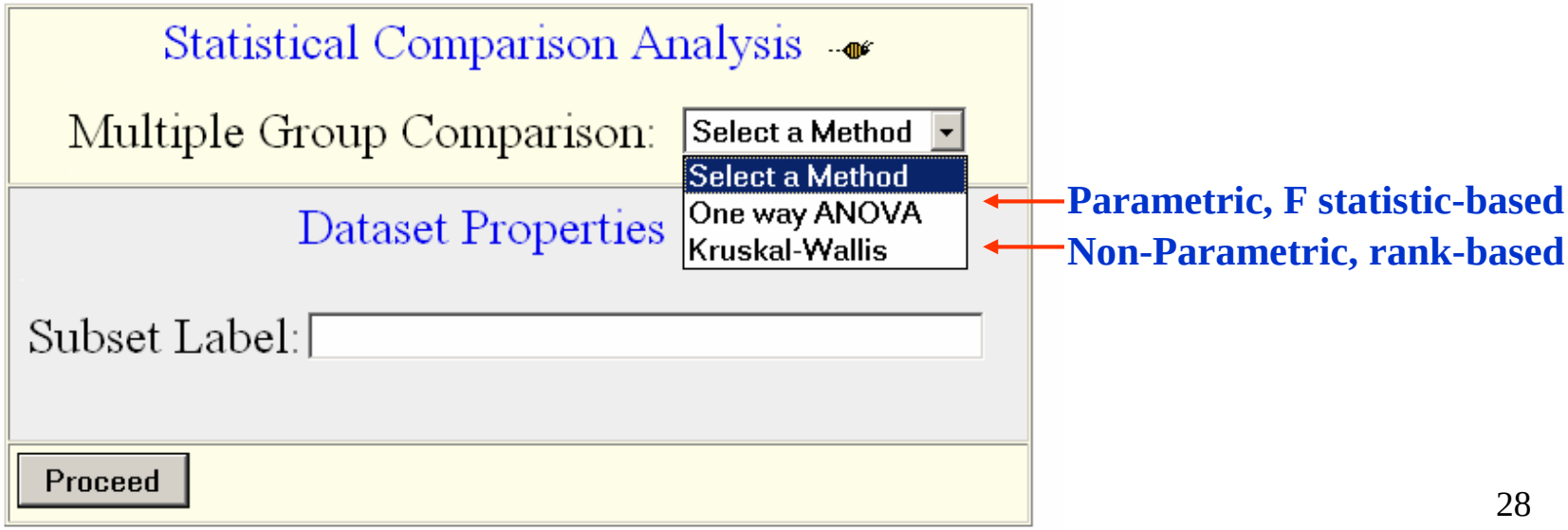
vs.

$$H_1 : \mu_i \neq \mu_j; \quad \text{for some } 1 \leq i \neq j \leq k$$

mAdb Multiple-Group Test



Multiple group analysis automatically selected for a > 2 group dataset

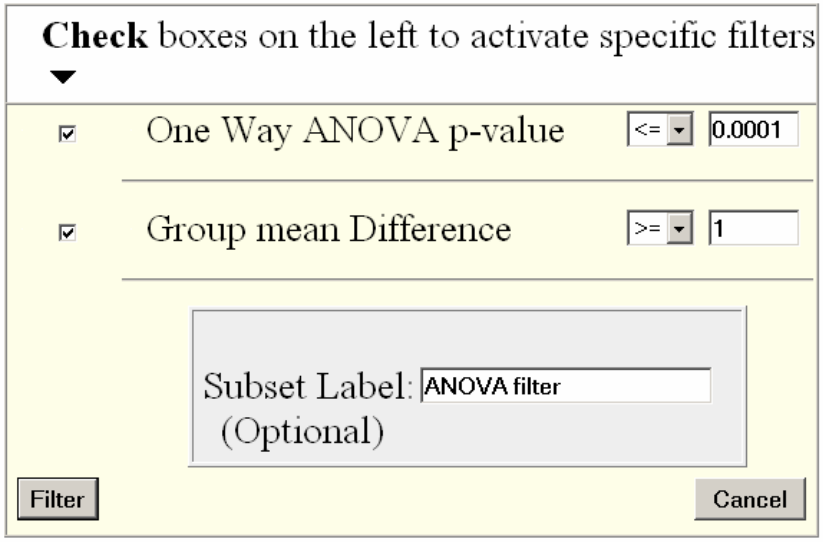


ANOVA Results and Filtering

p-Value	Difference	Groups
9.6276e-22	4.11	A-B
3.488e-20	2.99	D-C
2.5008e-19	3.59	A-B
2.5733e-18	2.59	A-D
1.4459e-17	2.76	D-A
5.7703e-17	2.89	A-B
8.728e-17	3.14	D-B
1.3957e-16	3.95	C-A
4.1114e-16	4.03	A-B
1.4464e-15	3.76	A-B
2.369e-15	3.1	D-B
7.4515e-15	3.32	A-B
8.187e-15	2.76	A-C
2.5078e-14	4.1	A-B
2.5526e-14	5.68	D-B

← Group Pair for Max Mean Difference

↑ Maximum Difference between Group Means



Hands-on Session 4

- Lab 9
- Total time: 10 minutes

Multiple Comparison

- Statistical problems with large-scale experiments
  - Many null hypotheses are tested simultaneously in microarray, one for each probe.
  - Although p-value cut off (α) of 0.01 is significant in a conventional single-variable test, a microarray experiment for 20,000 gene probes would identify 20,000 x 0.01 = 200 genes just by chance!

Multiple Comparison Correction

- False Discovery Rate (FDR)

	Not Rejected	Rejected	Total
H_0 True	$m_0 - R_0$	R_0	m_0
H_1 True	$m_1 - R_1$	R_1	m_1
Total	$m - R$	R	m

m: # hypothesis/genes
R_0: # false positive
R: # significant hypothesis

Probability of false-positive discovery (False Discovery Rate):

$$FDR = E(\frac{R_0}{R} \mid R > 0) \times \Pr(R)$$

Significance Analysis of Microarrays (SAM)

- http://www-stat.stanford.edu/~tibs/SAM/index.html
- Goal is to select a fairly large number of differentially expressed genes (R), accepting some falsely significant genes (R_0), as long as the FDR is low. i.e. R_0 is relatively small compared to R.
- For one or two groups, SAM computes a t-like statistic d(i) for each probe i (i=1,2…n), measuring the relative difference between the group means.
- For more groups, SAM computes a F-like statistic.

SAM for 2 groups

The "relative difference" *d(i)* in gene expression for two groups *I* and *U* of repeated samples is:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

$\bar{x}_I(i)$: average expression level for gene *i* in group I,

$\bar{x}_U(i)$: average expression level for gene *i* in group U,

s(*i*) : standard deviation of repeated measurements,

s_o : the fudge factor that reduces the "relative differences" of the genes with a small *s(i)*, such as low expressed genes (noise) and genes with similar expression levels.

