# Exploring Class Prediction for Leukemia Gene Expression Data


**Alex Smith**


**CAMDA 2000: December 18th, 2000**


**with**

**Jaya Satagopan**

**Mithat Gonen**

**Colin B. Begg**

**Memorial Sloan-Kettering Cancer Center, New York**

# ABSTRACT:

**An increasingly common objective in the analysis of genetic microarray data is to investigate the association between genomic profiles and disease class or outcome (for example, tumor or tissue type). A clinical goal of such efforts would be the ability to predict disease class based solely upon a sample's gene expressions. To accomplish this, we must first select a subset of genes from among all those considered, with the optimal subset being that which best predicts disease class using as few genes as possible.**

**In a recent article Golub et al (1999) analyzed gene expression data from a training set of 38 (27 ALL, 11 AML) and a test set of 34 (20 ALL, 14 AML) leukemia patients for class discovery and prediction. Approximately 1400 genes were found to be highly expressed in ALL or AML. An arbitrary total of 50 genes from among these that were most highly associated with disease type were then used for prediction. The aim of our analysis is to investigate more efficient prediction strategies.**

**Using a two-step procedure, we first selected candidate genes based upon their association with leukemia type using the training set. Next, discriminant functions were generated using the training set for gene subsets of increasing size. The subset providing the maximum classification rate on the test set was then declared optimal.**

**We explored two methods for candidate gene selection. In the first, two-sample t-statistics were calculated for each gene. Genes were then ranked based on the absolute value of these statistics. In the second, genes were selected using stepwise discrimination, where a new gene was chosen based on its association with leukemia type after adjusting for information provided by the genes already selected. While the possible number of candidate genes considered under the t-statistic method can be arbitrary, the maximum number under stepwise discrimination will be limited by the number of samples.**

**In the optimal subset selection step, Fisher's classification functions were developed from the training set on every increasing gene subset size. These were then used to classify the samples in the corresponding test set. The optimal subset was the one providing the maximum classification rate.**

**While all 38 training samples were obtained from adult bone marrow, some test samples came from peripheral blood or pediatric patients. To ensure homogeneity, we derived new training and test sets randomly from the pooled set of all 72 samples, assigning 36 samples to each training and test set. Our results are based on 100 such resamplings.**

**Maximum average classification rates across the 100 test sets were observed to be 91% with the 5 top genes selected by t-statistic method and 88% with the 4 top genes selected by stepwise discrimination. The protein zyxin was selected as the top gene in 45 of the 100 resampled data sets. Classifying all the resampled test data sets using zyxin alone provided an average rate of 92% (range: 78% - 100%). Further, zyxin correctly classified 91% of the 34 patients from the original test set.**

**In conclusion, reanalysis of the leukemia data using these alternative methods provides empirical evidence that the predictive information is contained in a very small subset of the genes.**

# Golub’s Goals:

- **Examine clustering methods for “Class Discovery”**
- **Develop an algorithm for “Class Prediction”**
  - Create a metric to measure gene-class association
  - Determine a cut-off for significant genes
  - Create a weighted-voting prediction scheme
  - Select top 50 genes, and classify test set samples

# Our Goals:

- Examine more efficient methods for “Class Prediction”

# Our Steps:

- Create 100 resampled training and test sets from the original 72 samples to increase homogeneity between sets

