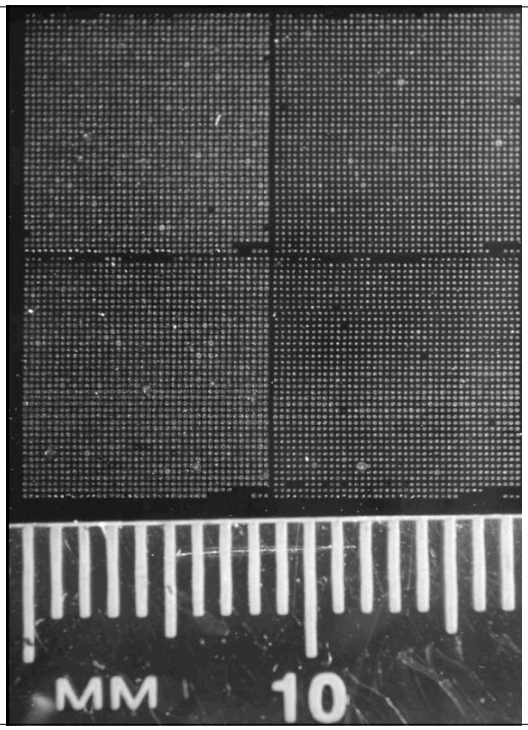
The practical art of analyzing whole-genome expression data (using spotted DNA microarrays)

