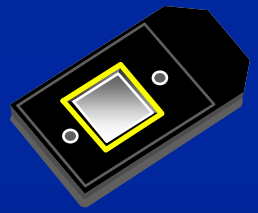
Graphical Exploration of Gene Expression Data by Spectral Map Analysis

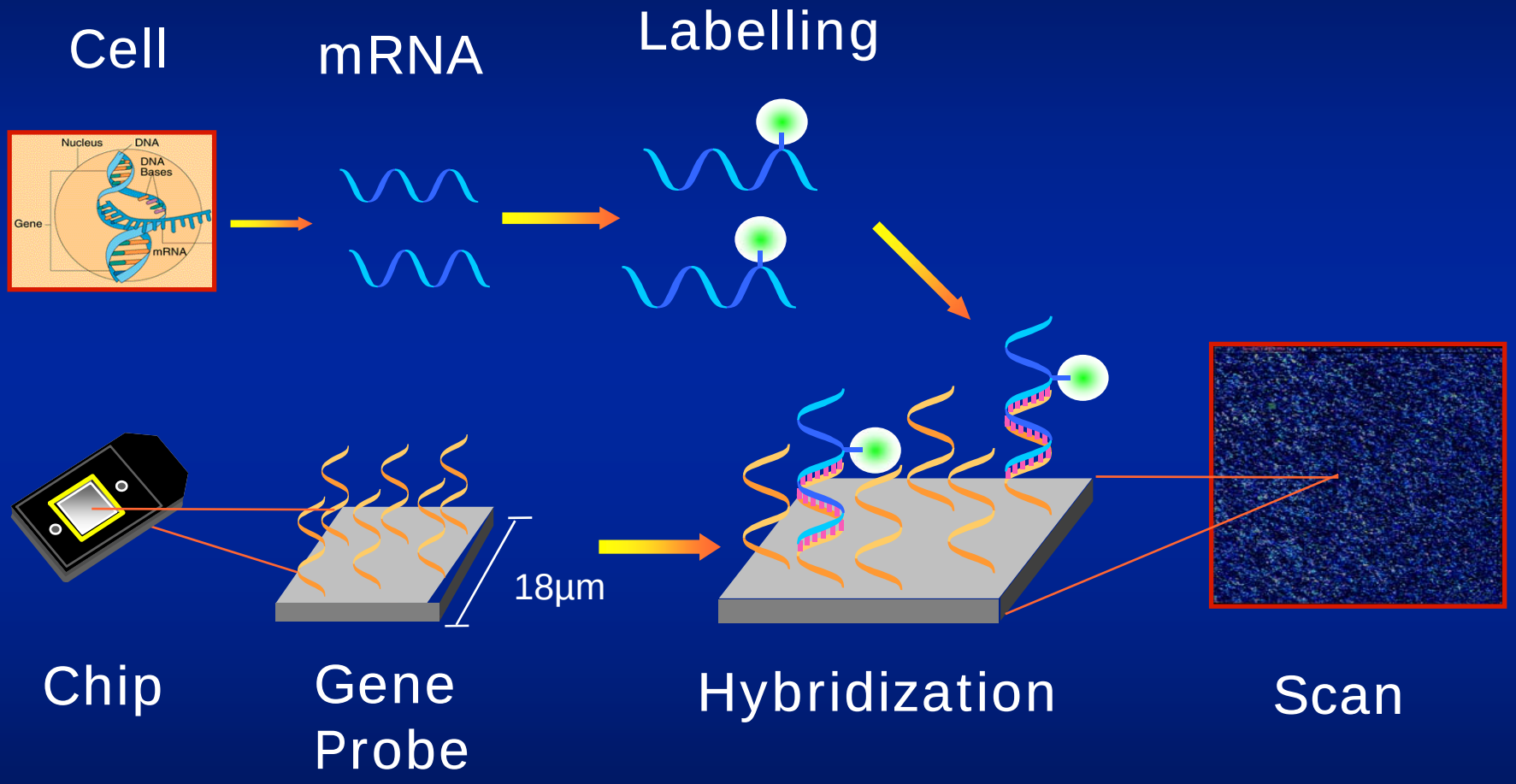


L. Wouters, H. Göhlmann, L. Bijnens, P. Lewi

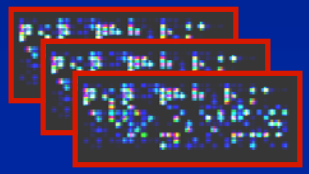
Outline

- Short introduction to microarray technology
- Exploratory multivariate data analysis
- Multivariate projection methods
  - Principal component analysis
  - Correspondence factor analysis
  - Spectral map analysis
- Case data: Leukemia data (Golub, 1999)
- Comparison of multivariate projection methods

