

Analysis of Gene Expression Data

Jianlin Cheng, PhD

Department of Computer Science
Informatics Institute



2011

Outline

- Introduction to gene expression and DNA microarray
- Data normalization
- Analysis of differential gene expression
- Clustering of gene expression data
- Classification of gene expression data

The Dramatic Consequences of Gene Regulation in Biology



Anise swallowtail, *Papilio zelicaon*

- Same genome** →
Different tissues
- Different physiology
 - Different proteome
 - Different expression pattern

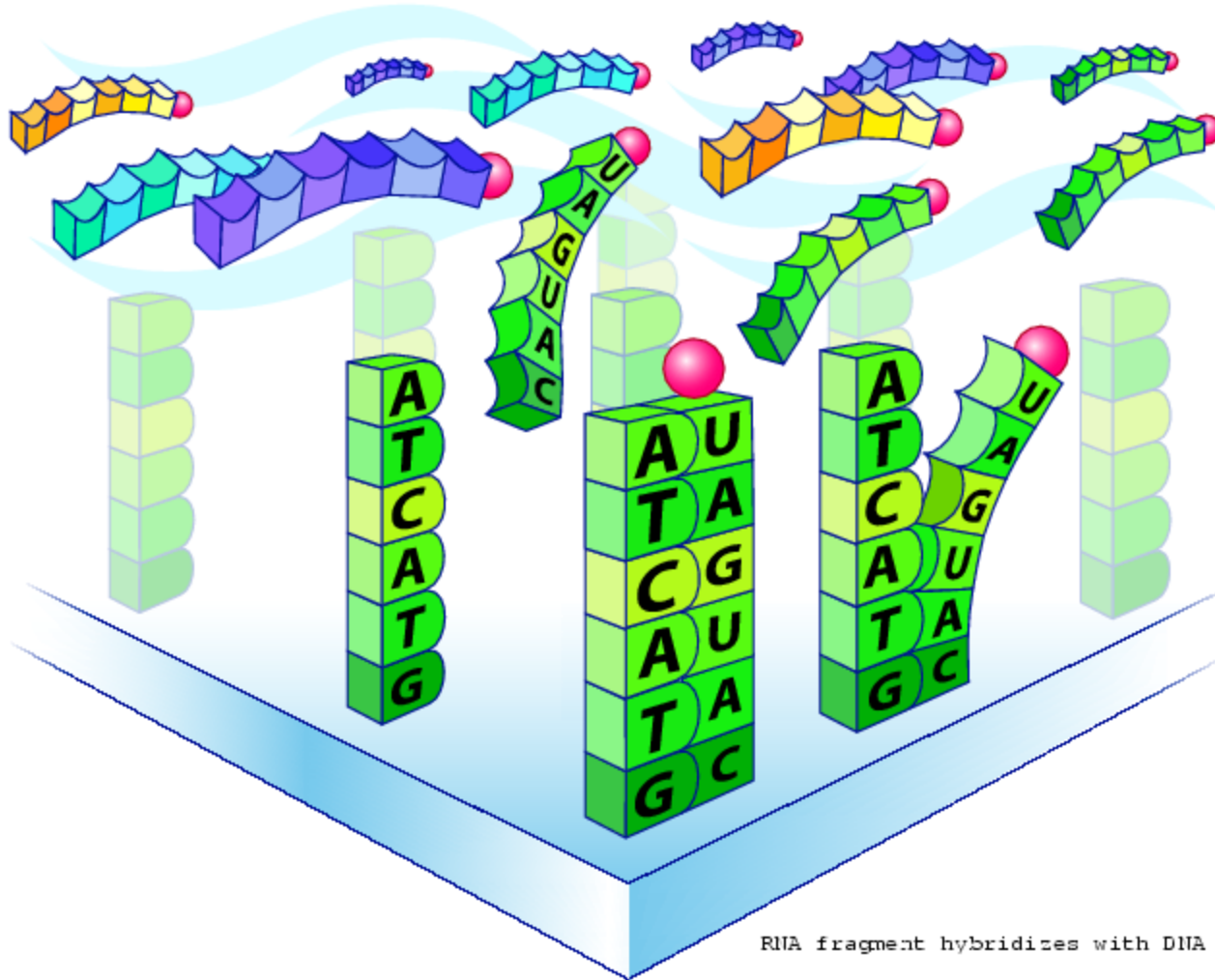


Gene Expression Measurement

- mRNA expression represents dynamic aspects of cell
- mRNA expression can be measured by DNA Microarrays
- mRNA is isolated and labeled with fluorescent protein
- mRNA is hybridized to the target; level of hybridization corresponds to light emission which is measured with a laser
- DNA Microarray can measure the expression of thousands of genes at the same time (high throughput)

GeneChip® Hybridization

RNA fragments with fluorescent tags from sample to be tested

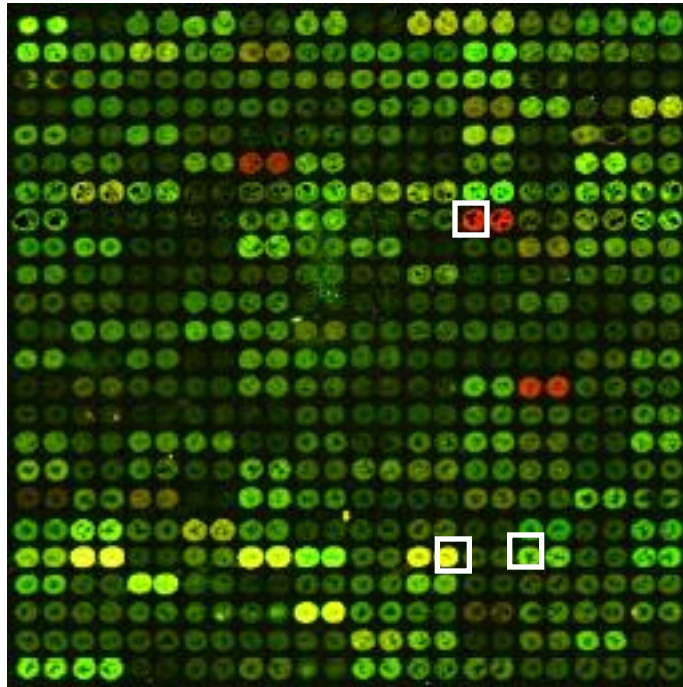




RNA fragment hybridizes with DNA on GeneChip

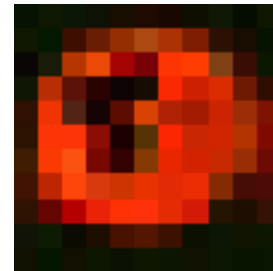
Image courtesy of Affymetrix.

