10.555
Bioinformatics: Principles, Methods and Applications
MIT, Spring term, 2002

Lecture 11

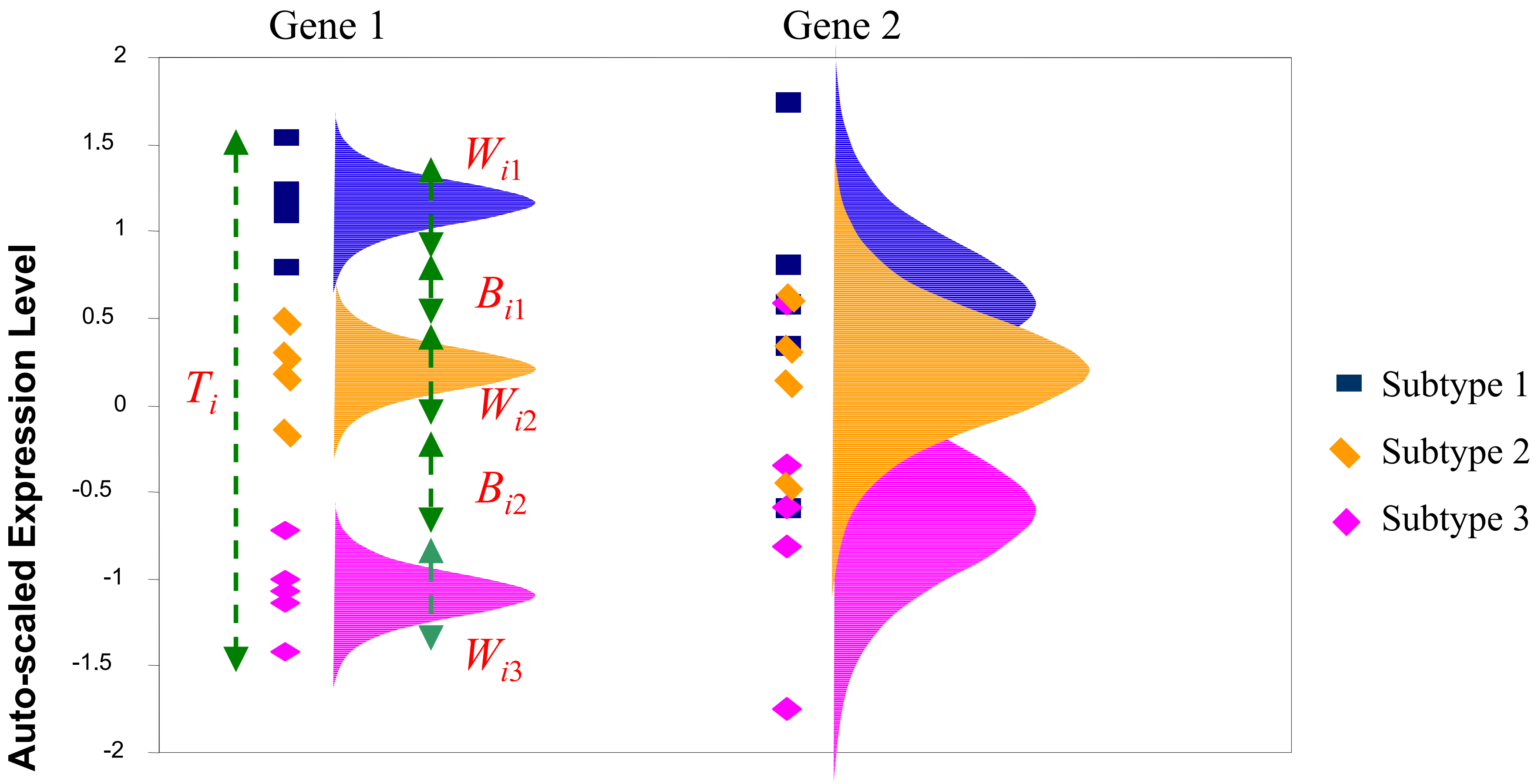
- Identification of discriminatory genes
- Dimensional reduction - Projection methods
- Discriminatory gene expression *patterns*

Analysis of Microarray Data

3. Analysis of *Static* expression data

- Statistical methods
- Decision trees
- Projection methods

Statistical methods: Identification of discriminatory genes



$$W_i = W_{i1}+W_{i2}+W_{i3}$$

$$B_i = B_{i1}+B_{i2}$$

Discriminatory gene
(Potential Disease-related gene)

Non-discriminatory gene

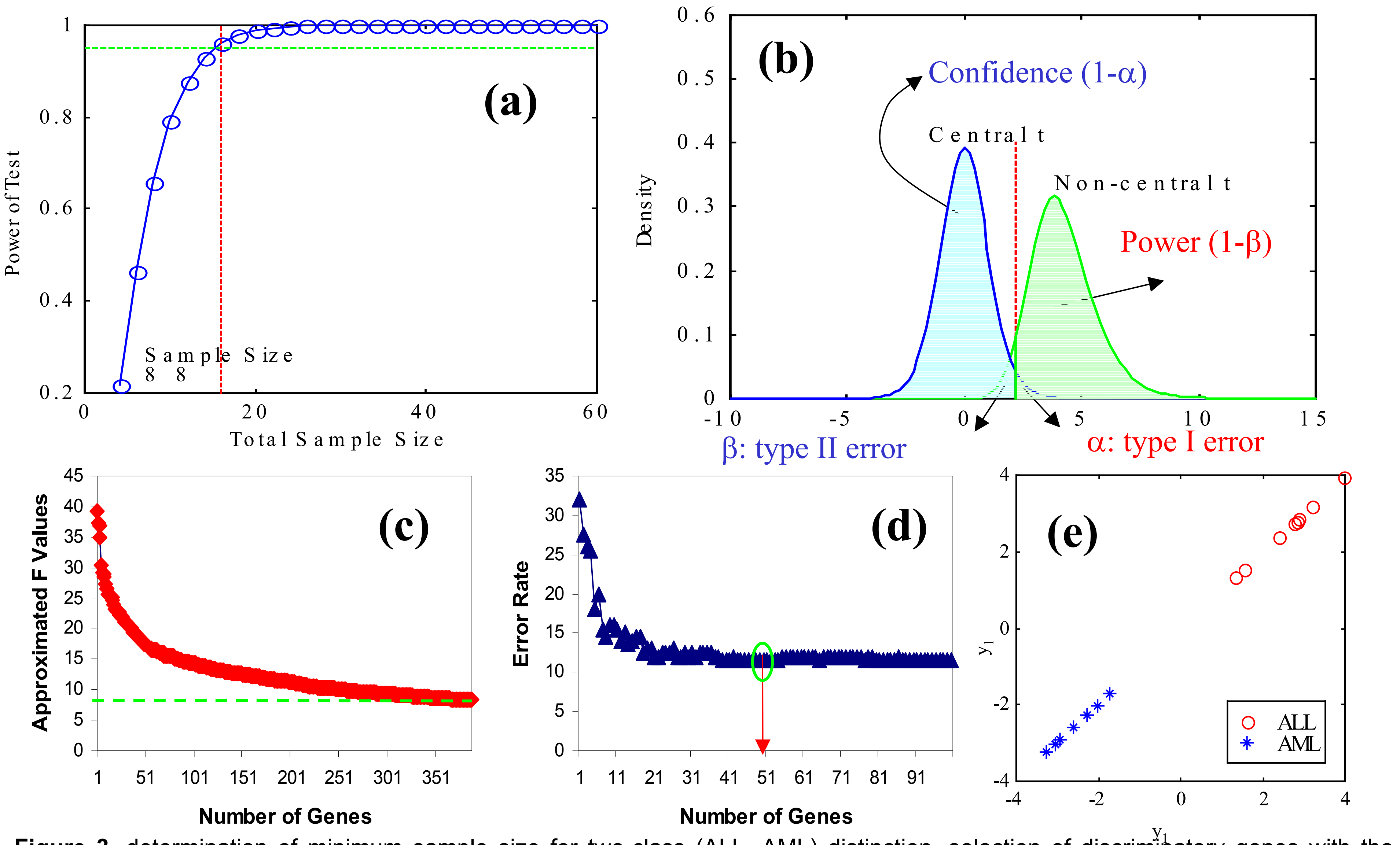
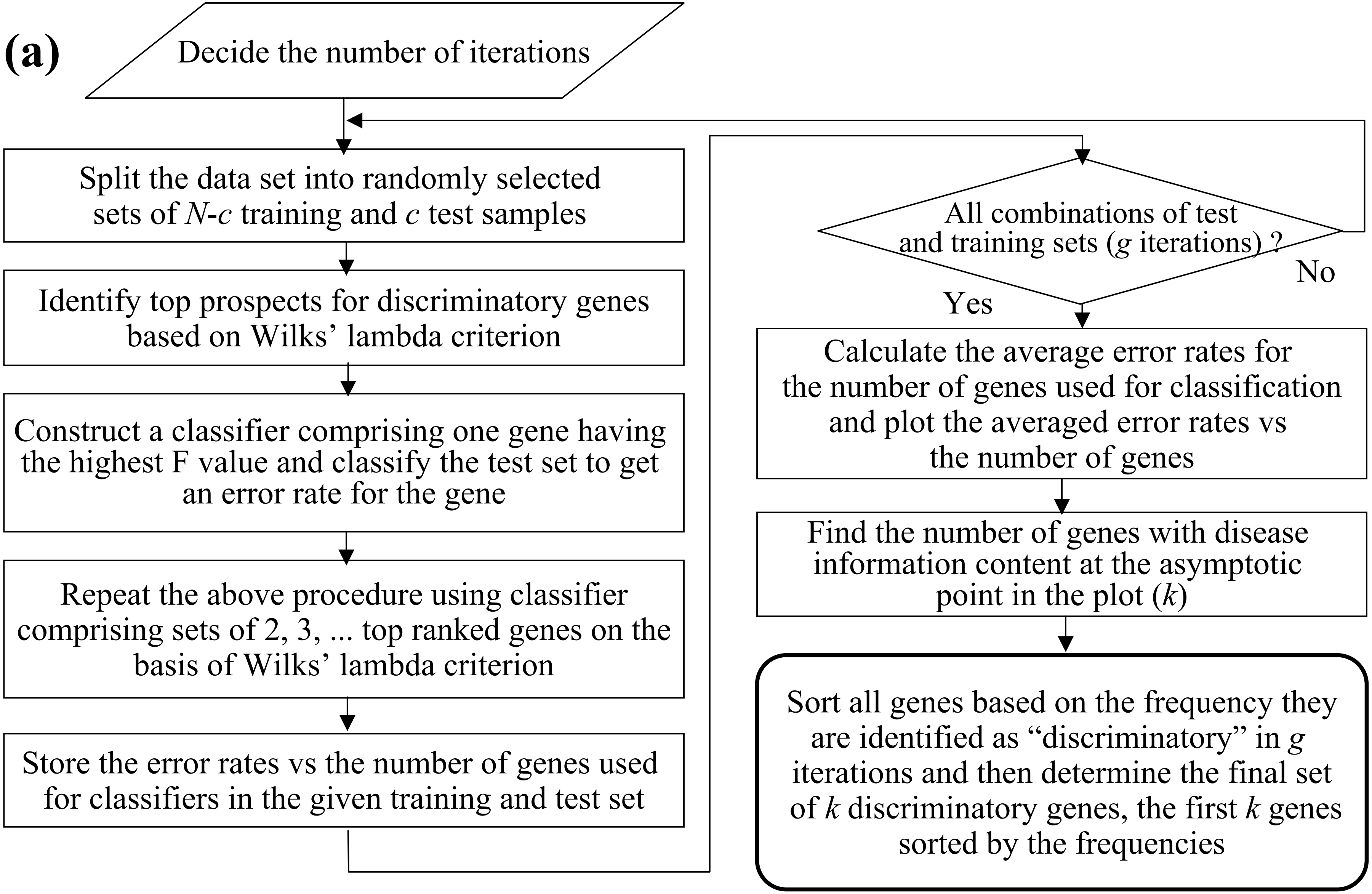


Figure 3. determination of minimum sample size for two-class (ALL, AML) distinction, selection of discriminatory genes with the estimated sample sizes of two classes, and FDA projection. (a) Power plot versus sample size showing how to determine the sample size required for two class distinction (8 from each class). (b) The distributions of H_0 and H_1 for the determined sample size. (c) Univariate F statistic values of the initial 388 discriminatory genes with a threshold ($F_{0.01(1,18)}$ = 8.2854) in randomly selected 8 ALL and 8 AML samples out of the entire data set. (d) Leave-one-out cross-validation applied to estimate the classification error rates and then to select the 50 most discriminatory genes with the same samples. (e) Separation of the 8 ALL and 8 AML samples in the two-dimensional FDA projection space defined discriminant axes of the 50 discriminatory genes.