Introduction to DNA Microarray Experiments

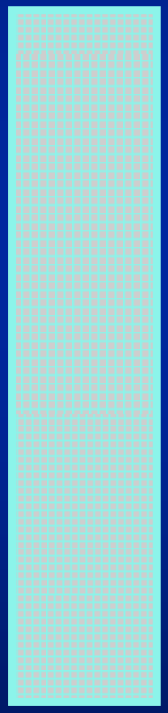


Structure of Microarray Data



Y =

p samples



n genes

- $p \ll n$
  (e.g. $p = 38$, $n = 5327$)

Exploratory Multivariate Analysis
&
Gene Expression Data

- Supervised learning:
  Which genes discriminate known groups of subjects ?
- Unsupervised learning:
  1. Do gene expression profiles reveal groups of patients (class discovery) ?
  2. Which genes correlate with these groups ?

  - clustering methods
  - projection methods

Multivariate Projection Methods
&
Gene Expression Data

- Principal Component Analysis
  Lefkovits, I., et al. (1988)
  Hilsenbeck S.G., et al. (1999)

- Correspondence Factor Analysis
  Fellenberg, K., et al. (2001)

- Spectral Map Analysis

Key Steps in Multivariate Projection Methods
Lewi, P.J. (1993)

- Weighting by marginal means or other
- Re-expression
  replace each table element by its logarithm
- Closure (row, column, or double)
  divide each table element by marginal totals
- Centering (row, column, or double)
  subtract from each table element marginal mean
- Normalization (row, column, or global)
  divide each table element by standard deviation
- Factorization
  generalized SVD
- Projection of scores and loadings with proper scaling
  biplot

Principal Component Analysis

- Pearson (1901), Hotelling (1933)
- Optional log-transformation
- Constant weighting of row and columns
- Column centering
- Column normalization
- Centering and normalization handle tables asymmetrically: R-mode - Q-mode analysis

Correspondence Factor Analysis

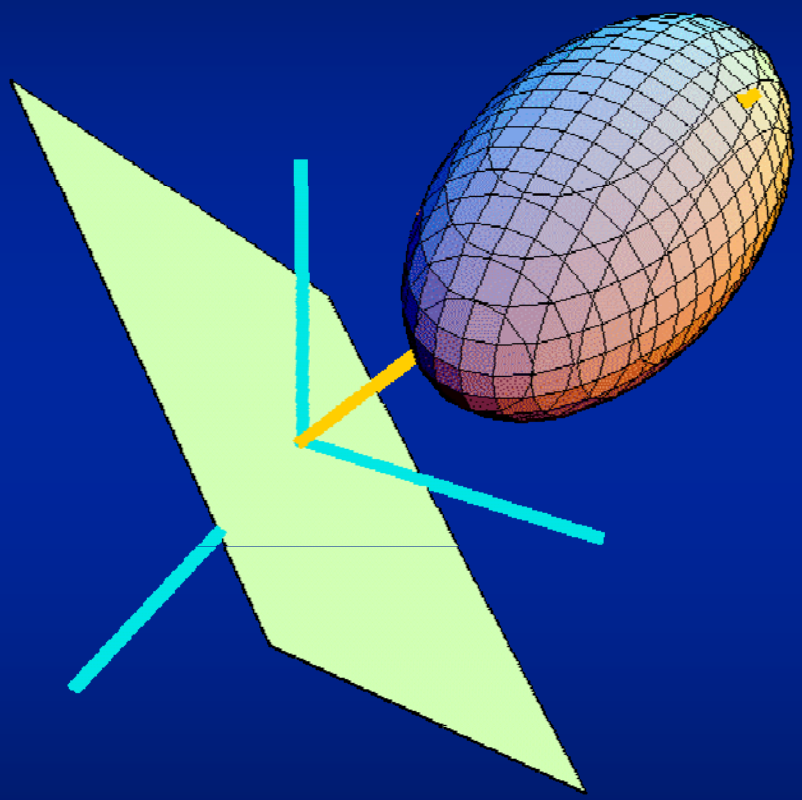
- Benzécri (1973), Greenacre (1984)
- Weighting of row and columns by marginal row and column totals
- Double closure (rank reduction)
- Double centering
- Global normalization
- Distances are chi-square related