Permutation & the Expected *d* Values

Group I	Group U
a1	b1
a2	b2
a3	b3
a4	b4

Group I	Group U
b1	a1
a2	b2
a3	b3
a4	b4

Group I	Group U
b1	a1
a2	b2
b3	a3
a4	b4

n: the number of hybridized signals (gene probes)
k: the number of permutations

Permutation 1: $d_1(1) \le \ldots \le d_1(n)$

……

Permutation p: $d_p(1) \le \ldots \le d_p(n)$

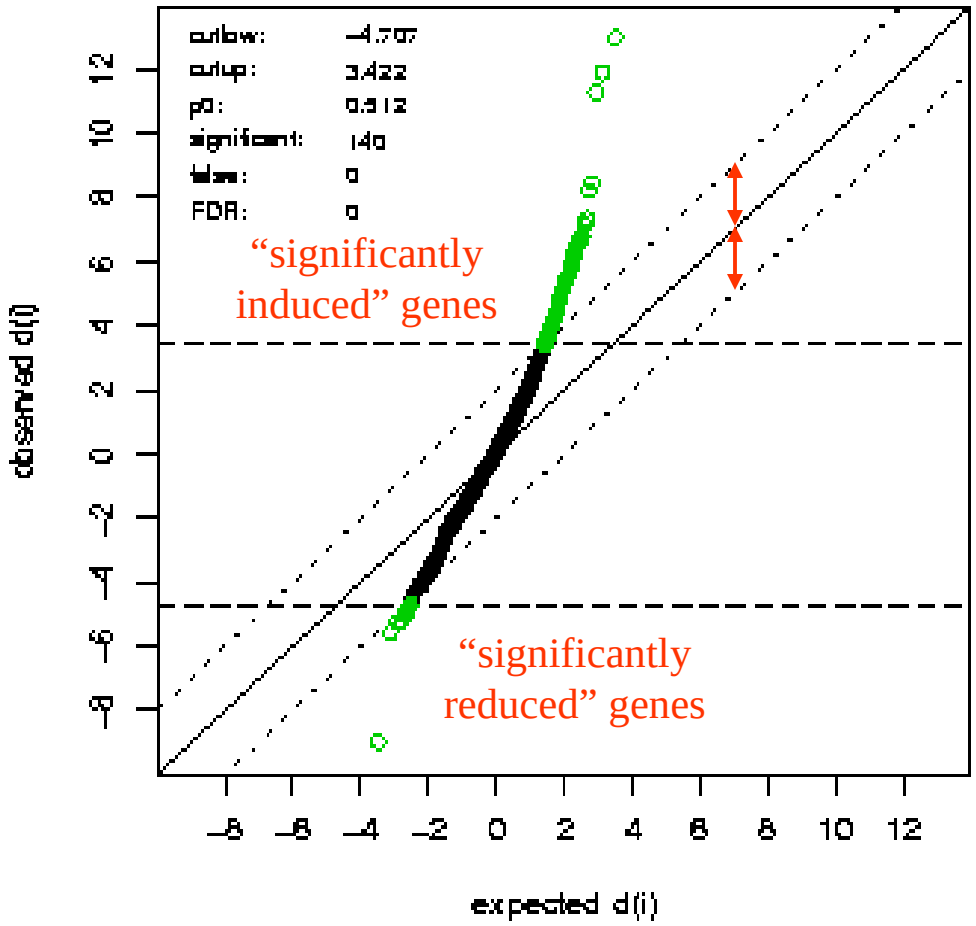
……

Permutation k: $d_k(1) \le \ldots \le d_k(n)$

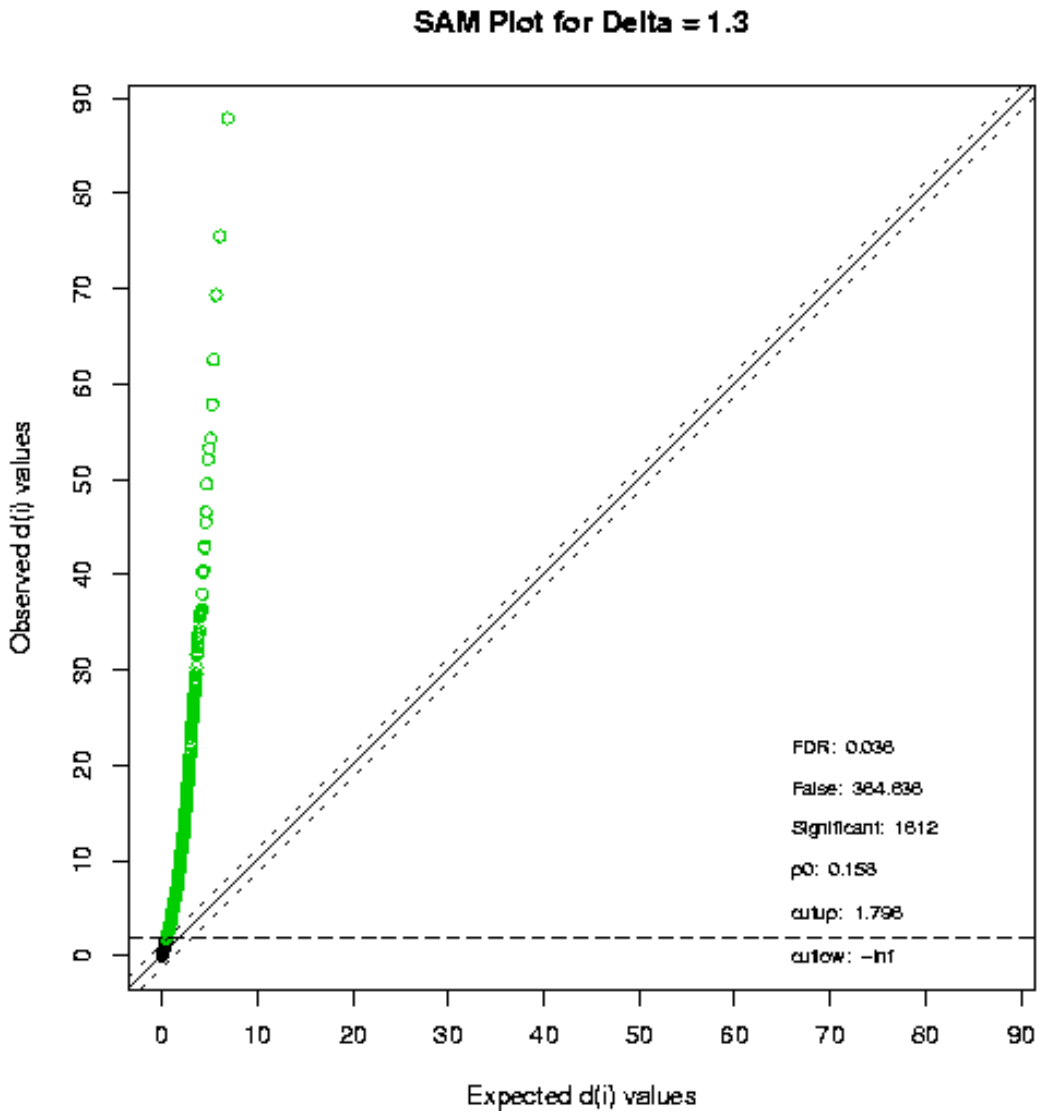
$$\bar{d}(i) = \frac{1}{k}\sum_{i=1}^{k} d_p(i)$$