(a)

Decide the number of iterations

Split the data set into randomly selected sets of *N*-*c* training and *c* test samples

Identify top prospects for discriminatory genes based on Wilks' lambda criterion

Construct a classifier comprising one gene having the highest F value and classify the test set to get an error rate for the gene

Repeat the above procedure using classifier comprising sets of 2, 3, ... top ranked genes on the basis of Wilks' lambda criterion

Store the error rates vs the number of genes used for classifiers in the given training and test set

All combinations of test and training sets (*g* iterations) ?

No

Yes

Calculate the average error rates for the number of genes used for classification and plot the averaged error rates vs the number of genes

Find the number of genes with disease information content at the asymptotic point in the plot (*k*)

Sort all genes based on the frequency they are identified as "discriminatory" in *g* iterations and then determine the final set of *k* discriminatory genes, the first *k* genes sorted by the frequencies

Figure 2. (a) Leave one out cross-validation (LOOCV) algorithm, where *N* is the total number of samples and *c* is the number of classes, so that one sample from each class is included in the test.

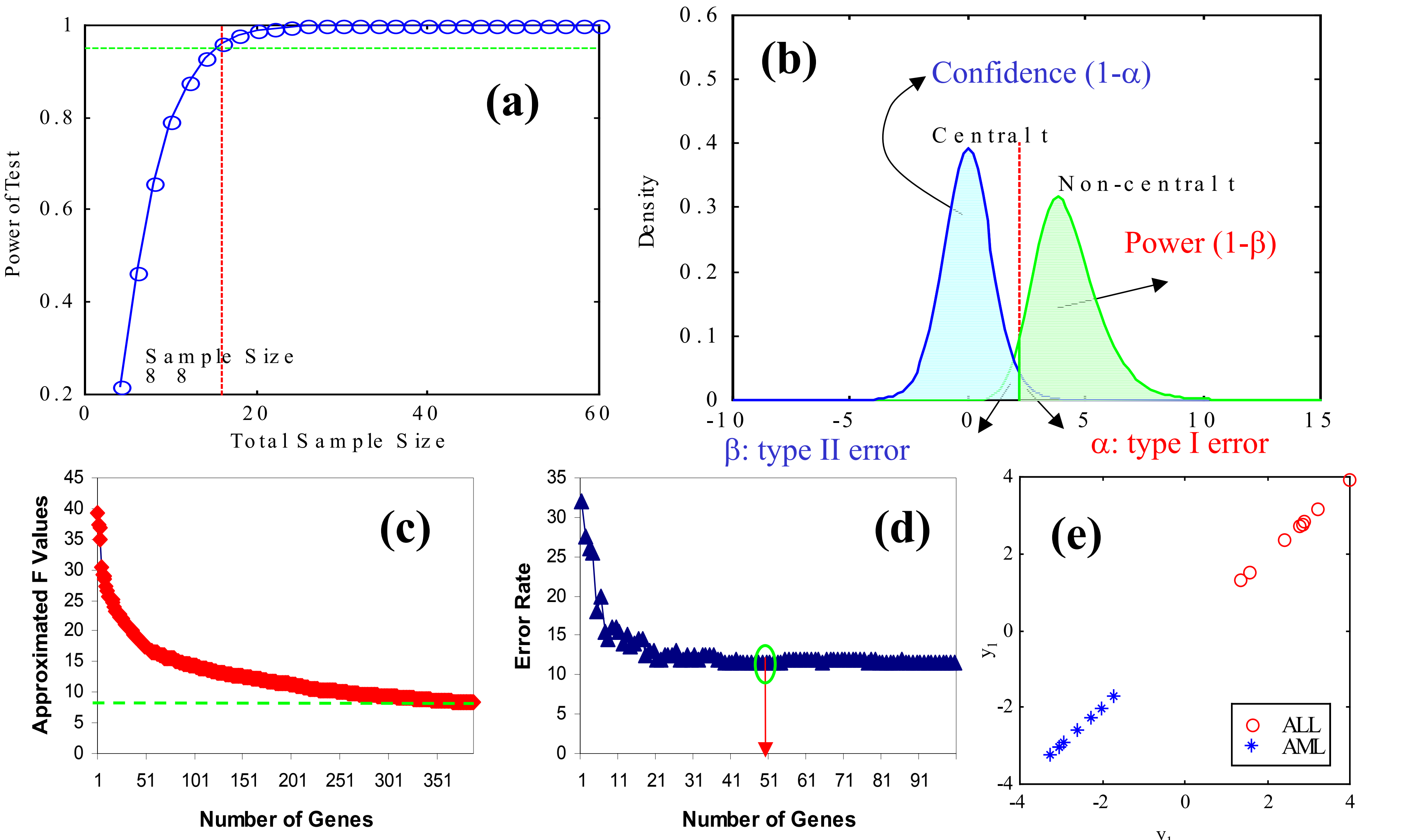


Figure 3. determination of minimum sample size for two-class (ALL, AML) distinction, selection of discriminatory genes with the estimated sample sizes of two classes, and FDA projection. (a) Power plot versus sample size showing how to determine the sample size required for two class distinction (8 from each class). (b) The distributions of H_0 and H_1 for the determined sample size. (c) Univariate F statistic values of the initial 388 discriminatory genes with a threshold ($F_{0.01(1,18)}$ = 8.2854) in randomly selected 8 ALL and 8 AML samples out of the entire data set. (d) Leave-one-out cross-validation applied to estimate the classification error rates and then to select the 50 most discriminatory genes with the same samples. (e) Separation of the 8 ALL and 8 AML samples in the two-dimensional FDA projection space defined discriminant axes of the 50 discriminatory genes.

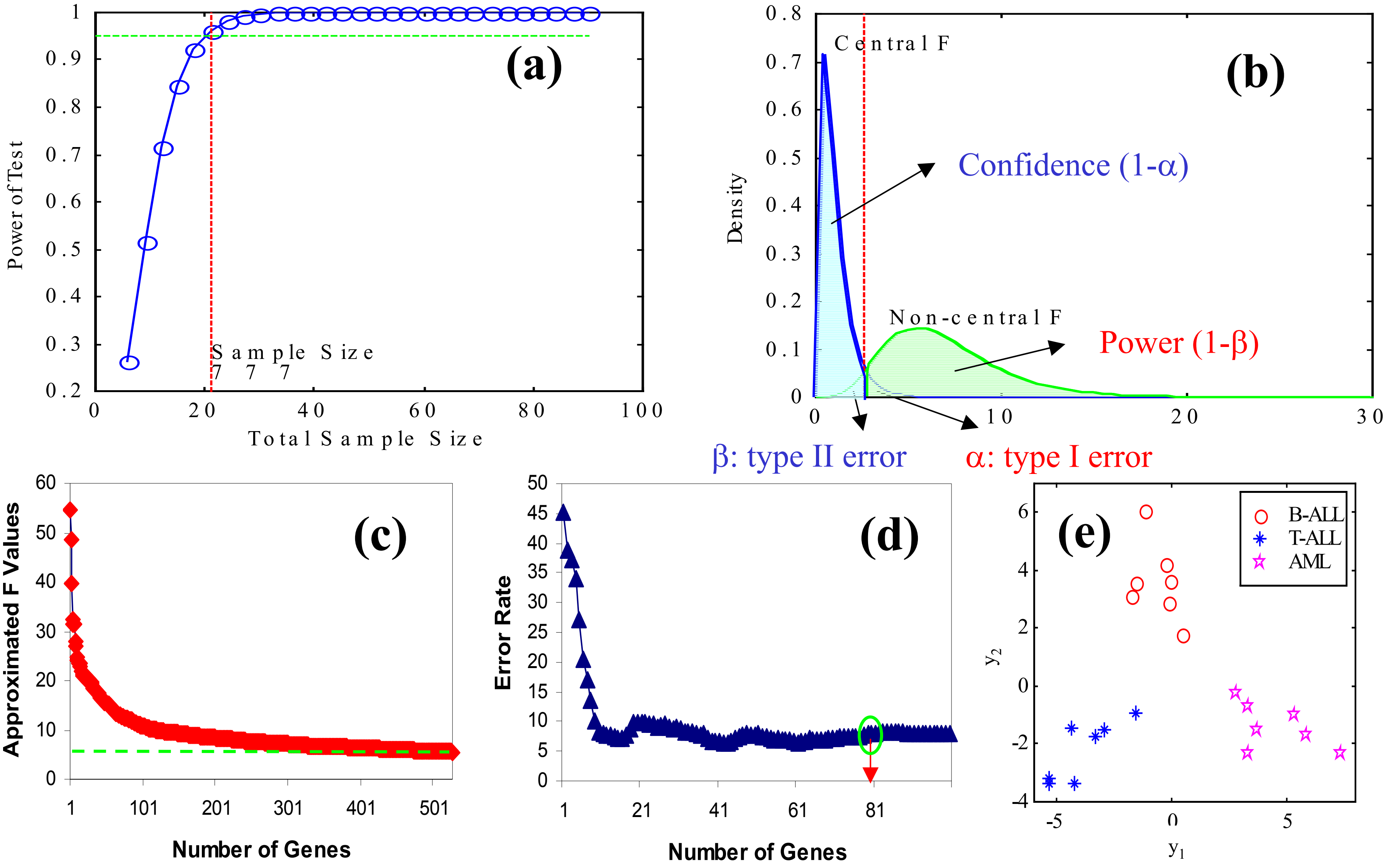


Figure 4. determination of minimum sample size for the three-class (B-ALL, T-ALL, AML) distinction, selection of discriminatory genes with the estimated sample sizes of three classes, and FDA projection. (a) Power plot versus sample size showing how to determine the minimum sample size (7 from each class). (b) The distributions of H_0 and H_1 for the determined sample size. (c) Univariate F statistic values of the initial 527 discriminatory genes with a threshold ($F_{0.01(2,26)}$ = 5.5263) in randomly selected 7 B-ALL, 7 T-ALL and 7 AML samples out of the entire data set. (d) Leave-one-out cross-validation applied to estimate the classification error rates and then to select the 80 most discriminatory genes with the same samples. (e) Separation of the 7 B-ALL, 7 T-ALL and 7 AML samples in the two-dimensional FDA projection space defined discriminant axes of the discriminatory 80 genes.