Rainer Breitling, 2005

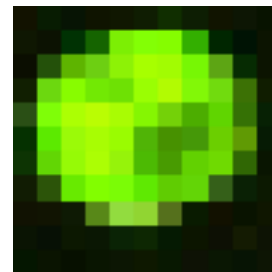
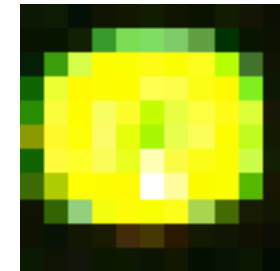
Microarray Images -> Differential Expression



 Reference cDNA
 Experimental cDNA

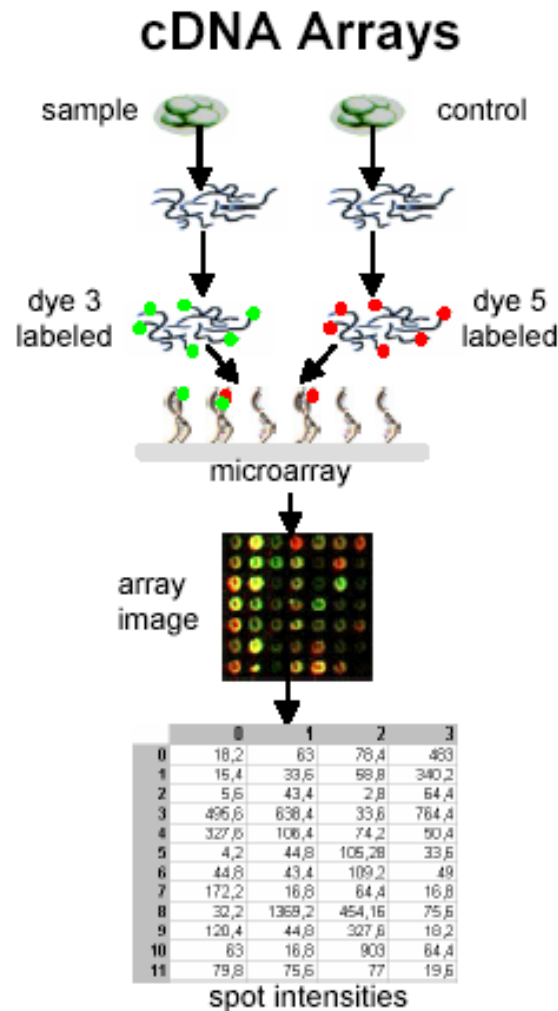


Upregulated

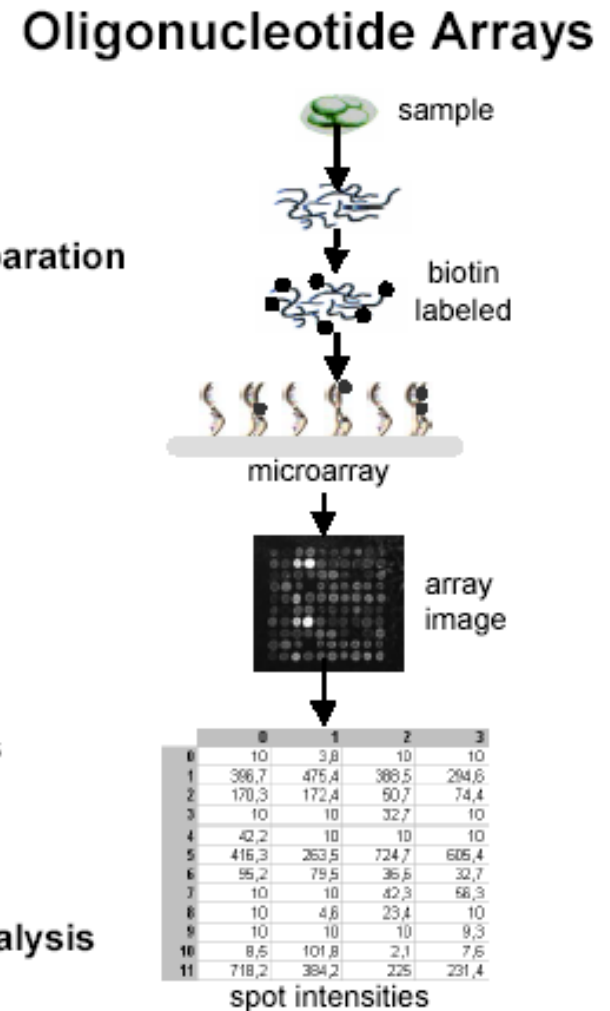


Downregulated

Microarray Experiment



- (1) Cell selection
- (2) RNA/DNA preparation
- (3) Hybridization
- (4) Array scan
- (5) Image analysis
- (6) Expression analysis



Data Extraction

One Color

- Calculate ratio of red to green fluorescence
- Convert to \log_2 and round to integer

Two-Color

- Calculate $\log R$ and $\log G$.

Microarray Data Example

Time Points

Genes

		1	2	3
		$\log_2.t_0$	$\log_2.t_{0.5}$	$\log_2.t_2$
1		-0.40	-0.91	-1.60
2		-0.99	-0.07	-0.83
3		-0.22	-0.49	-0.28
4		-0.31	-0.01	-0.09
5		-0.48	1.31	0.36
6		-0.38	0.35	0.60
7		-0.41	-0.49	-0.54
8		-0.46	-2.72	-3.16
9		-0.15	0.06	0.13
10		0.12	-0.67	-0.77
11		-0.03	-1.87	-2.58
12		0.31	0.02	-1.64
13		-0.06	-0.22	0.17
14		-0.03	-0.23	0.02
15		-0.12	0.11	-0.01
16		-0.21	-0.66	-0.30
17		-0.40	1.66	1.13
18		-0.58	0.25	0.72
19		-0.77	-0.05	1.11
20		-0.28	0.43	-0.57

Data Mining Challenges

- Too few experiments (samples), usually < 100
- Too many rows (genes), usually $> 1,000$
- Model needs to be explainable to biologists

Four Main Problems

1. Data pre-processing (normalization)
2. Identify differentially expressed genes in normal and non-normal situations.
3. Clustering genes according to expression data
4. Use gene expression data to classify samples (e.g., diagnosis of cancer)

Outline

- Introduction to gene expression and DNA microarray
- **Data normalization**
- Analysis of differential gene expression
- Clustering
- Classification
- Databases and software

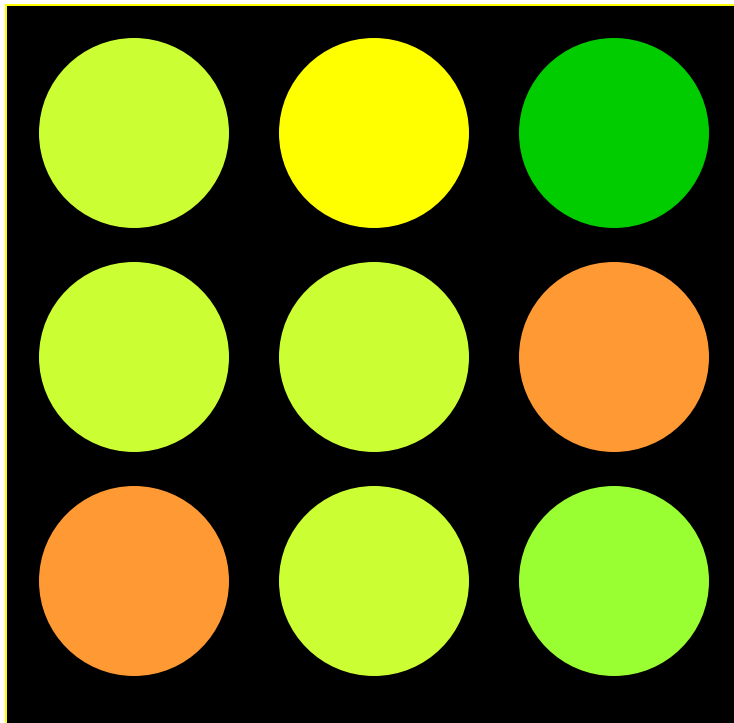
Microarray data analysis: normalization

The main goal of data preprocessing is to remove the systematic bias in the data as completely as possible, while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription.

