Bioinformatics and Biomarker Discovery Part 3: Examples

Limsoon Wong

28 August 2008

National University of Singapore

Outline

- **ALL**
  - Gene expression profile classification
  - Beyond diagnosis and prognosis
- **WEKA**
  - Breast cancer
  - Dermatology
  - Pima Indians
  - Echocardiogram
  - Mammography

Gene Expression Profile Classification

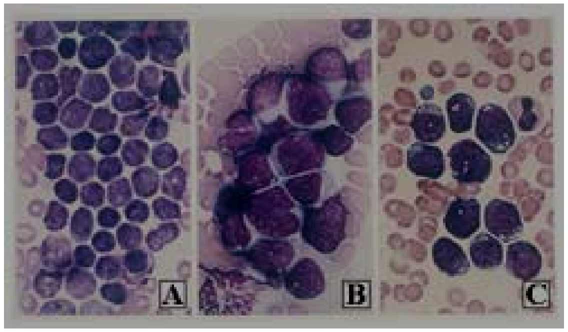
Diagnosis of Childhood Acute Lymphoblastic Leukemia and Optimization of Risk-Benefit Ratio of Therapy

NUS
National University
of Singapore

Childhood ALL

- **Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse

- **The subtypes look similar**



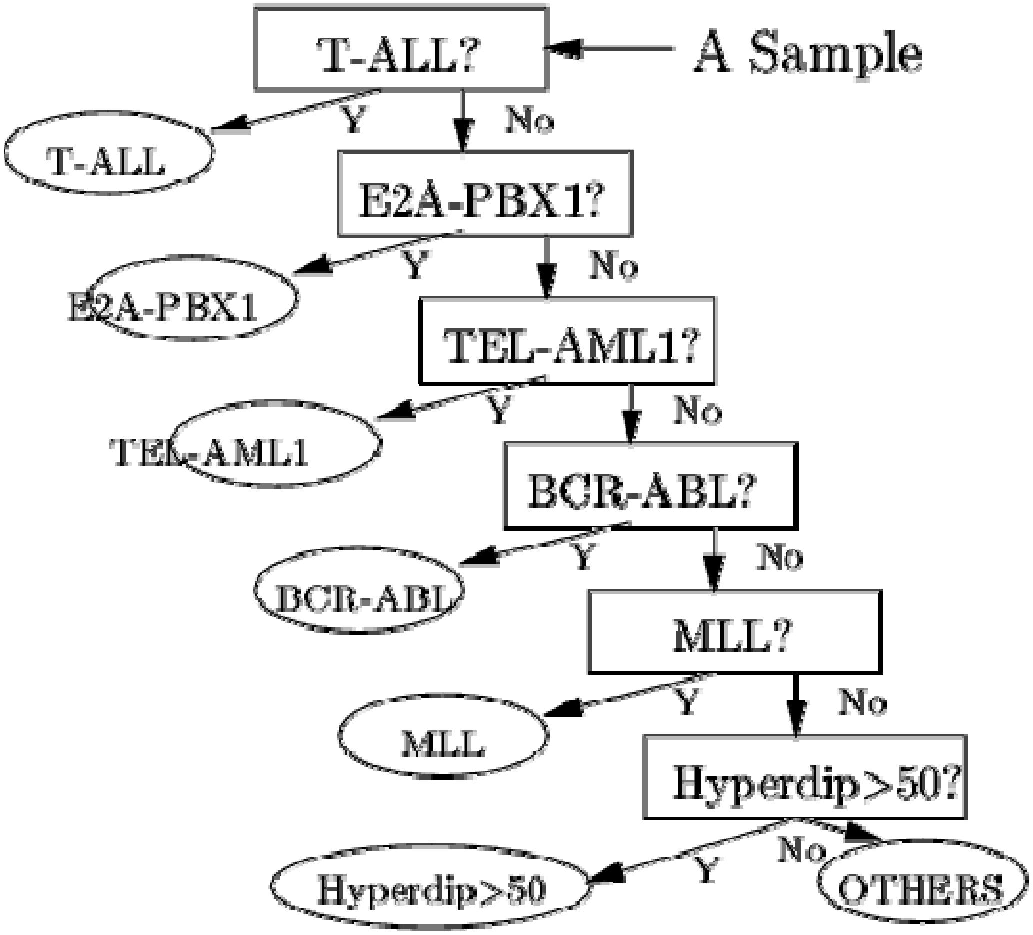
- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- **Unavailable in most ASEAN countries**