Classification using Decision trees

3. Analysis of *Static* expression data

- Statistical methods
- Classification using Decision trees
- Projection methods

Data Analysis and Pattern Classification

Problem-1: Consider N samples and M genes with their corresponding expression levels, e_i , where i = 1, …, M. M_1 of these tissues are characterized as "Healthy", while the other M_2 are labeled as "Pathological". ***Find the set of discriminatory genes*** whose expression levels can diagnose the state, i.e. healthy or pathological, of a new sample tissue.

Feature Space: The space of expression levels for the M genes, i.e. $FS = \{e_1 , e_2 , e_3 , \ldots, e_{M-1} , e_M \}$

Class: A set of genes characterized by the same label, e.g. C_1 = "Healthy" and C_2 = "Pathological".

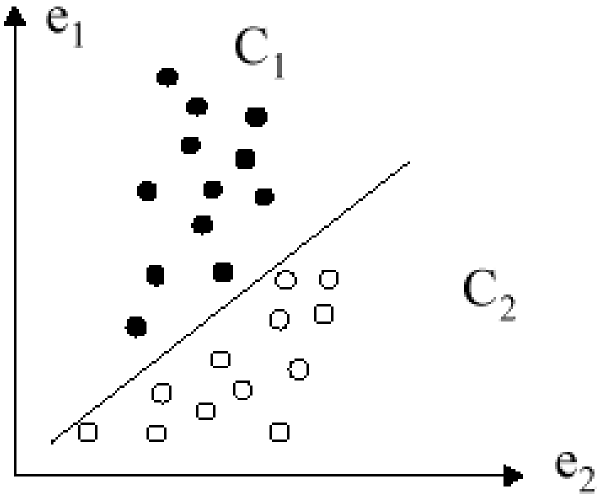
Pattern: The specific M-tuple of expression levels, which characterizes a tissue as belonging to a specific class, i.e. $\mathbf{p}^{(2)} = \{e^{(2)}_1 , e^{(2)}_2 , e^{(2)}_3 , \ldots, e^{(2)}_{M-1} , e^{(2)}_M \}$, Pattern for "Pathological" Tissues.

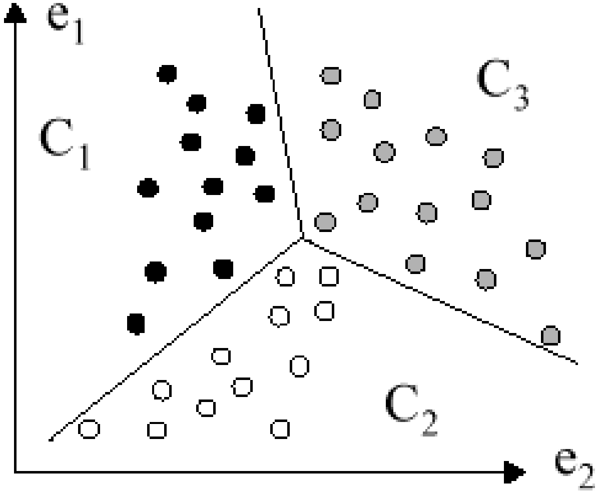
Data Analysis and Pattern Classification

Pattern Classification: The process through which the feature space, FS, is partitioned into K exclusive regions, FS_i i = 1, 2, …, K. Thus,

$$FS^{(i)} \cap FS^{(j)} = 0 \quad \text{and} \quad \cup_{\text{i=1-K}} FS^{(i)} = FS$$

Discriminant Functions: $d\,(\mathbf{p}) = d\,(e_1 , e_2 , e_3 , \ldots, e_{M-1} , e_M)$ define the partition of the feature space into the K regions.



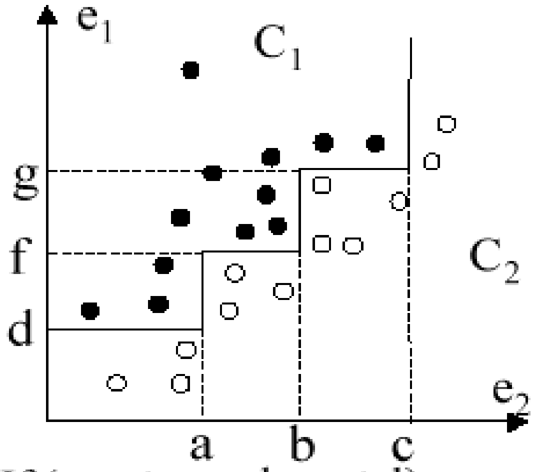


Data Analysis and Pattern Classification

Approaches: Stochastic or Deterministic

Logic-Based:

e.g. Decision Trees.



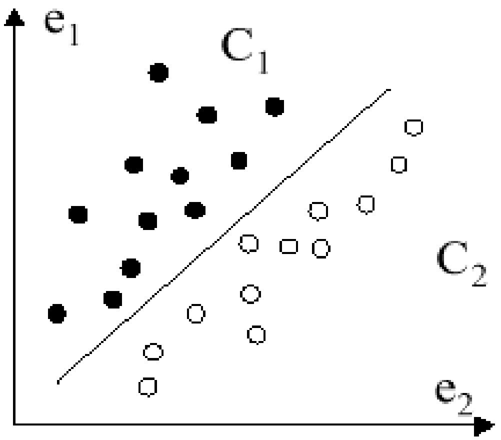
If ($e_2 < a$ and $e_1 < d$) or

($a < e_2 < b$ and $d < e_1 < f$),…

Then Class C_2

Linear:

e.g. PCA

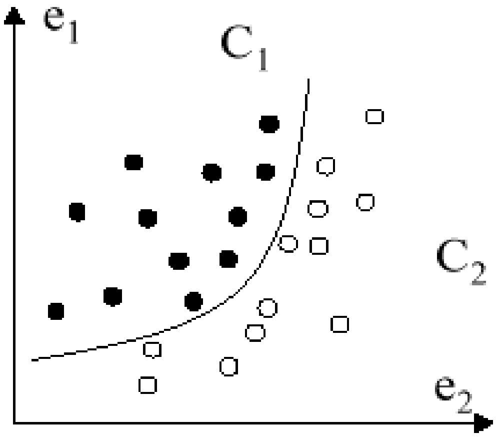


If = $k\, e_1 + l\, e_2 > 0$

Then Class C_1

Non-Linear:

e.g. Neural Networks



If = $f\,(e_1,e_2) > 0$

Then Class C_1

Data Analysis and Pattern Discovery

- **Problem-2:** Consider N samples and M genes with their corresponding expression levels, e_i , where i = 1, …, M. "Discover" the patterns in gene expression levels which are common in a number of samples, i.e. ***find the groups of samples***, each of which is characterized by a common pattern in gene expression and define this common pattern of gene expression levels for each group of samples.

- **Problem-3:** Consider one type of sample and the gene expression levels for M genes over a period of L time points. "Discover" the patterns in gene expression levels, which are common for a particular group of genes, and ***cluster the genes*** with similar patterns into the same group.

Data Analysis and Pattern Classification

Training:

The process through which one determines the discriminant functions, using past examples of "pattern" -"class" associations, i.e. associations between pattern $\mathbf{p}^{(i)} = \{e^{(i)}_1 , e^{(i)}_2 , e^{(i)}_3 , \ldots, e^{(i)}_{M-1} , e^{(i)}_M \}$ and Class $C^{(i)}$

Types of Problems:

- Static: when the gene expression levels represent the expression at a single time.
- Dynamic, or Time-Dependent: when the expression levels are measured over a period of time at various time intervals.
  - Equal sampling intervals.
  - Unequal sampling intervals.

Data Analysis and Pattern Classification

- **Issues to Resolve:**
  - Labeling the various samples
  - *Representation:*Selecting the distinguishing features for classification; particularly important for time-dependent data, e.g. do you use the values, or the time derivatives of expression levels for classification?
  - Selecting the form of the discriminant function
  - Do you have statistically "enough" data for training?
  - Do you have enough data for testing?
  - What is the "noise" in your measurements?
  - What is the sensitivity of the generated discriminant function?
  - What is the robustness of the resulting classification scheme?

Information Theory:Decision Trees in Pattern Classification

Let N be the total number of examples (e.g. samples) and M_i the number of samples in each of the K classes.
The Shannon entropy provides a measure of the information content in the data set,

$$I(M_1, M_2, \ldots, M_K) = \Sigma_{i=1\text{-}K} (M_i/M) \log_2 (M_i/M)$$

- If all examples belong in the same class then $I = 0$.
- The smaller the entropy the less variety of classes (more order) in the data set.

Split the data into two groups G_1 and G_2 with $M^{(1)}$ and $M^{(2)}$ examples (samples) in each group. Compute the information content for each group and for the whole set of examples.

Decision Trees in Pattern Classification

$$I(M_1, M_2, \ldots, M_K) = I^{(1)}(M_1^{(1)}, M_2^{(1)}, \ldots, M_K^{(1)}) + I(M_1^{(2)}, M_2^{(2)}, \ldots, M_K^{(2)})$$

$$= \sum_{i=1}^{K} \frac{M_i^{(1)}}{M^{(1)}} \log_2\left(\frac{M_i^{(1)}}{M^{(1)}}\right) + \sum_{i=1}^{K} \frac{M_i^{(2)}}{M^{(1)}} \log_2\left(\frac{M_i^{(2)}}{M^{(2)}}\right)$$

If all the examples in group G_1 belong to class C_1 and all the examples in group G_2 belong to the class C_2, then,

$$M_1^{(1)} = M^{(1)} \quad and \quad M_2^{(1)} = M_3^{(1)} = \ldots = M_K^{(1)} = 0$$