Audrey P. Gasch, PhD
Mike Eisen lab
Lawrence Berkeley Lab

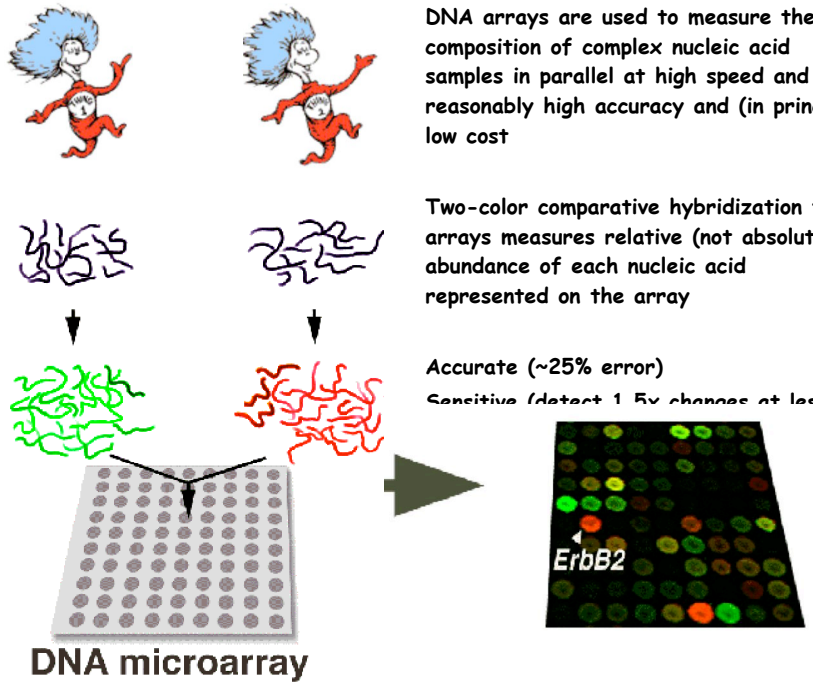
0.15	0.314	0.77	0.07	0.496	0.432	1.682	0.195	0.15	0.314	0.77	0.07	0.496	0.432	1.682	0.195
0.13	-2.158	0.18	-1.45	0.581	-0.296	-1.45	2.679	0.13	-2.158	0.18	-1.45	0.581	-0.296	-1.45	2.679
-1.141	-0.287	-0.97	-0.08	0.011	-0.222	1.737	-0.38	-1.141	-0.287	-0.97	-0.08	0.011	-0.222	1.737	-0.38
-0.1	-0.192	-0.11	-0.38	0.817	0.038	0.284	0.983	-0.1	-0.192	-0.11	-0.38	0.817	0.038	0.284	0.983
-0.247	0.137	-0.34	0.31	0.695	0.591	1.547	0.186	-0.247	0.137	-0.34	0.31	0.695	0.591	1.547	0.186
0.358	-0.649	0.29	-0.44	0.581	0.389	0.916	0.494	0.358	-0.649	0.29	-0.44	0.581	0.389	0.916	0.494
0.701	0.04	1.06	0.38	0.305	1.296	1.658	0.674	0.701	0.04	1.06	0.38	0.305	1.296	1.658	0.674
0.15	-0.005												.149	1.025	-1.409
-0.124	-0.44												.007	1.983	-0.44
0.314	0.15												496	0.195	1.682
-2.158	0.13												581	2.679	-1.45
-0.287	-1.141												011	-0.38	1.737
-0.192	-0.1												817	0.983	0.284
0.137	-0.247	0.31	-0.34	0.591					-0.247	0.31	-0.34	0.591	0.695	0.186	1.547
-0.649	0.358	-0.44	0.29	0.389					0.358	-0.44	0.29	0.389	0.581	0.494	0.916
0.04	0.701	0.38	1.06	1.296					0.701	0.38	1.06	1.296	0.305	0.674	1.658
-0.005	0.15	0.29	0.23	0.149					0.15	0.29	0.23	0.149	0.62	-1.409	1.025
