Molecular Classification of Cancer

Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub, D. K. Slonim & Others

1999

Big Picture in 1999

- **The Need for Cancer Classification**
  - Cancer classification very important for advances in cancer treatment.
  - Cancers of Identical grade can have widely variable clinical courses
- **Focus on improving cancer treatment by:**
  - Targeting specific therapies to pathogenetically distinct tumor types
  - To maximize efficacy
  - To minimize toxicity

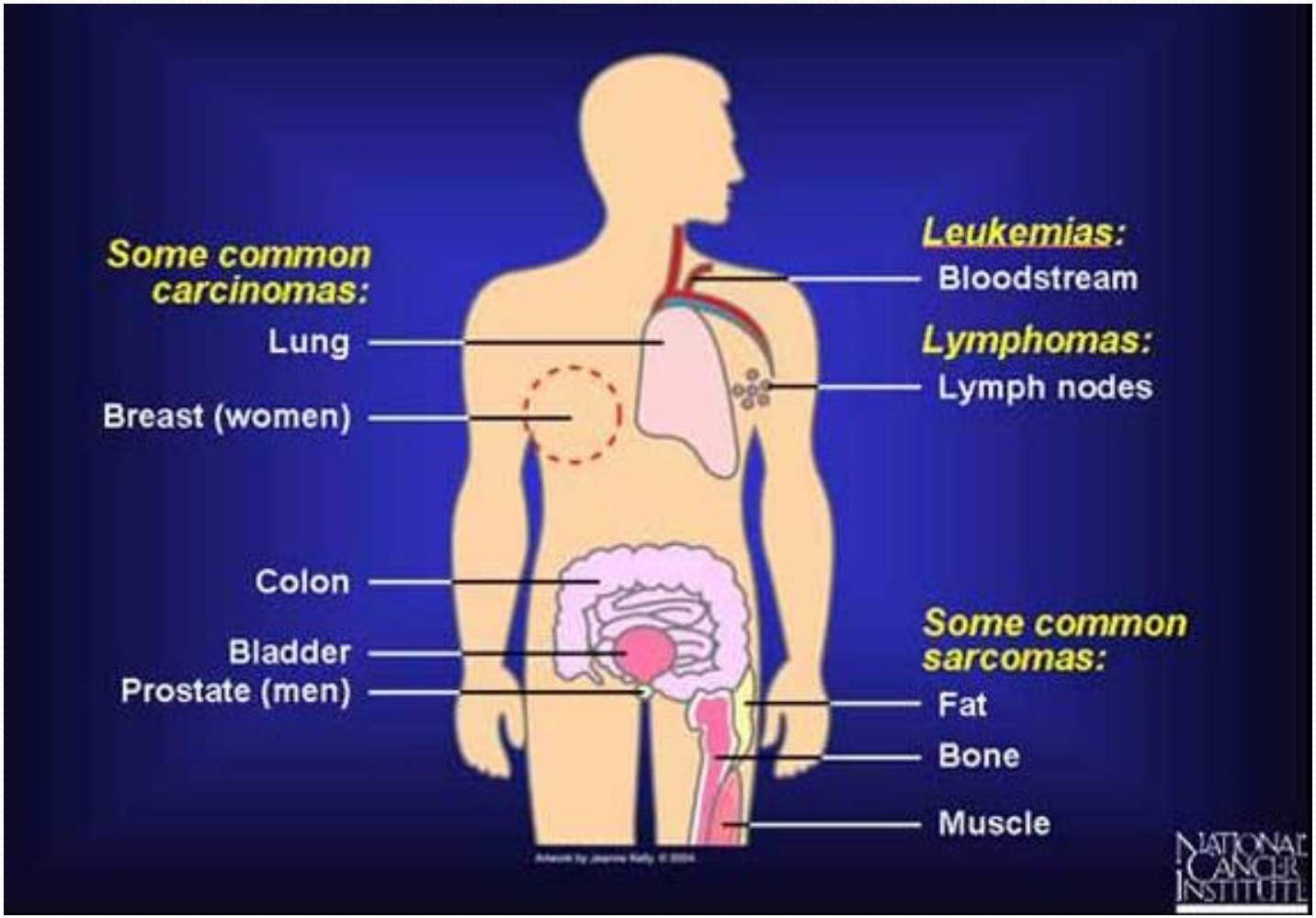
Big Picture in 1999

- **Cancer classification based on:**
  - Morphological appearance.
  - Enzyme-based histochemical analyses.
  - Immunophenotyping.
  - Cytogenetic analysis.
- **Methods had serious limitations:**
  - Tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy
  - Some of these differences have been explained by dividing tumors into sub-classes
  - In other tumors, important sub-classes may exist but are yet to be defined
- **Classification historically relied on specific biological insights**

Executive Summary

- A **generic approach** to cancer classification based on Gene Expression Monitoring by DNA microarrays
- Applied to human **Acute Leukemias** as a test case
- A **Class Discovery procedure** automatically discovered the distinction between AML and ALL without prior knowledge.
- An automatically derived **Class Predictor** to determine the class of new leukemia cases.
- **Bottom-line: A general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.**