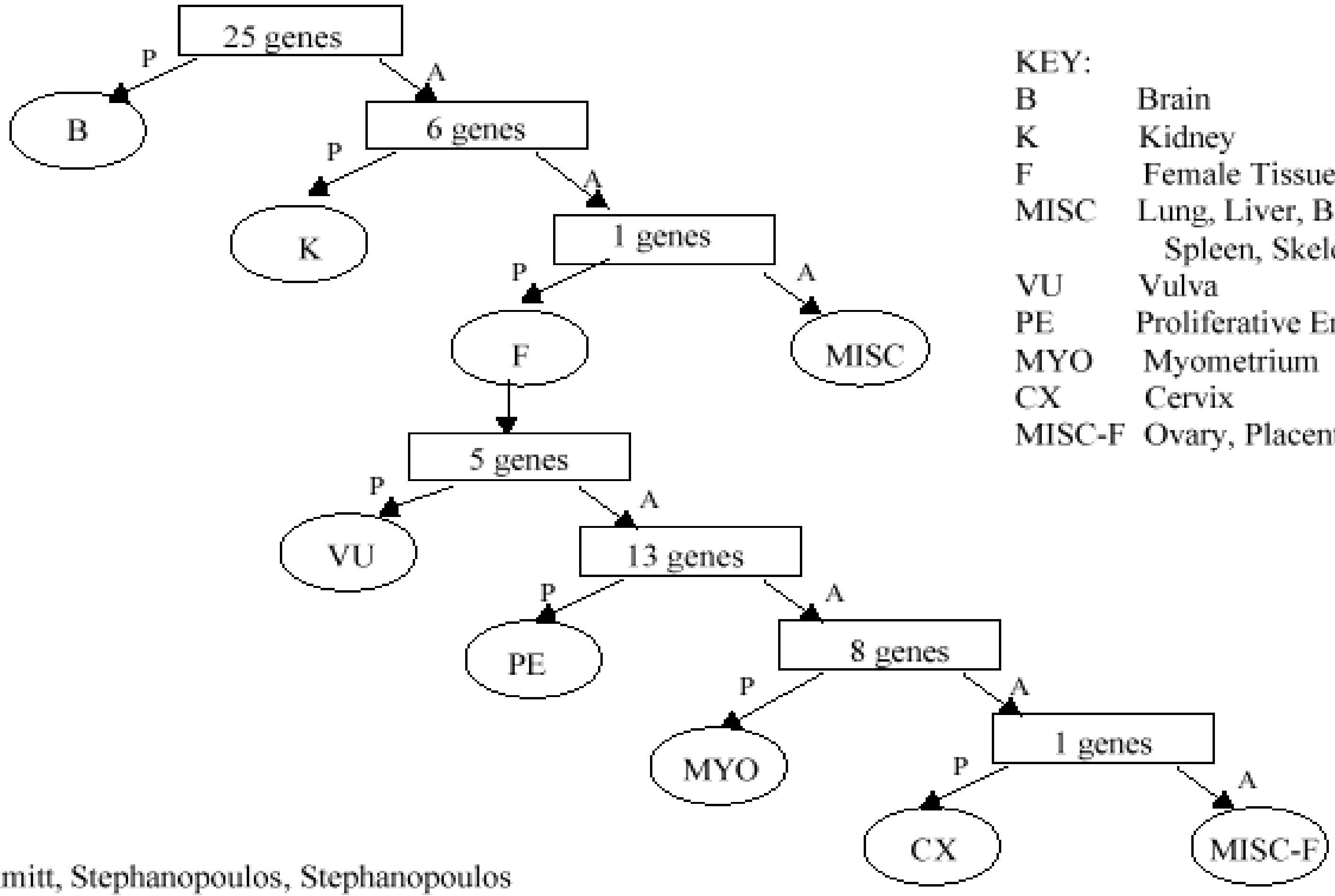
$$and$$

$$M_2^{(2)} = M^{(2)} \quad and \quad M_1^{(2)} = M_3^{(2)} = \ldots = M_K^{(2)} = 0$$

and $I^{(1)} = I^{(2)} = 0$, leading to the total $I = 0$.

Therefore, find the genes and their expression levels, which if were used to group the tissues into the K classes would "Minimize I "

Discriminating Tree for the Tissues



KEY:	
B	Brain
K	Kidney
F	Female Tissues
MISC	Lung, Liver, Blood, Spleen, Skeletal Muscle
VU	Vulva
PE	Proliferative Endometrium
MYO	Myometrium
CX	Cervix
MISC-F	Ovary, Placenta

Misra, Schmitt, Stephanopoulos, Stephanopoulos
BioInformatics and Metabolic Engineering Laboratory,
MIT, 12/3/99

Dimensional reduction. Projection methods

Why?

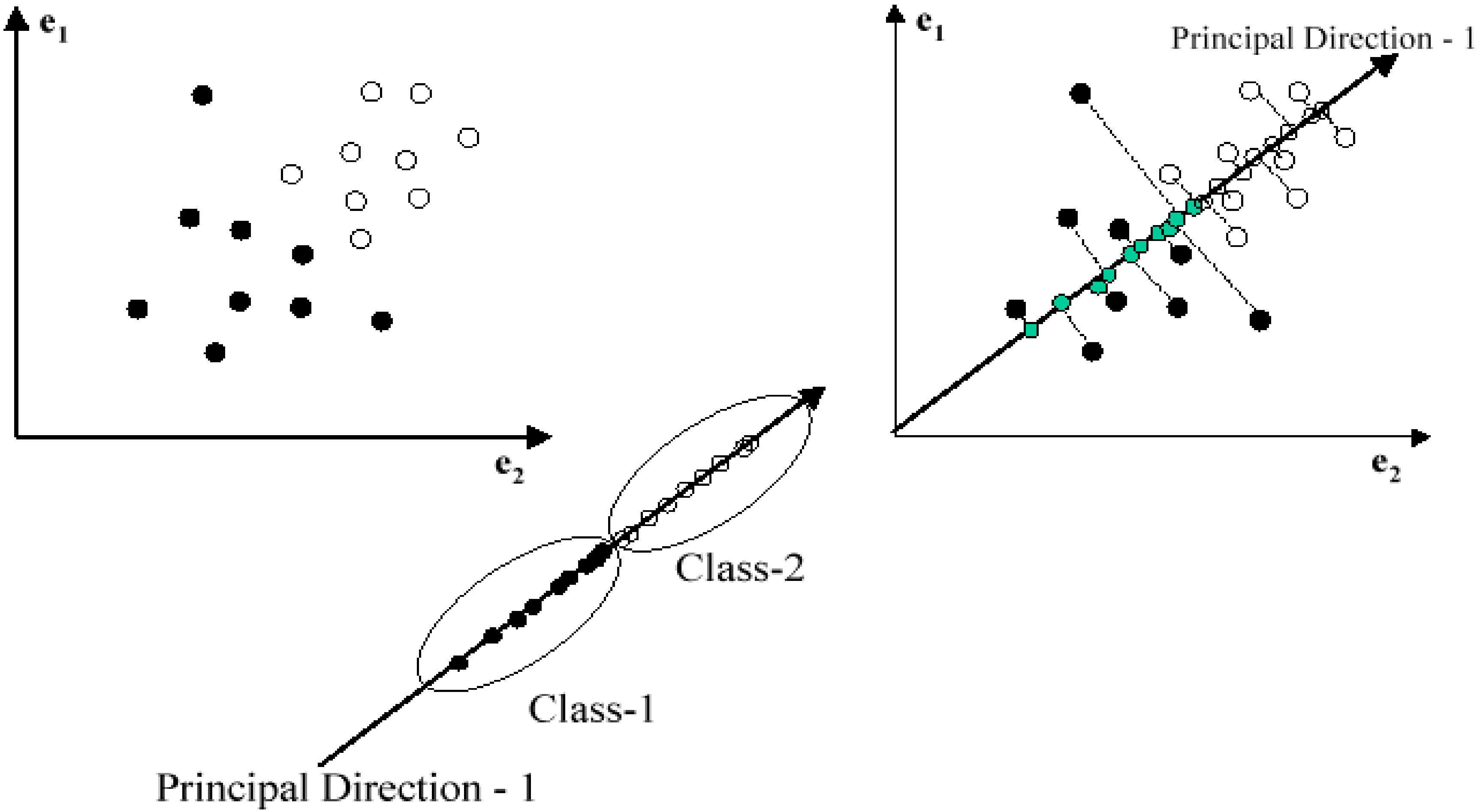
- Visualize data in fewer dimensions
- Class discovery
- Class separation
- Modeling

How?

- Identify projections that minimize information loss in the lower dimensional space

Principal Component Analysis

Basic Idea



A. Principal Component Analysis

- Matrix of measurements

$$\begin{array}{l|ccccc} & t=1 & t=2 & t=3 & \ldots & t=k \\ \hline \text{gene-1} & g_{1,1}- g_{1,aver} & g_{1,2} - g_{1,aver} & g_{1,3} - g_{1,aver} & \ldots\ldots & g_{1,k} - g_{1,aver} \\ \text{gene-2} & g_{2,1} - g_{2,aver} & g_{2,2} - g_{2,aver} & g_{2,3} - g_{2,aver} & \ldots\ldots & g_{2,k} - g_{2,aver} \\ \text{gene-3} & g_{3,1} - g_{3,aver} & g_{3,2} - g_{3,aver} & g_{3,3} - g_{3,aver} & \ldots\ldots & g_{3,k} - g_{3,aver} \\ \ldots\ldots & & & & & \\ \text{gene-n} & g_{n,1} - g_{n,aver} & g_{n,2} - g_{n,aver} & g_{n,3} - g_{n,aver} & \ldots\ldots & g_{n,k} - g_{n,aver} \end{array} = \mathrm{A}$$

- Then,

$$\mathbf{P}^T \Lambda \mathbf{P} = \mathbf{P}^T \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ldots & \\ & & & \lambda_k \end{bmatrix} \mathbf{P} = \mathbf{A}^T \mathbf{A}$$

where,

Λ *is the matrix of eigenvalues of* $(\mathbf{A}^T \mathbf{A})$*, and*

$\mathbf{P}$ *is the matrix with columns the eigenvectors of* $(\mathbf{A}^T \mathbf{A})$

A. Principal Component Analysis

- Projection of *gene-i* expression along the *j-th principal component*

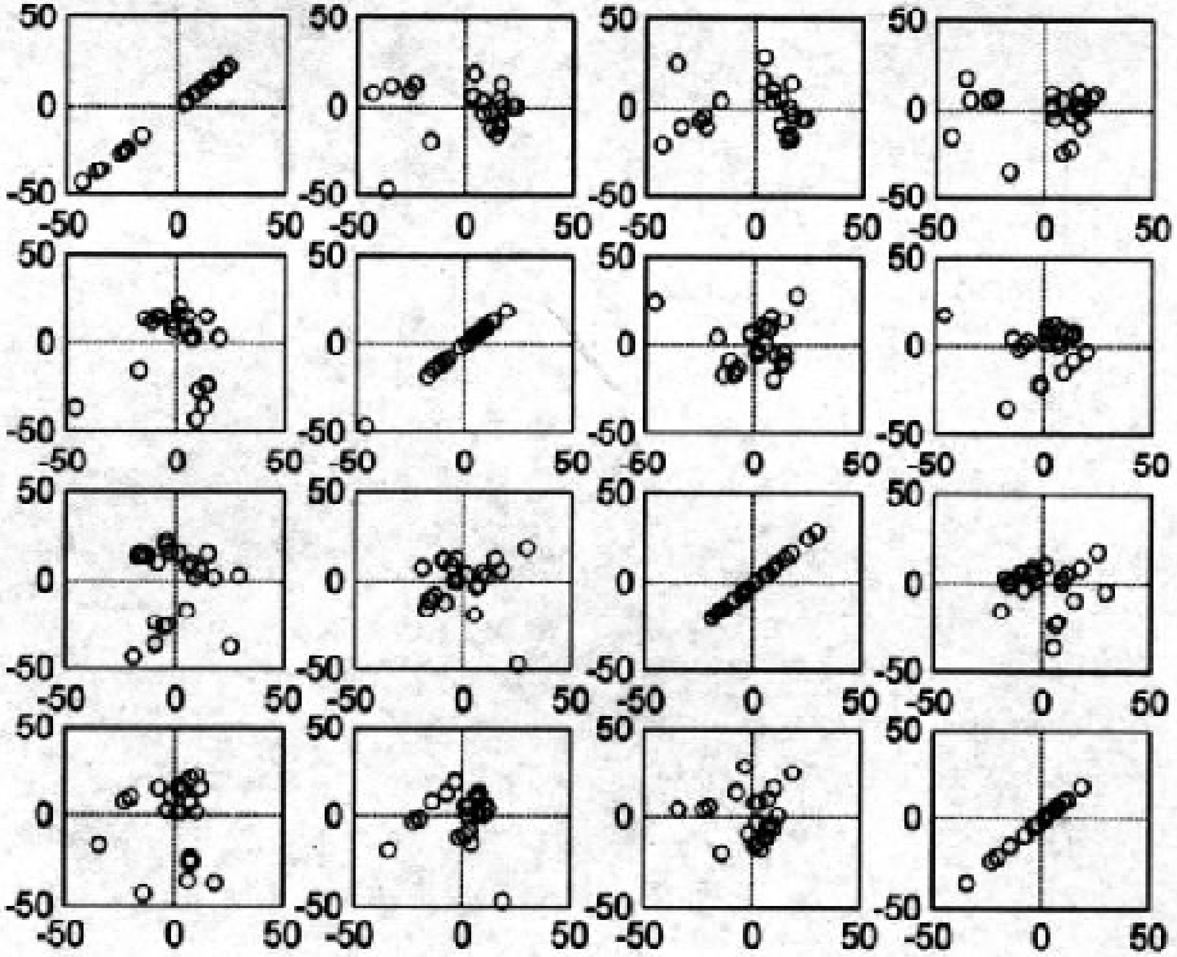
$$g^*_{ij} = \Sigma_{\mathrm{t=1\text{-}k}}\, g_{it}\, v_{tj}$$

where g_{it} is the gene expression at time t, and
v_{tj} is the I-th component of the j-th eigenvector

- The variance accounted for by each of the components is related to its associated eigenvalue. Consequently, the eigenvectors with larger eigen-values are the ones containing most of the information. Eigenvectors with small eigen-values are uninformative.