-0.44	-0.124	-0.25	0.05	0.007	0.195	-0.44	1.983	-0.44	-0.124	-0.25	0.05	0.007	0.195	-0.44	1.983
0.15	0.314	0.77	0.07	0.496	0.432	1.682	0.195	0.15	0.314	0.77	0.07	0.496	0.432	1.682	0.195
0.13	-2.158	0.18	-1.45	0.581	-0.296	-1.45	2.679	0.13	-2.158	0.18	-1.45	0.581	-0.296	-1.45	2.679
-1.141	-0.287	-0.97	-0.08	0.011	-0.222	1.737	-0.38	-1.141	-0.287	-0.97	-0.08	0.011	-0.222	1.737	-0.38
-0.1	-0.192	-0.11	-0.38	0.817	0.038	0.284	0.983	-0.1	-0.192	-0.11	-0.38	0.817	0.038	0.284	0.983
-0.247	0.137	-0.34	0.31	0.695	0.591	1.547	0.186	-0.247	0.137	-0.34	0.31	0.695	0.591	1.547	0.186
0.358	-0.649	0.29	-0.44	0.581	0.389	0.916	0.494	0.358	-0.649	0.29	-0.44	0.581	0.389	0.916	0.494
0.701	0.04	1.06	0.38	0.305	1.296	1.658	0.674	0.701	0.04	1.06	0.38	0.305	1.296	1.658	0.674
0.15	-0.005	0.23	0.29	0.62	0.149	1.025	-1.409	0.15	-0.005	0.23	0.29	0.62	0.149	1.025	-1.409
-0.124	-0.44	0.05	-0.25	0.195	0.007	1.983	-0.44	-0.124	-0.44	0.05	-0.25	0.195	0.007	1.983	-0.44



Spotted DNA Arrays

Pre-synthesized DNA (e.g. from PCR or oligo synthesis)

Robotically deposited on treated glass microscope slides in regular array



DNA arrays are used to measure the composition of complex nucleic acid samples in parallel at high speed and with reasonably high accuracy and (in principle) low cost