Spectral Map Analysis

- Lewi (1976)
- Constant weighting of row and columns, or weighting by marginal row and column totals
- Logarithmic re-expression
- Double centering (rank reduction)
- Global normalization
- Distances are contrasts (log-ratio)
- Areas of symbols are proportional to marginal row-and column-totals or to a selected column

Spectral Map Analysis
Double Centering

- On log-basis related to double closure
- Treats row- and column-space equivalently
- Geometrically results in a projection of the data on a hyperplane orthogonal to the line of identiy
- Number of dimensions is reduced by one
- Effectively removes a “size” component from the data



Properties of Gene Expression Data

- Presence of extreme high values:

  logarithmic re-expression

- Importance of level of gene expression, differences at lower levels are less reliable:

  weighting for marginal totals

- Extreme rectangular shape of matrices, e.g. 7000 rows, 20 columns:
  - assymetric factor scaling (downplay variability among genes)
  - label only most extreme genes

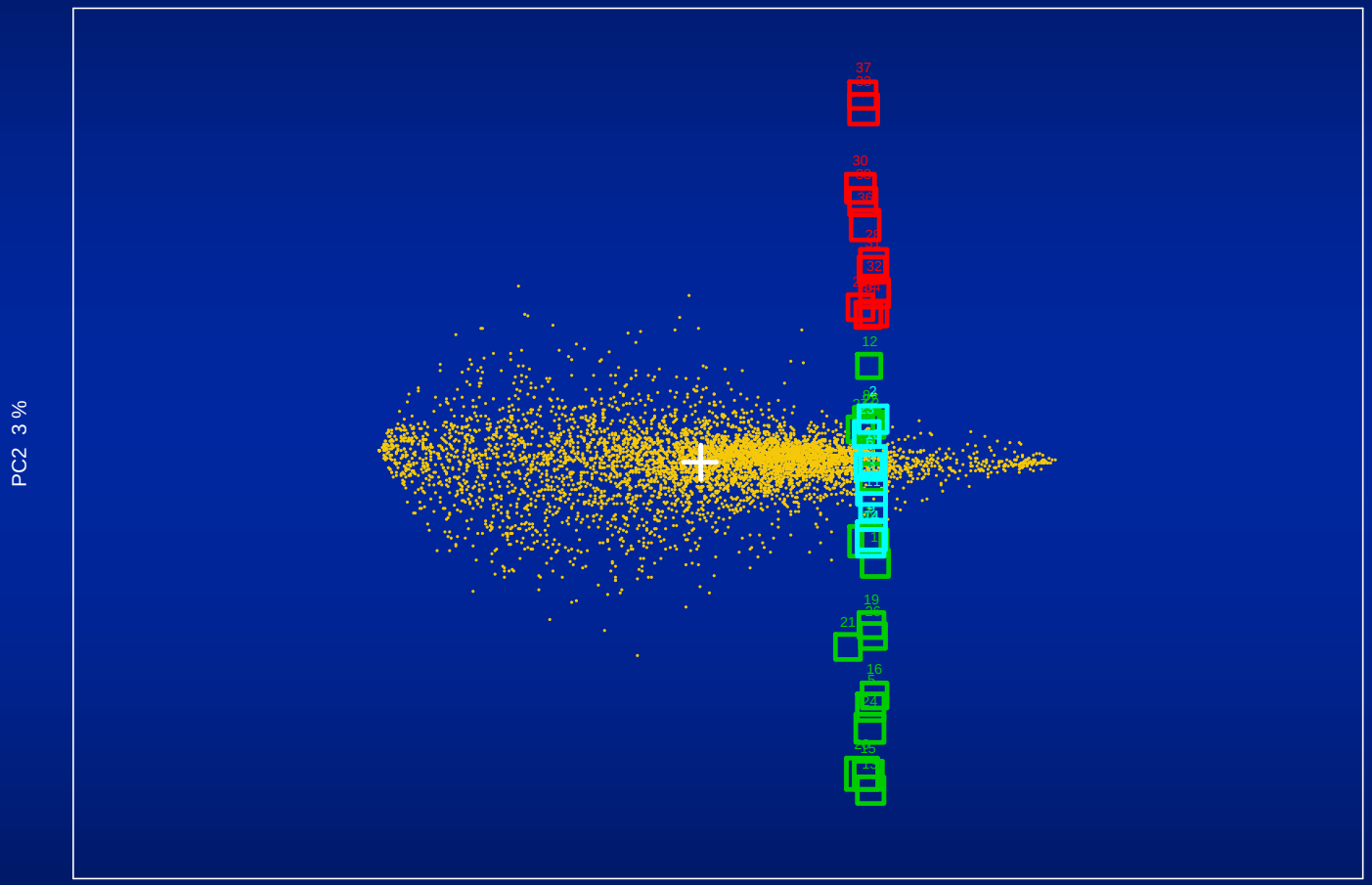
Case Study
MIT Leukemia Data (Golub, 1999)