- Keep a small number of eigenvectors reproducing the desired amount of variance in the data

Example-1: PCA; 24 Tissues and 7,000 Genes



Example: Principal Component Analysis

(Chu S., et al., *Science*, 282, p. 699-705, 1998)

- Yeast. 6118 genes. 7 time points
- Summary of the results; Keep 2 or 3 principal components.

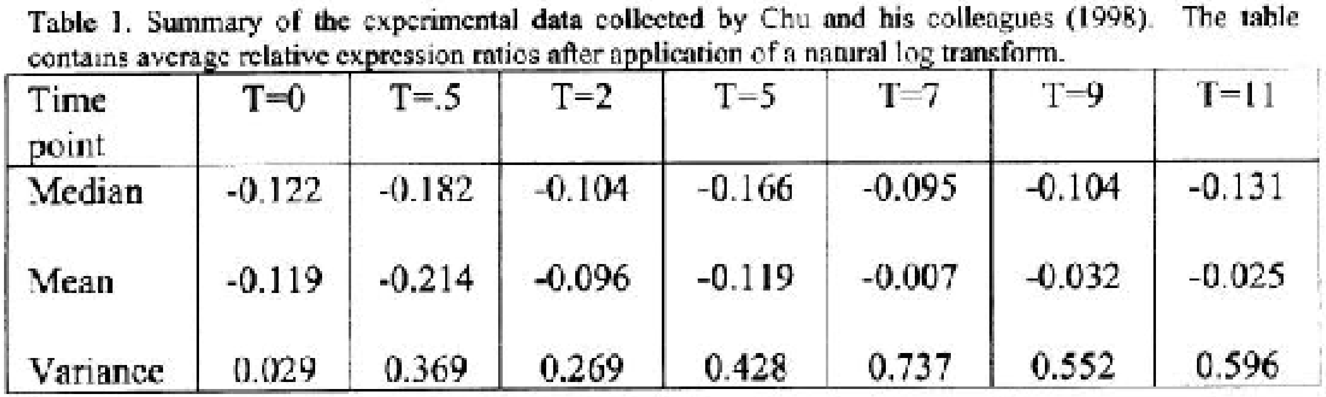


Table 1. Summary of the experimental data collected by Chu and his colleagues (1998). The table contains average relative expression ratios after application of a natural log transform.

Time point	T=0	T=.5	T=2	T=5	T=7	T=9	T=11
Median	-0.122	-0.182	-0.104	-0.166	-0.095	-0.104	-0.131
Mean	-0.119	-0.214	-0.096	-0.119	-0.007	-0.032	-0.025
Variance	0.029	0.369	0.269	0.428	0.737	0.552	0.596

Example: Principal Component Analysis

- Recovery of Variance

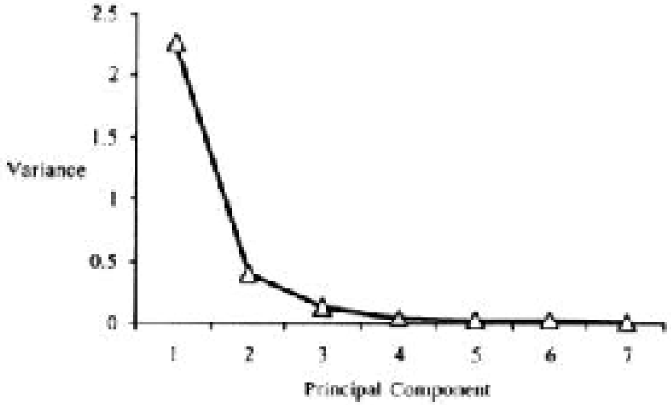


Figure 1. Plot of eigenvalues of the principal components. Most of the variance in the sporulation data set is contained in the first two principal components

- Interpreting the principal components

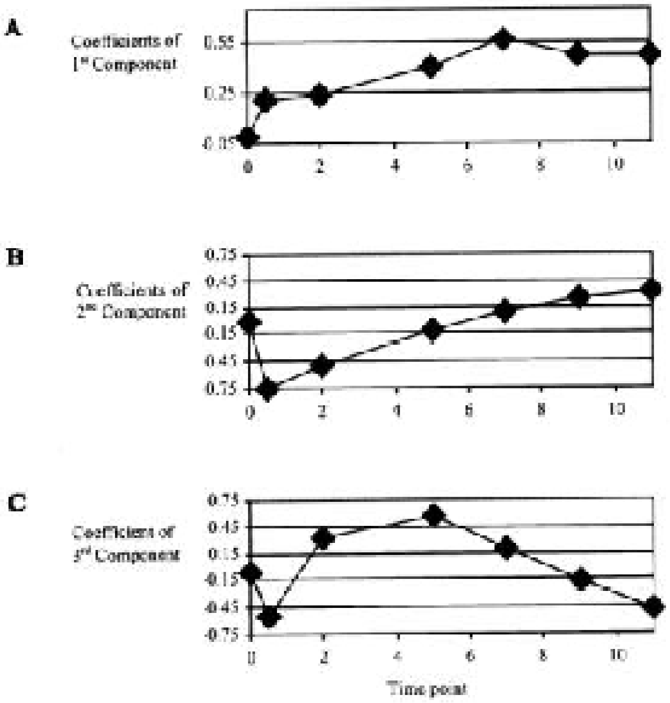


Figure 2. Plots of the coefficients of the first three principal components. Each coefficient indicates the weight of a particular experiment in the principal component. The first principal component has all positive coefficients, indicating a weighted average. The second principal component has negative values for the early time points and positive values for the later time points, indicating a measure of change in expression. The third coefficient captures information about the concavity in the expression pattern over time.

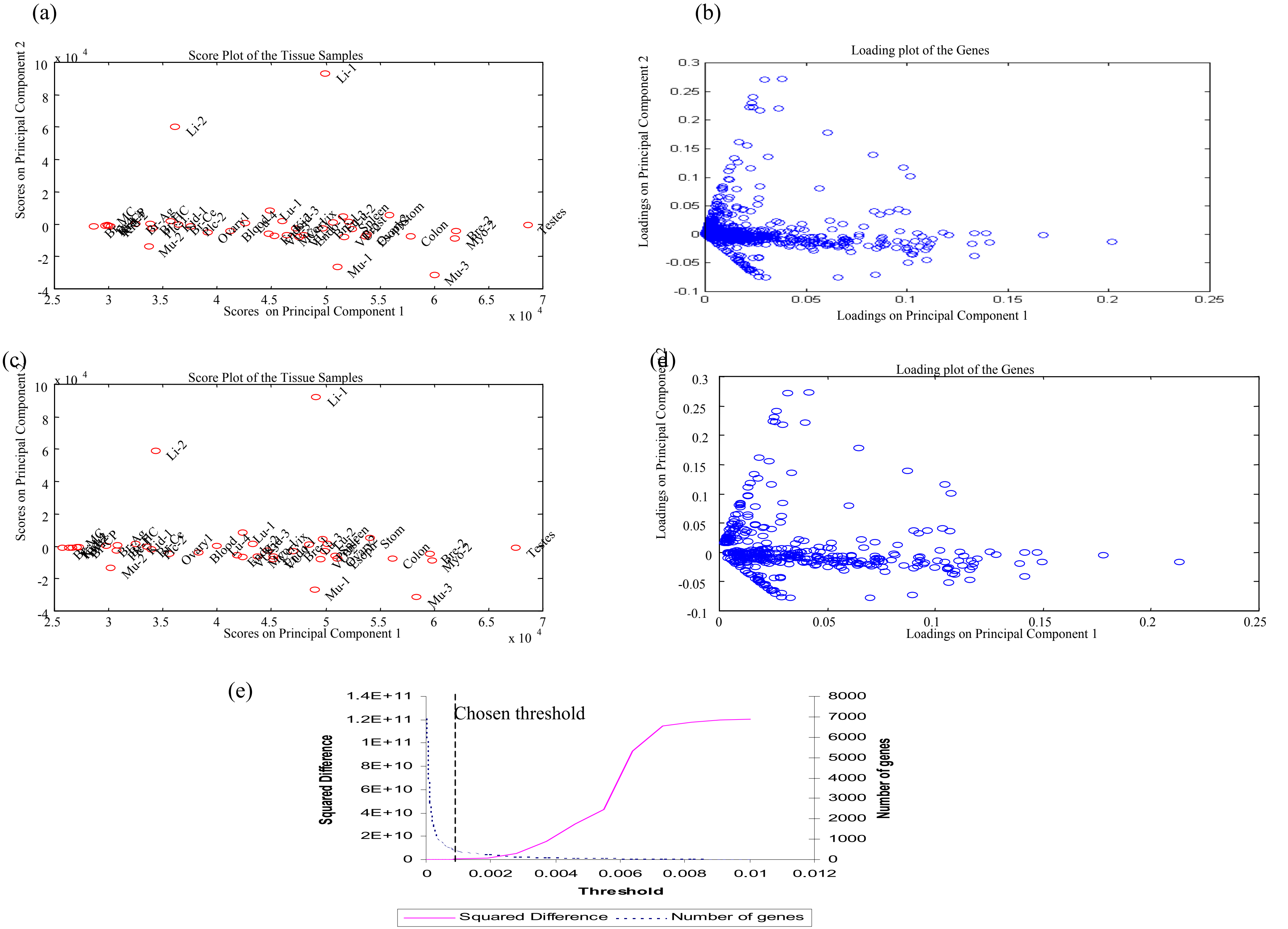
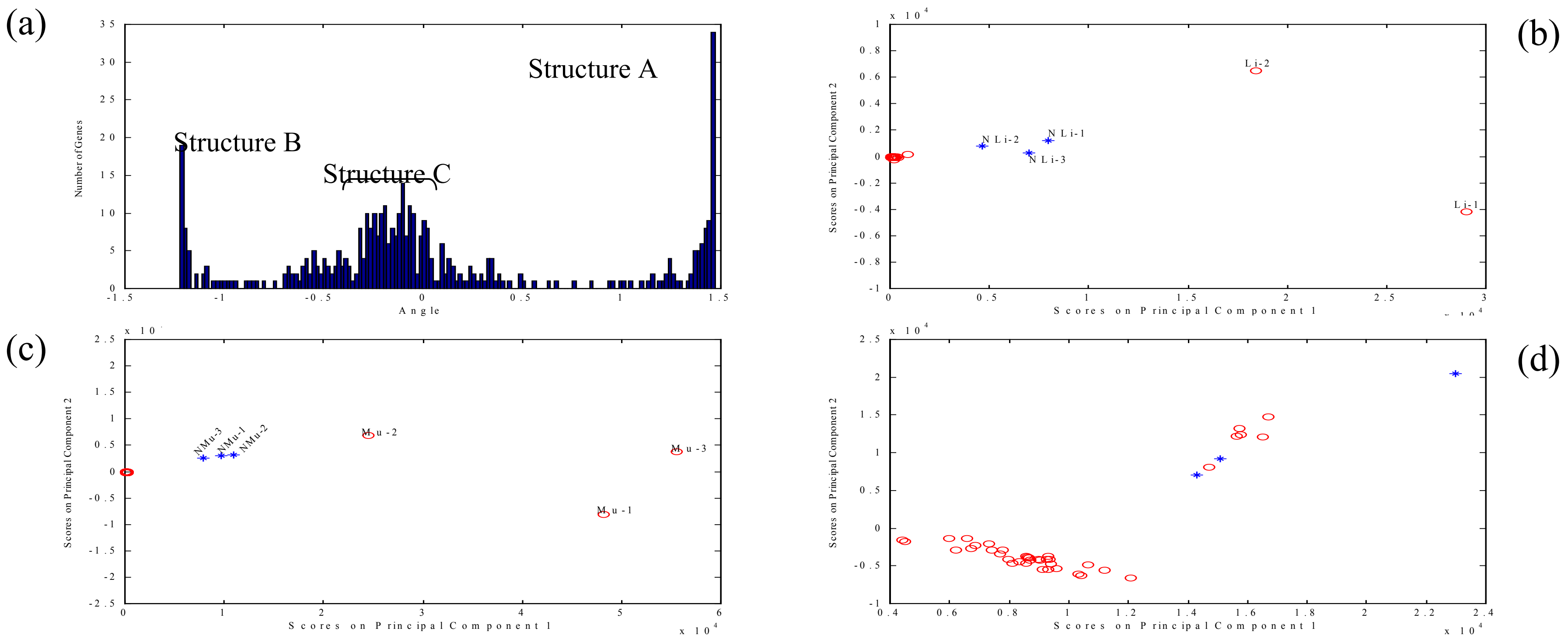