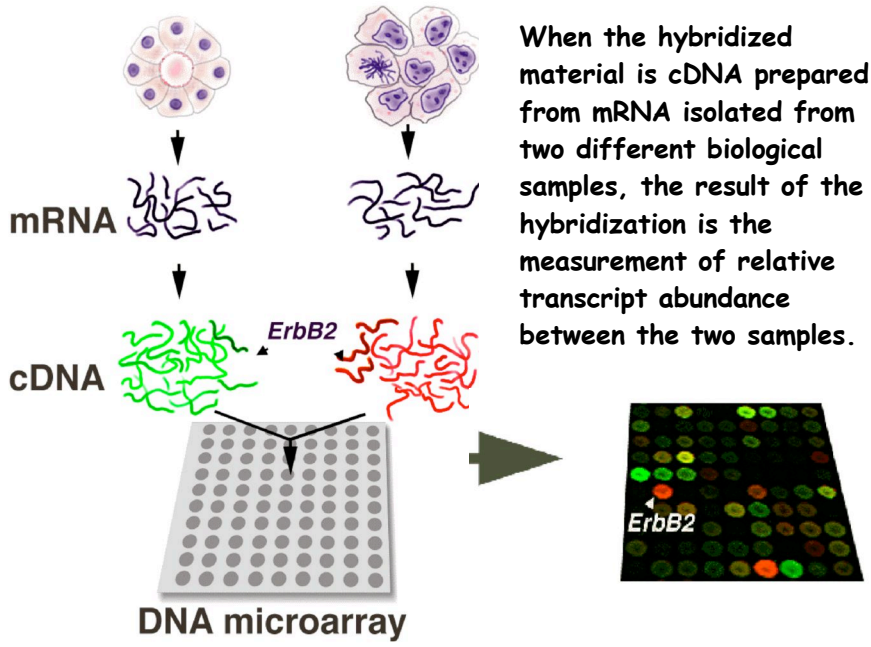
Two-color comparative hybridization to arrays measures relative (not absolute) abundance of each nucleic acid represented on the array

Accurate (~25% error)

Sensitive (detect 1.5x changes at less

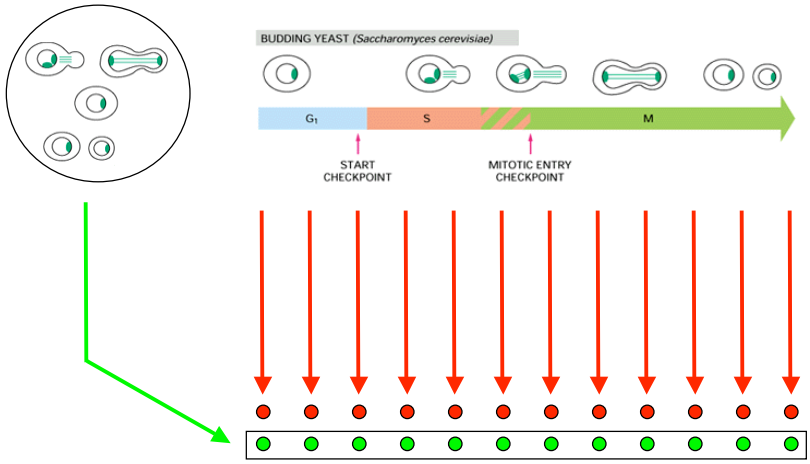
Microarrays can be used for any procedure that can take advantage of complementary DNA hybridization

- Gene expression analysis
- RNA localization
- RNA turnover
- RNA splicing and processing
- Chromatin & RNA immunoprecipitation
- Genotyping/Karyotyping
- Population analysis
- Population genomics
- Diagnostics/classification

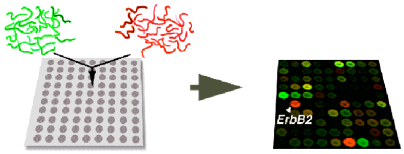


When the hybridized material is cDNA prepared from mRNA isolated from two different biological samples, the result of the hybridization is the measurement of relative transcript abundance between the two samples.