- Select and rank promising genes from each training set

- Determine number of genes giving best test set classification

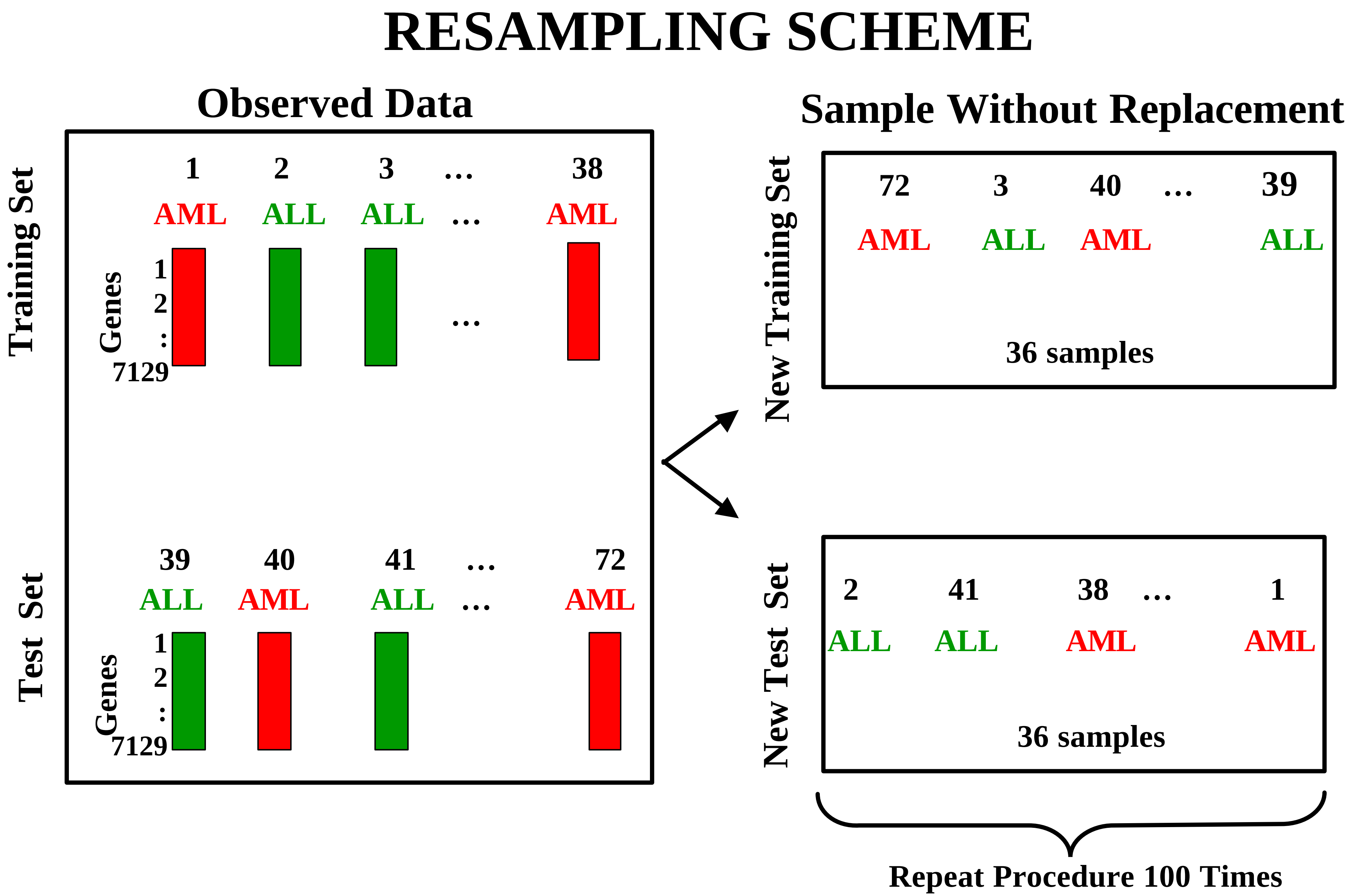
# Leukemia Data
## (7129 genes, standardized for each sample)

### Training Set

- 38 samples (27 ALL, 11AML)
- All samples taken from bone marrow
- All adult leukemia samples
- All samples collected and analyzed in same lab

### Test Set

- 34 samples (20 ALL, 14 AML)
- 24 bone marrow, 10 peripheral blood
- Some adult, some childhood leukemia samples
- Samples analyzed in different labs



RESAMPLING SCHEME
Observed Data
Training Set
1 2 3 … 38
AML ALL ALL … AML
Genes
1
2
:
7129
…
Test Set
39 40 41 … 72
ALL AML ALL … AML
Genes
1
2
:
7129
Sample Without Replacement
New Training Set
72 3 40 … 39
AML ALL AML ALL
36 samples
New Test Set
2 41 38 … 1
ALL ALL AML AML
36 samples
Repeat Procedure 100 Times