- Training sample:
  - 38 bone marrow samples
  - 27 acute lymphoblastic leukemia ALL (T-lineage, B-lineage)
  - 11 acute myeloid leukemia AML
  - 6817 genes (5327 used)
  - arbitrary choice of 50 informative genes to discriminate between ALL and AML
- Validation sample:
  - 34 bone marrow samples
  - 20 ALL (T-lineage, B-lineage), 14 AML

Class Discovery & Gene Detection

- Use case study to evaluate & validate three multivariate projection methods:
  - Ability to discover (known) classes
  - Identify genes related to these classes
  - Quantify relationships

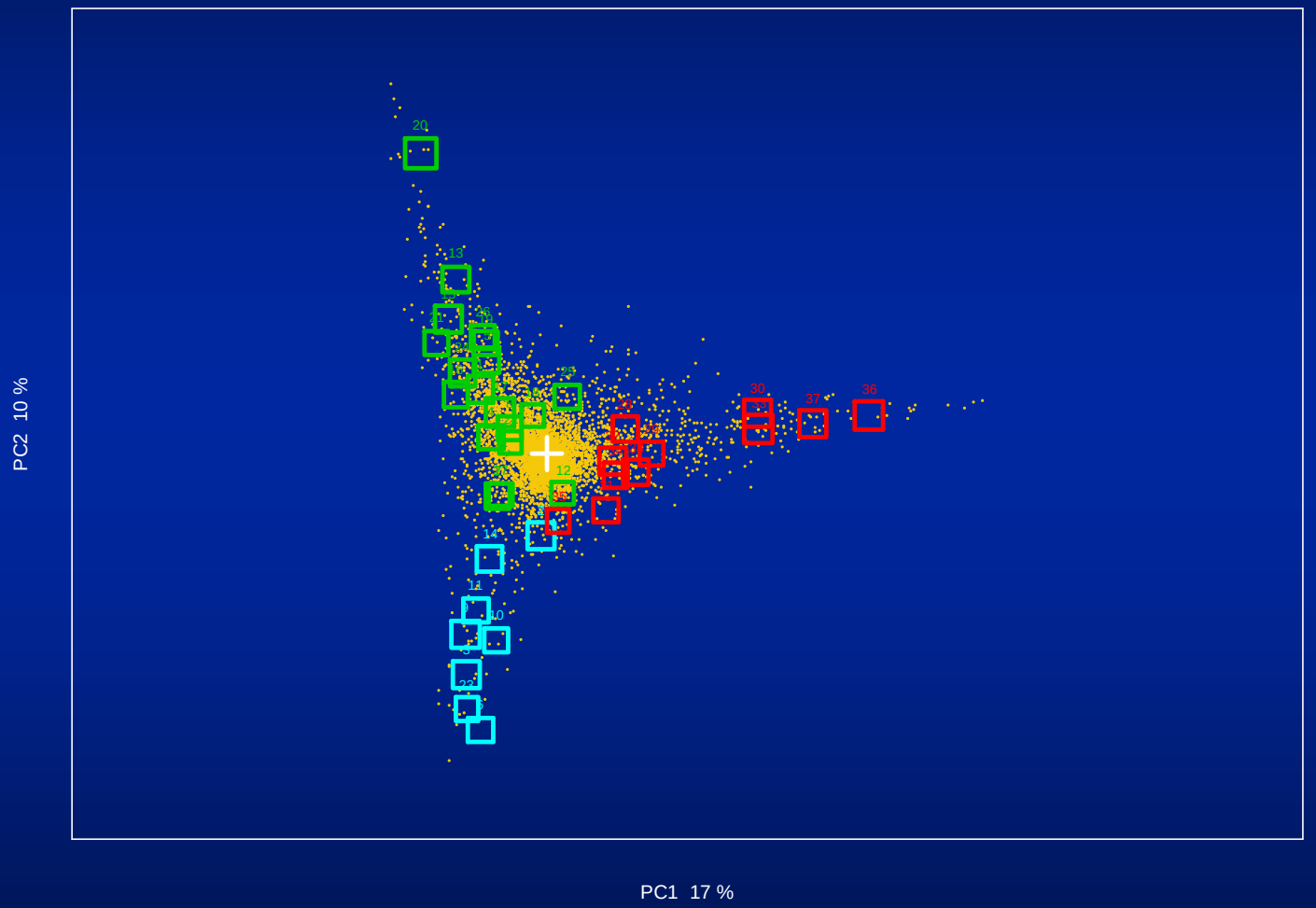
Principal Component Analysis



Principal Component Analysis
Conclusions

- Results of PCA obscured by presence of size component, i.e. overall level of expression of genes
- Poor performance in class discovery
- Useless for detecting genes related to classes