Expected relative difference for gene *i* (*i*=1,2,…n)

SAM Plot for Delta = 2



SAM Plot Multiple Groups



Calculating FDR

- Order the observed d statistics for all n genes so that $d_o(1) \leq \ldots \leq d_o(i) \ldots \leq d_o(n)$.
- Plot the observed d_o vs. expected d_e
- Select a cutoff value *delta*
- Significant genes (R): $| d_o - d_e | \geq delta$
- False genes from a permutation (R_{0p}) : $| d_p - d_e | \geq delta$
- Estimate false discovery (R_0): median of R_{0p}
- Estimate FDR: R_0 / R

Data for SAM in mAdb

- You can run SAM on data with 1, 2, or more groups
- Experimental design requirements are the same as those for t-test or ANOVA
- Note: SAM assumes that most of the genes in your dataset are NOT changed. So it is recommended that you run SAM on a larger dataset, instead of a small set with mostly significant genes.

mAdb SAM Data

Redisplay

☑ Show Array Details at the top of the page

Background Color - None - Contrast 2

Limiting display to to 25 genes

☑ Show Data Values ☑ Use Names in Column Heading

☐ Apply log2 transform ☐ Use Description in Column Heading

☑ Show Gene Symbols ☐ Show Map Information

☐ Show UniGene Cluster ☐ Show BioCarta Pathways

☐ Show KEGG Pathways

☐ Show GO Tier 2 Component ☐ Show GO Tier 3 Component

☐ Show GO Tier 2 Function ☐ Show GO Tier 3 Function

☐ Show GO Tier 2 Process ☐ Show GO Tier 3 Process

☑ Show Gene Description ☐ Show GO Terms

Save a Feature Property List (used with the Feature Properties Filtering tool).

Records 1 to 25 of 2308 total records displayed.

Data for Subset: **bl and nb**
from Dataset: **Small, Round Blue Cell Tumors (SRBCTs), Nature Medicine Vol 7, Num 6, 601-673 (2001)**

Filter/Group by Array Property
63 arrays and 2308 genes in the input dataset
20 arrays and 2308 genes in the output dataset.
8 arrays assigned to Group A
12 arrays assigned to Group B
Filter/Group by Array Property:
Group A: Array/Set Name Contains 'bl'
Group B: Array/Set Name Contains 'nb'

A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B		
BL-C5	BL-C6	BL-C7	BL-C8	BL-C1	BL-C2	BL-C3	BL-C4	NB-C1	NB-C2	NB-C3	NB-C6	NB-C12	NB-C7	NB-C4	NB-C5	NB-C10	NB-C11	NB-C9	NB-C8	Well ID	Featu
0.2989	0.1856	0.1045	0.3178	0.1437	0.3493	0.3796	0.0683	1.2511	1.2422	0.7843	0.7208	1.7054	1.3452	0.6575	0.5909	1.2263	1.2744	0.9407	0.5555	1080460	IMAGI
0.0839	0.1283	0.0994	0.0494	0.0563	0.0557	0.0640	0.1203	0.2242	0.1277	0.1423	0.0817	0.2167	0.1268	0.0779	0.1264	0.1296	0.0573	0.1279	0.1944	1080461	IMAGI
1.0989	1.7574	0.2362	0.9711	1.0739	1.8981	1.3961	0.5926	1.4717	2.8900	1.1627	0.6389	1.5466	3.1923	1.3970	0.3217	1.2785	1.2974	1.8580	0.7071	1080462	IMAGI
1.3145	1.3695	1.2625	1.2685	0.1198	0.1243	0.3185	0.1137	0.1005	0.1199	0.1469	1.6185	1.7928	1.5470	0.9163	1.2627	1.1213	1.4351	1.3606	1.6350	1080463	IMAGI
0.3285	0.1284	0.1687	0.0573	0.3935	0.3372	0.4620	0.6383	0.4352	0.4861	0.2977	0.1188	0.1924	0.1024	0.0945	0.1382	0.1177	0.0674	0.1523	0.1829	1080464	IMAGI
0.7530	0.5325	0.9698	1.0432	2.3396	2.0050	2.1145	1.7212	2.8457	1.3993	2.5561	1.3040	0.9871	0.6740	0.8526	1.1709	1.7376	1.5479	1.3387	1.6884	1080465	IMAGI
3.0222	4.8113	4.6305	3.7375	3.3334	4.5251	3.3524	3.8142	3.5181	2.9483	5.7054	5.4201	5.6752	4.1266	4.8610	4.5579	4.0917	6.9131	4.7579	6.3929	1080466	IMAGI
2.2284	1.1472	0.6647	0.5825	1.0947	2.2200	1.6359	1.2144	1.4148	0.8492	0.4446	0.6343	1.1375	0.7132	0.5911	0.5642	1.1463	0.6698	0.6328	0.6956	1080467	IMAGI
1.4646	2.8207	2.2148	1.2009	2.2681	1.4484	1.6515	1.8208	0.5277	2.2907	1.7167	1.0464	1.4179	2.7042	0.5633	0.7576	3.5107	2.8599	1.8068	0.7471	1080468	IMAGI
2.0438	2.6476	1.4568	1.6544	1.8761	2.7953	3.0725	1.8915	1.8990	1.4719	0.9198	1.4198	2.2044	1.8135	1.0141	1.0629	1.3173	1.1249	1.7079	1.1799	1080469	IMAGI
4.3938	4.5243	5.8249	5.6817	4.6666	5.2114	4.0503	4.6079	4.0354	3.6700	7.2208	5.0586	5.3212	4.6734	3.8197	4.2099	4.1700	5.8854	5.5536	6.8372	1080470	IMAGI

mAdb SAM

mAdb Dataset Display

Edit Data for Subset: **bl and nb groups**
from Dataset: **Small, Round Blue Cell Tumors (SRBCTs), Nature Medicine Vol 7, Num 6, 601-673 (2001)**

```
Filter/Group by Array Property
63 arrays and 2308 genes in the input dataset
20 arrays and 2308 genes in the output dataset.
8 arrays assigned to Group A
12 arrays assigned to Group B
Filter/Group by Array Property:
Group A: Array/Set Name Contains 'bl'
Group B: Array/Set Name Contains 'nb'
```

View the complete History.