# Using the Training Set to Select Promising Genes

## Two Selection Methods:

- **T-statistic**
- **Stepwise Discrimination (ANCOVA)**

# T-statistic

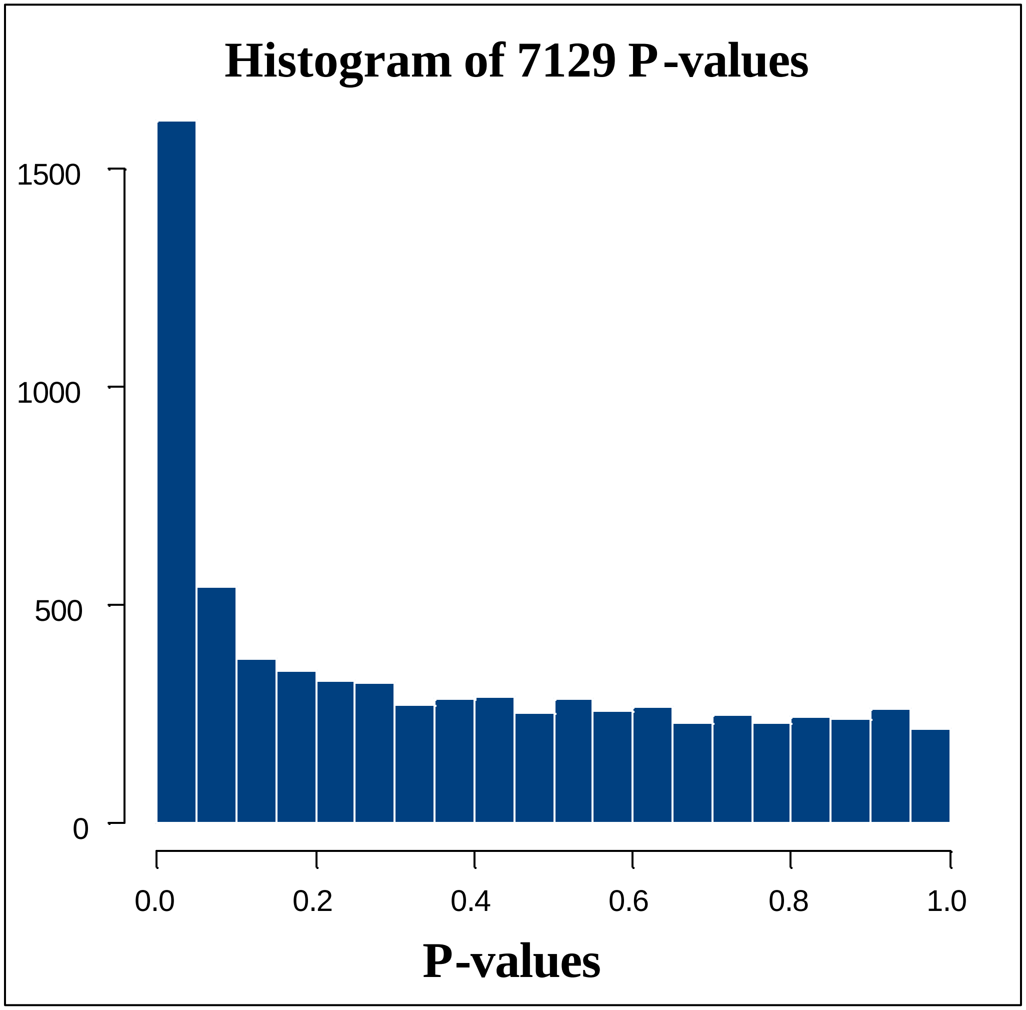
- **For every gene *k* (1 £ *k* £ 7129), compare mean expression in ALL and AML using a t-statistic:**

$$t_k = \frac{\overline{g}_{1k} - \overline{g}_{2k}}{\sqrt{s_k^2 \left(1/n_1 + 1/n_2\right)}},$$