Subtype Diagnosis by PCL

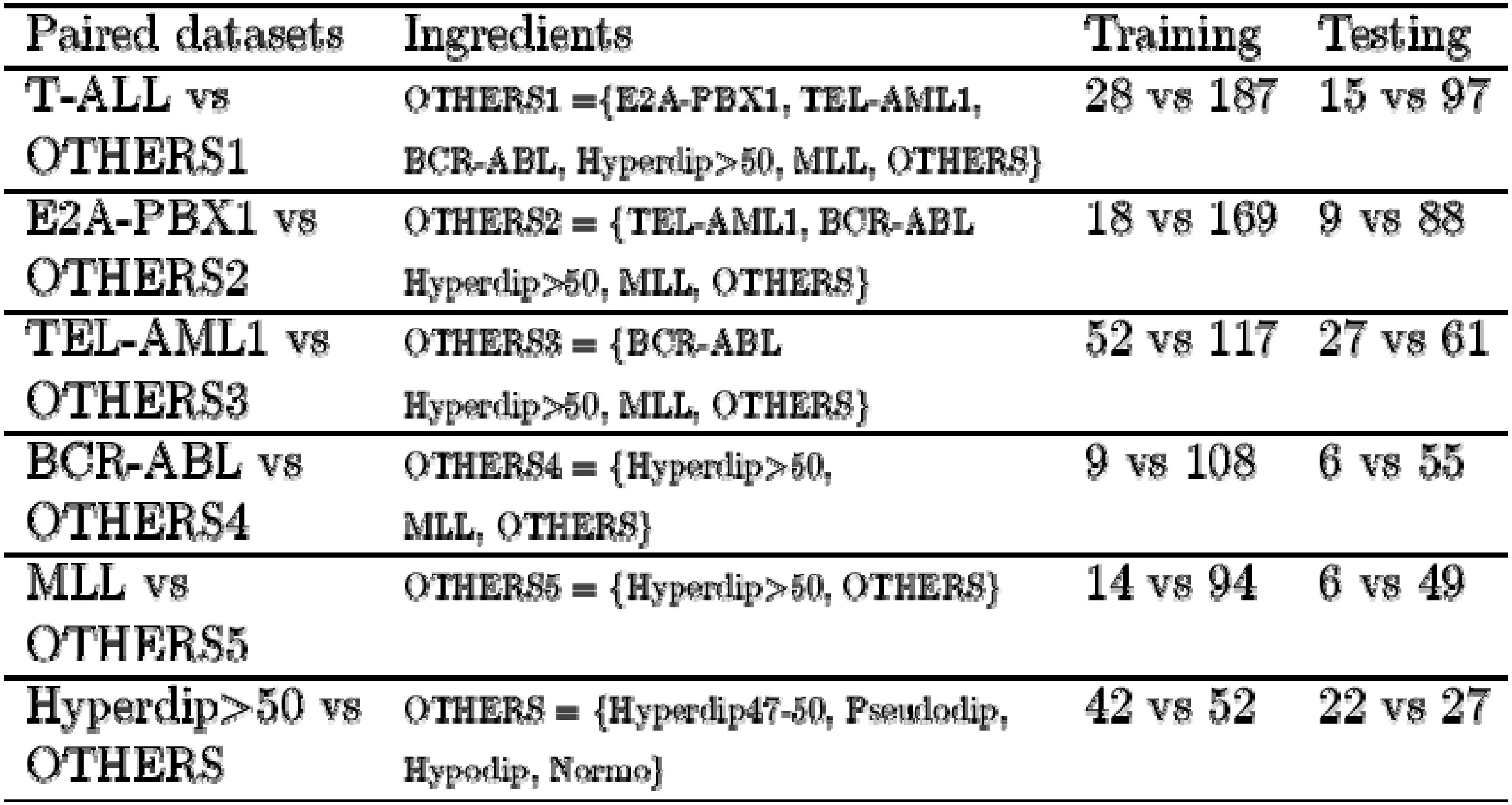
- Gene expression data collection
- Gene selection by $\chi 2$
- Classifier training by emerging pattern
- ~~Classifier tuning (optional for some machine learning methods)~~
- Apply classifier for diagnosis of future cases by PCL