Figure 1: Selection of relevant genes using the loadings on the principal components.



Identification of tissue-specific genes and validation using new samples.

(a) Histogram of the angles between the x-axis and the points defined by the two principal loadings of each gene shown in Fig. 1d. Three main features, corresponding to the linear structures shown in Fig. 1d can be discerned, and are labeled as A, B and C. (b) PCA projection of all samples using the genes in Structure A. The samples in the initial data set are represented by red circles, and the new samples by blue asterixes. The two liver samples in the initial data set (Li-1, Li-2) and the new liver samples (NLi-1, NLi-2, NLi-3) are separated from the other samples, all of which cluster at the origin. (c) Projection of all samples using the genes in structure B. The muscle samples in the initial data set (Mu-1, Mu-2, Mu-3) are separated from the other samples along PC1. All the other tissue samples cluster at the origin. The new muscle samples are also separated when projected using these genes (NMu-1, NMu-2, NMu-3). (d) Projection of all samples using the genes in structure C. The six brain samples in the initial data set, and the three new brain samples are separated from the other samples.

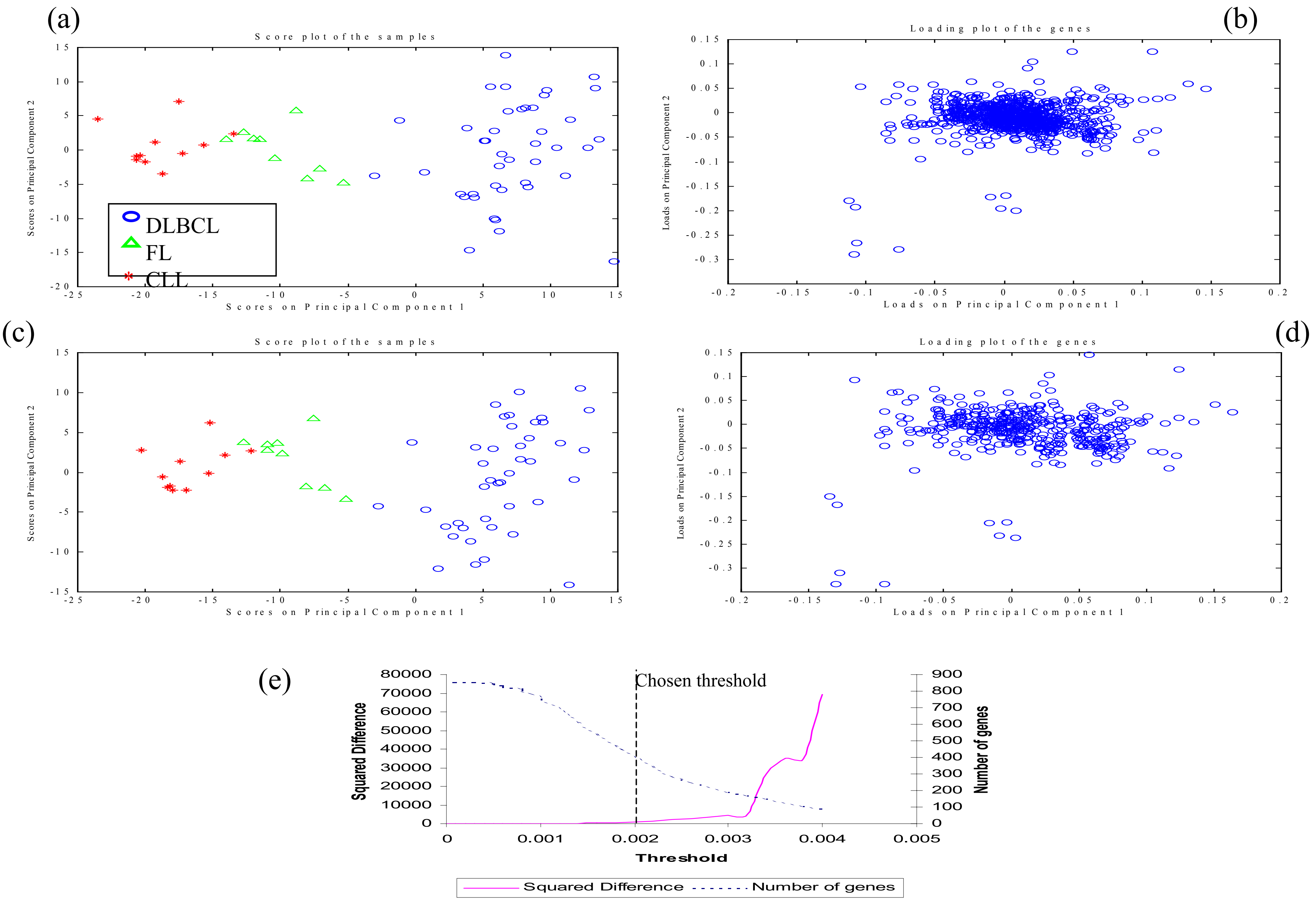


Figure 5.8 Projection of the Lymphoma data using PCA.

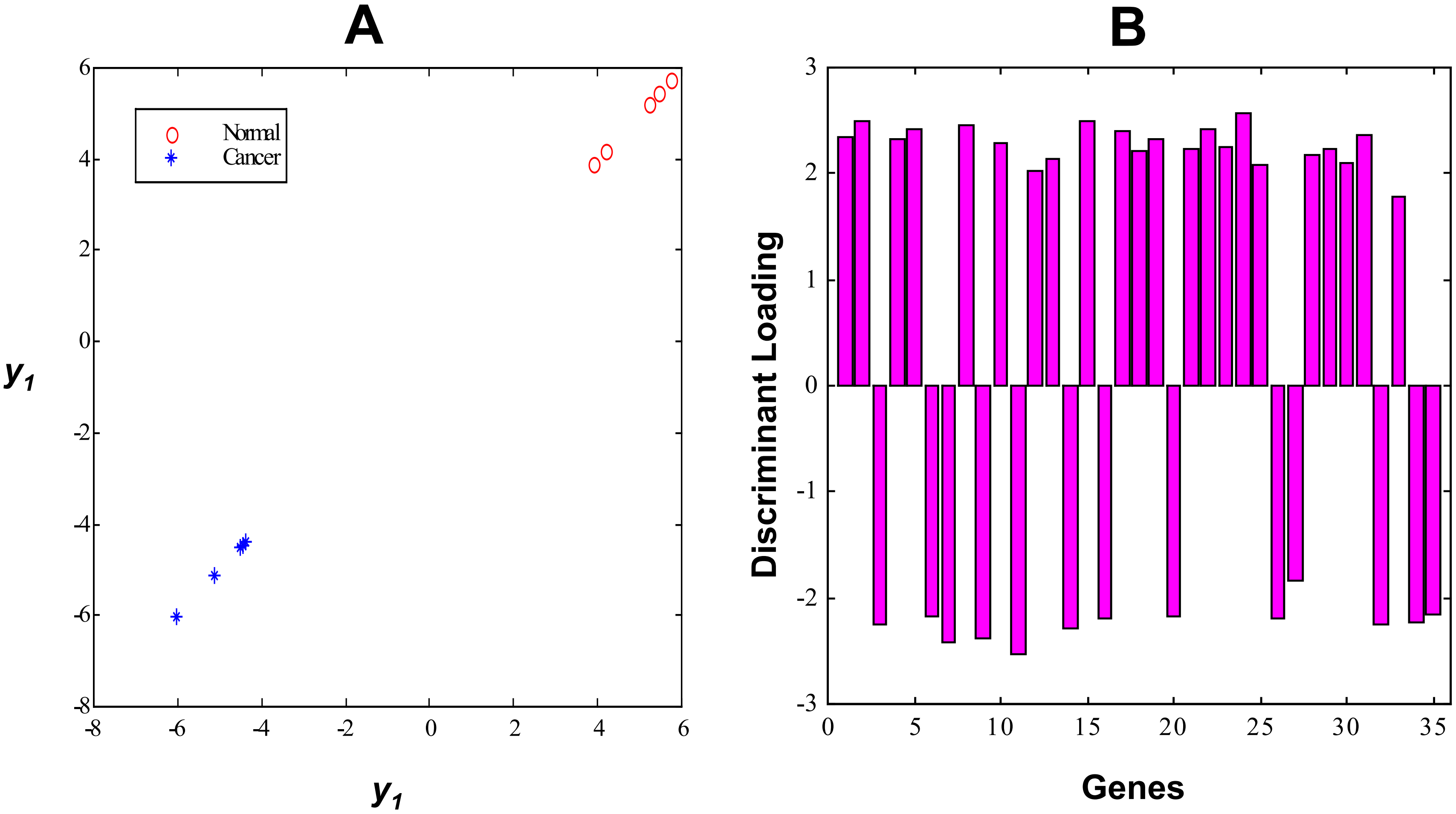


Figure 2. FDA projection of the expression phenotypes comprising 7070 genes measured in samples obtained from healthy individuals (5 samples) and patients with oral epithelium cancer (5 samples). (A) 35 discriminatory genes out of 7070 total genes allow FDA to clearly separate the two groups in one dimensional discrimination line. (B) Discriminant loading shows how 35 genes behave for separation in (A): positively co-regulated group includes NmU, aldehyde dehydrogenase 9 and 10, Her3, KIAA0089, diazepam binding inhibitor, monoamine oxidase B, crystallin alpha B, carboxylesterase 2, Wilm tumor-related protein, Zinc finger protein 273, MHC class I polypeptide related sequence A, Hpx-42, Lysophospholipase like, placental protein (PP11), cytochrome c oxidase subunit Vb, Cytochrome P4502C9 subfamily IIC, TF 20, FUT6, TYRO3, Keratin 4, and HLF. The negatively co-regulated group includes Ferritin, Urokinase plasminogen activator, Gro2 oncogene, 5T4 oncofetal trophoblast glycoprotein, HSP 90, Cathepsin L, Runt-related TF, Phospholipase A2, FAT tumor suppressor, macropain subunit zeta, CD38, TAL1 (SCL) and G-protein-coupled receptor (AZ3B). These two groups are anti-correlated with respect to each other.