Expand this Dataset.
Access Datasets in your Temporary area.

NEW Post a copy of this Dataset to other mAdb users.

Filtering/Grouping/Analysis Tools

Choose a Tool [SAM: Significance Analysis of Microarrays] and [Proceed]

Interactive Graphical Viewers

Choose a Viewer [MDS: MultiDimensional Scaling] and [View]

Dataset Retrieval & Display Options

[Retrieve] Dataset formatted for [Eisen Cluster]

[Redisplay] ☐ Show Array Details at the top of the page
Background Color [- None -] Contrast [2]
Limiting display to [to 25 genes]

mAdb SAM

mAdb SAM Options

SAM help

*** Notice ***

By default, any genes with missing values are removed for SAM analysis.
Currently you can chose to replace those missing values with row mean values.
A mAdb "*Missing Value Imputation*" tool is in final testing and is expected to be available soon, which offers more option for handling missing values.

Handling Missing Values: Remove

Number of permutations: 500

Use a fixed random seed (reproducible results): Yes

Continue Cancel

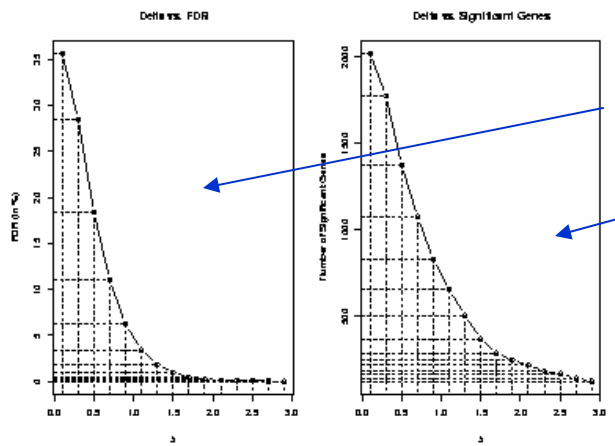
mAdb SAM Results I

Clicking on a Delta value to create a new subset or on a image icon to generate the corresponding SAM plot;
or input a Delta value at the bottom and Click "Create Subset".

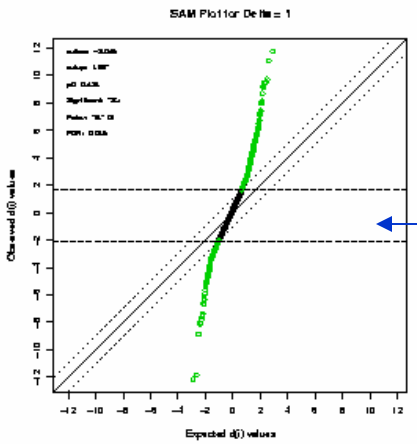
Delta	# of Sig. Genes	# of False Positives	FDR*
0.1	2020	719.87	0.3564
0.2	1946	646.87	0.3324
0.3	1776	505.57	0.2847
0.4	1605	376.30	0.2345
0.5	1374	252.76	0.1840
0.6	1194	171.35	0.1435
0.7	1077	118.13	0.1097
0.8	929	77.12	0.0830
0.9	829	51.30	0.0619
1.0	735	34.15	0.0465
1.1	654	22.35	0.0342
1.2	587	14.92	0.0254
1.3	504	9.16	0.0182
1.4	428	5.54	0.0129
1.5	370	3.53	0.0095
1.6	340	2.45	0.0072
1.7	287	1.40	0.0049
1.8	267	0.96	0.0036
1.9	245	0.64	0.0026
2.0	232	0.42	0.0018
2.1	218	0.28	0.0013
2.2	197	0.20	0.0010
2.3	182	0.13	0.0007
2.4	176	0.08	0.0005
2.5	169	0.07	0.0004
2.6	157	0.04	0.0003
2.7	142	0.03	0.0002
2.8	135	0.02	0.0002
2.9	125	0.02	0.0001

Create Subset

FDR: (# of False Positives)/(# of Sig. Genes)



Above as EPS, PDF, PNG



Above as EPS, PDF, PNG

SAM plot for a particular Delta

SAM result subset

Delta vs. FDR
Delta vs. Sig. genes

SAM plot for a default Delta

mAdb SAM Results II

☑ d.value ☑ Stand. Deviation
☑ q.value ☑ Fold Change

Save a Feature Property List (used with the Feature Properties Filtering tool).

Records 1 to 25 of 370 total records displayed.

d.value	Stand. Deviation	q.value	Fold Change	Well ID	Feature ID	Map	UniGene	Gene
-12.1298	0.2684	0	0.0518	1081305	IMAGE:183337	6p21.3	Hs.77522	HLA-DMA
-11.8486	0.3205	0	0.0384	1082374	IMAGE:840942	6p21.3	Hs.814	HLA-DPB1
11.7632	0.2149	0	12.3195	1081310	IMAGE:563673	5q31	Hs.74294	ALDH7A1
11.0799	0.2100	0	10.7428	1081326	IMAGE:784593	2q23.3	Hs.6838	ARHE
9.7225	0.2372	0	9.1553	1081886	IMAGE:504791	6p12.1	Hs.169907	GSTA4
9.5226	0.2314	0	8.3336	1082121	IMAGE:377048	2q12-q34	Hs.121576	MYO1B
9.5000	0.3766	0	18.2186	1082060	IMAGE:629896	5q13	Hs.103042	MAP1B
9.4193	0.3259	0	12.8741	1081201	IMAGE:812105	1q21	Hs.75823	AF1Q
9.2278	0.2293	0	7.7811	1082481	IMAGE:204545	2p13.1	Hs.8966	TEM8
9.1644	0.3089	0	14.6853	1080695	IMAGE:878280	4p16.1-p15	Hs.155392	CRMP1
-8.8426	0.2167	0	0.1633	1081617	IMAGE:814526	20q13.31	Hs.236361	RNPC1
8.6979	0.3444	0	11.1301	1081525	IMAGE:486110	3q25.1-q25.2	Hs.91747	PFN2
8.1327	0.3580	0	11.4990	1082603	IMAGE:308231	2q12-q34	Hs.121576	MYO1B
-8.1047	0.4404	0	0.0448	1082375	IMAGE:80109	6p21.3	Hs.198253	HLA-DQA1
8.1040	0.1935	0	4.6185	1082036	IMAGE:813742	16p12.1-p11.2	Hs.70500	KIAA0370
-8.0900	0.2531	0	0.1701	1080610	IMAGE:745343	2p12	Hs.1032	REG1A
7.9838	0.2486	0	6.2752	1081034	IMAGE:823886	17	Hs.296842	
-7.8279	0.3406	0	0.0956	1081295	IMAGE:241412	13q13	Hs.154365	ELF1
-7.5480	0.2898	0	0.1597	1080582	IMAGE:236282	Xp11.4-p11.21	Hs.2157	WAS

SAM d statistics (normalized distance)

Significance value (lowest FDR)

Average(B)/Average(A) (for 2-group only)

mAdb SAM Results III

- [x] d.value
- [x] q.value
- [x] Groups
- [x] Stand. Deviation
- [x] Max Group Mean Difference

Save a Feature Property List (used with the Feature Properties Filtering tool).