Childhood ALL Subtype Diagnosis Workflow

A tree-structured diagnostic workflow was recommended by our doctor collaborator

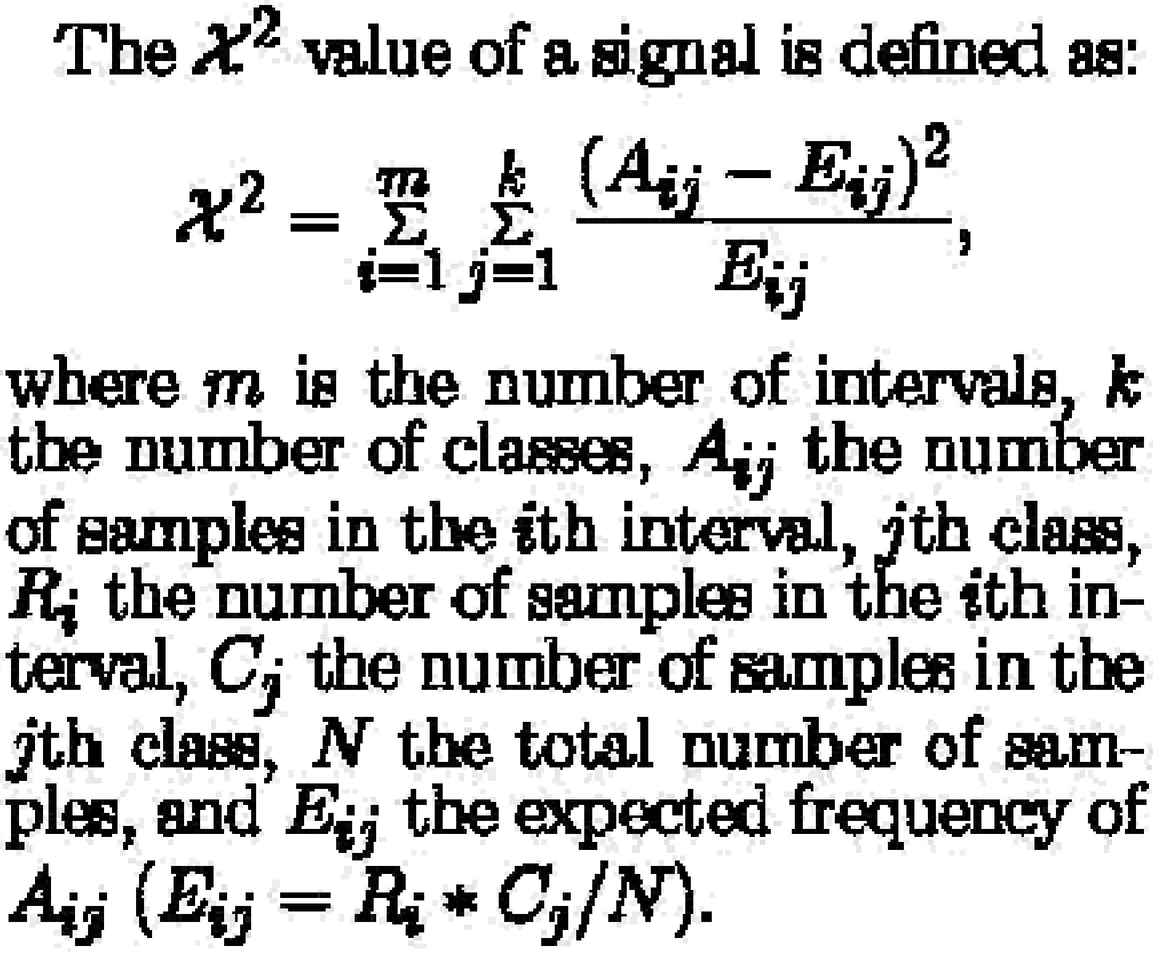


Training and Testing Sets



Paired datasets	Ingredients	Training	Testing
T-ALL vs OTHERS1	OTHERS1 ={E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip>50, MLL, OTHERS}	28 vs 187	15 vs 97
E2A-PBX1 vs OTHERS2	OTHERS2 = {TEL-AML1, BCR-ABL Hyperdip>50, MLL, OTHERS}	18 vs 169	9 vs 88
TEL-AML1 vs OTHERS3	OTHERS3 = {BCR-ABL Hyperdip>50, MLL, OTHERS}	52 vs 117	27 vs 61
BCR-ABL vs OTHERS4	OTHERS4 = {Hyperdip>50, MLL, OTHERS}	9 vs 108	6 vs 55
MLL vs OTHERS5	OTHERS5 = {Hyperdip>50, OTHERS}	14 vs 94	6 vs 49
Hyperdip>50 vs OTHERS	OTHERS = {Hyperdip47-50, Pseudodip, Hypodip, Normo}	42 vs 52	22 vs 27

