Golub '99: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.

- Identification of cancer subtypes is important for proper treatment
  - Acute myeloid leukemia (AML) vs. acute lymphoblastic leukemia (ALL)
- Classification used to be based primarily on morphological appearance of the tumor.
  - But tumors with similar histopathological appearance can follow different clinical courses and different response to therapy

Cancer Classification Challenges

- Gene discovery: identification of genes that differ from one tumor class to another
- Class prediction: assigning tumors to known classes.
- Class discovery: identification of new cancer classes.

Classification of Acute Leukemias

- Classification of acute leukemia
  - 1940s: Observation of subtle variability in clinical outcome
  - Subtle differences in clinical outcomes
  - 1960s: Some leukemias were periodic acid-Schiff positive (staining to detect glycogen, glycoproteins, glycolipids) → Lymphoid
  - Others were myeloperoxidase positive → Myeloid (bone marrow)
  - 1970s: Classification further validated by antibodies recognizing lymphoid and myeloid cell surface receptors.
- 1990s: Further subclassification
  - t(12;21)(p13;q22) chromosomal translocation occurs in 25% of patients with ALL
  - t(8;21)(q22;q22) occurs in 15% of patients with AML
- Classification ALL vs. AML well established
  - but no single test sufficient to establish diagnosis
  - requires expert analysis from specialized lab tests
  - Error prone

DNA microarrays as tool for cancer classification

- Previous microarrays focused on cell culture rather than primary patient samples
- Previous study by same researchers
  - Normal kidney vs. renal cell carcinoma
  - Morphological distinction is easier for that problem

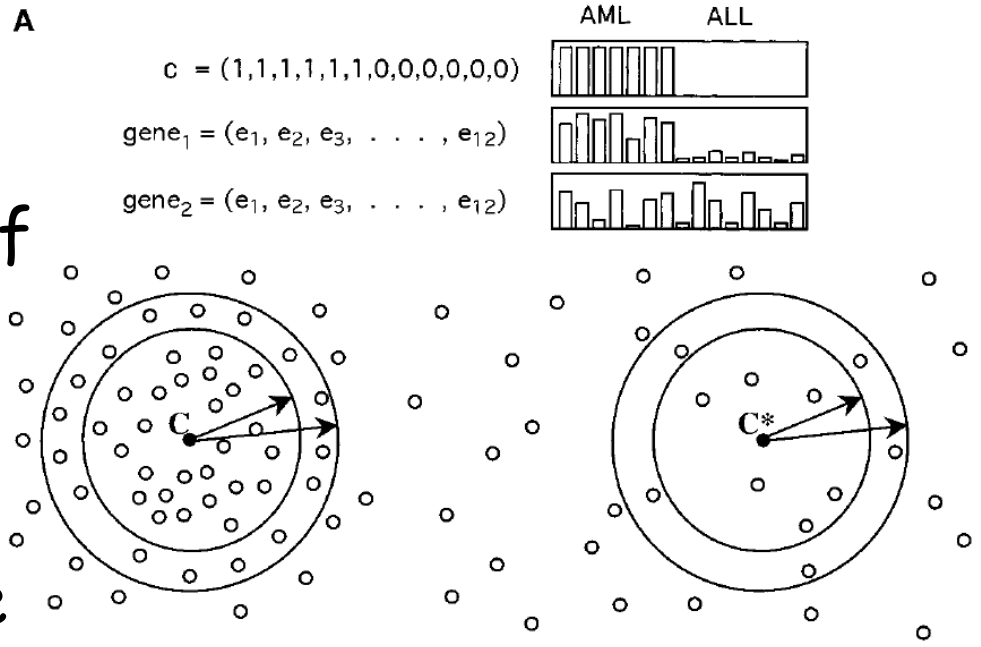
Data collection

- 38 bone marrow samples
  - 27 ALL, 11 AML
- Affymetrix chip containing probes for 6817 human genes