**where** $\overline{g}_{1k}$, $\overline{g}_{2k}$ **are mean expression levels of gene *k* in ALL and AML patients and** $s_k^2$ **is the pooled sample variance.**

- **Rank genes based on absolute t-statistic value.**
- **A candidate subset can be the top *K* genes.**

# T-statistics from a Resampled Training Set




| Alpha Level | Sig.Genes* |
|---|---|
| .05 | 1612 |
| .01 | 816 |
| .001 | 288 |
| .0001 | 113 |
| .00001 | 46 |
| (.05/7129) | 42 |

* P-values not corrected for multiple comparisons

# Stepwise Discrimination

- First gene is selected from an ANOVA model (equivalent to “top” gene found by t-statistic).

- Subsequent genes selected from an ANCOVA model, where previously selected genes are covariates

- Object: Select genes most strongly associated with class, given the information already provided by previously selected genes

# Stepwise Discrimination Procedure

- **Step 1:** For each gene individually, fit the ANOVA model

$$g_{ijk} = m_k + a_{ik} + e_{ijk} \quad \left(\text{where } a_{(\mathrm{ALL})k} + a_{(\mathrm{AML})k} = 0\right),$$

for group *i*, subject *j*, gene *k*; gene expression $g_{\mathrm{ijk}}$, gene mean 💻$_k$, error term ♑$_{\mathrm{ijk}}$

Select gene with the most significant ✂ effect above, and call it $\boldsymbol{g}_{(1)}$

- **Step 2:** Given first gene, fit each remaining gene with ANCOVA model

$$g_{ijk} = m_k + a_{ik} + b_{k(1)} g_{ij(1)} + e_{ijk}$$

where 👓$_{\mathrm{k(1)}}$ is the coefficient for the **covariate** gene selected in step 1

Select gene with most significant ✂ given first gene, and call it $\boldsymbol{g}_{(2)}$

- **Step *K*:** Select $K^{\mathrm{th}}$ gene, using model with *K*-1 covariate genes

$$g_{ijk} = m_k + a_{ik} + b_{k(1)} g_{ij(1)} + \ldots + b_{k(K-1)} g_{ij(K-1)} + e_{ijk}$$