Signal Selection by χ2

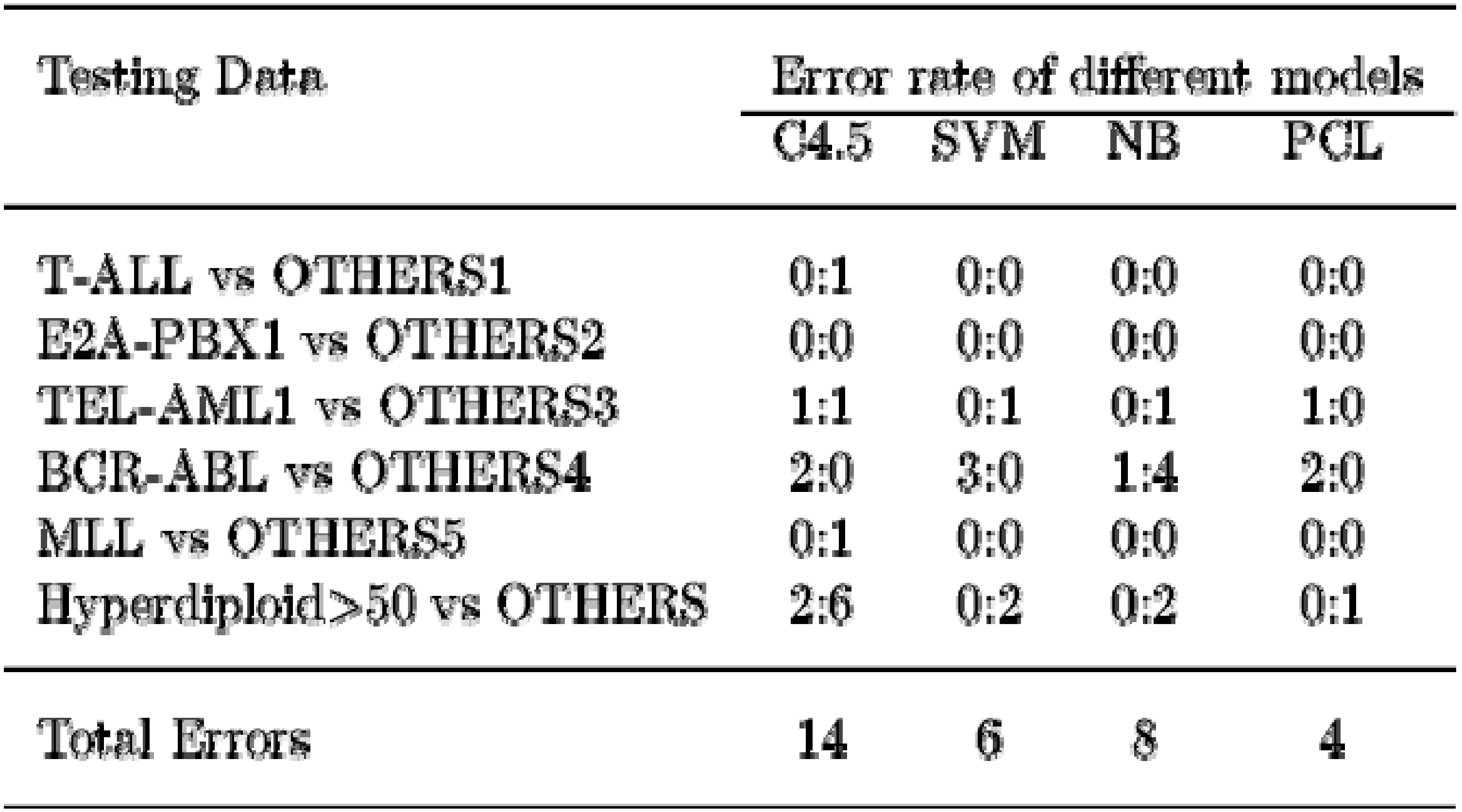


The χ^2 value of a signal is defined as:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

where m is the number of intervals, k the number of classes, A_{ij} the number of samples in the ith interval, jth class, R_i the number of samples in the ith interval, C_j the number of samples in the jth class, N the total number of samples, and E_{ij} the expected frequency of A_{ij} ($E_{ij} = R_i * C_j / N$).