Types of Cancer



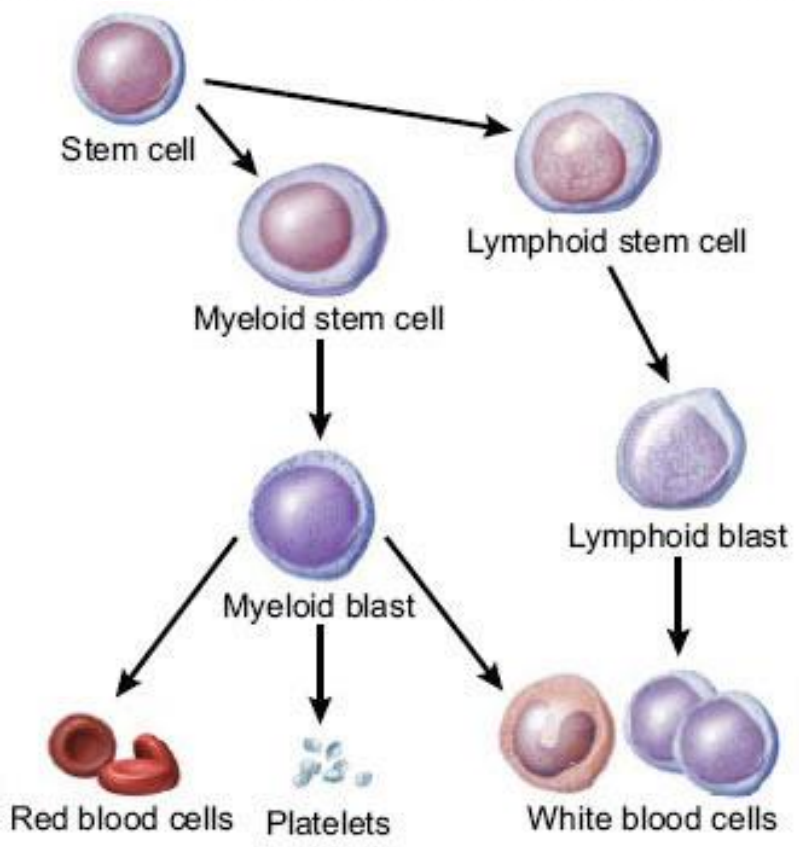
Leukemia

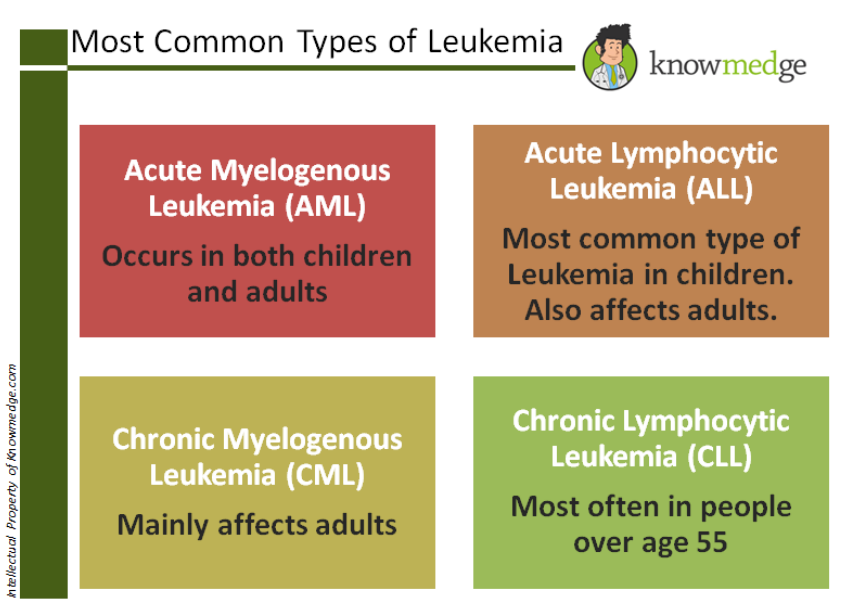
- Leukemia is Cancer of the Blood or Bone Marrow
- Characterized by abnormal production of WBC in the body

Classification of Leukemia

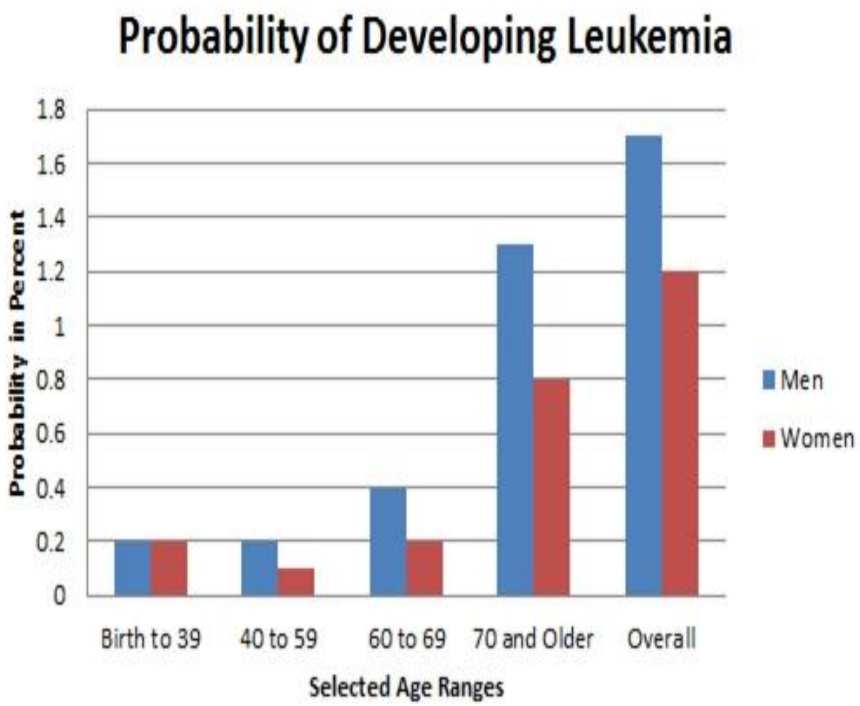
- **Acute vs Chronic**
  - **Chronic:** The abnormal cells are more mature (look more like normal white blood cells)
  - **Acute:** Abnormal cells are immature (look more like stem cells).
- **Myelogenous vs Lymphocytic**
  - **Myelogenous:** Leukemias that start in early forms of myeloid cells
  - **Lymphocytic:** Leukemias that start in immature forms of lymphocytes

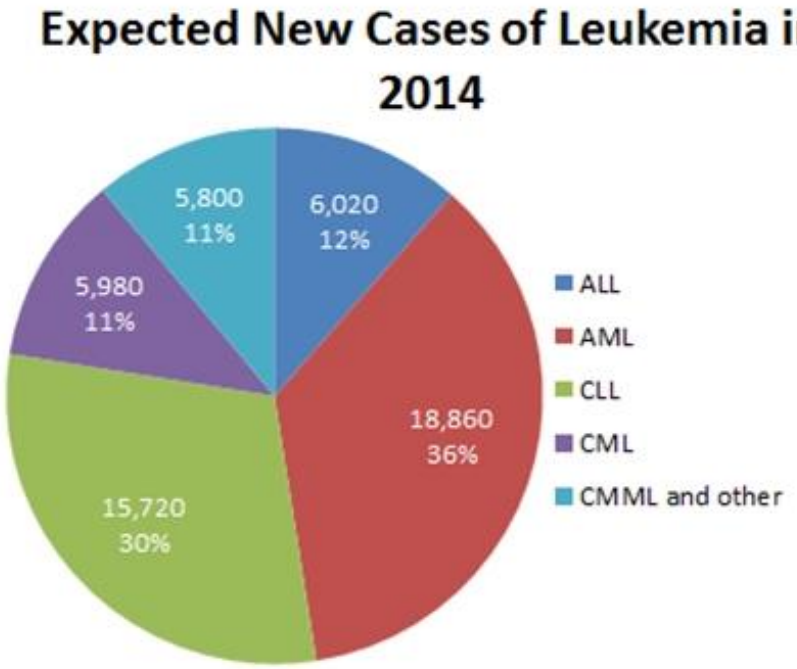
Classification of Leukemia





Some Statistics on Leukemia





More Background on Leukemia

- In 1999, no single test is sufficient to establish the diagnosis
- A combination of different tests in morphology, histochemistry and immunophenotyping used.
- Although usually accurate, leukemia classification remains imperfect and errors do occur

Problem

How do we categorize different types of Cancer so that we can increase effectiveness of treatments and decrease toxicity?

Motivation

No general approach for identifying new cancer classes (Class Discovery) or for assigning tumors to known classes (Class Prediction).

Idea / Intuition

Cancers can be automatically classified based on Gene Expression.

Objective

To develop a more systematic approach to cancer classification based on the simultaneous expression monitoring of thousands of genes using DNA microarrays with leukemia as test cases.

Gene Expression Monitoring

- **Gene Expression**
  - Process by which information from a gene is used in the synthesis of a functional gene product.
  - Products are typically proteins
  - In tRNA or snRNA genes, the product is a functional RNA.

Problem Breakdown

- **Class Prediction:** Assignment of particular tumor samples to already-defined classes (supervised learning).
- **Class Discovery:** Defining previously unrecognized tumor subtypes. (unsupervised learning).

Class Prediction

- How can we use an initial collection of samples belonging to known classes to create a class Predictor?
  - **Issue-1:** Are there genes whose expression pattern are strongly correlated with the class distinction to be predicted?
  - **Issue-2:** How do we use a collection of known samples to create a "class predictor" capable of assigning a new sample to one of two classes?
  - **Issue-3:** How do we test the validity of these class predictors?