Changes in transcript abundance during the yeast cell cycle

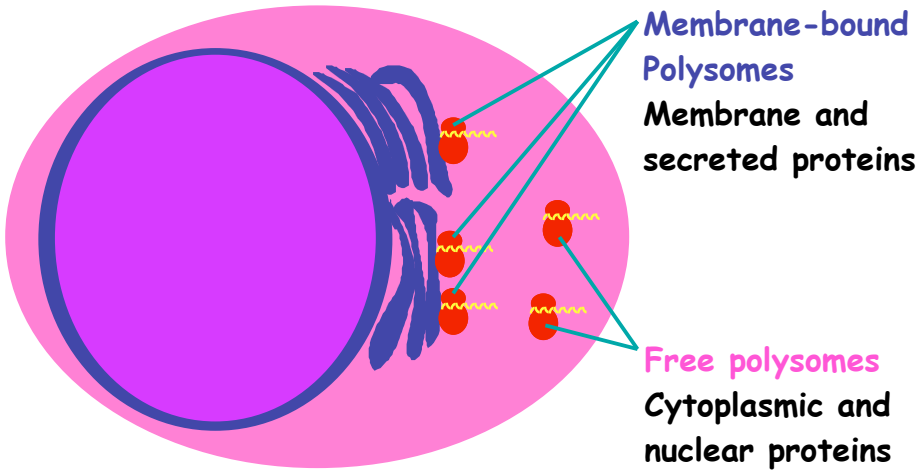


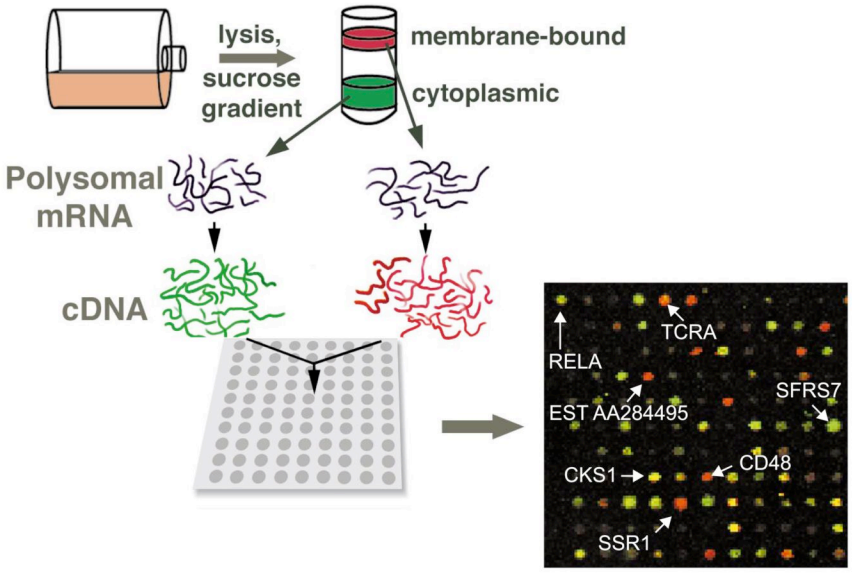
Spellman et al. 1998
Cho et al. 1998
Cho et al. 2001
Whitfield et al. 2001



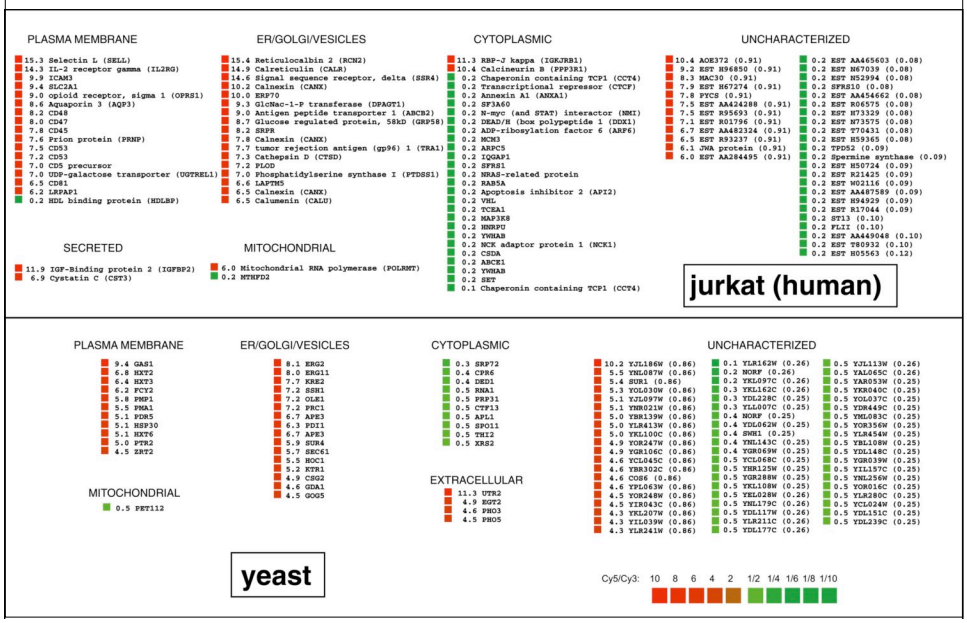
Use of cDNA Microarrays to Identify Genes Encoding Membrane and Secreted Proteins