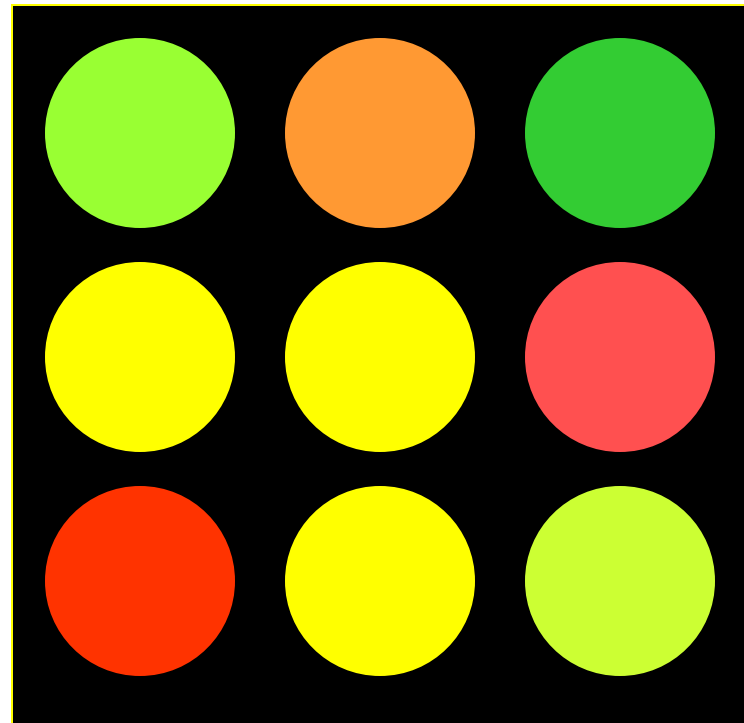
A basic assumption of most normalization procedures is that the average gene expression level does not change in an experiment.

Data normalization

Uncalibrated, red light under detected



Calibrated, red and green equally detected



Normalization: global

- Normalization based on a *global adjustment*


$$\log_2 R/G \rightarrow \log_2 R/G - c$$

- Common choices for $c =$ *median* or *mean* of log ratios for a particular gene set (e.g. all genes, or control or housekeeping genes)

Gene expression data example

Data on m genes for n samples

		mRNA samples					
		sample1	sample2	sample3	sample4	sample5	...
Genes	1	0.46	0.30	0.80	1.51	0.90	...
	2	-0.10	0.49	0.24	0.06	0.46	...
	3	0.15	0.74	0.04	0.10	0.20	...
	4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	5	-0.06	1.06	1.35	1.09	-1.09	...



Gene expression level of gene i in mRNA sample j

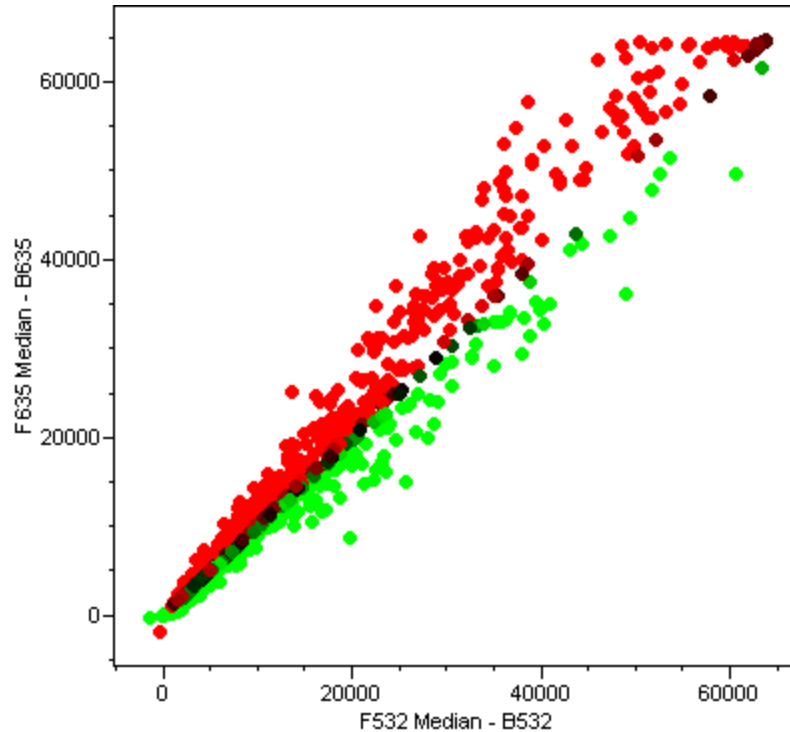
$$= (\text{normalized}) \text{Log}(\text{Red intensity} / \text{Green intensity})$$

Outline

- Introduction to gene expression and DNA microarray
- Data normalization
- **Analysis of differential gene expression**
- Clustering
- Classification
- Inference of gene regulatory networks
- Databases and software

Scatter plots

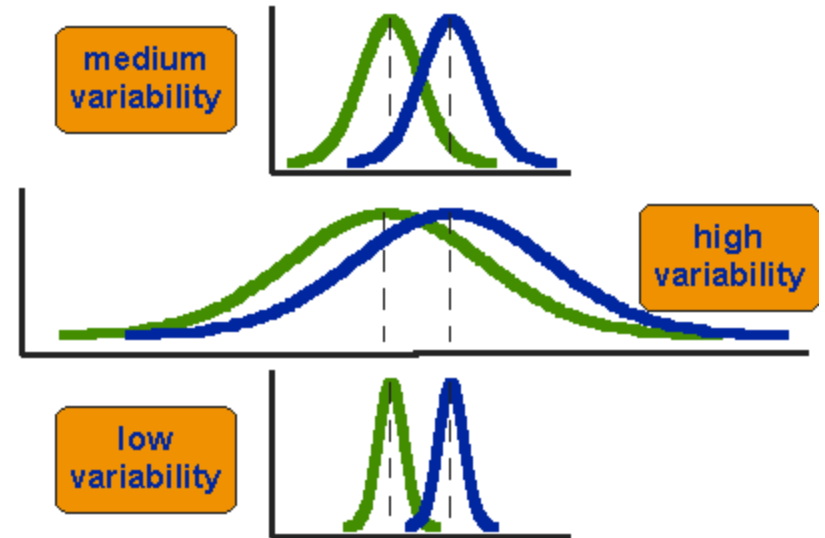
classical scatter plot



Differentially expressed genes are higher (or lower) in one of the samples

t-test = statistical significance of observed difference

- requires independent experimental replication
- assumes the data are identically normally distributed