Data: Biological Samples

- **Primary samples:**
  - 38 bone marrow samples (27 ALL, 11 AML)
  - obtained from acute leukemia patients at diagnosis
- **Independent samples:**
  - 34 leukemia samples (24 bone marrow, 10 peripheral blood samples)

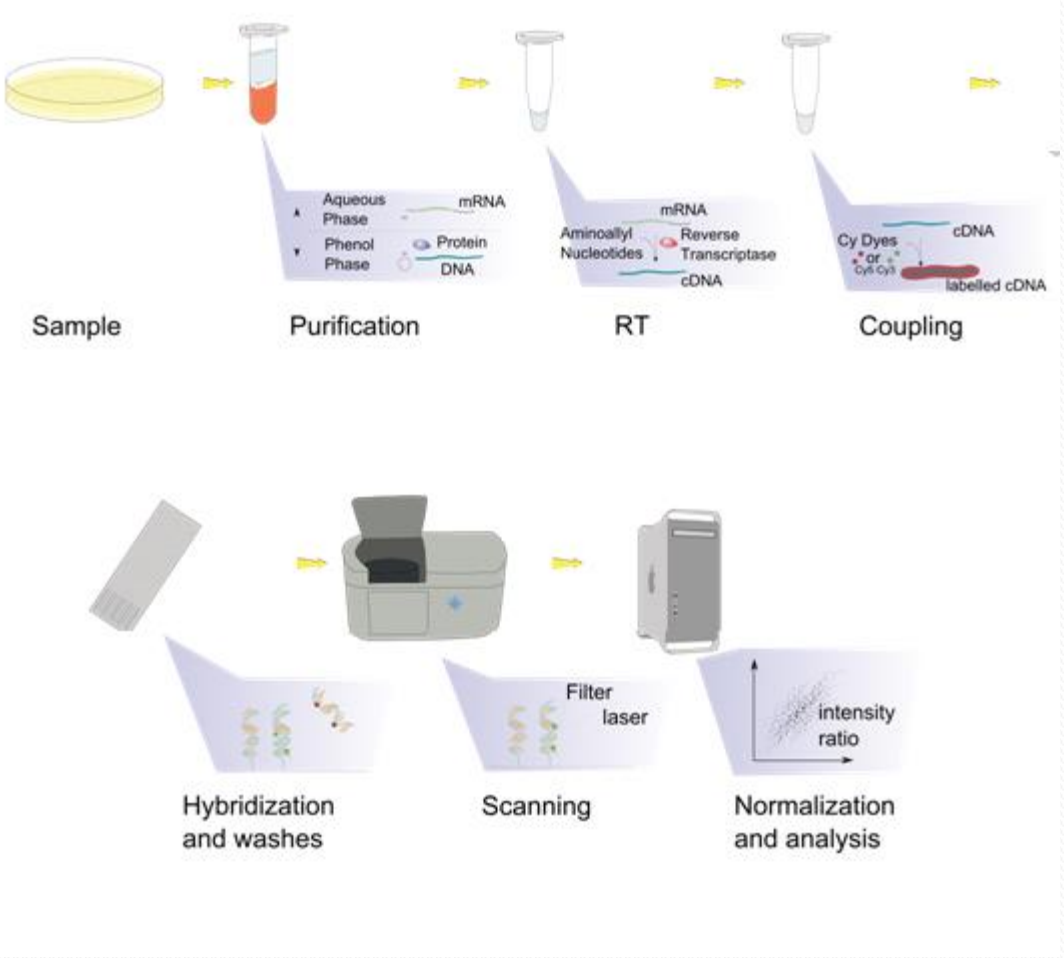
Process: Use DNA Microarrays

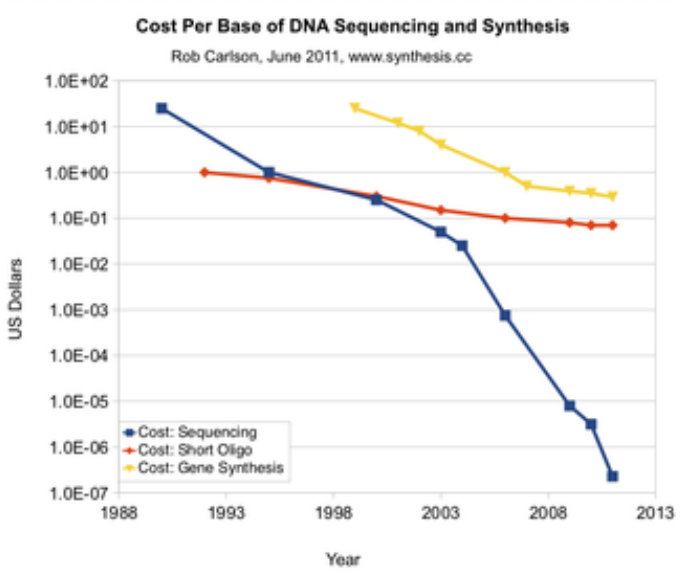
- MicroArrays contained probes for 6817 human genes
- RNA prepared from cells was hybridized to high-density oligonucleotide MA
- Samples were subjected to a priori quality control standards regarding the amount of labeled RNA and the quality of the scanned microarray image.

About DNA Microarrays

- Also known as DNA chip or biochip
- Collection of microscopic DNA spots attached to a solid surface.
- Used to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome.

DNA MicroArrays





Issue-1: Are there strong correlations?

Issue-1: Are there genes whose expression pattern are strongly correlated with the class distinction to be predicted?

- **Use Neighborhood Analysis**
  - **Objective:** To establish whether the observed correlations were stronger than would be expected by chance
  - Defines an "idealized expression pattern" corresponding to a gene that is uniformly high in one class and uniformly low in the other
  - Tests whether there is an unusually high density of genes "nearby" (or similar to) this idealized pattern, as compared to equivalent random patterns.
- **Why do we want to start with informative genes?**
  - To be readily applied in a clinical setting
  - Highly instructive

Neighborhood Analysis

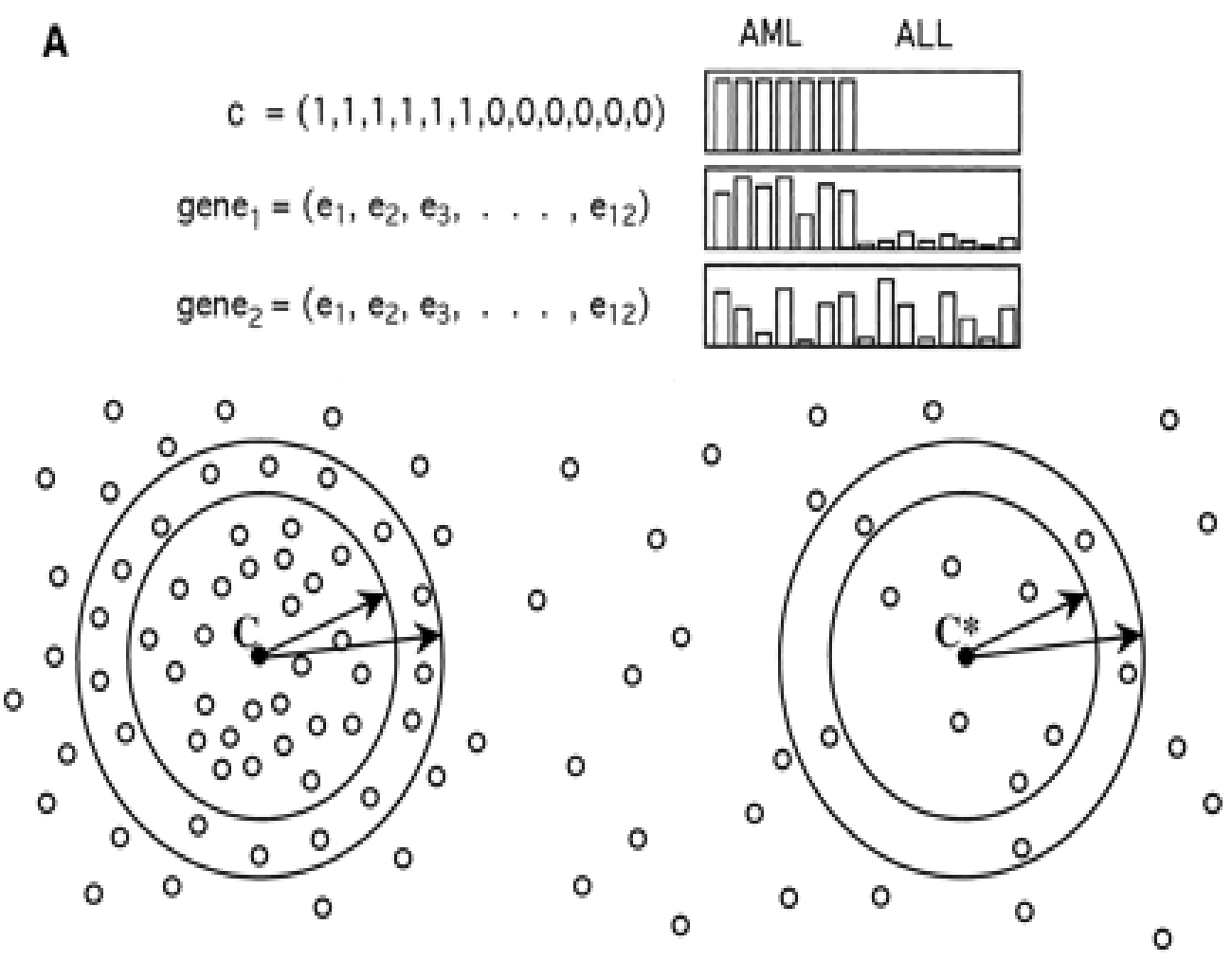
1. $v(g) = (e_1, e_2, \ldots, e_n)$
2. $c = (c_1, c_2, \ldots, c_n)$
3. Compute the correlation between v(g) and c.
    1. *Euclidean distance*
    2. *Pearson correlation coefficient.*
    3. $P(g,c) = [\mu_1(g) - \mu_2(g)]/[\sigma_1(g) + \sigma_2(g)]$