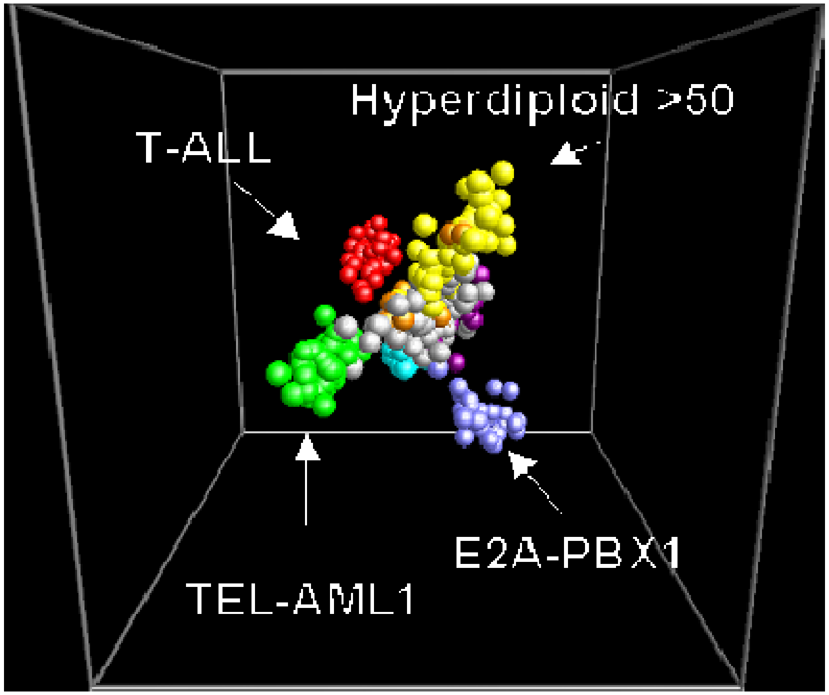
Accuracy of Various Classifiers

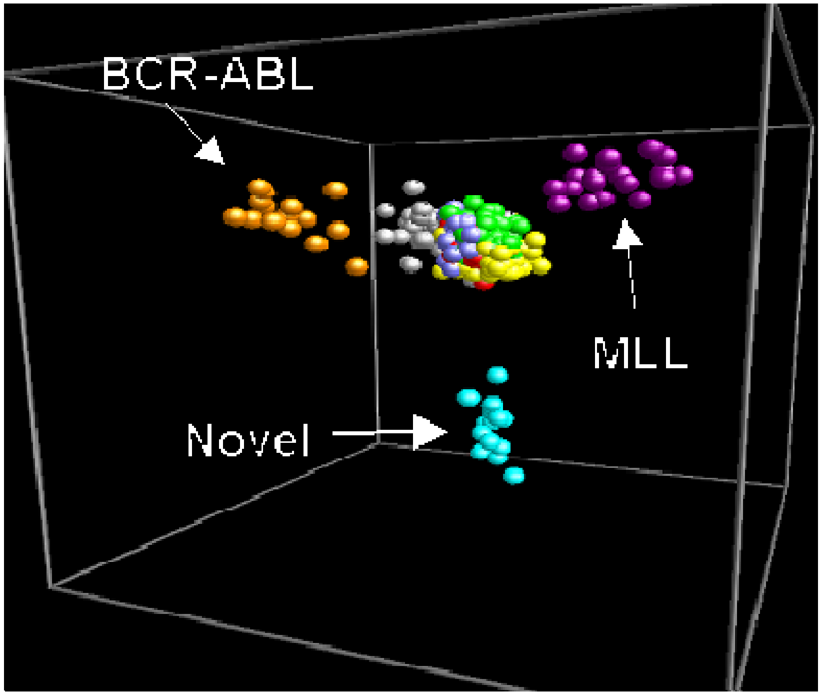


Testing Data	Error rate of different models			
	C4.5	SVM	NB	PCL
T-ALL vs OTHERS1	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS2	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS3	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS4	2:0	3:0	1:4	2:0
MLL vs OTHERS5	0:1	0:0	0:0	0:0
Hyperdiploid>50 vs OTHERS	2:6	0:2	0:2	0:1
Total Errors	14	6	8	4