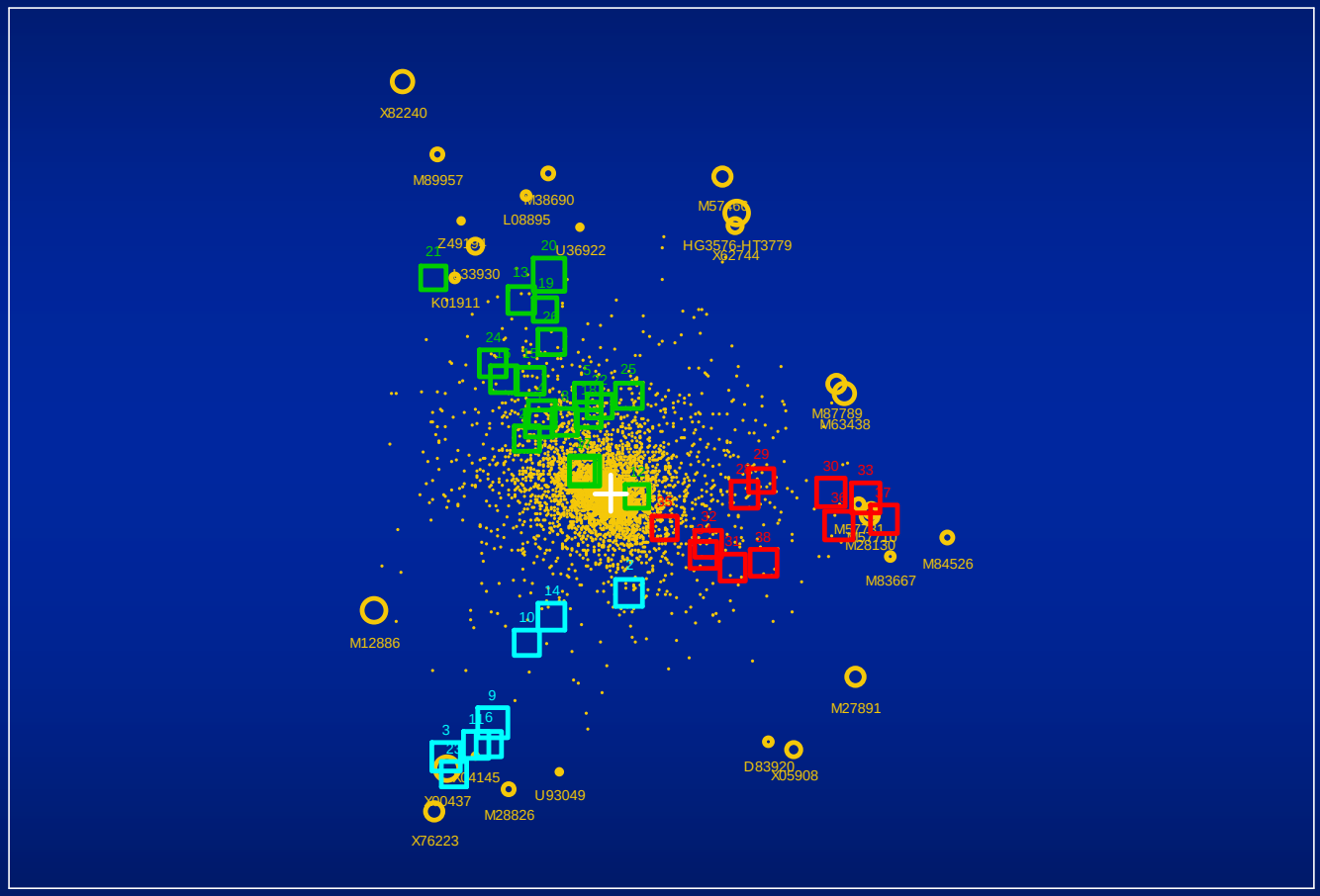
Correspondence Factor Analysis



Correspondence Factor Analysis Conclusions

- Excellent performance in class discovery
- Difficult to detect genes related to classes
- Difficult to interpret distances between genes, samples (chi-square values)