V(g) = expression vector, with e_i denoting the expression level of gene g in ith sample

C=vector of idealized expression pattern. c_i = +1 or 0 based on i-th sample belonging to class 1 or 2

P(g,c) = Measure of Signal-to-noise ratio

Neighborhood Analysis



Results of Neighborhood Analysis

- Neighborhood Analysis showed that roughly **1100 genes** of the 6,817 genes were more highly correlated with the AML-ALL class distinction than would be expected by chance
- Suggested that classification could indeed be based on expression data.

Results of Neighborhood Analysis

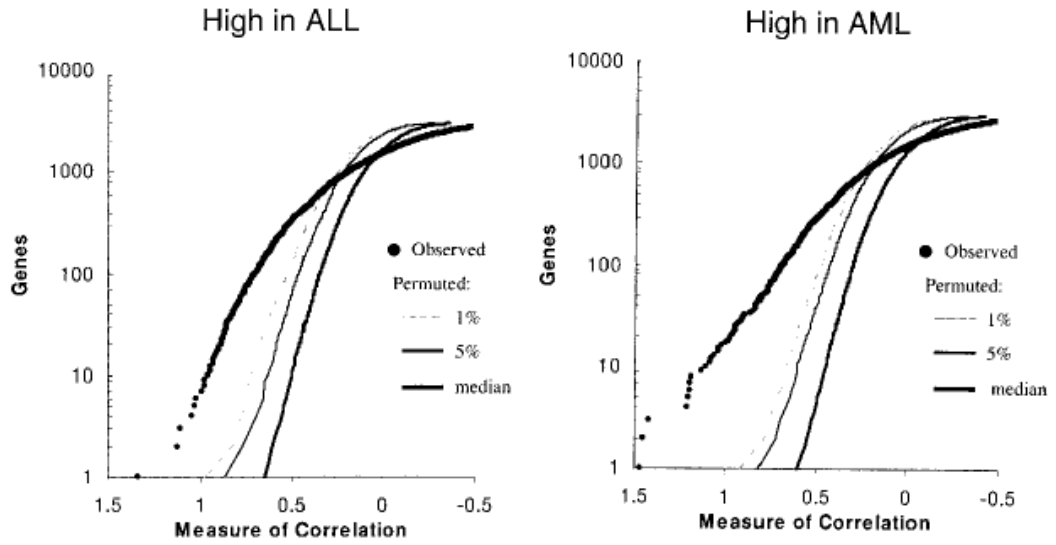


Fig. 2. Neighborhood analysis: ALL versus AML. For the 38 leukemia samples in the initial data set, the plot shows the number of genes within various "neighborhoods" of the ALL-AML class distinction together with curves showing the 5 and 1% significance levels for the number of genes within corresponding neighborhoods of the randomly permuted class distinctions (*16*, *17*). Genes more highly expressed in ALL compared to AML are shown in the left panel; those more highly expressed in AML compared to ALL are shown in the right panel. The large number of genes highly correlated with the class distinction is apparent. In the left panel (higher in ALL), the number of genes with correlation $P(g,c) > 0.30$ was 709 for the AML-ALL distinction, but had a median of 173 genes for random class distinctions. $P(g,c) = 0.30$ is the point where the observed data intersect the 1% significance level, meaning that 1% of random neighborhoods contain as many points as the observed neighborhood around the AML-ALL distinction. Similarly, in the right panel (higher in AML), 711 genes with $P(g,c) > 0.28$ were observed, whereas a median of 136 genes is expected for random class distinctions.

Issue-2: Building a Predictor

Issue-2: How do we use a collection of known samples to create a "class predictor" capable of assigning a new sample to one of two classes?

- Use a set of informative genes to build the predictor
- They chose 50 genes most closely correlated with AML-ALL distinction in the known samples.
  - Why 50? Why not 20 or 100?
  - Predictors with 10 to 200 genes all gave 100% accurate classification
  - 50 seemed like a reasonably robust against noise but small enough to be readily applied in a clinical setting