Task1: Gene discovery

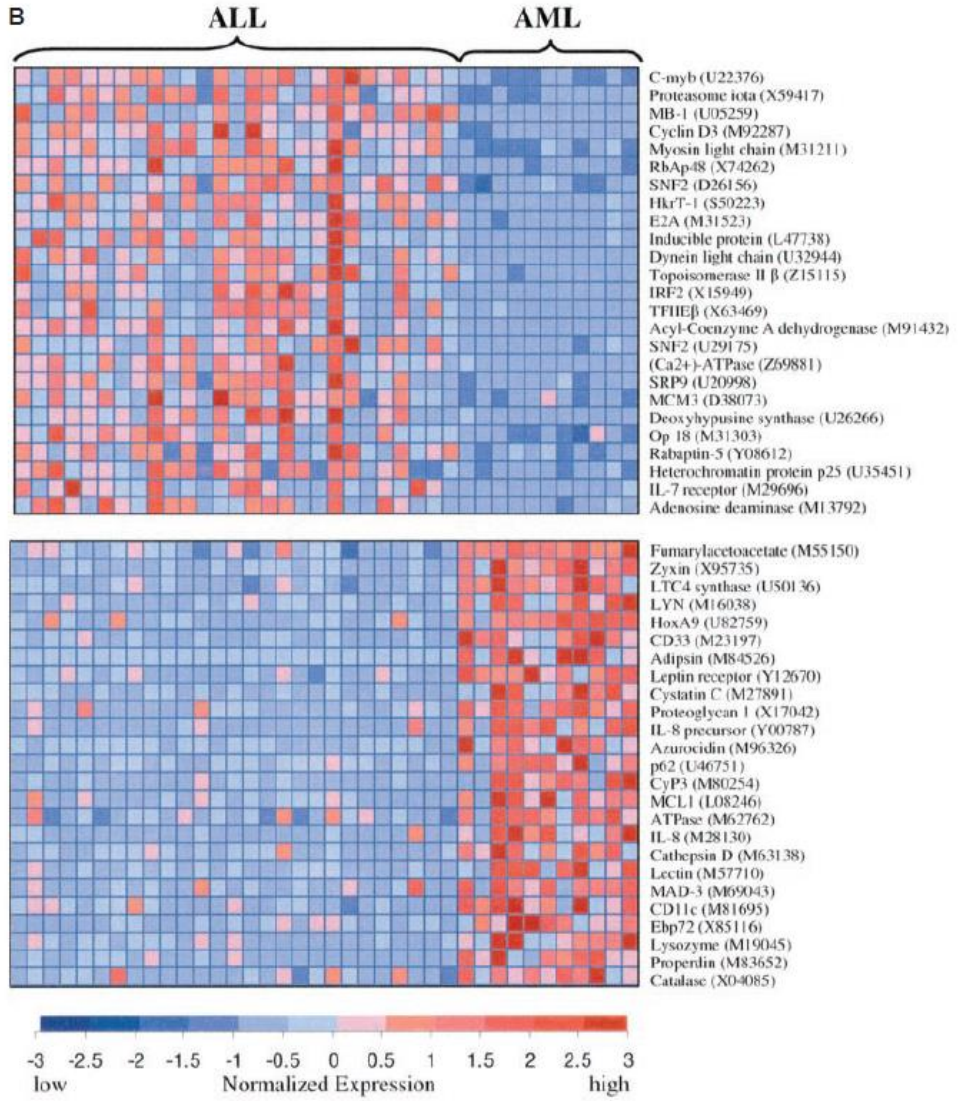
- Are there genes with expression patterns strongly correlated with class distinction?

- "Neighborhood Analysis"
- Analyze distribution of correlations with "idealized expression pattern" c.
- Compare with distribution by chance (shuffle samples in c)



Task1: Gene discovery

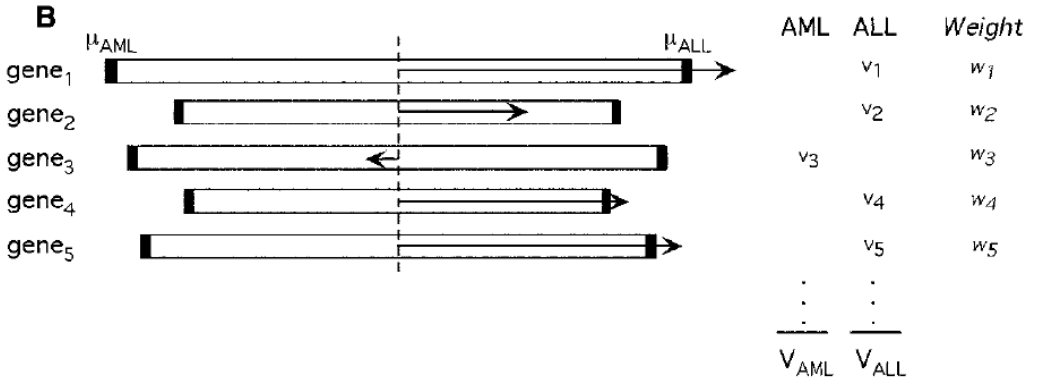
- 50 genes most correlated with class distinction