Spectral Map Analysis
Marginal Means Weighted



AML

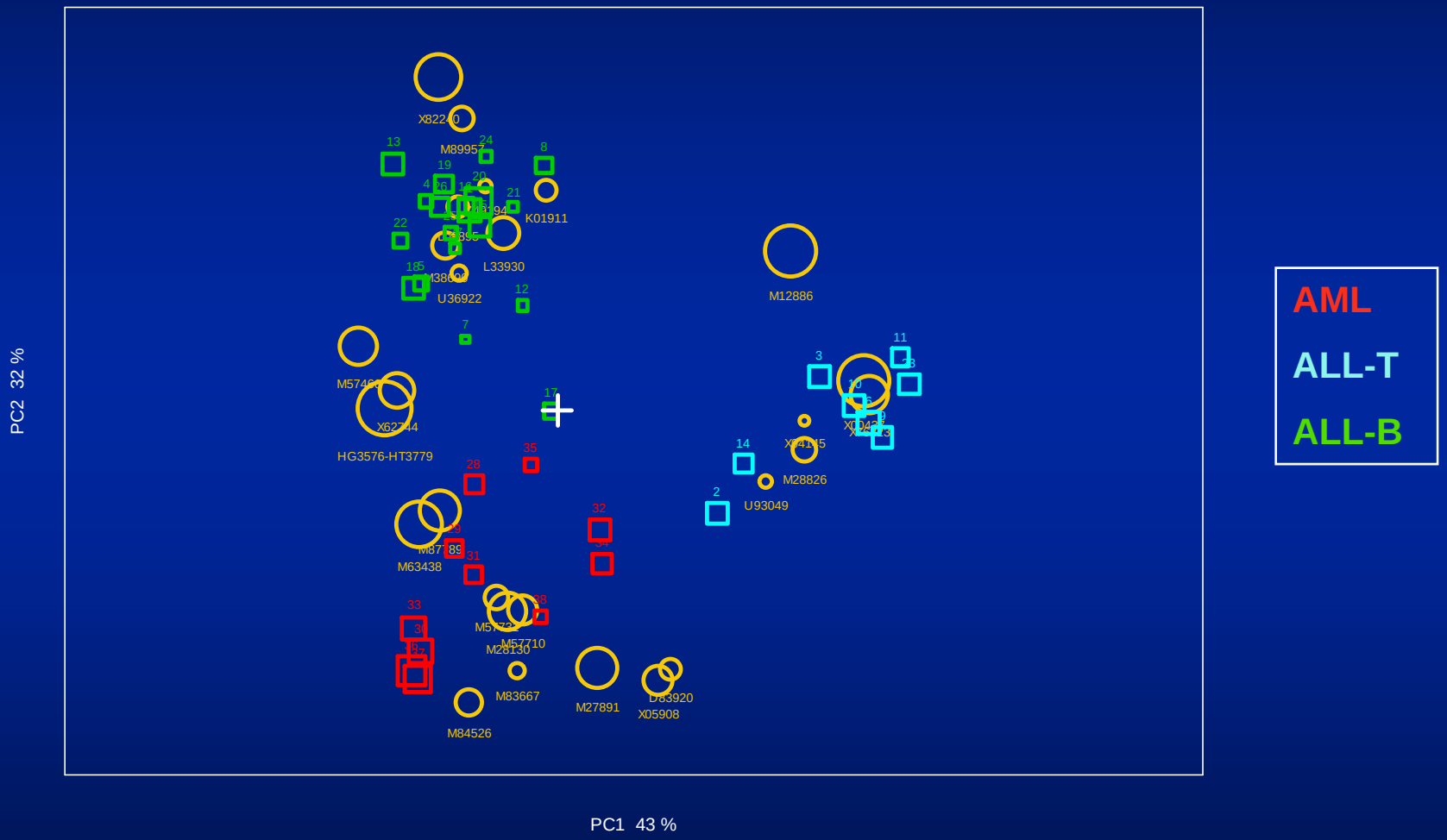
ALL-T

ALL-B

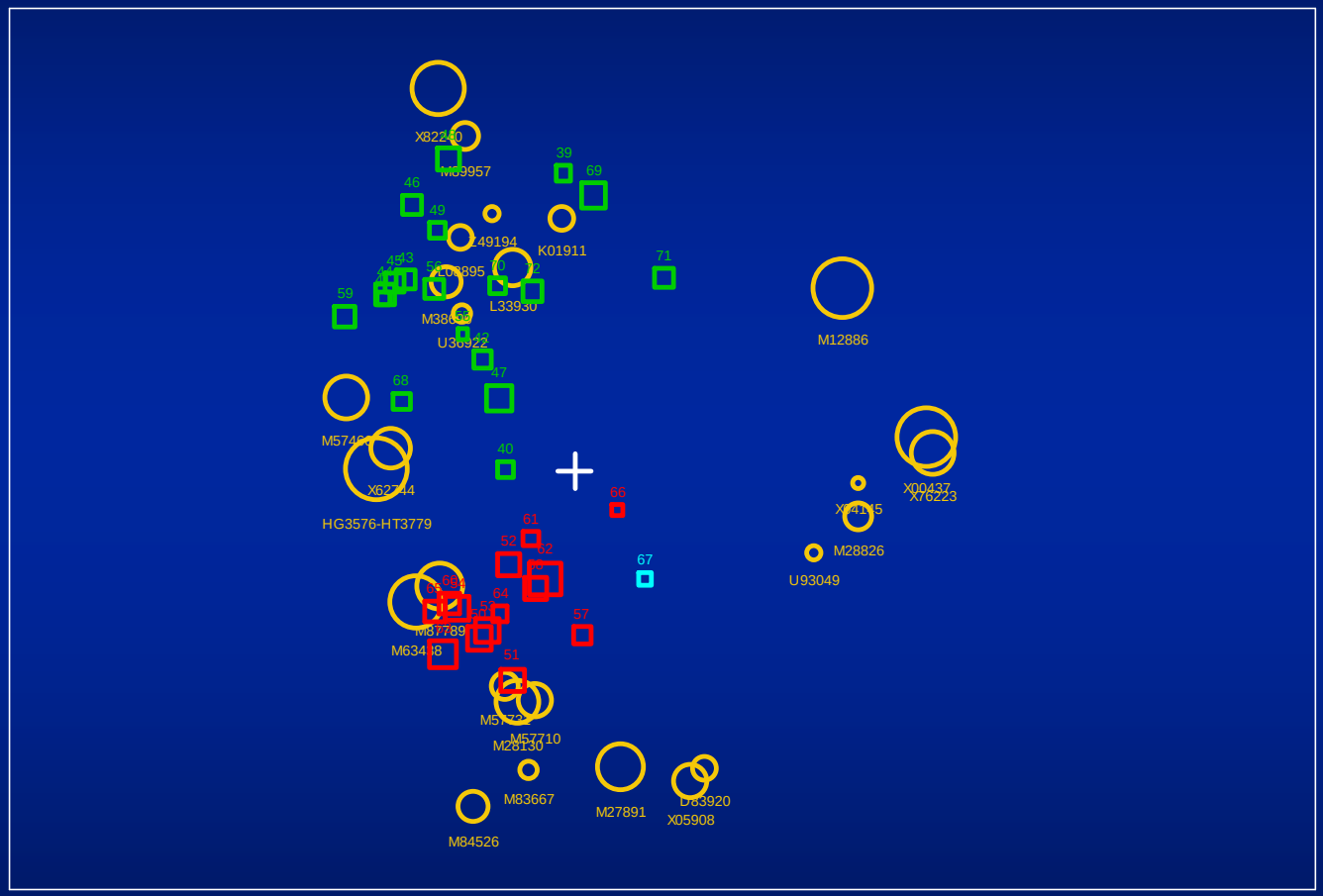
Spectral Map Analysis
Conclusions

- Excellent performance in class discovery
- Allows to detect genes related to classes
- Distances between objects are contrasts (ratios)
- Suggests data reduction