Diehn et al. 2000

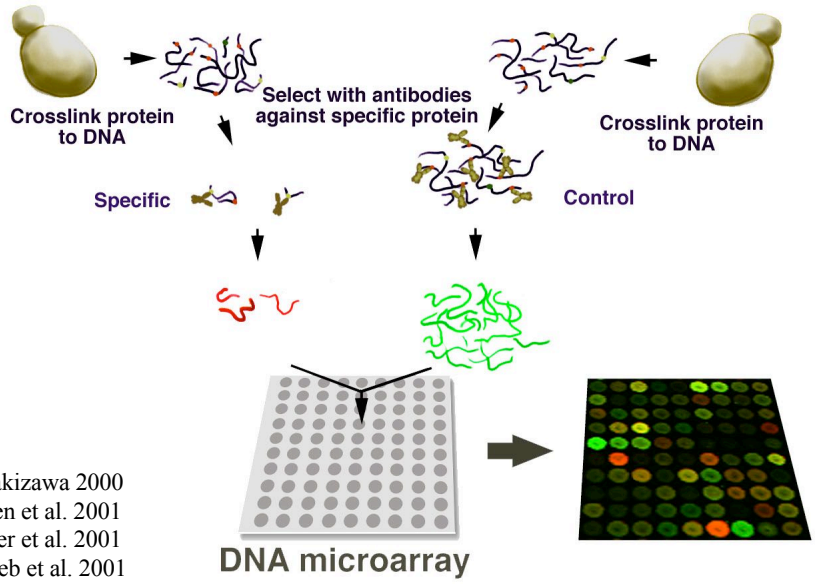




Free- and membrane-bound transcripts in human and yeast cells

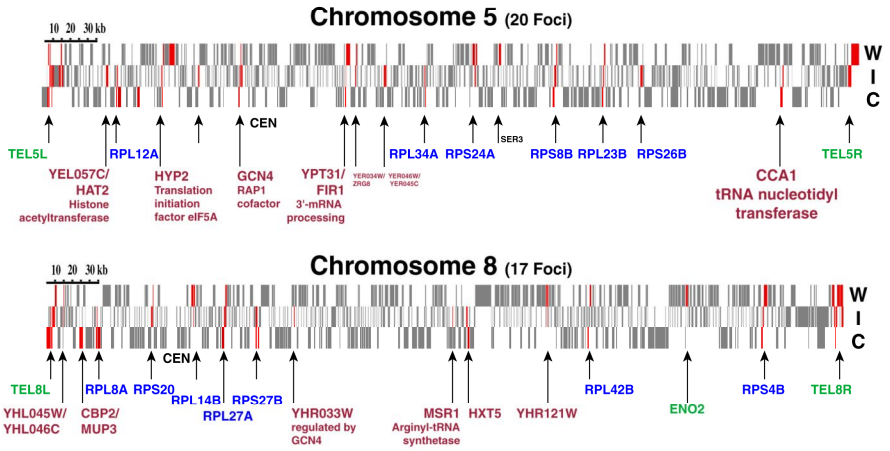


Mapping the distribution of specific proteins on the genome

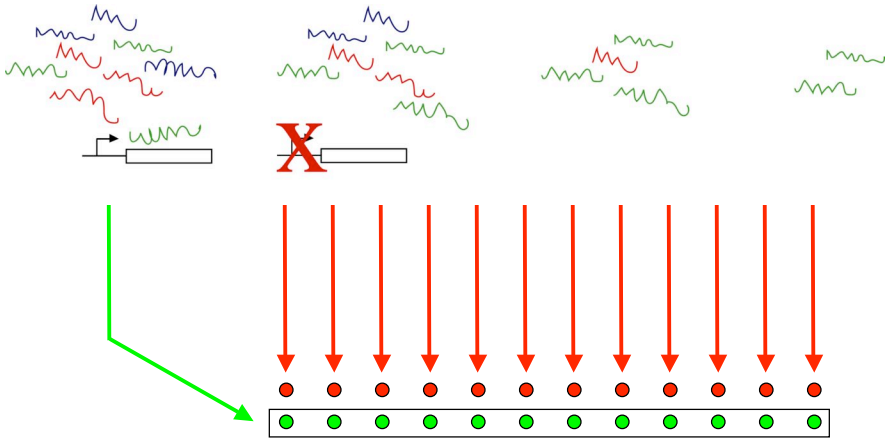


Takizawa 2000
Ren et al. 2001
Iyer et al. 2001
Lieb et al. 2001
Simon et al. 2001
Lee et al. 2002

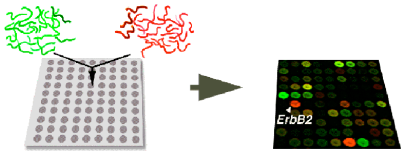
RAP1 is Localized to Telomeres and 263 Internal Foci, Potentially Affecting the Regulation of 358 Genes