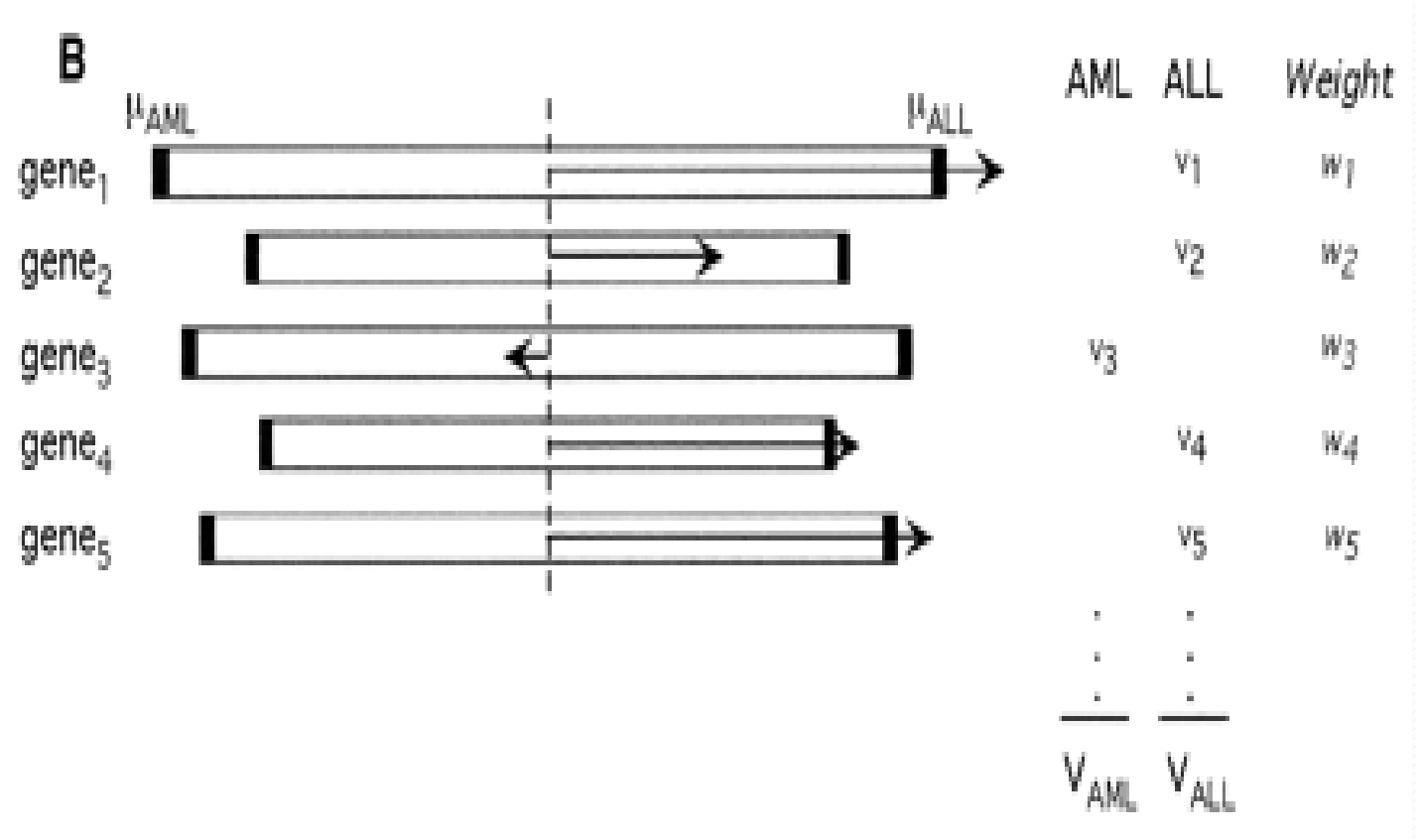
Class Predictor via Gene Voting

- Developed a procedure that uses a fixed subset of “informative genes”
- Makes a prediction on basis of the expression level of these genes in a new sample
- Each informative gene casts a “weighted vote” for one of the classes
- The magnitude of each vote dependent on the expression level in the new sample and the degree of that gene's correlation with the class distinction
- Votes were summed to determine the winning class
- “Prediction Strength” (PS), a measure of the margin of victory that ranges from 0 to 1
- The sample was assigned to the winning class if PS exceeded a predetermined threshold, and was otherwise considered uncertain.

Class Predictor via Gene Voting

1. Parameters (a_g, b_g) are defined for each informative gene
2. $a_g = P(g,c)$
3. $b_g = [\mu_1(g) + \mu_2(g)]/2$
4. $v_g = a_g(x_g - b_g)$
5. $V_1 = \sum | V_g |$; for $V_g > 0$
6. $V_2 = \sum | V_g |$; for $V_g < 0$
7. $PS = (V_{win} - V_{lose})/(V_{win} + V_{lose})$
8. The sample was assigned to the winning class for PS > threshold.

Class Predictor via Gene Voting



Issue-3: Validation of Class Predictors

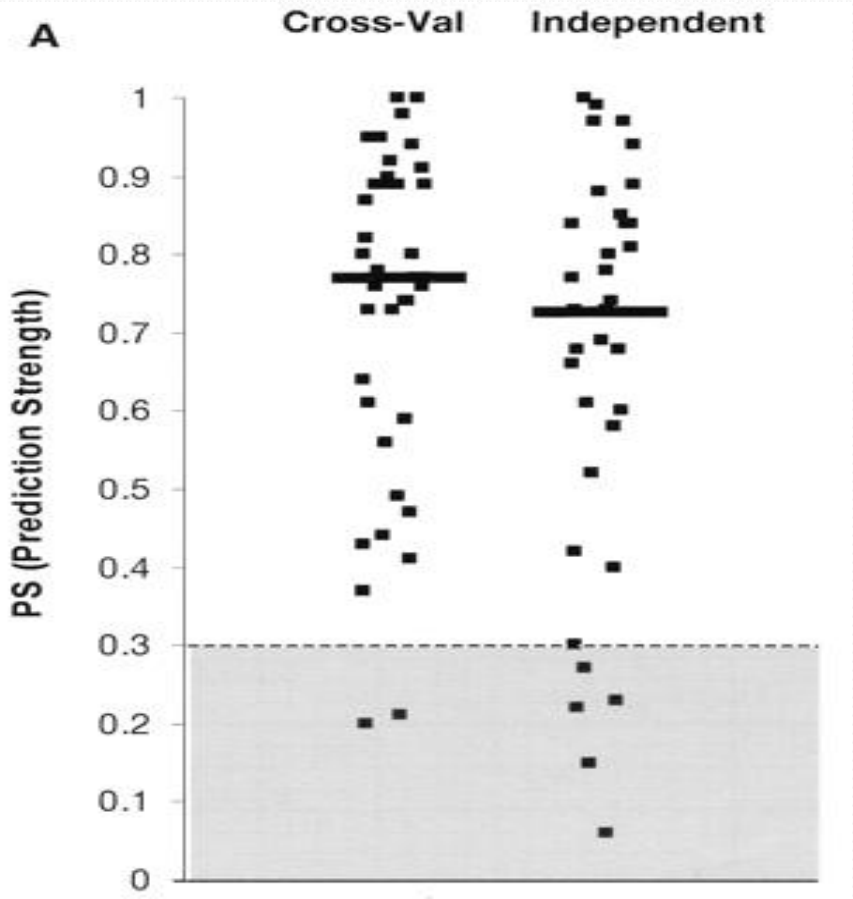
Issue-3: How do we test the validity of the class predictors?

- **Two-step validation:**
  - Cross-Validation (Leave-one-out)
  - Independent Sample Validation

Results of Validation of Class Predictors

- **Initial Samples:**
  - **36 of the 38 samples** as either AML or ALL and two as uncertain
  - All 36 samples agree with clinical diagnosis
- **Independent Samples:**
  - **29 of 34 samples** are strongly predicted with 100% accuracy
  - Average PS was lower for samples from one lab that used a different protocol
  - Should standardize of sample preparation in clinical implementation.