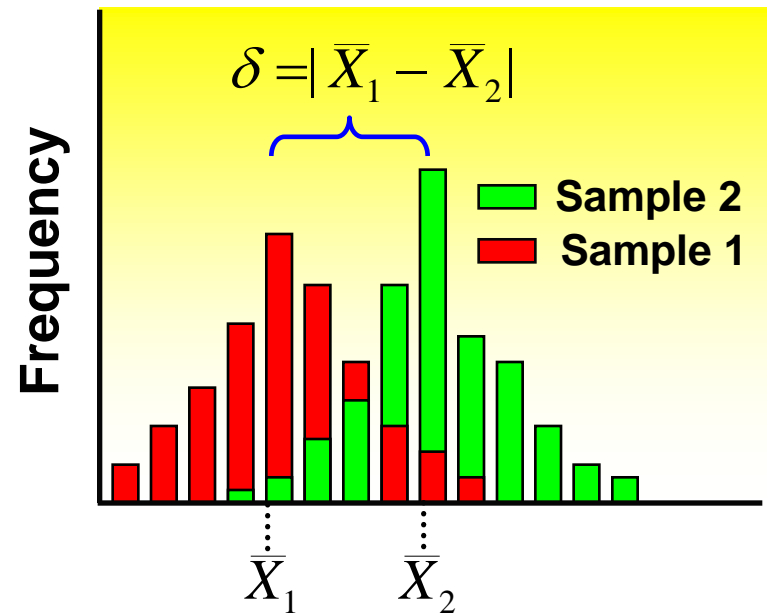


$$t = \frac{\text{difference of means}}{\text{variability}}$$

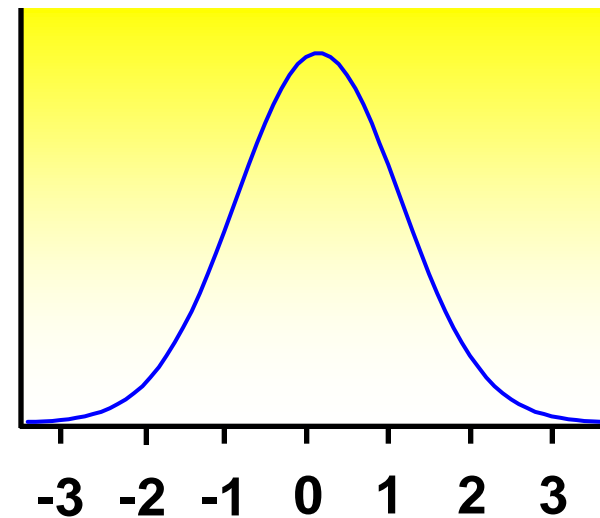
$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}}$$

Testing an intrinsic hypothesis

- Two samples with mean expression that differ by some amount δ .
- If $H_0 : \delta = 0$ is true, then the expected distribution of the test statistic t is



Probability



$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

The result of “differential expression” statistical analysis

Fold-Change	Gene Symbol	Gene Title
1	TNFAIP6	tumor necrosis factor, alpha-induced protein 6
2	THBS1	thrombospondin 1
3	SERPINE2	serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2
4	PTX3	pentaxin-related gene, rapidly induced by IL-1 beta
5	THBS1	thrombospondin 1
6	CXCL10	chemokine (C-X-C motif) ligand 10
7	CCL4	chemokine (C-C motif) ligand 4
8	SOD2	superoxide dismutase 2, mitochondrial
9	IL1B	interleukin 1, beta
10	CCL20	chemokine (C-C motif) ligand 20
11	CCL3	chemokine (C-C motif) ligand 3
12	SOD2	superoxide dismutase 2, mitochondrial
13	GCH1	GTP cyclohydrolase 1 (dopa-responsive dystonia)
14	IL8	interleukin 8
15	ICAM1	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor
16	SLC2A6	solute carrier family 2 (facilitated glucose transporter), member 6
17	BCL2A1	BCL2-related protein A1
18	TNFAIP2	tumor necrosis factor, alpha-induced protein 2
19	SERPINB2	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 2
20	MAFB	v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian)

Outline

- Introduction to gene expression and DNA microarray
- Data normalization
- Analysis of differential gene expression
- **Clustering**
- Classification
- Inference of gene regulatory networks
- Databases and software

Clustering Goals

- Identify gene classes / gene correlations / gene functions
- Support biological analysis / discovery (pathways, regulatory sites)
- Hierarchical clustering

Two Components of Clustering Algorithms

- Similarity / Distance Measures
- Clustering Methods

Similarity / Distance Measures

Pearson correlation

(looks for similarity in shape of the response profile, not the absolute values)

$$r = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

Euclidean distance

takes absolute expression level into account

$$d = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Manhattan (or city-block) **distance**

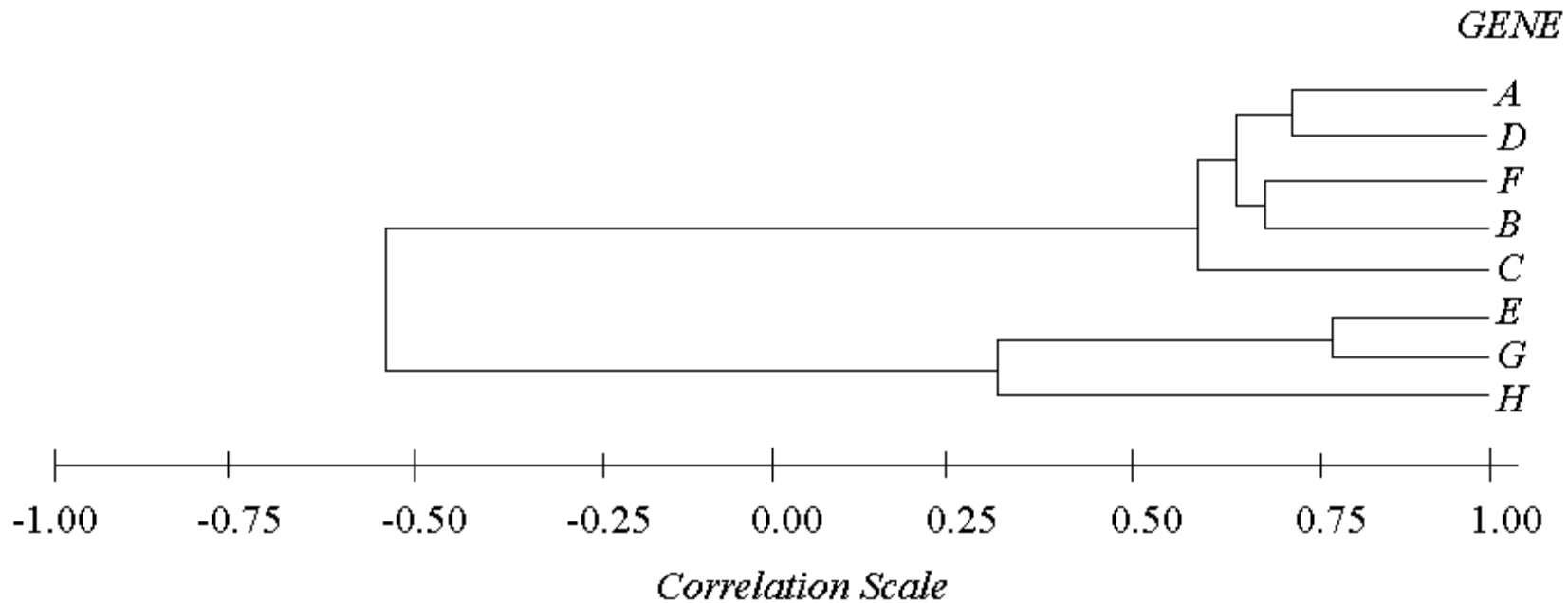
$$d = \sum_{i=1}^n |x_i - y_i|$$

Hierarchical Clustering

- The first algorithm used in gene expression data clustering (Eisen et al., 1998)
- Algorithm
 - Assign each data point into its own cluster (node)
 - Repeat
 - Select two closest clusters are joined. Replace them with a new parent node in the clustering tree.
 - Update the distance matrix by computing the distances between the new node with other nodes.
 - Until there is only one node (root) left.