# Comparison of Methods

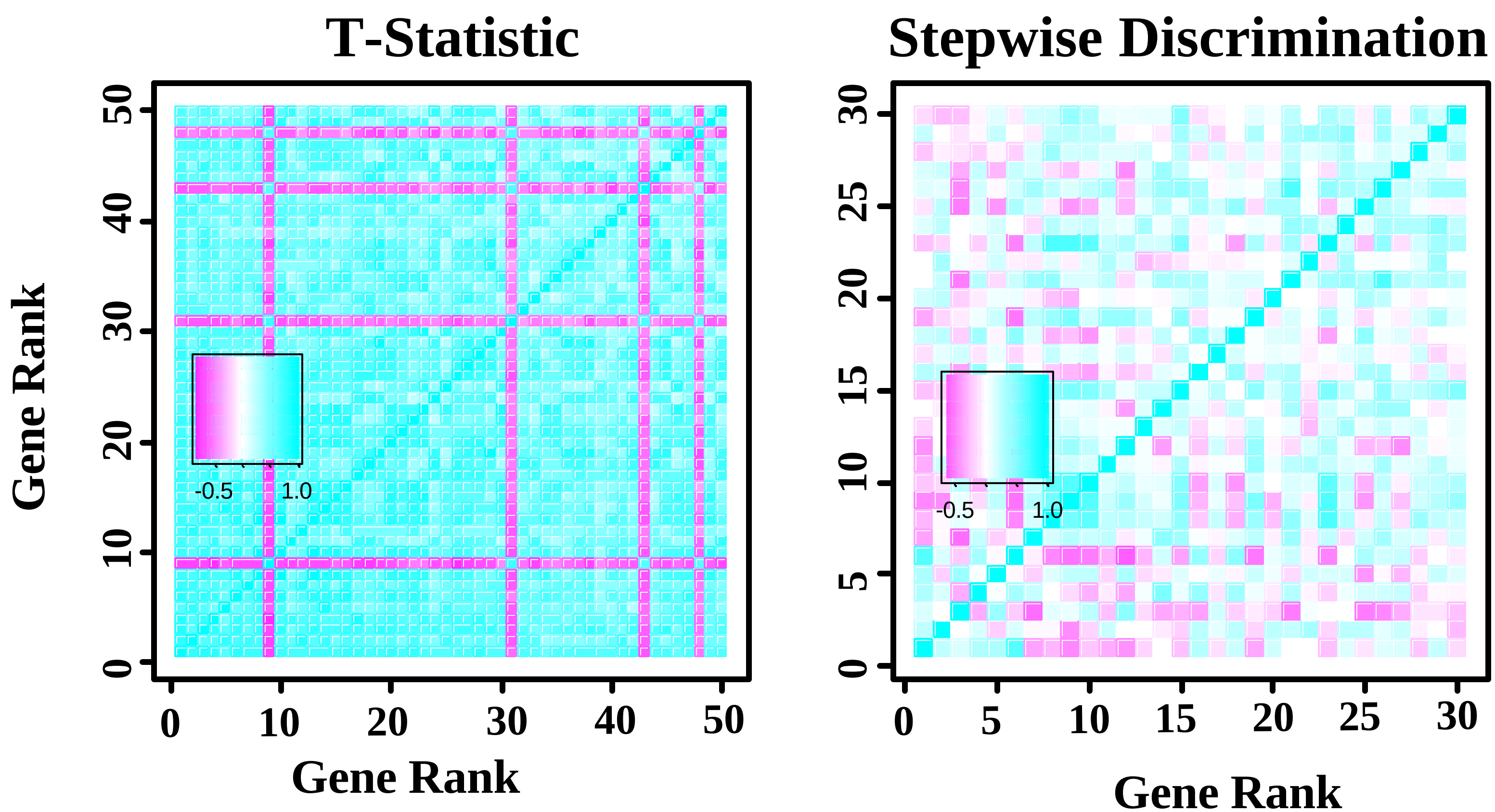
## T-statistic

- Computationally simple
- Compares two groups
- No limit on maximum genes selected
- Selected genes will often be highly correlated

## Stepwise Discrimination

- Computationally intensive
- Compares two or more groups
- Maximum number of genes selected limited by degrees of freedom
- Less likely to select correlated genes

# CORRELATION AMONG TOP GENES IN ONE RESAMPLED TRAINING SET



# Using the Test Set for Classification

Select top genes in training set using either selection method

↓

Create discriminant function from training set

↓

Classify each sample in the test set

↓

Determine the proportion of correct classifications

↓

Repeat last three steps for top 1, 2, . . ., $K$ genes

↓

Observe the number of genes leading to maximum classification rate

## Classification Using Fisher's Discriminant Function

- **Create *K*-gene discriminant function from training set:**

$$d_K = \frac{1}{2}\left(\bar{g}_{1K} - \bar{g}_{2K}\right)' S_K^{-1}\left(\bar{g}_{1K} + \bar{g}_{2K}\right)$$