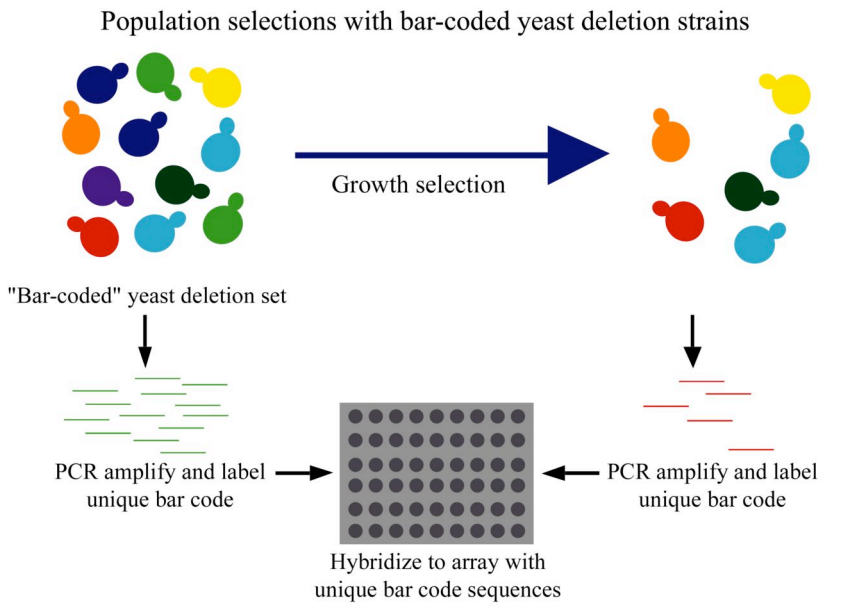


Measuring mRNA turnover rates using cDNA microarrays

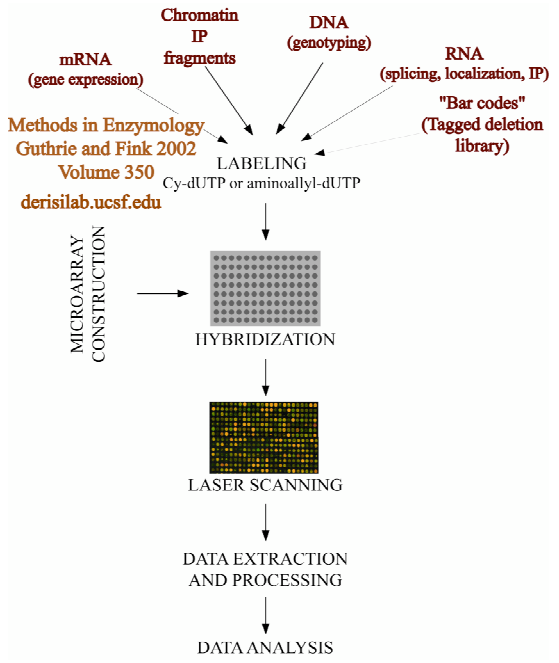


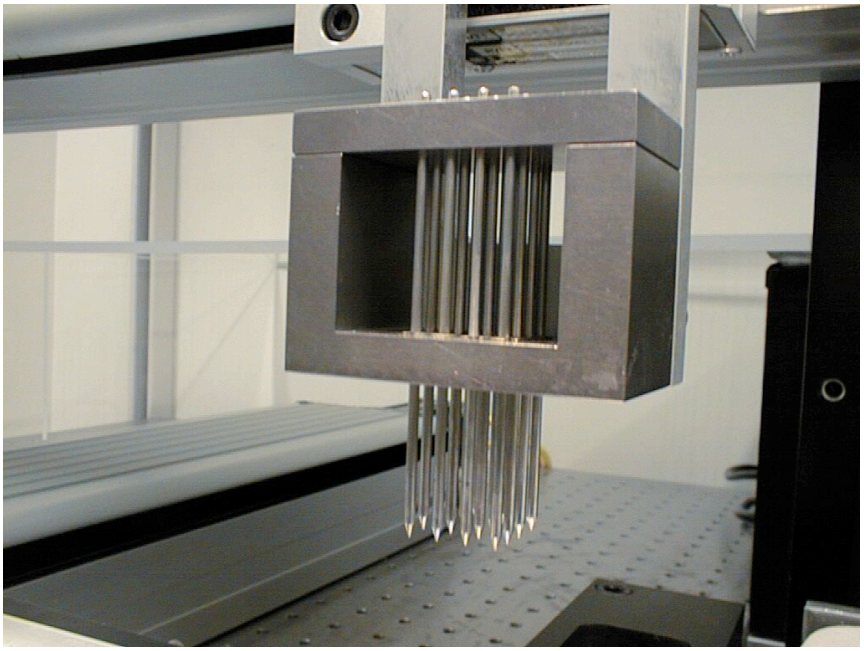
Wang et al. 2002
Lam et al. 2002
Fan et al. 2002
Bernstein et al. 2002
Gutierrez et al. 2002