Hierarchical Clustering

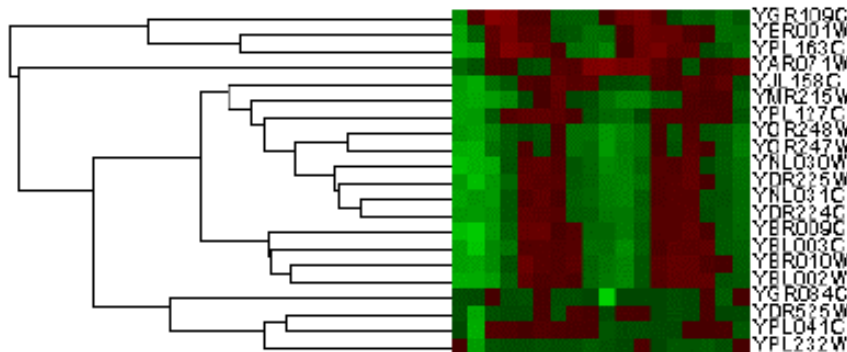
Combine most similar genes into agglomerative clusters,
build tree of genes



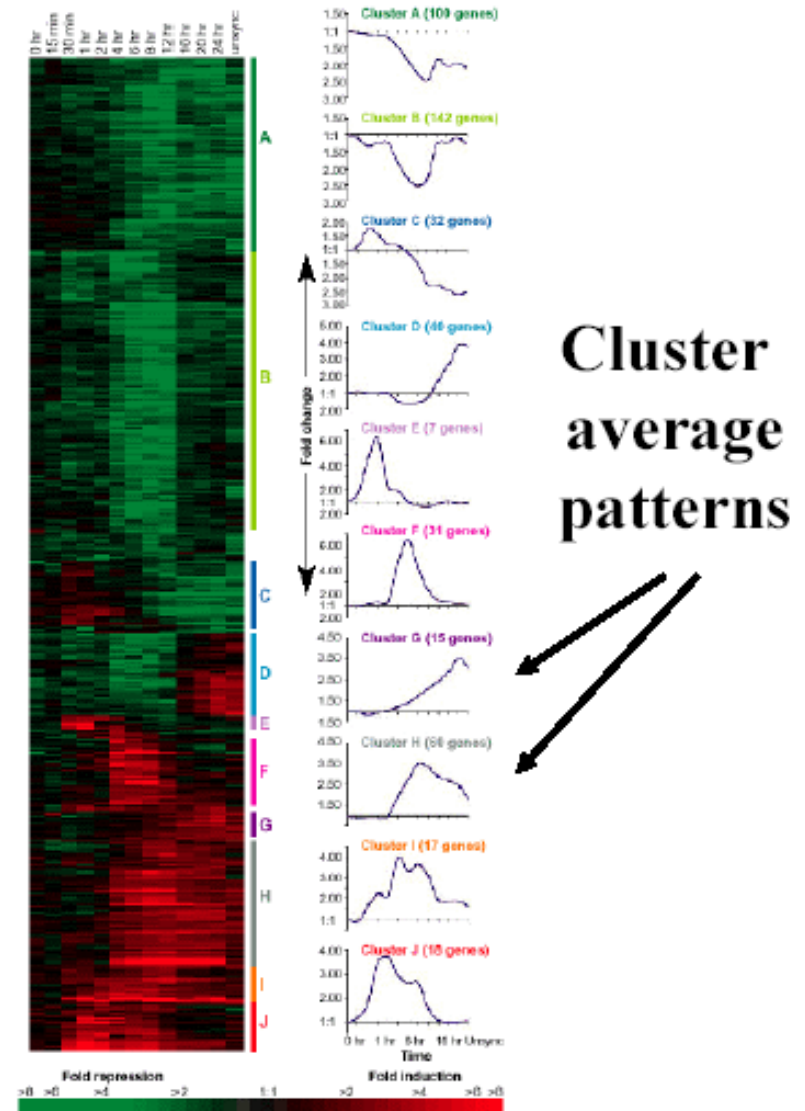
Rainer Breitling, 2005

Iyer et al., Science, Jan 1999:

Genes from functional classes are clustered together.