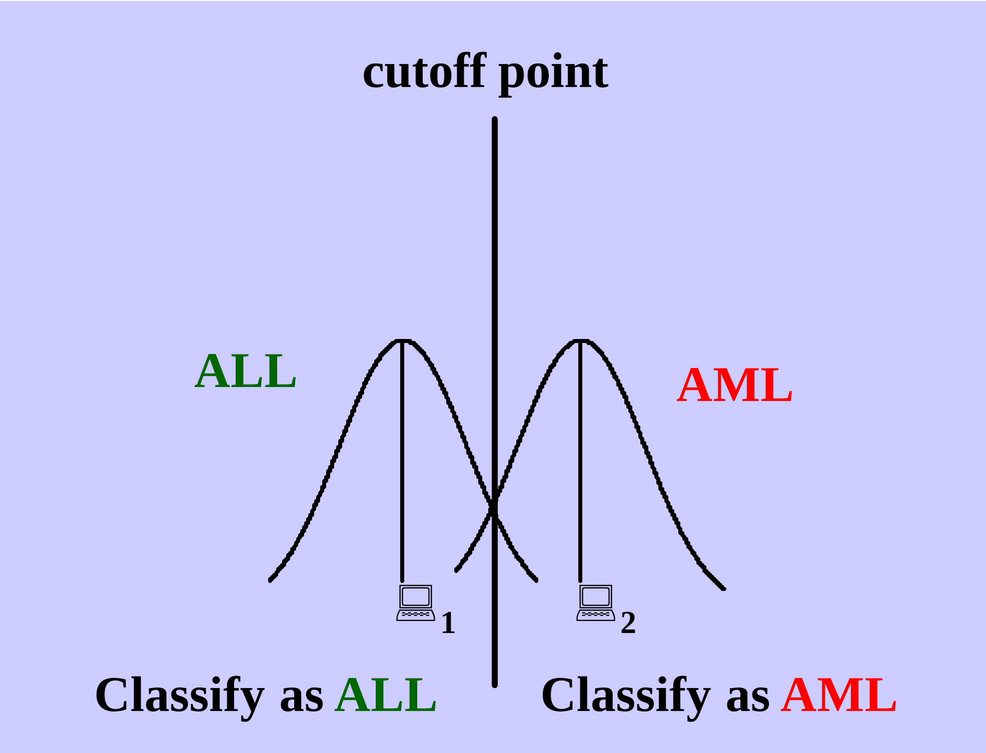
$\bar{g}_{1K}, \bar{g}_{2K}$ : **top *K*-gene mean vectors of AML, ALL**

$S_K^{-1}$ : **pooled covariance matrix**

- **Classify test sample *j* as AML if**

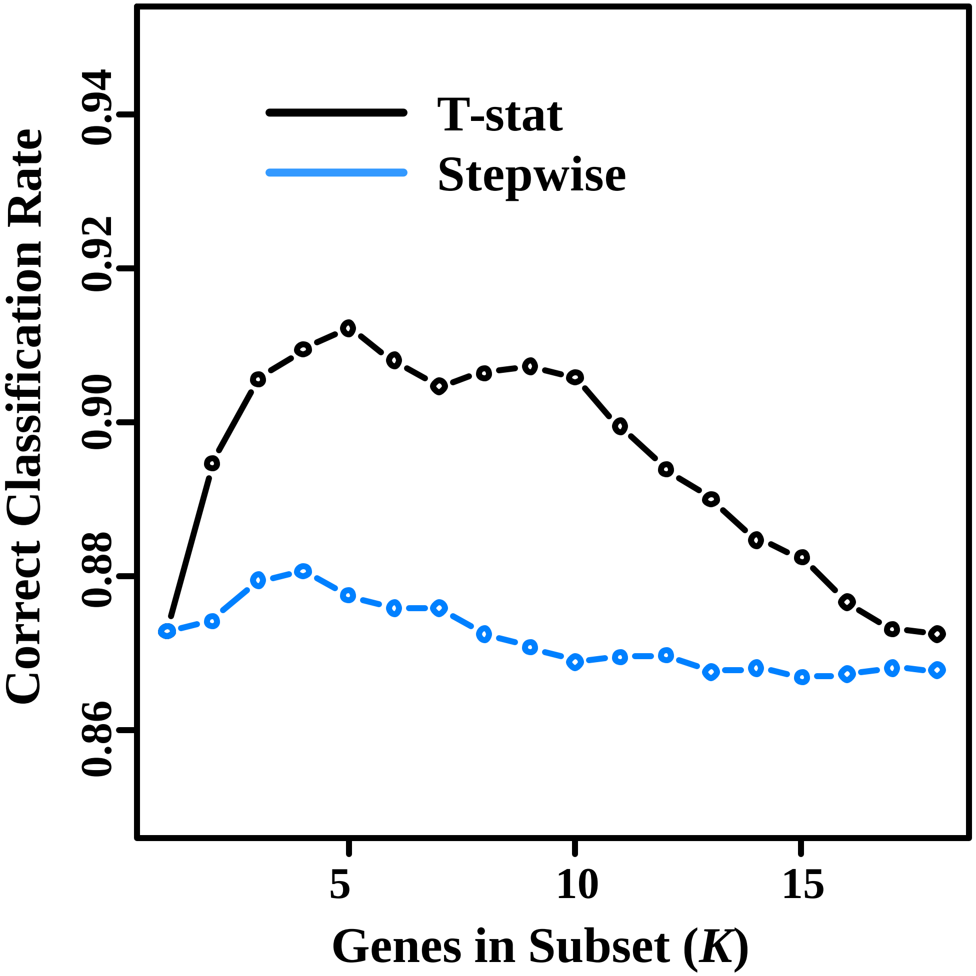
$$\left(\bar{g}_{1K} - \bar{g}_{2K}\right)' S_K^{-1} g_j \geq d_K \quad \textbf{(otherwise ALL)}$$