Weighted Spectral Map Analysis
0.5 % Most Extreme Genes



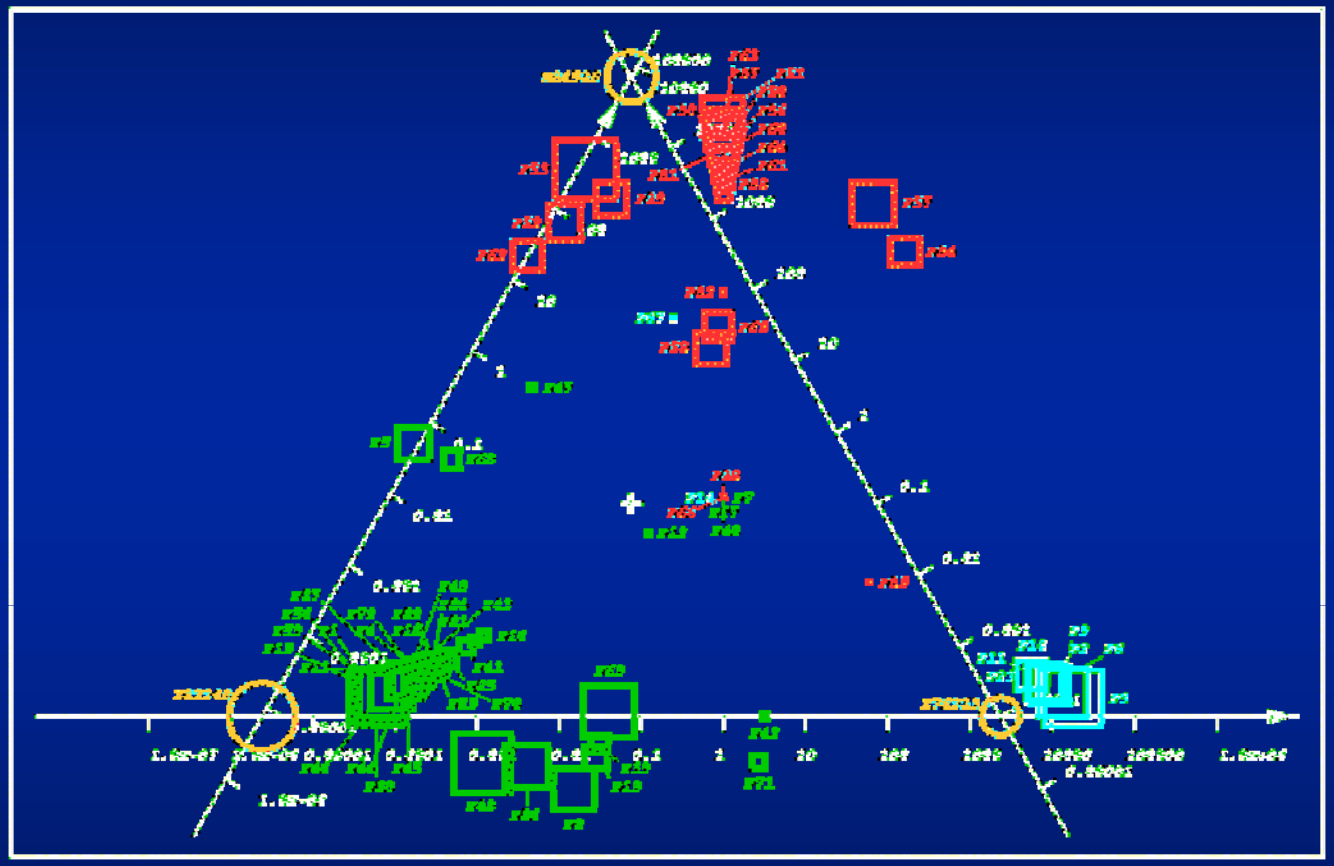
Positioning of Validation Set



Golub Data Set Conclusion

- Weighted SMA allows to reduce data from 5327 genes to 27 genes without loss of important information
- Positioning validation data indicates 27 genes are also predictive for new cases
- Further data reduction possible ?
- Quantify samples for gene specificity ?

Spectral Mapping as a Tool for Quantitative Differential Gene Expression



AML

ALL-T

ALL-B

Conclusions

- Weighted SMA outperforms the other projection methods in:
  - Visualizing data
  - Class detection of biological samples
  - Identifying important genes, confirmed by literature search
  - Reducing the data
  - Identifying & interpreting relations between genes and subjects
  - Quantifying differential gene expression

SPM Library

- Principal component analysis

  PCA<-spectralmap(DATA,center="column",normal="column")
  plot(PCA,scale="uvc",label.tol=c(1,1),col.group=GRP)

- Correspondence analysis

  CFA<spectralmap(DATA,logtrans=False,row.weight="mean",
  col.weight="mean",closure="double")
  plot(CFA,scale="uvc",label.tol=c(1,1),col.group=GRP)

- Spectral map analysis, weighted by marginal totals

  SPM<-spectralmap(DATA,row.weight="mean",col.weight="mean")
  plot(SPM,scale="uvc",col.group=GRP)

References

- Wouters, L., Göhlmann, H.W., Bijnens, L., Kass, S.U., Molenberghs, G., Lewi, P.J. (2003). Graphical Exploration of Gene Expression Data: A Comparative Study of Three Multivariate Methods. Biometrics 59, 1131-1139,

- SPM-library availability: http://www.datascope.be