Cluster dendrogram

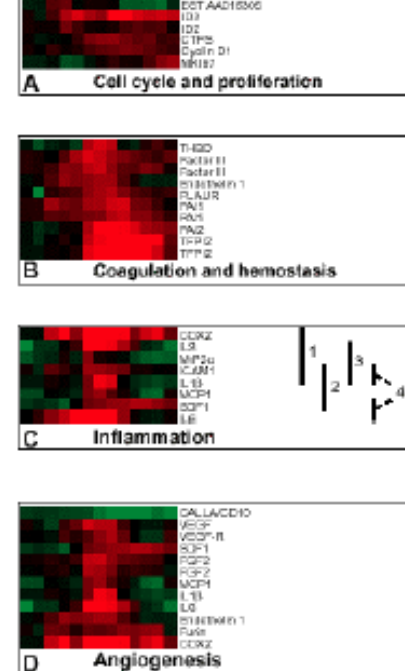
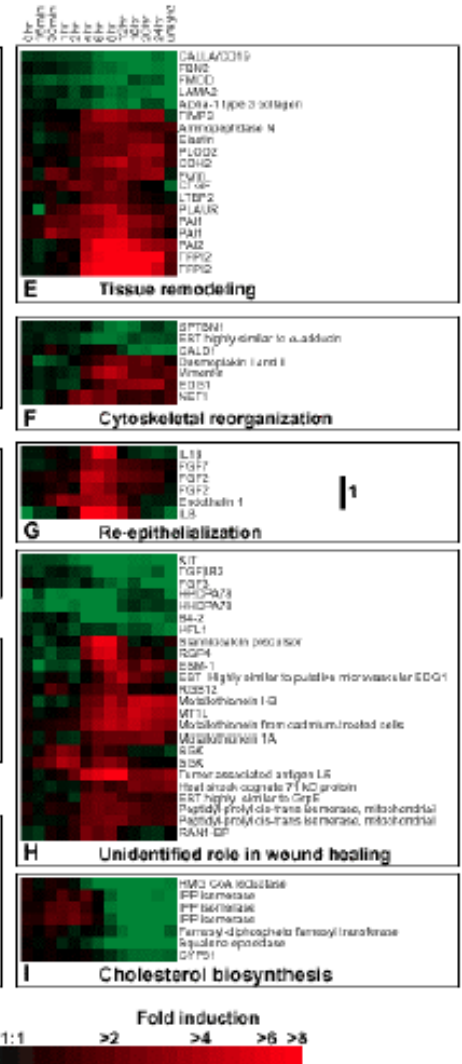
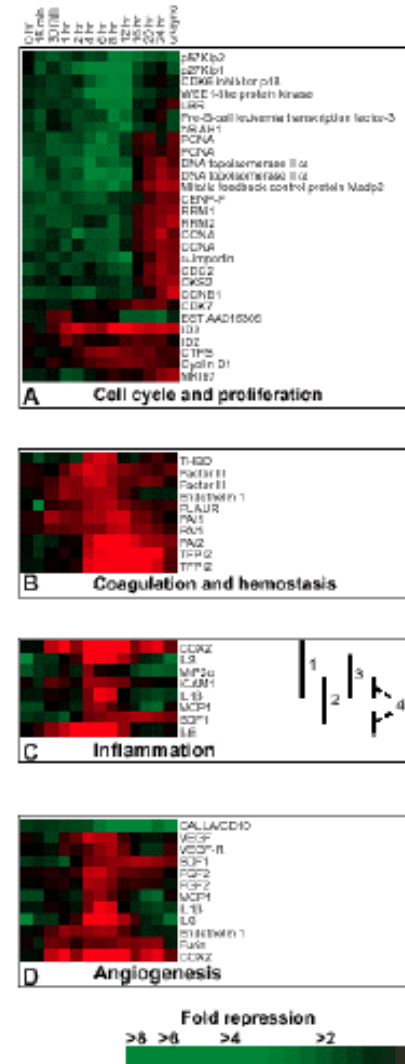
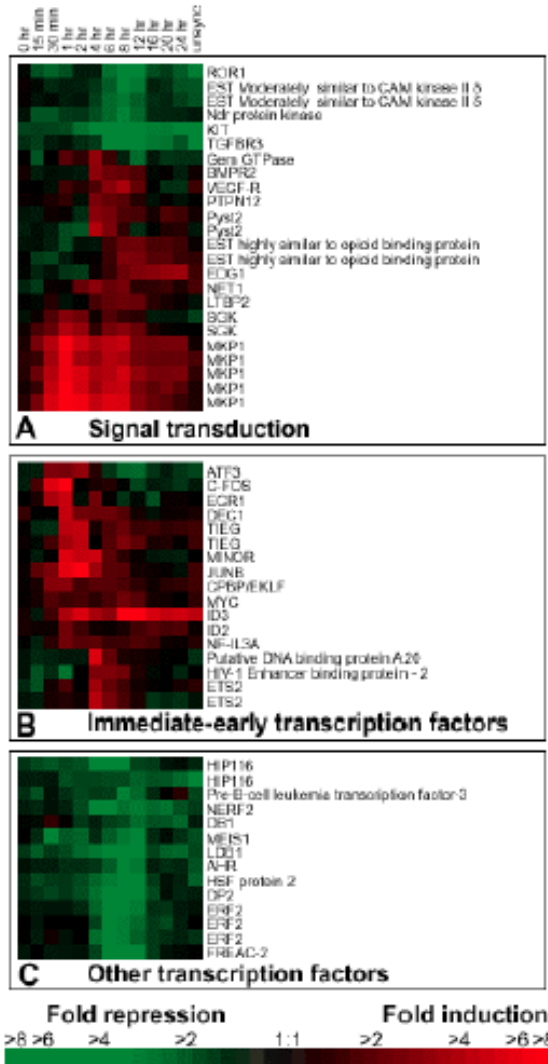


Cluster average patterns

Iyer et al.,
Science,
Jan 1999:

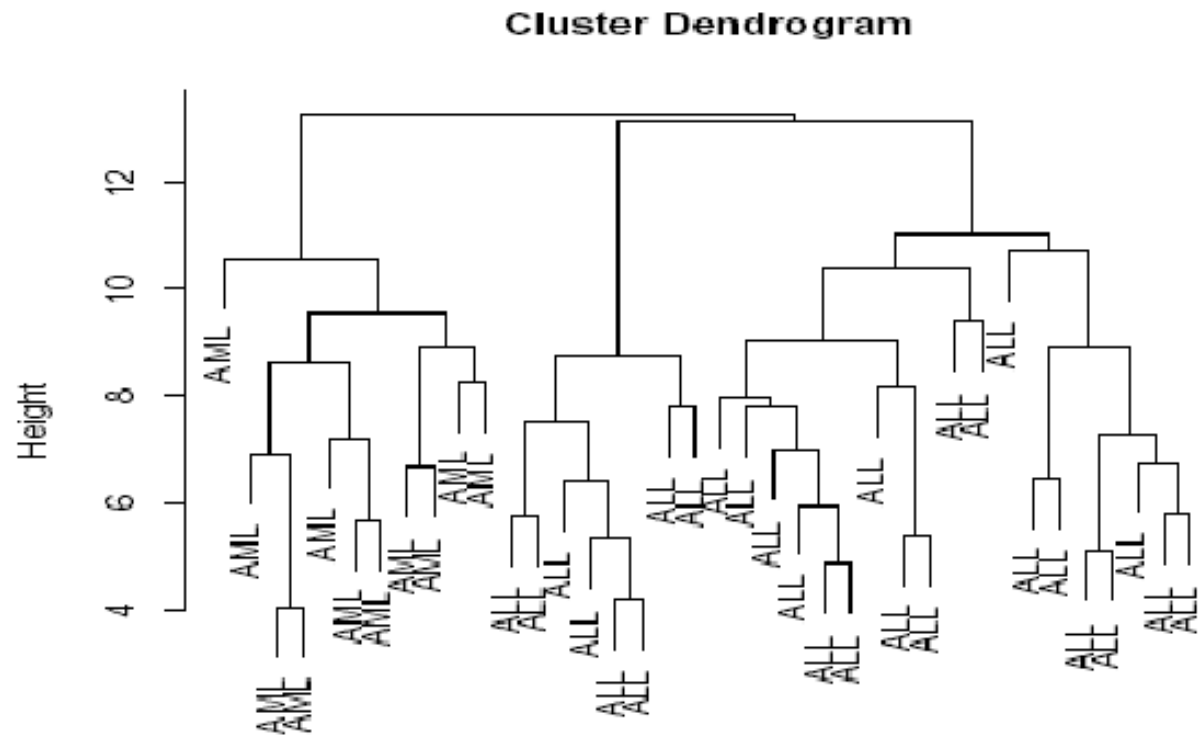
Genes from
functional
classes are
clustered
together
(sometimes!).

Careful
interpretation
necessary!