Validation of Class Predictors

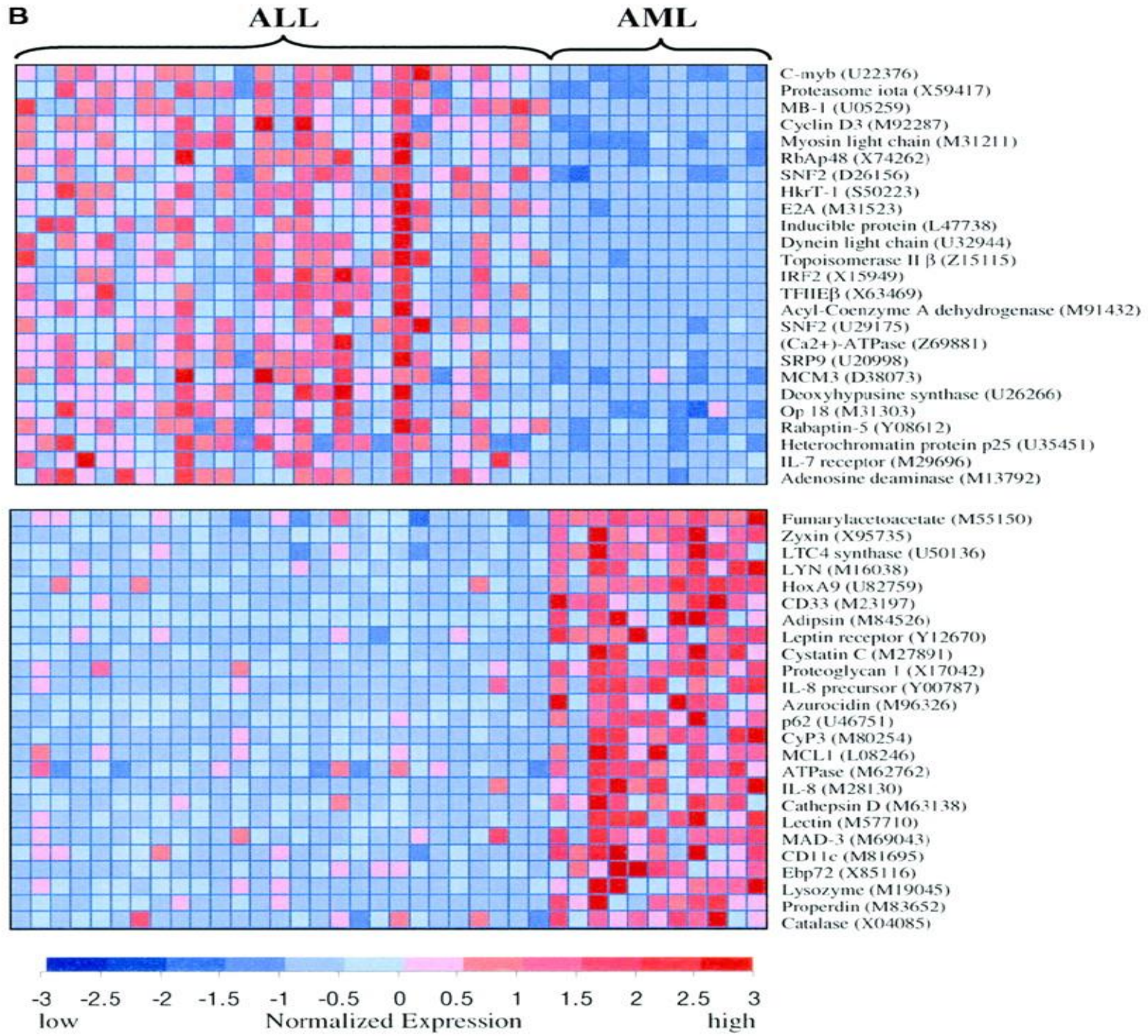


Prediction Strengths were quite high:

- Median PS = 0.77 in cross-validation
- Media PS = 0.73 in independent test

A Look at the Set of 50 Genes

- The list of informative genes used in the predictor was highly instructive
- Some genes, including CD11c, CD33, and MB-1, encode cell surface proteins useful in distinguishing lymphoid from myeloid lineage cells.
- Others provide new markers of acute leukemia subtype. For example, the leptin receptor, originally identified through its role in weight regulation, showed high relative expression in AML.
- Together, these data suggest that genes useful for cancer class prediction **may also provide insight** into cancer pathogenesis and pharmacology.


B
ALL
AML
C-myb (U22376)
Proteasome iota (X59417)
MB-1 (U05259)
Cyclin D3 (M92287)
Myosin light chain (M31211)
RbAp48 (X74262)
SNF2 (D26156)
HkrT-1 (S50223)
E2A (M31523)
Inducible protein (L47738)
Dynein light chain (U32944)
Topoisomerase II β (Z15115)
IRF2 (X15949)
TFIIEβ (X63469)
Acyl-Coenzyme A dehydrogenase (M91432)
SNF2 (U29175)
(Ca2+)-ATPase (Z69881)
SRP9 (U20998)
MCM3 (D38073)
Deoxyhypusine synthase (U26266)
Op 18 (M31303)
Rabaptin-5 (Y08612)
Heterochromatin protein p25 (U35451)
IL-7 receptor (M29696)
Adenosine deaminase (M13792)
Fumarylacetoacetate (M55150)
Zyxin (X95735)
LTC4 synthase (U50136)
LYN (M16038)
HoxA9 (U82759)
CD33 (M23197)
Adipsin (M84526)
Leptin receptor (Y12670)
Cystatin C (M27891)
Proteoglycan 1 (X17042)
IL-8 precursor (Y00787)
Azurocidin (M96326)
p62 (U46751)
CyP3 (M80254)
MCL1 (L08246)
ATPase (M62762)
IL-8 (M28130)
Cathepsin D (M63138)
Lectin (M57710)
MAD-3 (M69043)
CD11c (M81695)
Ebp72 (X85116)
Lysozyme (M19045)
Properdin (M83652)
Catalase (X04085)
-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3
low
Normalized Expression
high

When Does This Methodology Work Best?

- Can be applied to any measurable distinction among tumors
- Importantly, such distinctions could concern a **future clinical outcome**
- Ability to predict response to chemotherapy:
  - Among the 15 adult AML patients who had been treated and for whom long-term clinical follow-up was available.
  - No evidence of a strong multigene expression signature was correlated with clinical outcome (This could reflect the relatively small sample size).
  - single most highly correlated gene out of the 6817 genes was the homeobox gene HOXA9, which was over-expressed in patients with treatment failure
  - Further clinical trials needed to test the hypothesis that HOXA9 expression plays a role in predicting AML outcome.

Class Discovery

- If the AML-ALL distinction was not already known, could it have been discovered simply based on gene expression?
- Issues in Class Discovery:
  - Cluster tumors based on Gene Expression
  - Determining whether putative classes produced are meaningful i.e. whether they reflect true structure in the data rather than simply random aggregation.

Class Discovery

- Clustering for class discovery (Unsupervised)
- Self-organizing maps (SOMs) technique:
  - User specifies the number of clusters to be identified.
  - SOM finds an optimal set of "centroids" around which the data points appear to aggregate.
  - It then partitions the data set, with each centroid defining a cluster consisting of the data points nearest to it.

Video on Clustering

K-Means Clustering:
https://www.youtube.com/watch?v=_aWzGGNrcic

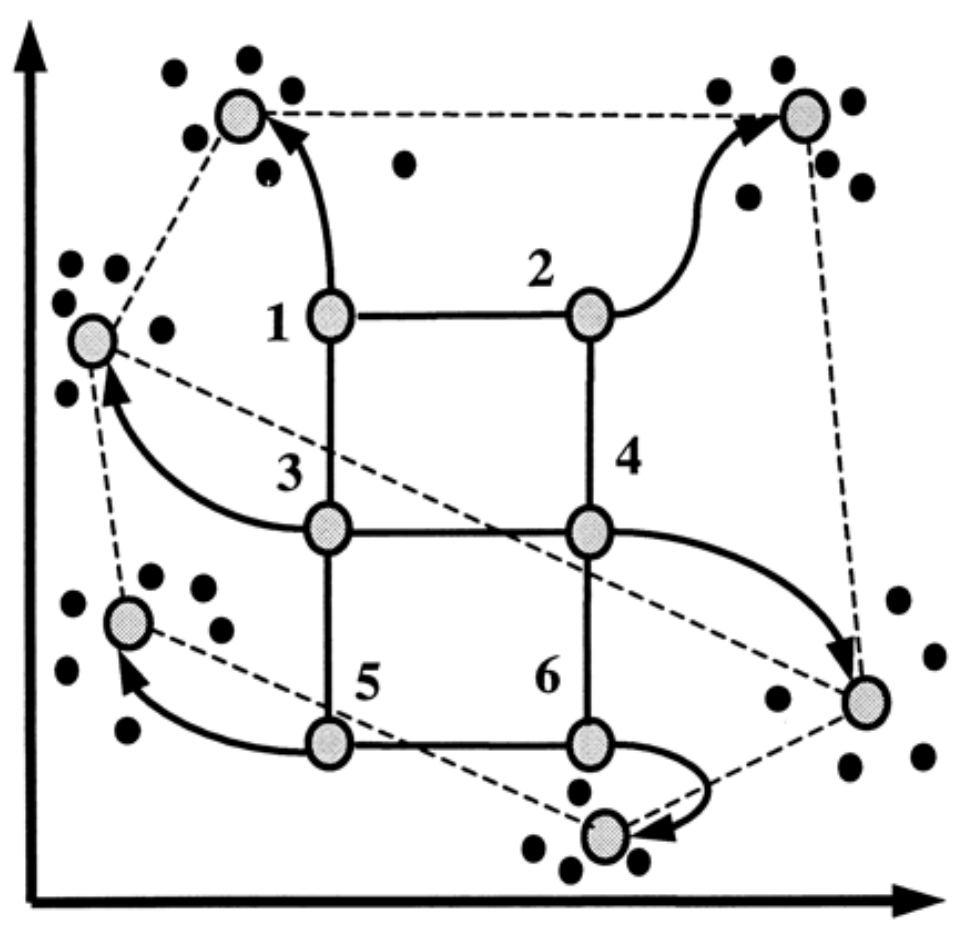
SOM:

https://www.youtube.com/watch?v=H9H6s-x-0YE

Self Organizing Map (SOM)

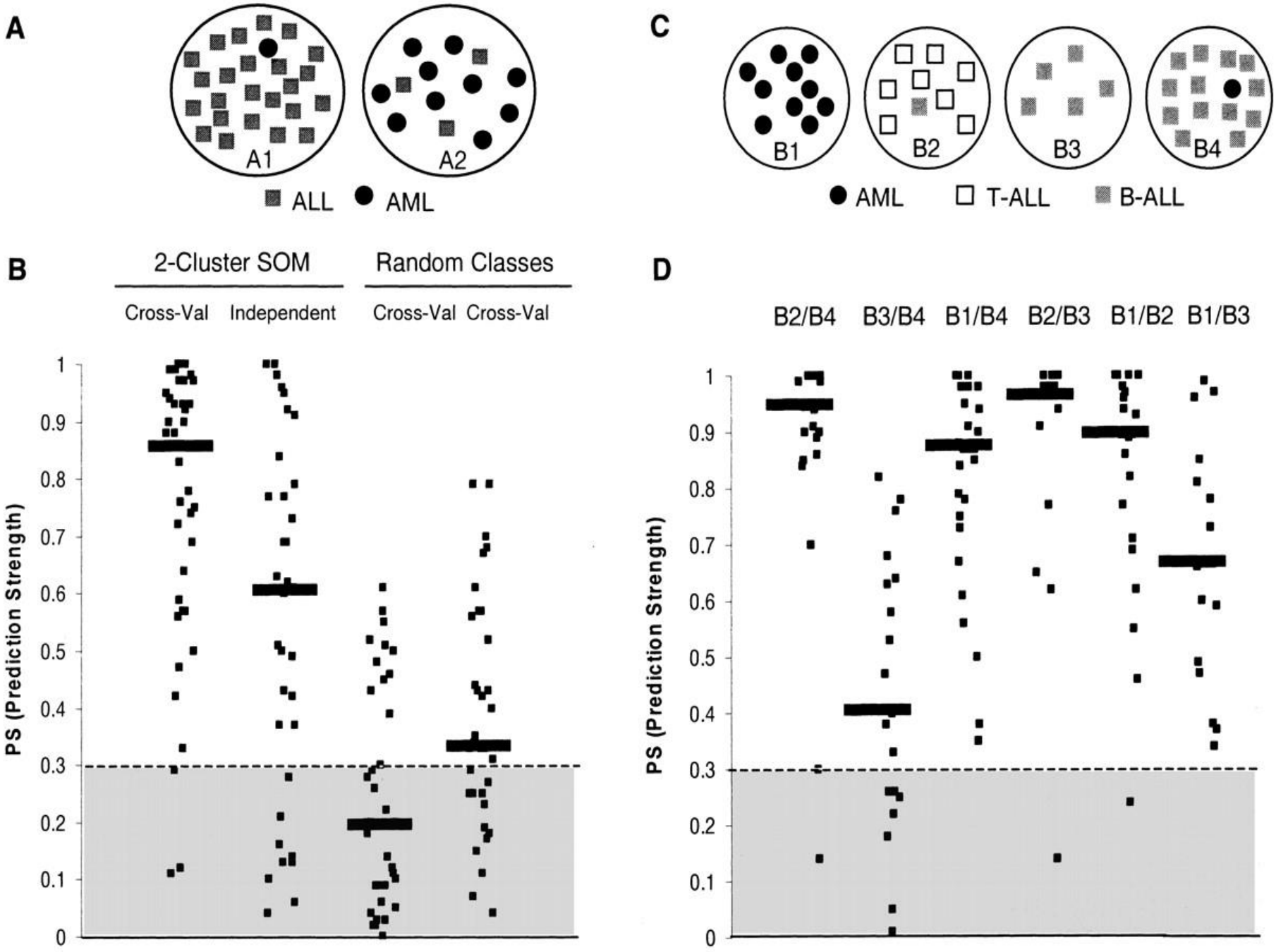
- SOM is a mathematical cluster analysis for recognizing and classifying features in complex, multidimensional data (similar to K-mean approach)
  - Chooses a geometry of “nodes”
  - Nodes are mapped into K-dimensional space, initially at random
  - Iteratively adjust the nodes
- Adjusting the Nodes:
  - Randomly select a data point P
  - Move the nodes in the direction of P
  - The closest node N_p is moved the most
  - Other nodes are moved depending on their distance from N_p in the initial geometry

Self Organizing Map (SOM)



Results of Two-Cluster Analysis

- Two-cluster SOM was applied to automatically group the 38 initial leukemia samples into two classes on the basis of the expression pattern of all 6817 genes.
- Clusters were evaluated by comparing them to the known AML-ALL classes
  - Class A1 contained mostly ALL (24 of 25 samples)
  - Class A2 contained mostly AML (10 of 13 samples)
  - SOM was thus quite effective at automatically discovering the two types of leukemia.


A
A1
A2
ALL
AML
C
B1
B2
B3
B4
AML
T-ALL
B-ALL
B
2-Cluster SOM
Random Classes
Cross-Val
Independent
Cross-Val
Cross-Val
PS (Prediction Strength)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
D
B2/B4
B3/B4
B1/B4
B2/B3
B1/B2
B1/B3
PS (Prediction Strength)

Discovering New Classes

Issue: How do we evaluate such putative clusters if the "right" answer were not already known?

- **Idea:** Class discovery can be tested using Class Prediction
- **Intuition:** If putative classes reflect true structure, then a class predictor based on these classes should perform well.
- **Discussion:** Is this Reasonable? Is it possible that the putative classes perform well even if they do not reflect true structure?