The classifiers are all applied to the 20 genes selected by $\chi 2$ at each level of the tree

Visualization by PCA

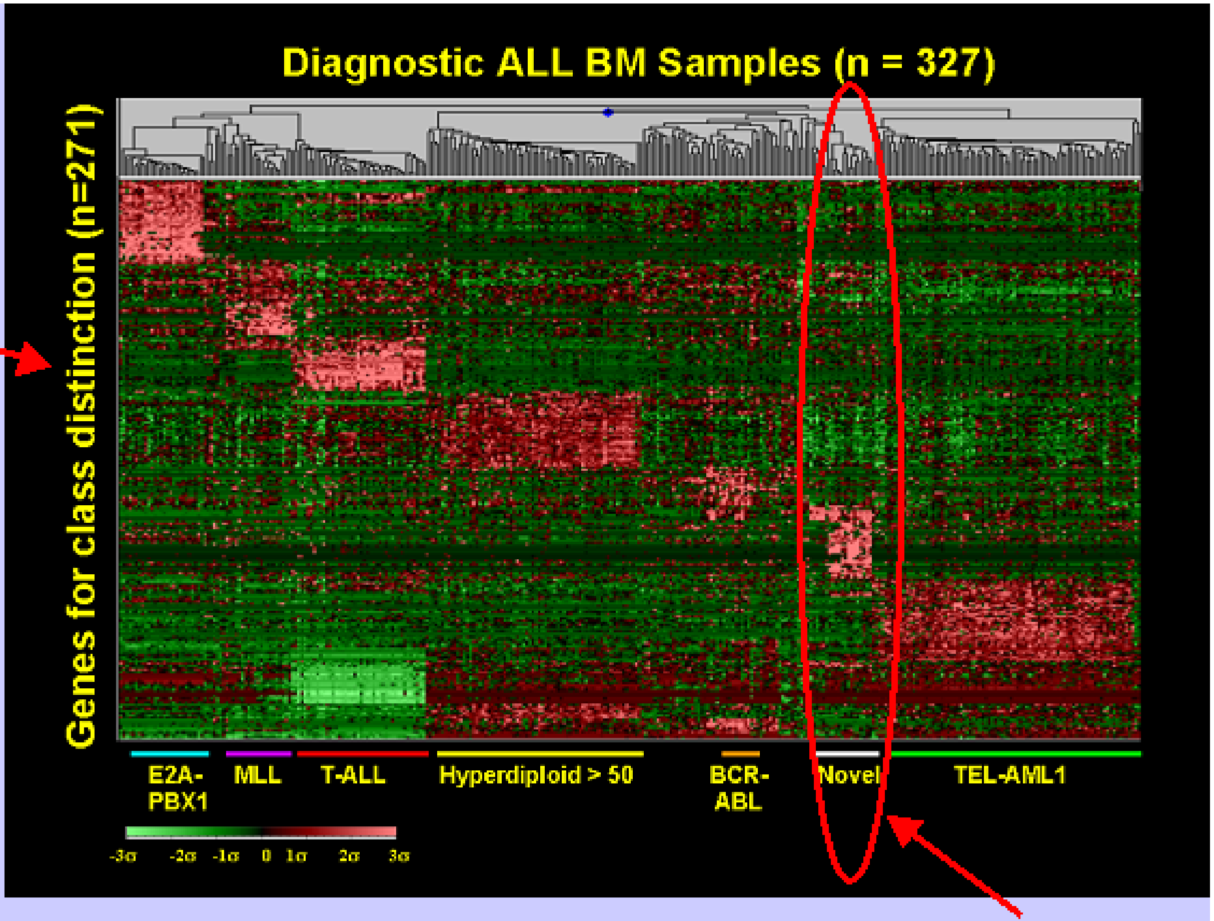




Obtained by performing PCA on the 20 genes chosen for each level

Visualization by Clustering

Genes selected by χ2

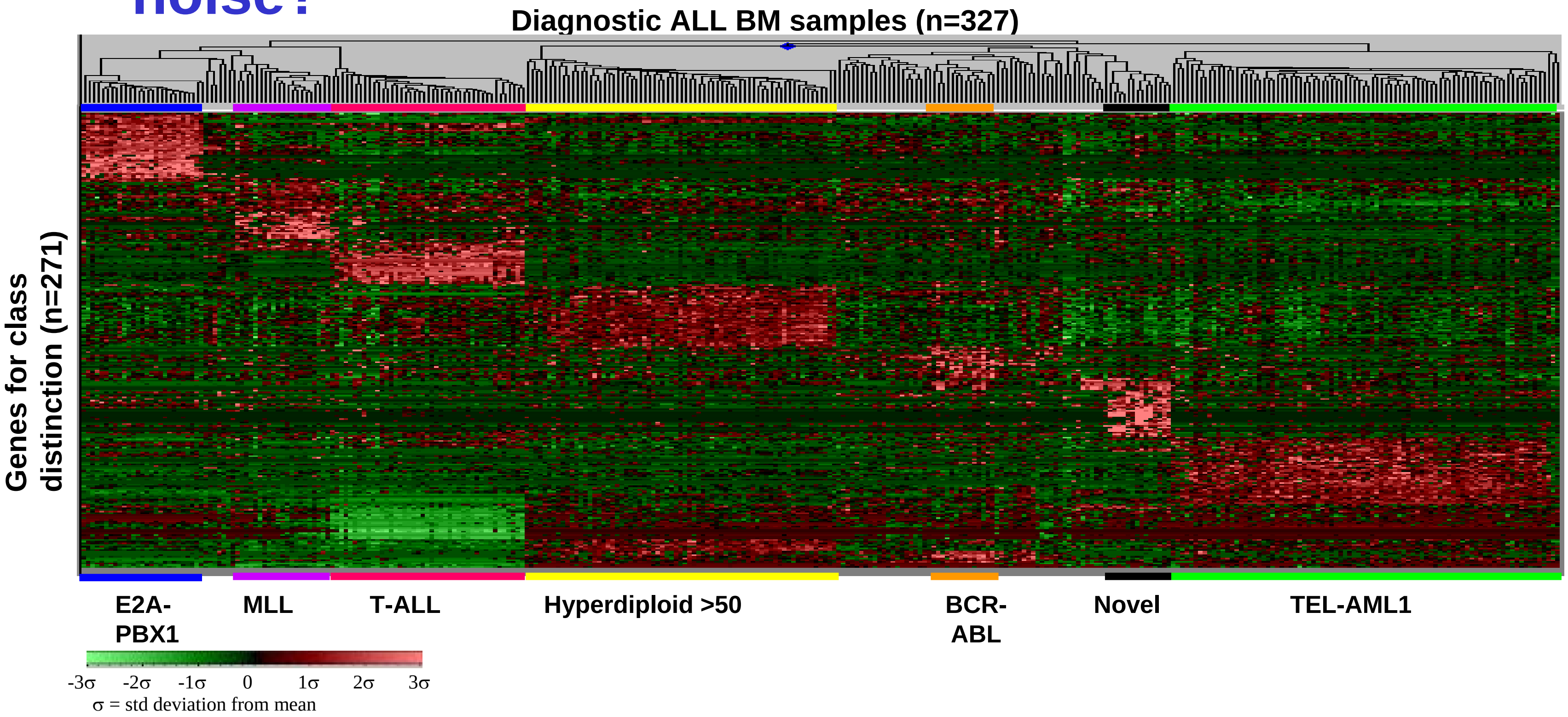


New subtype discovered

Beyond Disease Diagnosis & Prognosis

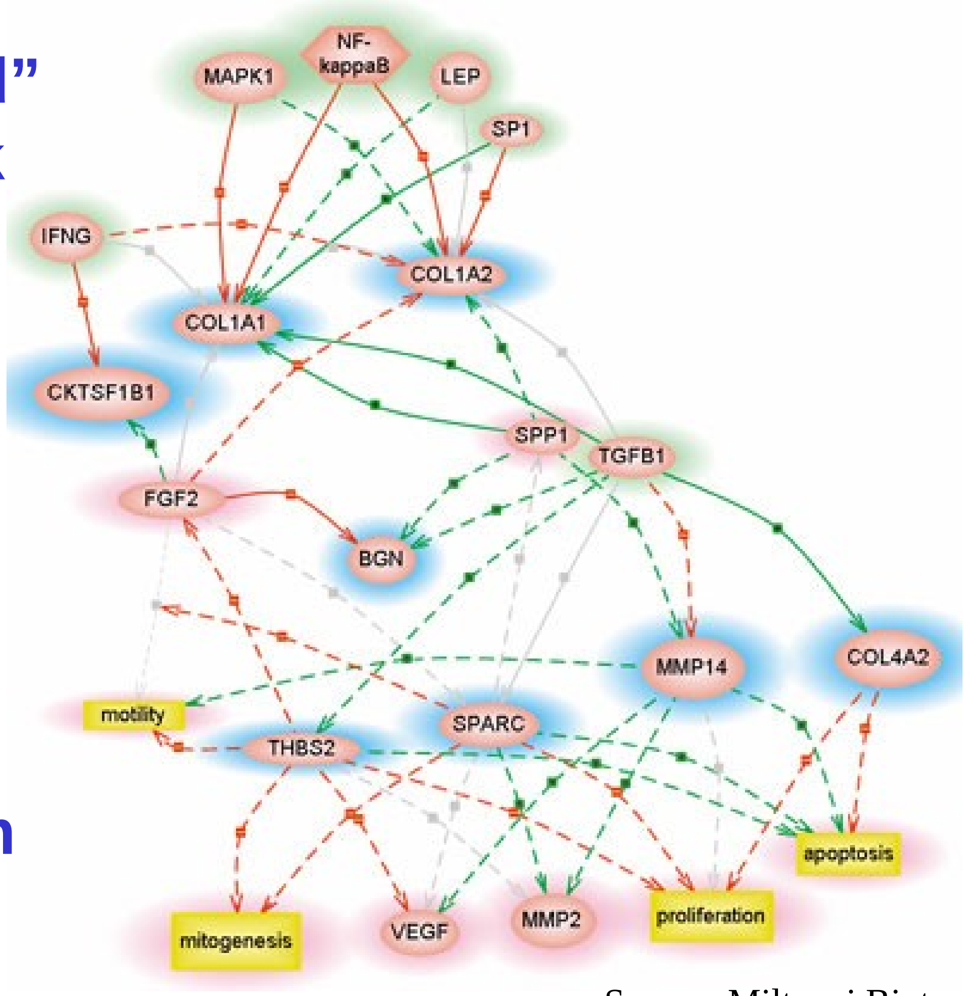
Beyond Classification of Gene Expression Profiles

- **After identifying the candidate genes by feature selection, do we know which ones are causal genes, which ones are surrogates, and which are noise?**



Gene Regulatory Circuits

- **Genes are “connected” in “circuit” or network**
- **Expr of a gene in a network depends on expr of some other genes in the network**
- **Can we “reconstruct” the gene network from gene expression and other data?**



Source: Miltenyi Biotec

Hints to extend reach of prediction

- Each disease subtype has underlying cause