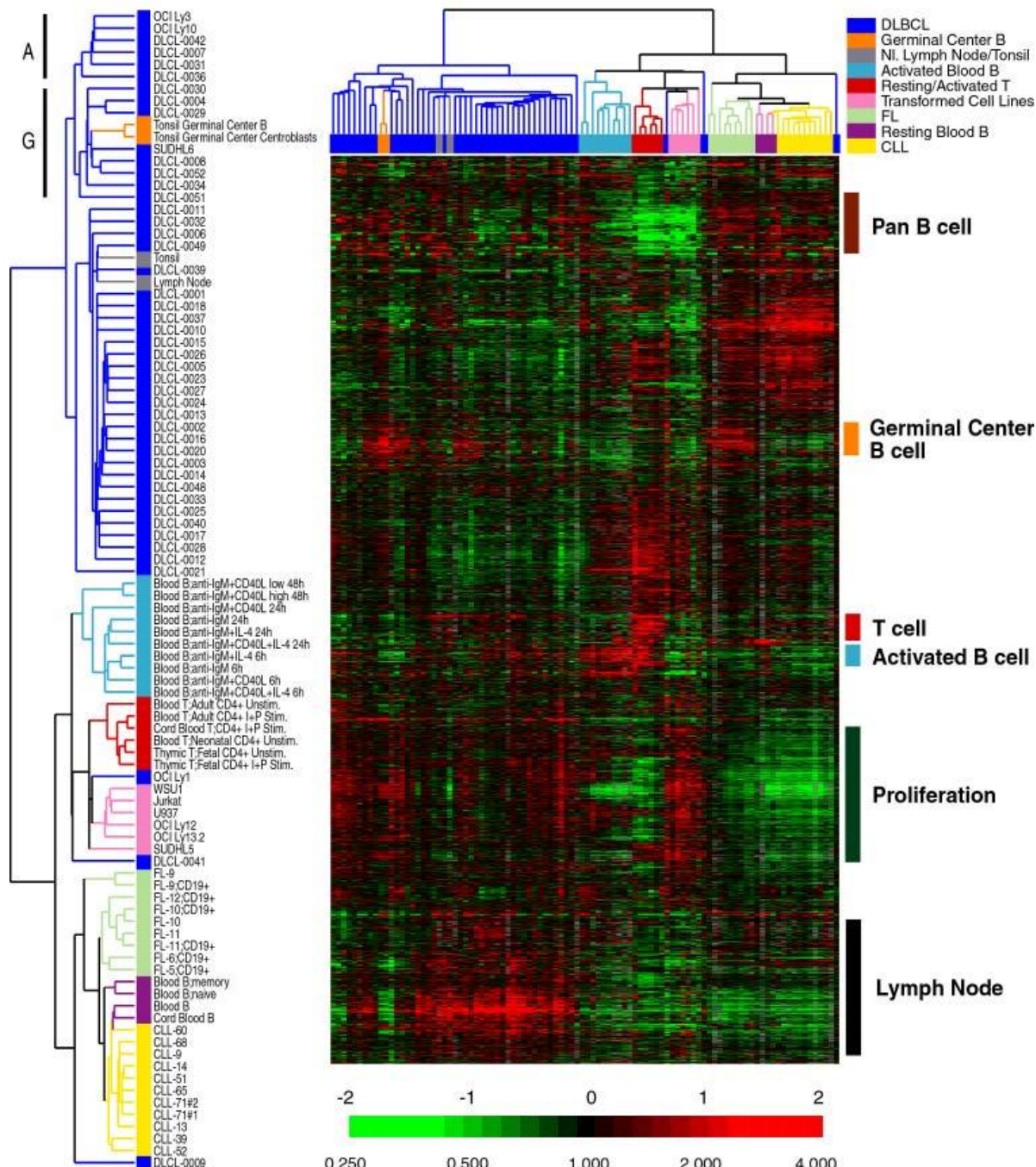


Golub et al.: Leukemia dataset, <http://www.genome.wi.mit.edu/MPR>

25 acute myeloid leukemia (AML),
47 acute lymphoblastic leukemia (ALL)
9 T-cell and 38 B-cell.



Shows perfect separation.

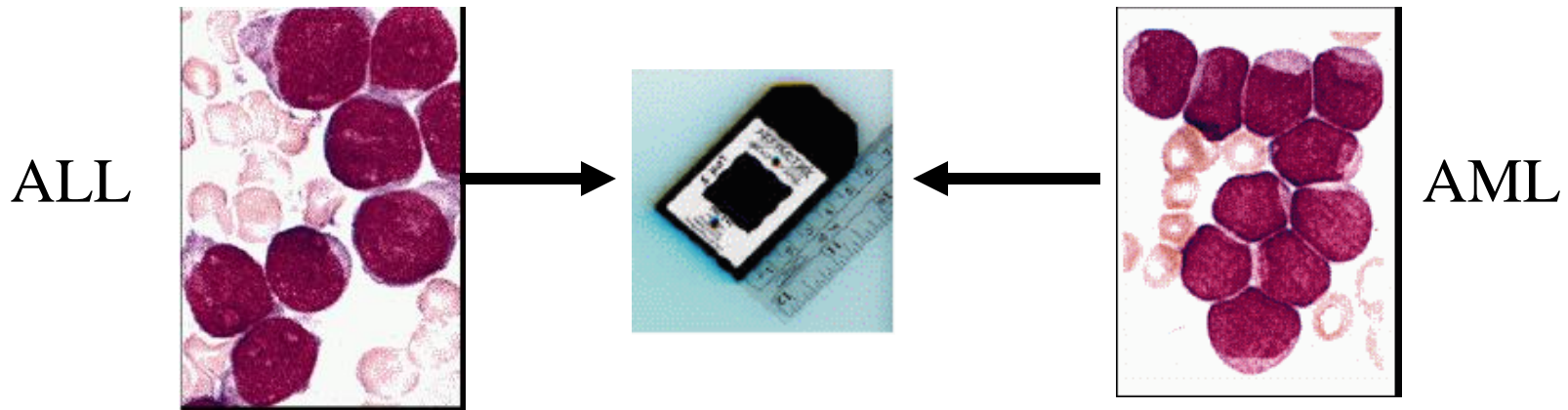


Two-way clustering of genes (y-axis) and cell lines (x-axis) (Alizadeh et al., 2000)