Records 1 to 25 of 400 total records displayed.

d.value	Stand. Deviation	q.value	Max Group Mean Difference	Groups	Well ID	Feature ID
87.8794	0.5766	0	4.1071	A-B	1081848	IMAGE:770394
75.5112	0.5097	0	2.9854	D-C	1082414	IMAGE:784224
69.3372	0.6445	0	3.5930	A-B	1080705	IMAGE:377461
62.5424	0.4836	0	2.5858	A-D	1082413	IMAGE:814260
57.8456	0.5291	0	2.7609	D-A	1081462	IMAGE:796258
54.2733	0.4645	0	2.8916	A-B	1081004	IMAGE:1435862
53.2386	0.5035	0	3.1403	D-B	1081653	IMAGE:859359
52.0802	1.2099	0	3.9545	C-A	1082509	IMAGE:295985
49.4837	1.1140	0	4.0322	A-B	1080566	IMAGE:365826
46.5782	1.0812	0	3.7594	A-B	1081778	IMAGE:866702
45.4725	0.4809	0	3.1012	D-B	1080460	IMAGE:21652
42.9725	0.8917	0	3.3179	A-B	1082104	IMAGE:52076
42.7721	0.5400	0	2.7641	A-C	1081301	IMAGE:810057
40.4288	1.1435	0	4.1011	A-B	1082167	IMAGE:43733
40.3929	2.6457	0	5.6842	D-B	1080646	IMAGE:296448

SAM d statistics (normalized distance)

Significance value (lowest FDR)

Group pair with max difference

Hands-on Session 5

- Lab 10
- Total time: 10 minutes

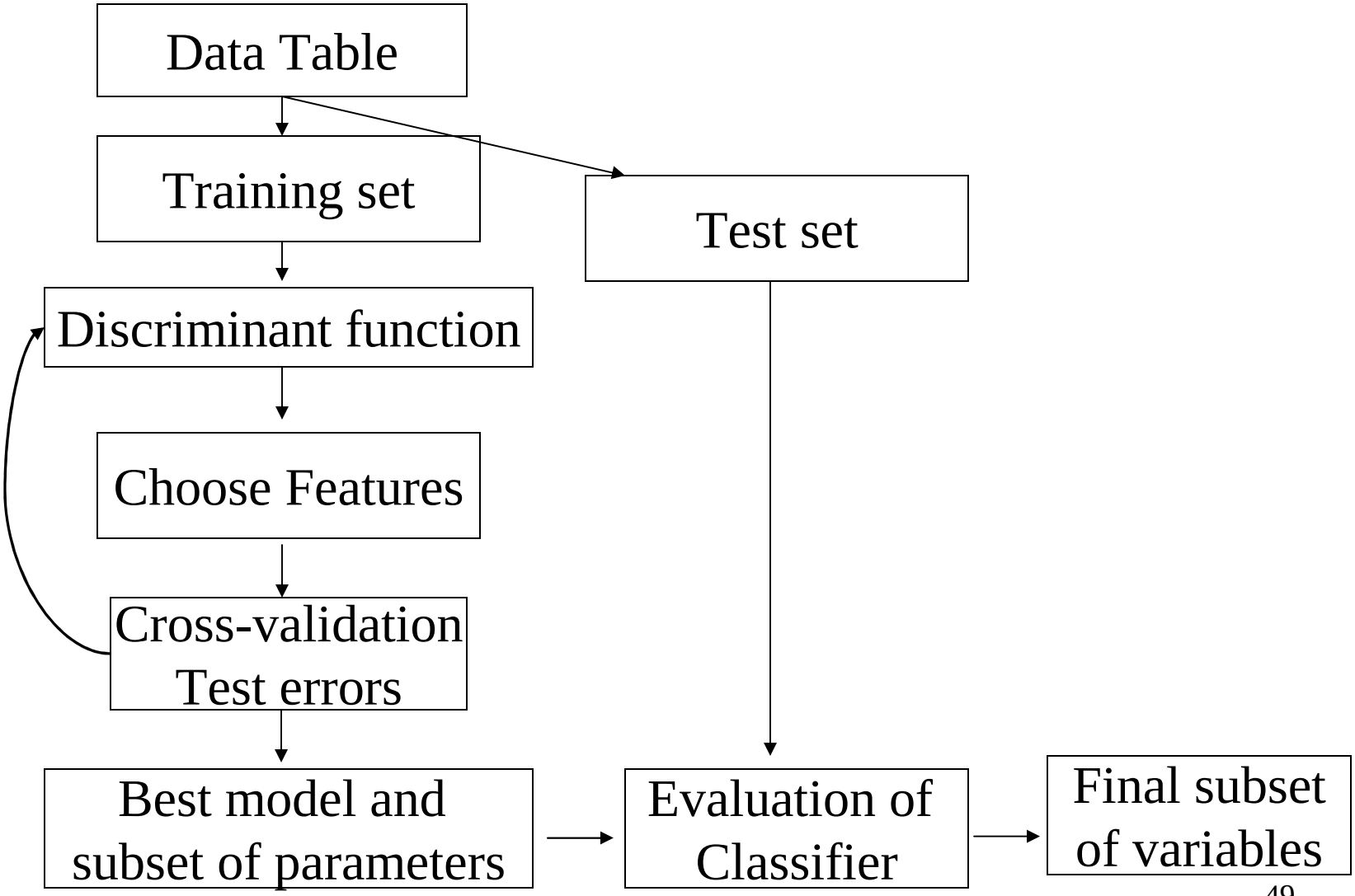
Agenda

1. mAdb system overview
2. mAdb dataset overview
3. mAdb analysis tools for dataset
   - Class Discovery - clustering, PCA, MDS
   - Class Comparison - statistical analysis
     - t-test
     - One-Way ANOVA
     - Significance Analysis of Microarrays - SAM
   - Class Prediction - PAM

Class Prediction
Supervised Model for Two or More Classes

- Prediction Analysis for Microarrays (PAM)
- http://www-stat.stanford.edu/~tibs/PAM
- Provides a list of significant genes whose expression characterizes each class
- Estimates prediction error via cross-validation
- Imputes missing values in dataset

Design of the PAM algorithm



Calculating the Discriminant Function

For each gene i, a centroid (mean) is calculated for each class k.