Process & Results (Two Cluster)

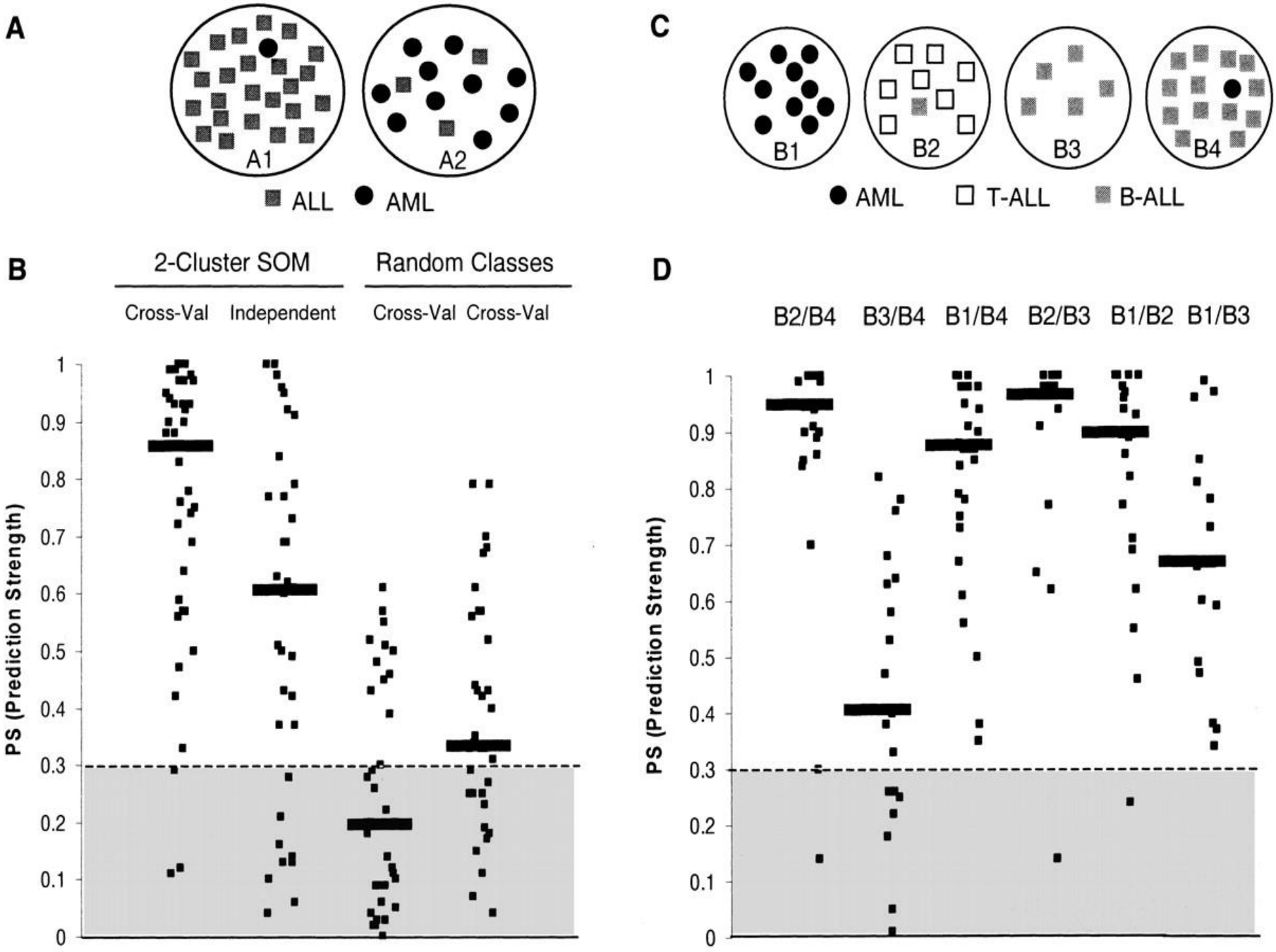
- **Clusters A1 and A2 were evaluated:**
  - Constructed predictors to assign new samples as “Type A1" or “Type A2“
- **Cross-Validation:**
  - Predictors that used a wide range of different numbers of informative genes performed well
  - Cross-validation thus not only showed high accuracy, but actually refined the SOM-defined classes except for the subset of samples accurately classified
  - Similar analysis on random clusters yielded predictors with poor accuracy in cross-validation

Process & Results (Two Cluster)

- **Independent Set Validation:**
  - Median PS was 0.61, and 74% of samples were above threshold
  - High PS indicates that the structure seen in the initial data set is also seen in the independent data set
  - Predictors from random clusters consistently yielded low PS on independent data set
- **Conclusion:**
  - A1-A2 distinction can be seen to be meaningful, rather than simply a statistical artifact of the initial data set
  - Results show that the AML-ALL distinction could have been automatically discovered and confirmed without previous biological knowledge

Process & Results (Four Cluster)

- SOM divides the samples into four clusters
- Largely corresponded to AML, T-lineage ALL, B-lineage ALL & B-lineage ALL
- Four-cluster SOM thus divided the samples along another key biological distinction
- Evaluated classes by constructing class predictors. The four classes could be distinguished from one another, with the exception of B3 versus B4
- The prediction tests thus confirmed the distinctions corresponding to AML, B-ALL, and T-ALL
- Suggested that it may be appropriate to merge classes B3 and B4, composed primarily of B-lineage ALL


A
A1
A2
ALL
AML
C
B1
B2
B3
B4
AML
T-ALL
B-ALL
B
2-Cluster SOM
Random Classes
Cross-Val
Independent
Cross-Val
Cross-Val
PS (Prediction Strength)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
D
B2/B4
B3/B4
B1/B4
B2/B3
B1/B2
B1/B3
PS (Prediction Strength)

Conclusion

- Technique for creating class predictors
- These class predictors could be adapted to a clinical setting, with appropriate steps to standardize the protocol for sample preparation.
- Such a test supplementing rather than replacing existing leukemia diagnostics;
- Class predictors can be constructed for known pathological categories and provide diagnostic confirmation or clarify unusual cases.
- The technique of class prediction can be applied to distinctions relating to future clinical outcome, such as drug response or survival.
- Class prediction provides an unbiased, general approach to constructing such prognostic tests.

Conclusion

- In principle, the class discovery techniques discovered here can be used to identify fundamental subtypes of any cancer.
- In general, such studies will require careful experimental design to avoid potential experimental artifacts--especially in the case of solid tumors.
- Various approaches could be used to avoid such artifacts;
- Class discovery methods could also be used to search for fundamental mechanisms that cut across distinct types of cancers.

WOS Citation Report

WEB OF SCIENCE™ THOMSON REUTERS

Search | Return to Search Results | My Tools | Search History | Marked List

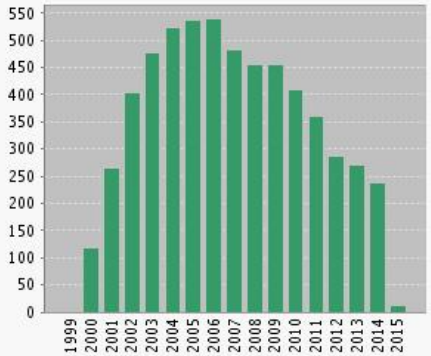
Citation Report: 5838

(from Web of Science Core Collection)

For: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. ...More

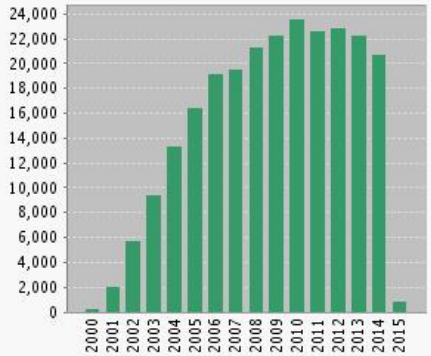
This report reflects citations to source items indexed within Web of Science Core Collection. Perform a Cited Reference Search to include citations to items not indexed within Web of Science Core Collection.

Published Items in Each Year



The latest 20 years are displayed.

Citations in Each Year



The latest 20 years are displayed.

Results found:	5838
Sum of the Times Cited [?] :	242977
Sum of Times Cited without self-citations [?] :	211502
Citing Articles [?] :	139721
Citing Articles without self-citations [?] :	134486
Average Citations per Item [?] :	41.62
h-index [?] :	201

Mayo 50 Oncogene Panel