J. Pevsner, 2005

A Sample Classification Example

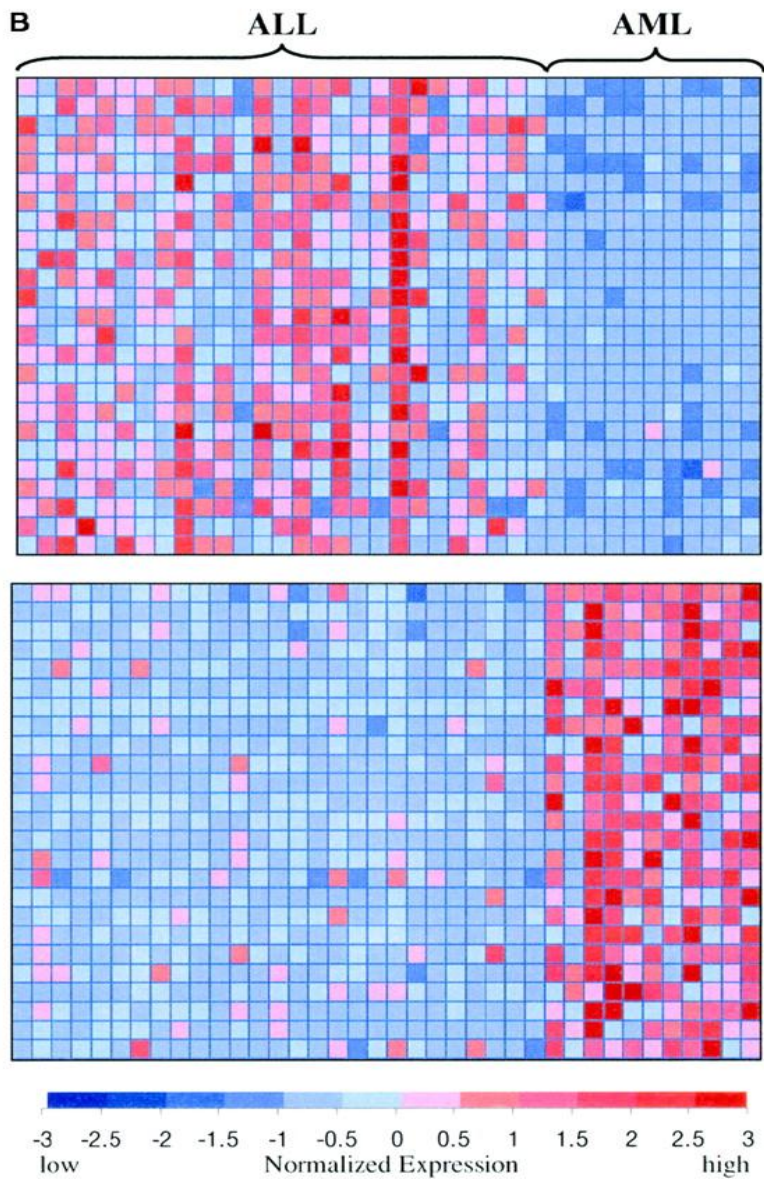
- Leukemia: Acute Lymphoblastic (ALL) vs Acute Myeloid (AML), Golub et al, Science, v.286, 1999
 - 72 examples (38 train, 34 test), about 7,000 genes
 - Gene expression values are features



Visually similar, but genetically very different

Results on the Test Data

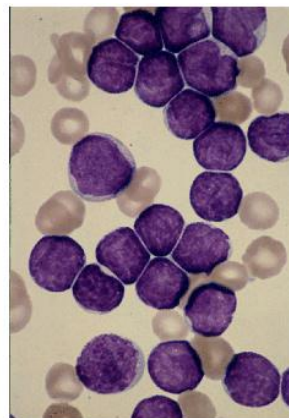
- Select genes (Feature selection)
- Best neural net model used 10 genes per class
- Evaluation on test data (34 samples) gives 1 or 2 errors (94-97% accuracy) using most classification methods



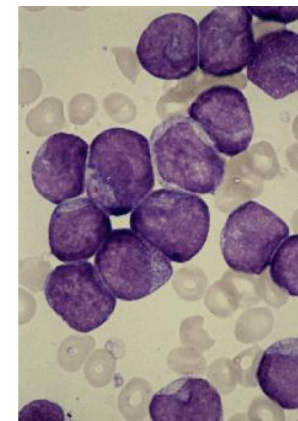
Classical study of cancer subtypes

Golub et al. (1999)

identification of diagnostic genes



ALL
acute lymphoblastic leukemia
(lymphoid precursors)



AML
acute myeloid leukemia
(myeloid precursor)

Rainer Breitling, 2005

Major Public Gene Expression Databases

- 3D-GeneExpression Database
- ArrayExpress
- BodyMap
- ChipDB
- ExpressDB
- Gene Expression Omnibus (GEO)
- Gene Expression Database (GXD)
- Gene Resource Locator
- GeneX
- Human Gene Expression Index (HuGE Index)
- RIKEN cDNA Expression Array Database (READ)
- RNA Abundance Database (RAD)
- Saccharomyces Genome Database (SGD)
- Stanford Microarray Database (SMD)
- TissueInfo
- yeast Microarray Global Viewer (yMGV)

Comprehensive Software

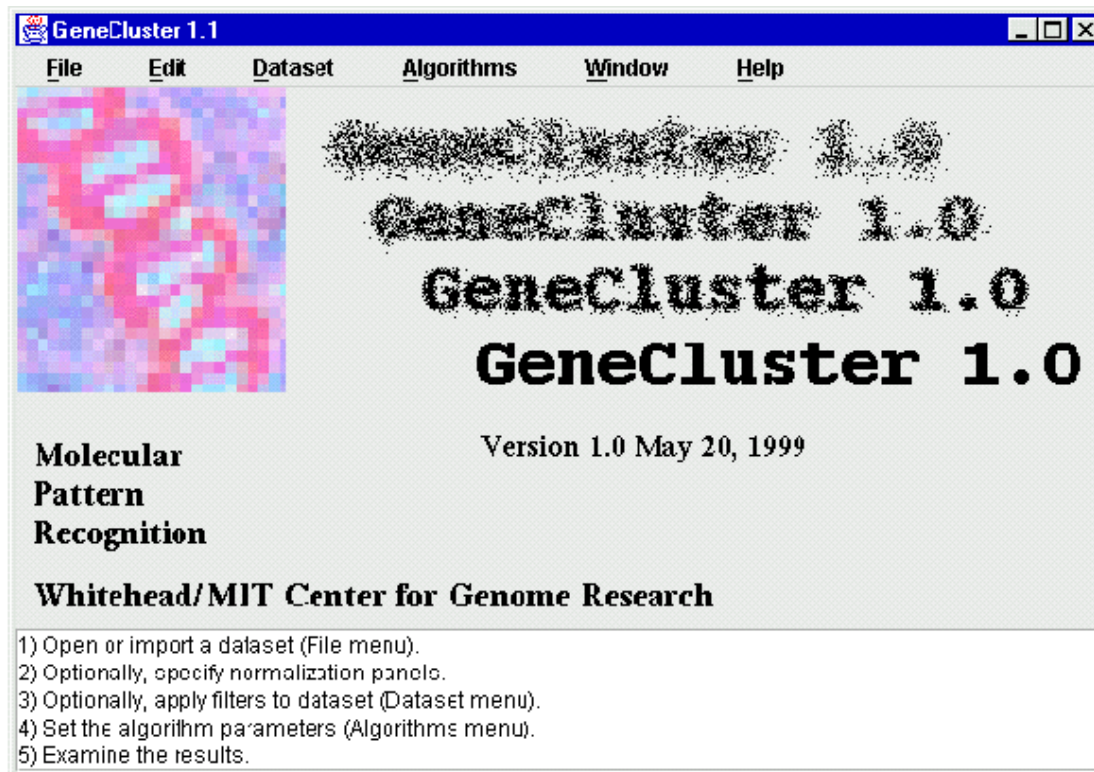
- Definition: Software incorporate many different analyses for different stage in a single package.
- Examples
 - **Cluster (Mike Eisen, LBNL)**
 - GeneMaths (Applied Maths)
 - GeneSight (Biodiscovery)
 - GeneSpring (Silicon Genetics)

Specific Analysis Software

- Definition: Software performing a few/ one specific analysis
- Examples
 - GeneCluster (Whitehead Institute Centre for genome research)
 - INCLUSive - INtegrated CLustering, Upstream Sequence retrieval and motif Sampler (Katholieke Universiteit Leuven)
 - SAM – Significance Analysis of Microarrays (Stanford University)

GeneCluster

- GeneCluster – performing normalization, filter and SOM



Y. F. Leung, 2005

Inclusive

- INCLUSive - INtegrated CLustering, Upstream Sequence retrieval and motif Sampler
- SAM – finding statistical significant differentially expressed gene