Standardized centroid distance:
Class average of the gene expression value minus the overall average of the gene expression value, divided by a standard deviation-like normalization factor (NF) for that gene.

$$d_{ik} \textbf{ (centroid distance) = (class k avg – overall avg) / NF}$$

Creates a normalized average gene expression profile for each class.

Reducing the Feature Set

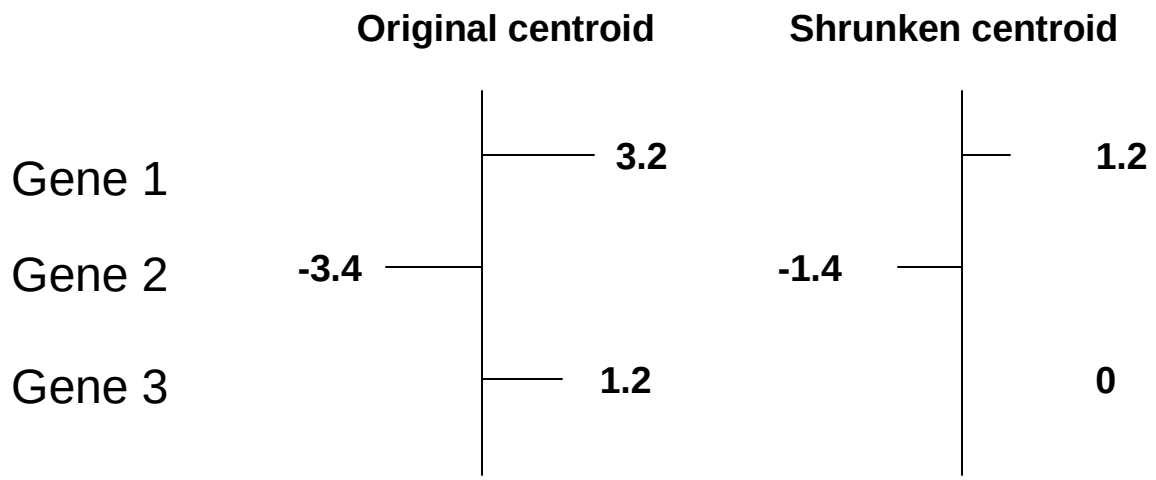
Nearest shrunken centroid:
To "shrink" each of the class centroids toward the overall centroid for all classes by a threshold we call Δ.

Soft threshold:
To move the centroid towards zero by Δ, setting it to zero when it hits zero.

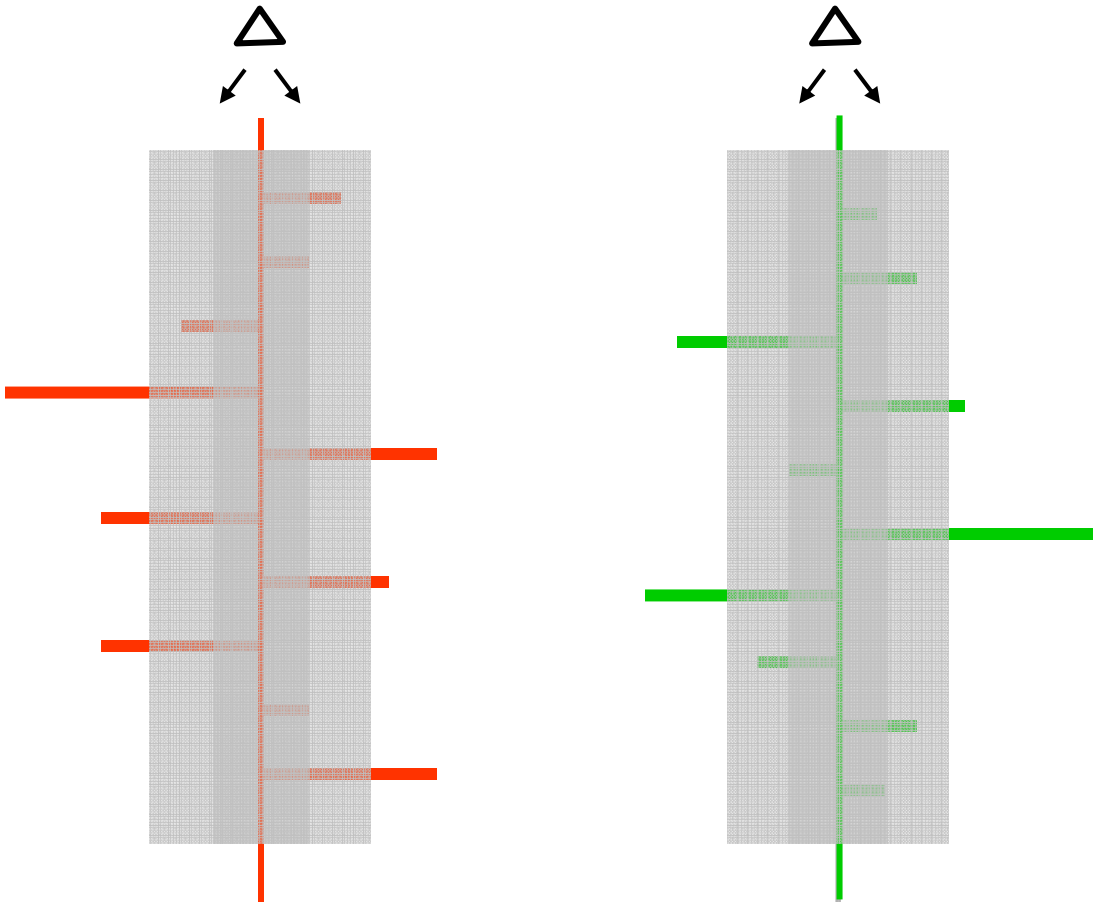
After shrinking the centroids, the new sample is classified by the usual nearest centroid rule, but using the shrunken class centroids.

Shrinking the Centroid

Threshold $\Delta = 2.0$:
a centroid of 3.2 would be shrunk to 1.2;
a centroid of -3.4 would be shrunk to -1.4;
and a centroid of 1.2 would be shrunk to 0.