⇒ There is a unifying biological theme for genes that are truly associated with a disease subtype.

- Uncertainty in reliability of selected genes can be reduced by considering molecular functions and biological processes associated with the genes

- The unifying biological theme is basis for inferring the underlying cause of disease subtype

Intersection Analysis

- Intersect the list of differentially expressed genes with a list of genes on a pathway

- If intersection is significant, the pathway is postulated as basis of disease subtype or treatment response

Exercise: What is a good test statistics to determine if the intersection is significant?

Caution:

- Initial list of differentially expressed genes is defined using test statistics with arbitrary thresholds
- Diff test statistics and diff thresholds result in a diff list of differentially expressed genes

⇒ Outcome may be unstable

Any Question?

NUS

National University of Singapore

References

- E.-J. Yeoh et al., “Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling”, *Cancer Cell*, 1:133--143, 2002
- H. Liu, J. Li, L. Wong. Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data. *Bioinformatics*, 21(16):3377--3384, 2005.
- L.D. Miller et al., “Optimal gene expression analysis by microarrays”, *Cancer Cell* 2:353--361, 2002
- J. Li, L. Wong, “Techniques for Analysis of Gene Expression”, *The Practical Bioinformatician*, Chapter 14, pages 319—346, WSPC, 2004
- D. Soh, D. Dong, Y. Guo, L. Wong. “Enabling More Sophisticated Gene Expression Analysis for Understanding Diseases and Optimizing Treatments”. *ACM SIGKDD Explorations*, 9(1):3--14, 2007

A Popular Software Package: WEKA

- **http://www.cs.waikato.ac.nz/ml/weka**
- **Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.**

Exercise: Download a copy of WEKA. What are the names of classifiers in WEKA that correspond to C4.5 and SVM?

Let’s try WEKA on …

- **Breast cancer**
- **Dermatology**
- **Pima Indians**
- **Echocardiogram**
- **Mammography**