Task2: Class Prediction

- Use 50 most informative genes
  - genes with highest correlation to class.
  - Other number of genes gave similar results.
- Leave-one-out cross-validation
- For a new sample, each gene casts a "weighted vote"
  - Weight depends on expression level in new sample and degree of correlation of that gene with class distinction.

- "Neigbhorhood Analysis"
- Analyze distribution of correlations with "idealized expression pattern" c.
- Compare with distribution by chance (shuffle samples in c)



Task2: Class Prediction

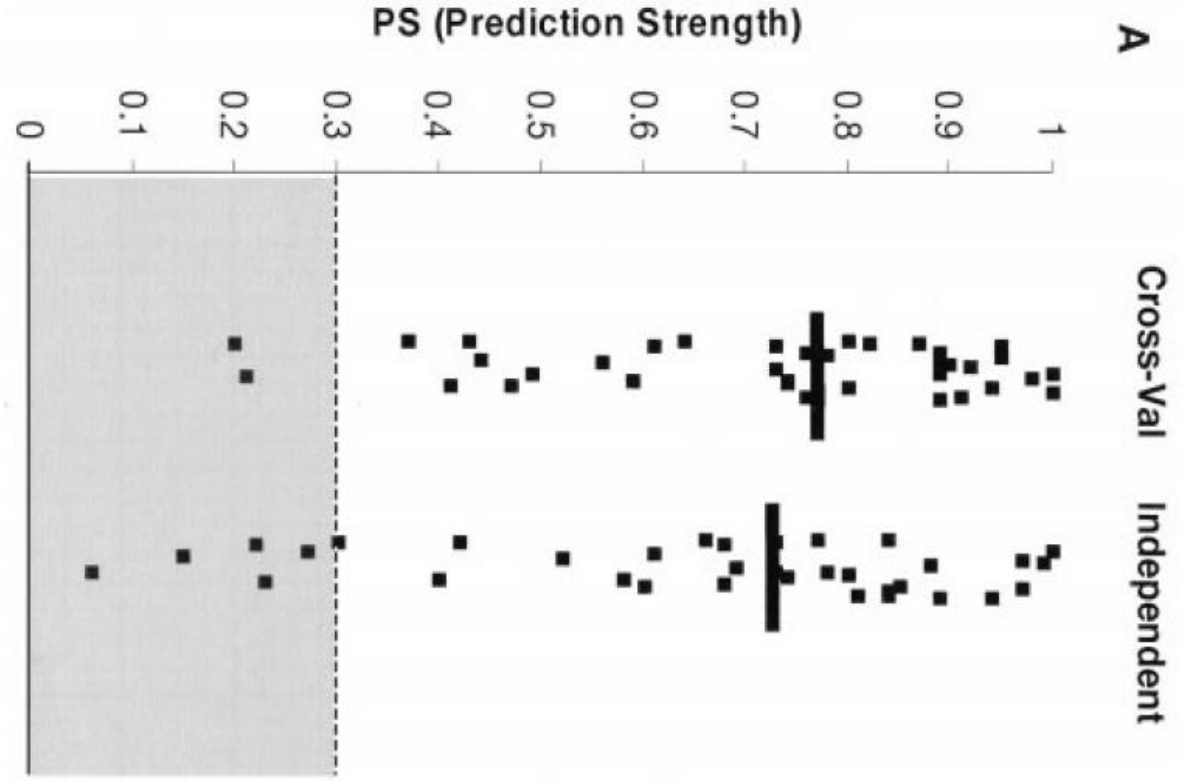


Fig. 3. (**A**) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (**B**) Genes

- "Prediction Strength": confidence in prediction
- Class labels not depicted here
- Median PS: 0.77 for cross-validation and 0.73 in independent

Task3: Class Discovery