Reduce Gene Number



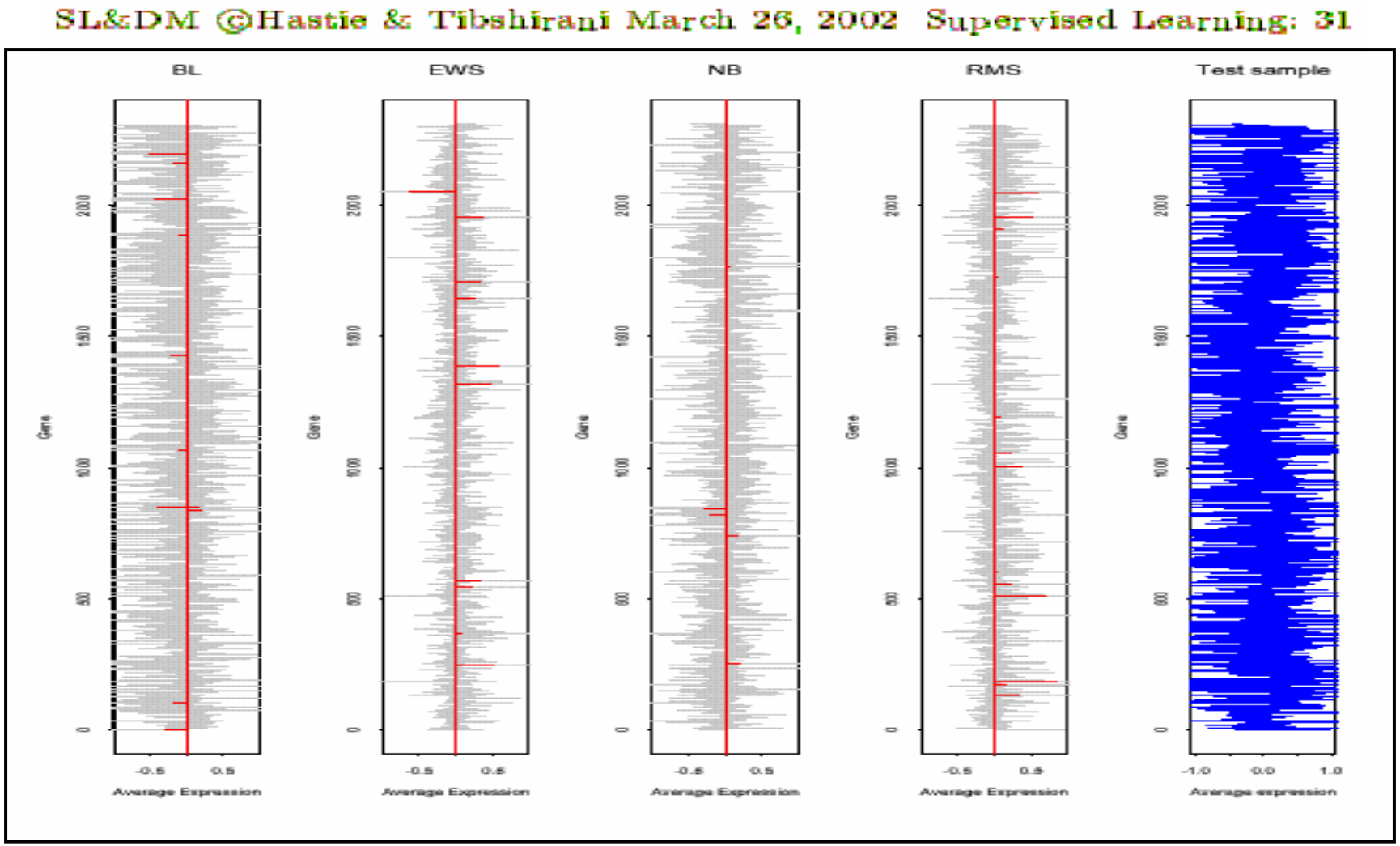
Group A

Group B

Sample

- 63 Arrays representing 4 groups
  - BL (Burkitt Lymphoma, n1=8)
  - EWS (Ewing, n2=23)
  - NB (neuroblastoma, n3=12)
  - RMS (rhabdomyosarcoma, n4=20)
- There are 2308 features (distinct gene probes)
- No missing values in array data sets
- Each group has an aggregate expression profile
- An unknown can be compared to each tumor class profile to predict which class it most likely belong

Class Centroids



Compare model with new tumor tissues to make diagnosis

Classifying an Unknown Sample

- Comparison between the gene expression profile of a new unknown sample and each of these class centroids.

- Classification is made to the nearest shrunken centroid, in squared distance.

K-fold Cross Validation

•The samples are divided up at random into K roughly equally sized parts.

Entire Data Set

50 Group A

25 Group B

25 Group C

K = 5

1	2	3	4	5
10 Group A	10 Group A	10 Group A	10 Group A	10 Group A
5 Group B	5 Group B	5 Group B	5 Group B	5 Group B
5 Group C	5 Group C	5 Group C	5 Group C	5 Group C

K-fold Cross Validation

For each part in turn, the classifier is built on the other K-1 parts then tested on the remaining part.

1 TRAIN	2 TRAIN	3 TRAIN	4 TRAIN	5 TEST

K-fold Cross Validation

1 TRAIN	2 TRAIN	3 TRAIN	5 TRAIN	4 TEST
1 TRAIN	2 TRAIN	4 TRAIN	5 TRAIN	3 TEST

etc….

Estimating Misclassification Error

- PAM estimates the predicted error rate based on misclassification error, which is calculated by averaging the errors from each of the cross validations.

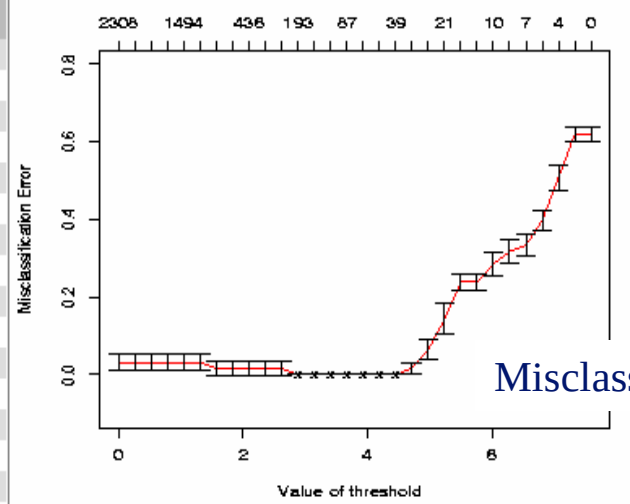
- The model with lowest Misclassification Error is preferred.

PAM Results

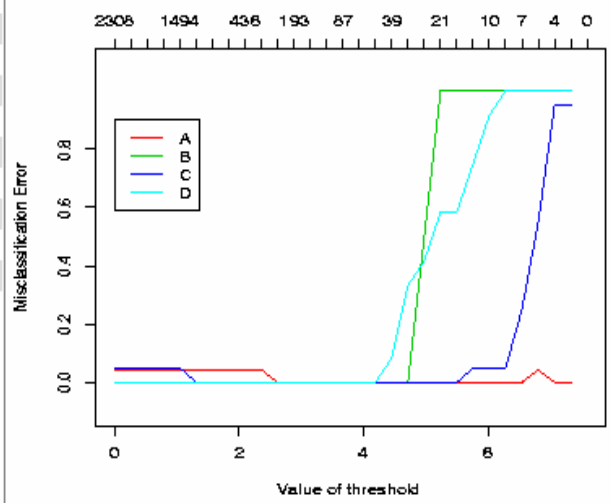
Clicking on a Delta value creates a new data Subset or enter a Delta value at the bottom and Click "Create Subset".