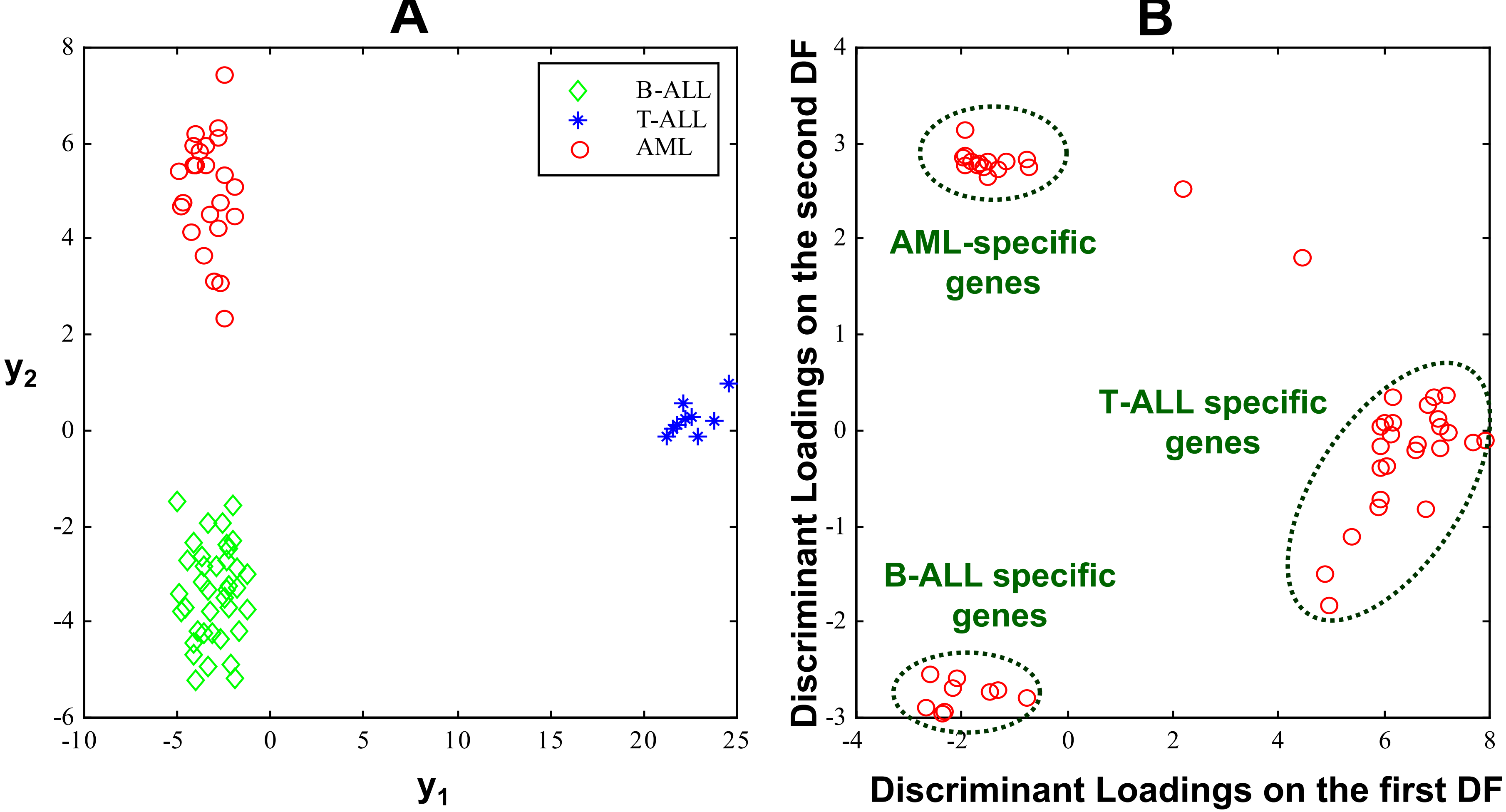


Figure 3. FDA projection of expression data obtained from patients with B-ALL, T-ALL, and AML. (A) Projection of the samples using 50 discriminatory genes allows FDA to clearly separate the three classes of leukemia expression phenotype in a 2-D discrimination space. The first DF distinguishes the T-ALL group from B-ALL and AML. The second DF separates B-ALL group from AML to complete the group separation. (B)The contributions of individual genes to the discrimination and their interactions are evident on plotting the discriminant loadings, where the genes are clustered into three groups, and show group-specific regulation patterns, except two genes between AML-specific gene group and T-ALL specific gene group. Ten of the 14 AML specific genes observed above are common with the 25 AML genes identified by Golub *et al*. (2000), 2 of the 25 T-ALL and 2 of the 9 B-ALL genes above are common with the 25 ALL genes identified by Golub *et al* (2000). The identity of these genes is provided in Supplementary Materials.

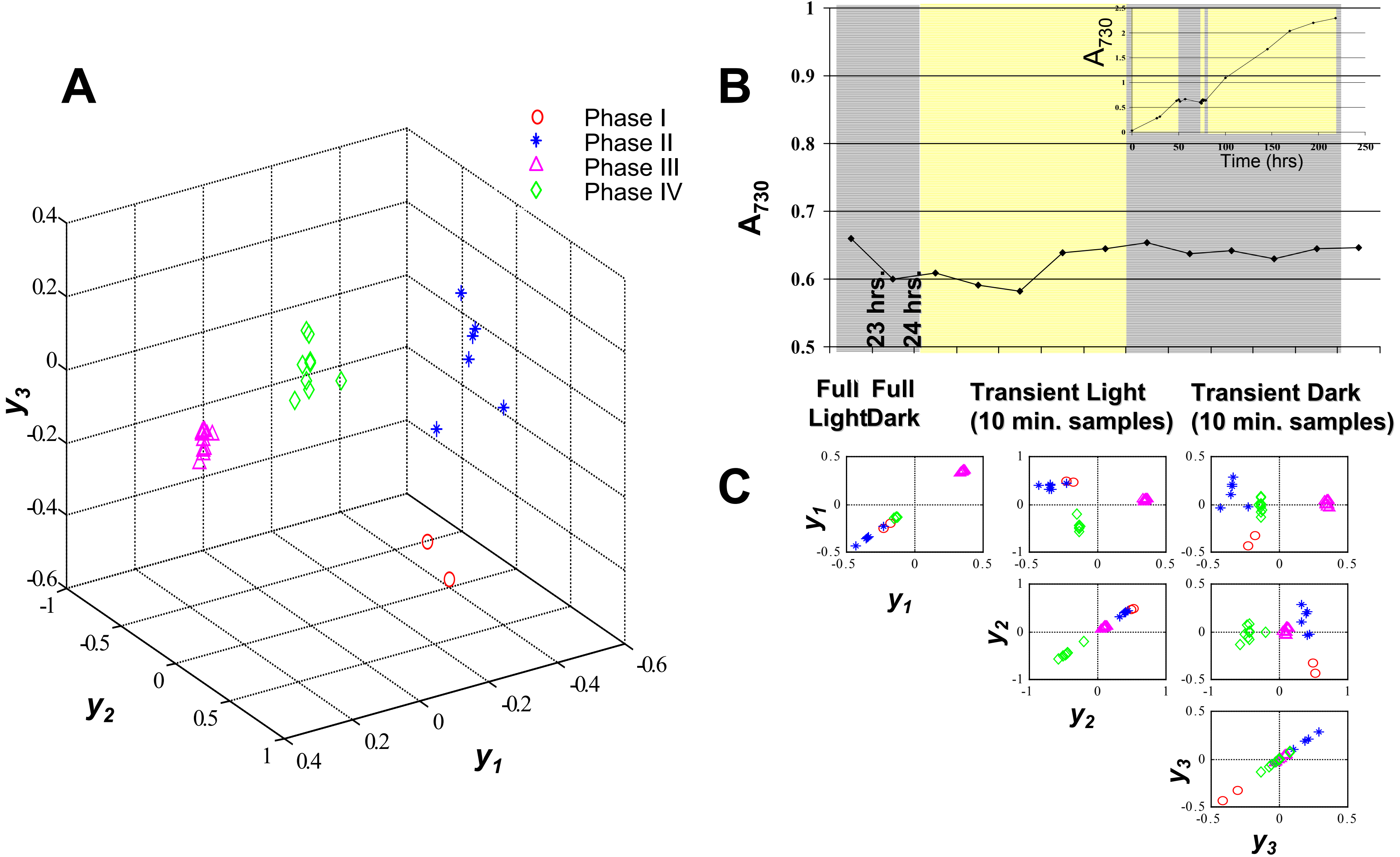


Figure 4. (A)Projection of the expression phenotypes of cultures of *Synechocystis* sp. PCC 6803 to a FDA-defined discrimination space. This photosynthetic bacterium was grown under conditions shown in (B) and the expression levels of 88 genes were measured by a DNA microarray at 29 time points spanning the entire course of the experiment. Of the 88 genes, 27 were identified as most discriminating of the four classes defined by the four different light conditions and their expression levels were projected to the FDA-defined space. It can be seen that the four phenotypic classes are clearly identified in the 3-dimensional FDA projection space. (C) The first DF shows the largest discrimination power separating all the groups, discriminating clearly Phase III from the others. The second DF separates Phase IV from Phases I and II, while the third DF is necessary to separate Phase I from II.

Use of microarray data in Drug Discovery

(Expression data from:
"Functional discovery via a compedium of expression profiles," Huges *et al.*, *Cell*, 102: 109-126, (2000))