**where $g_j$ is the vector of *K* specified genes in sample *j***




- **Calculate correct classification rates based on top *K* genes for increasing values of *K***

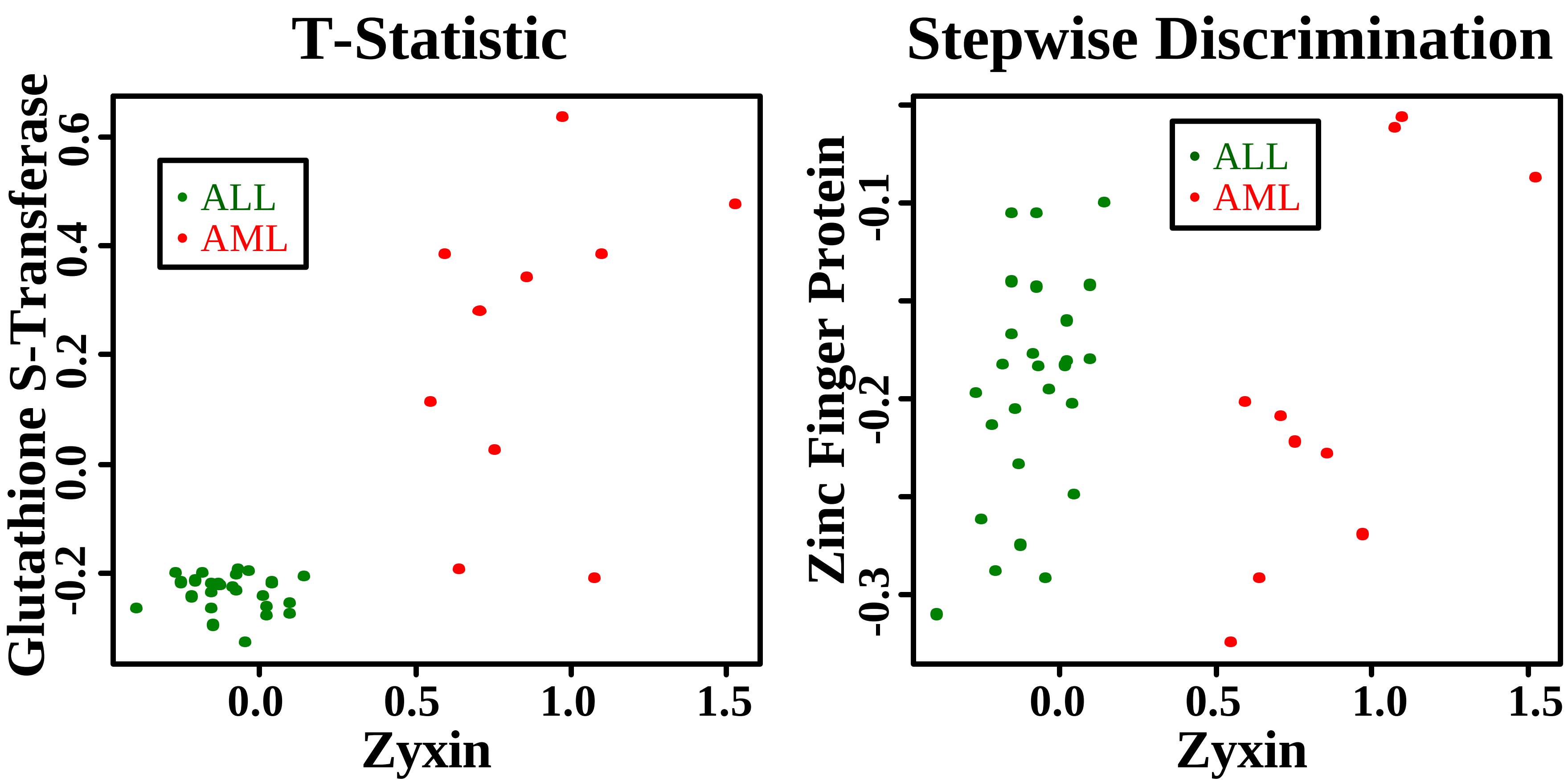
# Average Classification Rates of 100 Resampled Test Sets