Shrinkage Delta	# of Genes	Misclass. Error
0.000	2308	0.032
0.262	2289	0.032
0.524	2145	0.032
0.786	1878	0.032
1.048	1494	0.032
1.309	1137	0.032
1.571	853	0.016
1.833	609	0.016
2.095	436	0.016
2.357	330	0.016
2.619	244	0.016
2.881 **	**193**	**0.000**
3.143 **	**151**	**0.000**
3.404 **	**107**	**0.000**
3.666 **	**87**	**0.000**
3.928 **	**68**	**0.000**
4.190 **	**52**	**0.000**
4.452 **	**39**	**0.000**
4.714	32	0.016
4.976	23	0.063
5.238	21	0.143
5.499	16	0.238
5.761	11	0.238
6.023	10	0.286
6.285	9	0.317
6.547	7	0.333
6.809	5	0.397
7.071	4	0.508

Create Subset



Above as EPS, PDF, PNG



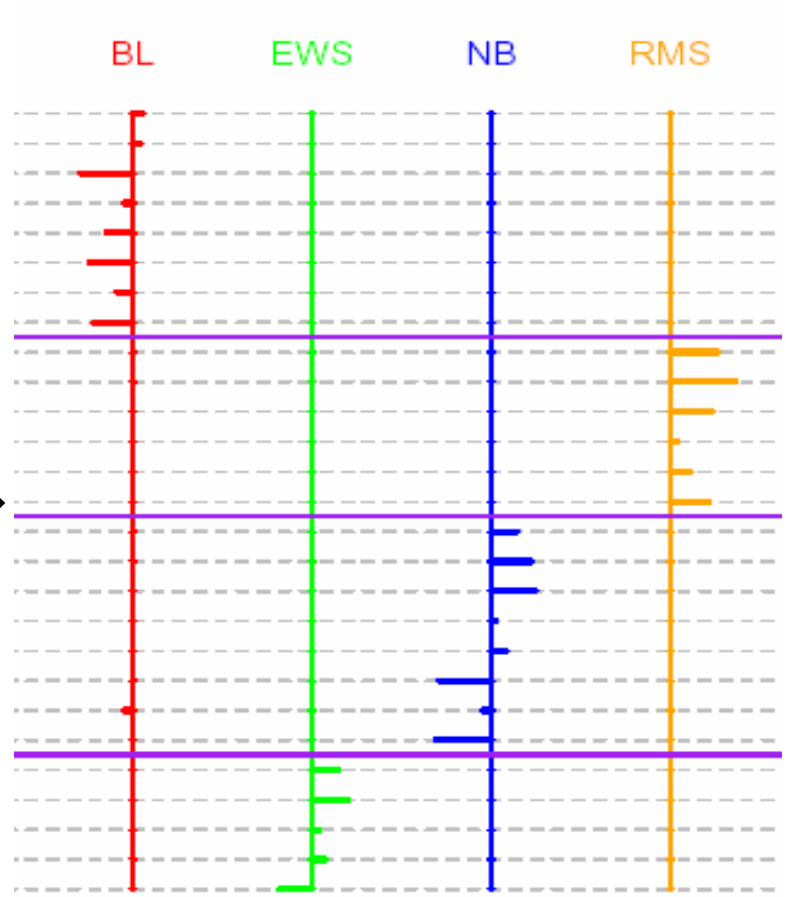
Link leads to the dataset with PAM model →

Create new model by fill in a new Delta value →

Misclassification error

Prediction Model for SRBCT

A Score	B Score	C Score	D Score
0.6092	-0.0866	0.0000	0.0000
0.0000	0.0000	0.0000	0.5862
-0.0696	0.0000	0.0000	0.5764
-0.5421	0.0000	0.0000	0.0000
0.5338	0.0000	0.0000	0.0000
0.0000	-0.5321	0.0000	0.0000
0.0000	0.0000	0.0000	0.4936
0.0000	-0.4873	0.0000	0.0000
0.0000	0.0000	0.0000	0.4821
0.0000	-0.4661	0.0000	0.0000
0.4380	0.0000	0.0000	0.0000
-0.0110	0.0000	0.0000	0.4269
0.0000	-0.4153	0.0000	0.0000
0.4086	0.0000	0.0000	0.0000
0.0000	0.0000	-0.3828	0.0000
0.3346	0.0000	0.0000	0.0000



PAM summary

- It generates models (classifiers) from microarray data with phenotype information
- It does automatic gene selection for each models.
- Misclassification errors are calculated with the data for model selection.
- Require adequate numbers of samples in each group

Hands-on Session 6

- Lab 11, Lab 12 (optional)
- Total time: 15 minutes

References

- Clustering
  - Eisen, et al, Cluster analysis and display of genome-wide expression patterns. *PNAS* 1998, 95:14863-14868.
  - Tavazoie, et al, Systematic determination of genetic network architecture. *Nat Genet* 1999, 22:281-285.
  - Sherlock, Analysis of large-scale gene expression data. *Brief Bioinform* 2001, 2(4):350-62.
- PCA
  - Yeung & Ruzzo, Principal component analysis for clustering gene expression data. *Bioinformatics* 2001, 17(9): 763-74.
- Statistical Analysis
  - Cui & Churchill, Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* 2003, 4:210
- SAM
  - Tusher, Tibshirani and Chu, Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 2001, 98: 5116-5121
- PAM
  - Tibshirani, et al, Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS* 2002, 99:6567-6572

Other Microarray Resources

- Statistical Analysis of Microarray Data & BRB Array Tools (NCI Biometrics Research Branch) class #410. Offered bimonthly; 4/8-9/08
- Partek, R, GeneSpring classes – training.cit.nih.gov
- Introduction to Principal Component Analysis and Distance Geometry class #407
- Clustering: How Do They Make Those Dendrograms and Heat Maps – class #406
- Microarray Interest Group
  - 1st Wed. seminar, 3rd Thu. journal club
  - To sign up: http://list.nih.gov/archives/microarray-user-l.html
- Class slides available on “Reference” page

mAdb Development and Support Team

- John Powell, Chief, BIMAS, CIT
- Lynn Young, Ph.D.
- Esther Asaki*
- Yiwen He, Ph.D.*
- Kathleen Meyer*
- Wenming Xiao, Ph.D.*

*SRA International contractor

http://madb.nci.nih.gov
http://madb.niaid.nih.gov

For assistance, remember:

madb_support@bimas.cit.nih.gov