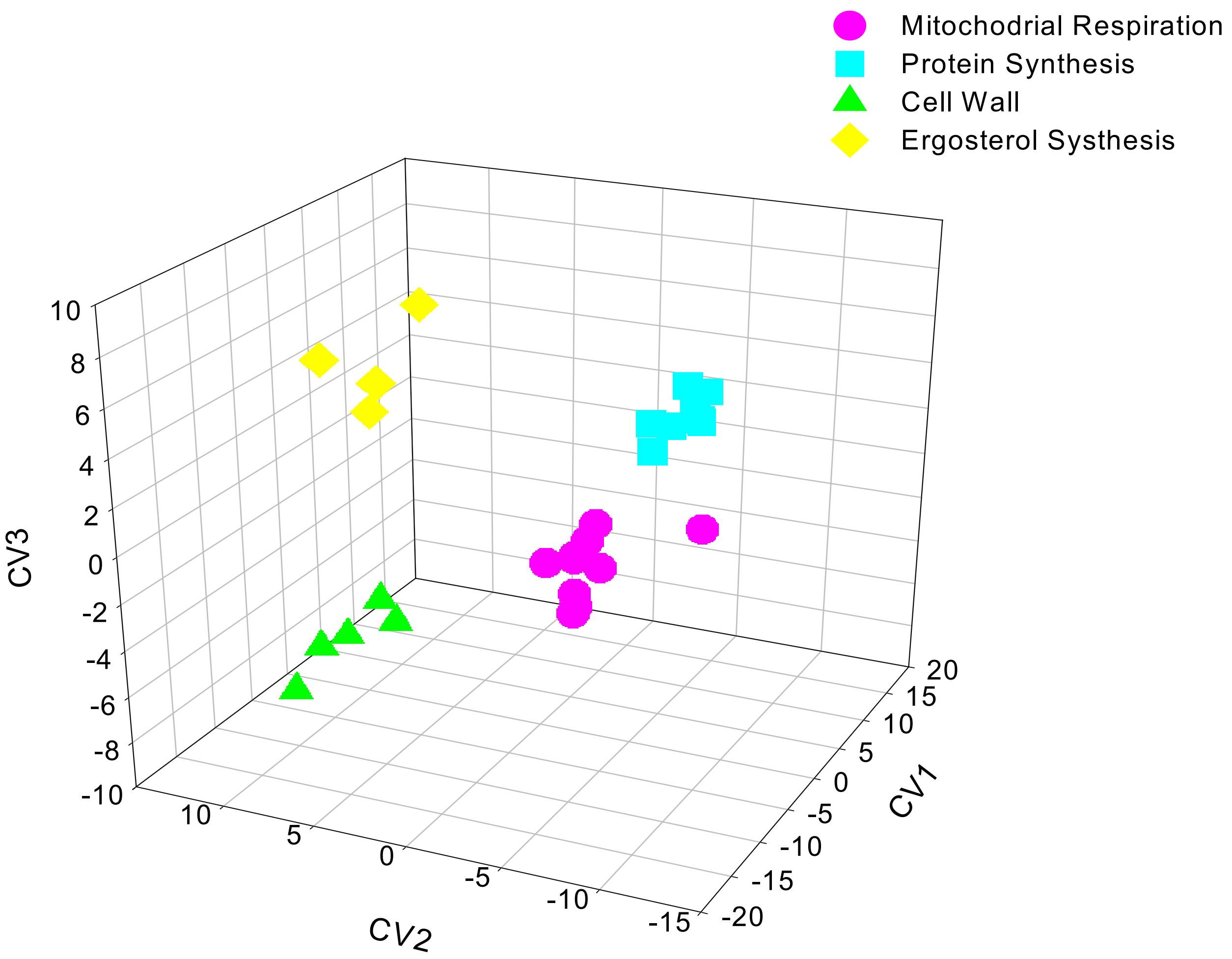
- Case study:

  45 single gene deletion yeast mutants were classified in the following 4 groups according to the effect that each gene deletion had on cell physiology:

  - Mitochondria respiration
  - Cell wall
  - Protein synthesis
  - Ergosterol synthesis

- Microarray gene expression data were collected for each mutant and projected in a CDA space to yield a well defined description of the physiology of each mutant

Drug Discovery (cont'd)



Use of microarray data in Drug Discovery

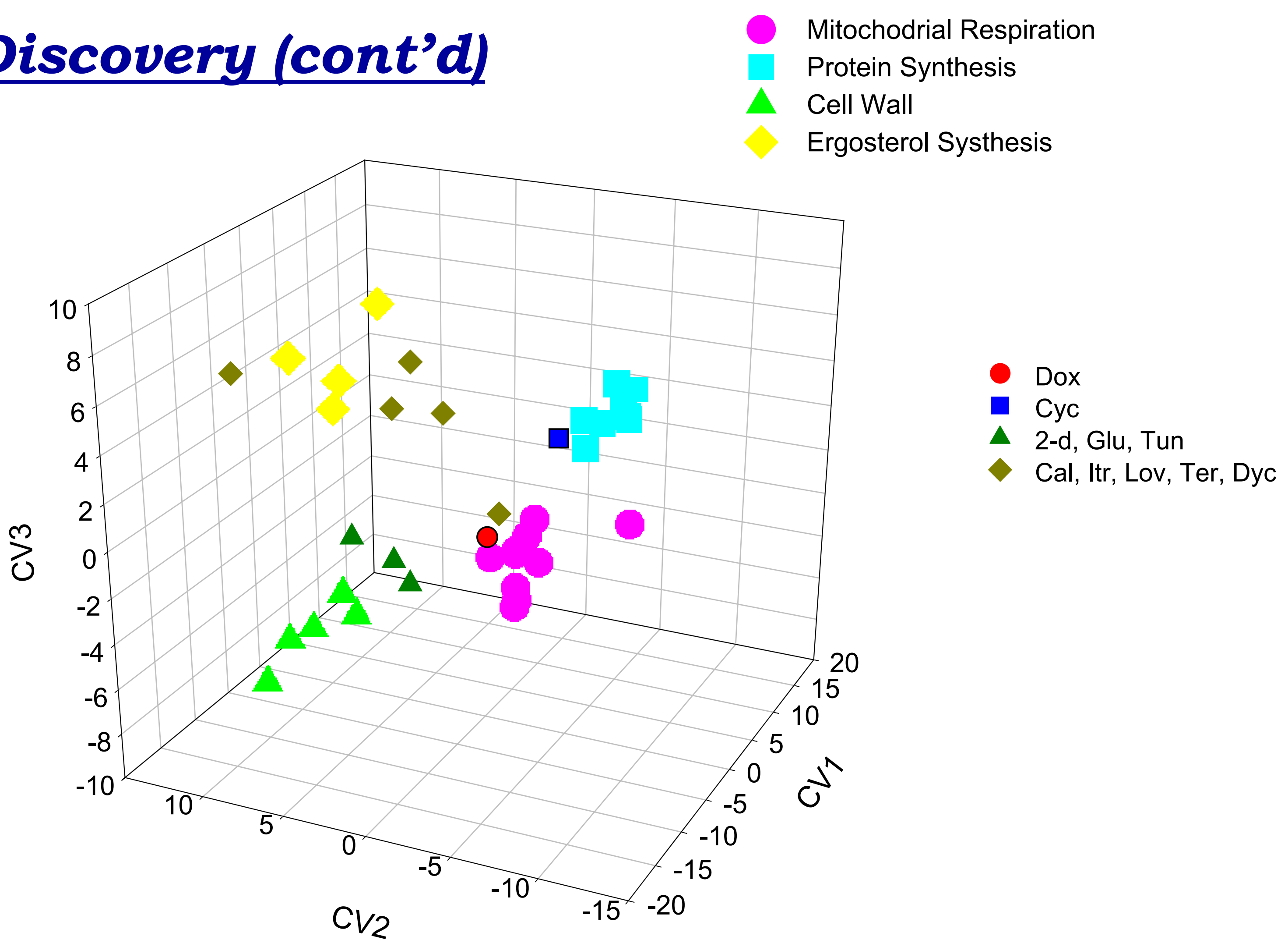
- Case study:

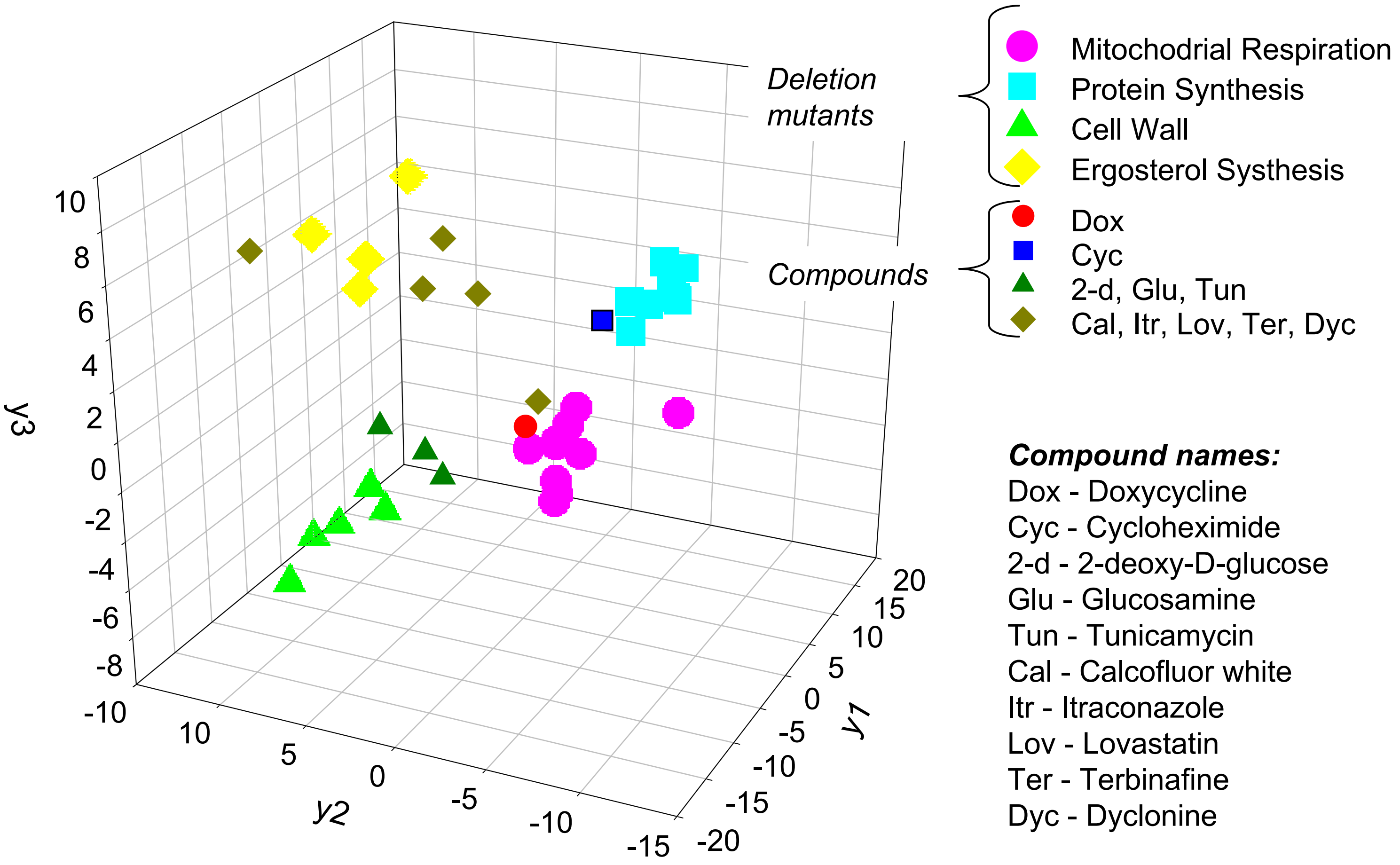
  45 single gene deletion yeast mutants were classified in the following 4 groups according to the effect that each gene deletion had on cell physiology:

  - Mitochondria
  - Cell membrane
  - .....
  - .....

- Microarray gene expression data were collected for each mutant and projected in a CDA space to yield a well defined description of the physiology of each mutant
- Then various drugs were tested as to their effect on the wild type as determined by the expression phenotype and its CDA projection

Drug Discovery (cont'd)





Compound names:

Dox	- Doxycycline
Cyc	- Cycloheximide
2-d	- 2-deoxy-D-glucose
Glu	- Glucosamine
Tun	- Tunicamycin
Cal	- Calcofluor white
Itr	- Itraconazole
Lov	- Lovastatin
Ter	- Terbinafine
Dyc	- Dyclonine

Figure 5. FDA projection of 27 yeast deletion mutant expression phenotype experiments grouped by the functionality of the eliminated gene. Four groups of related mutants have been distinguished using three DFs by projecting the expression levels of 200 of the most discriminating genes. The expression phenotypes obtained from the application of 10 chemical compounds to the wild-type yeast cultures are also projected into the FDA space defined by the mutants. The proximity in FDA space of these projections to those of the expression phenotype of the deletion mutant groups helps characterize the action of the compound on cell physiology. Note that one compound experiment (Cal) which appears incorrectly classified is actually in the center of the 3-D diagram, and not clearly associated with any of the groups shown. The classification suggested by the proximity of the projected phenotypes to the deletion mutants groups agrees with classification provided by Hughes *et al.* (2000).