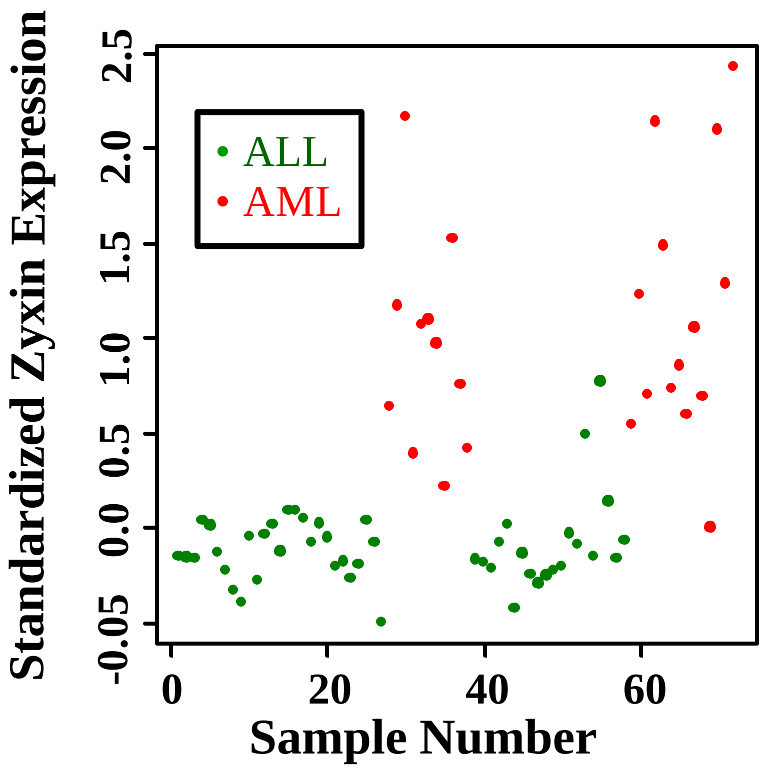

## Points of Interest

- **Max. Rate at 4-5 genes**
- **Rate Range: 87% -91%**
- **T-statistic performs slightly better than Stepwise**

# Scatterplots of the Top 2 Genes Selected by Each Method on a Resampled Training Set



# Characteristics of Zyxin




- **Selected as top gene in 55 of 100 resamplings.**
- **Average classification rate from 100 resamplings is 92% (range: 78% - 100%).**
- **91% classification rate on "original" test set of 34 patients.**

# Summary

- Maximum predictive information is contained in five or fewer genes by either method
- Genes selected by t-statistic achieved a higher classification rate for this data
- Classification rate of the protein zyxin alone:
  - 78% to 100% on resampled test sets
  - 91% on observed test set