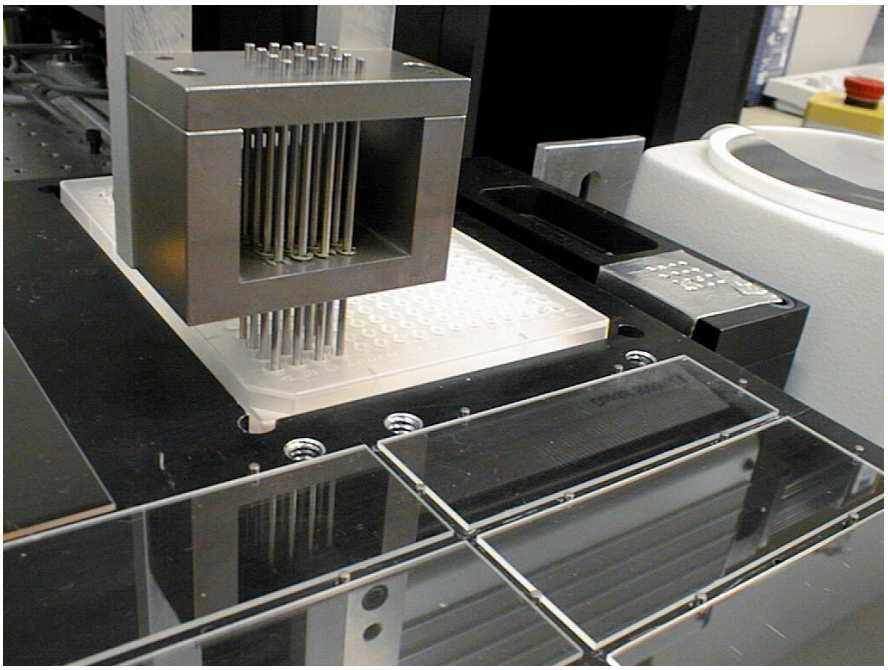


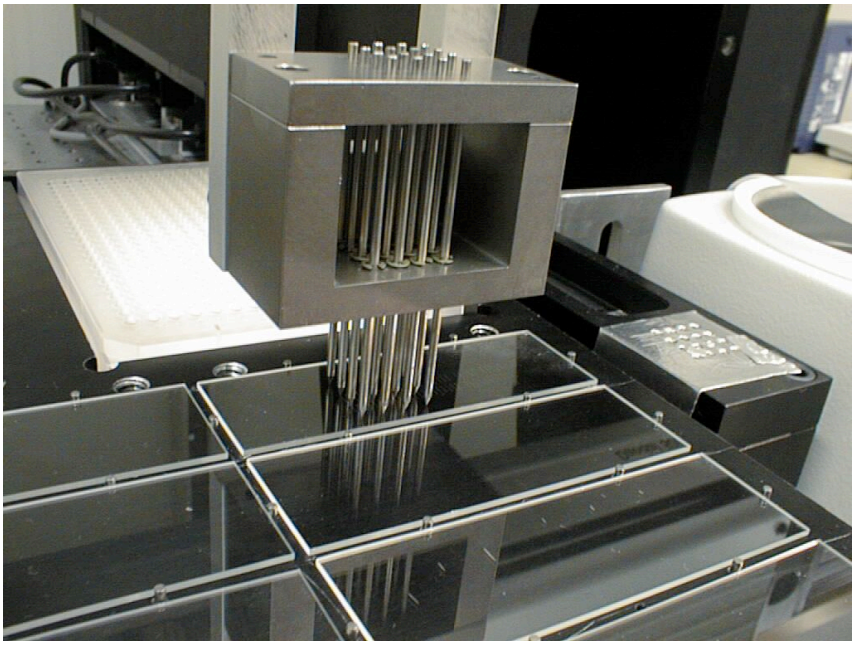


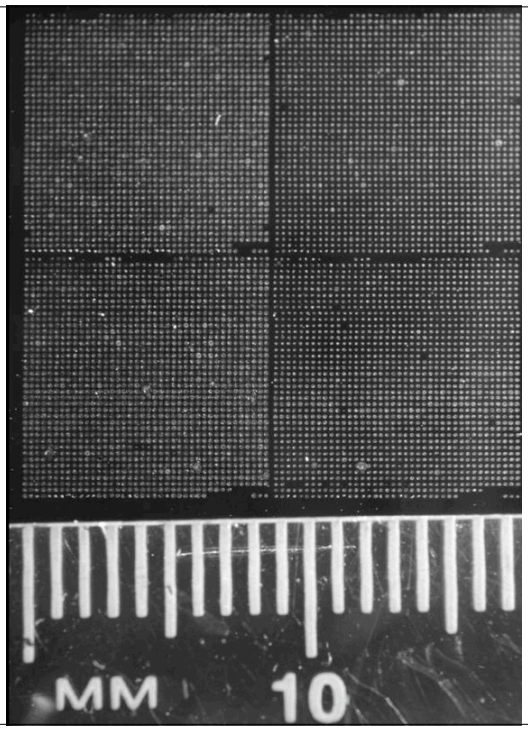
Winzeler et al. 2000; Birrell et al. 2001, 2002; Steinmetz et al. 2002a, 2002b; Giaever et al. 2002







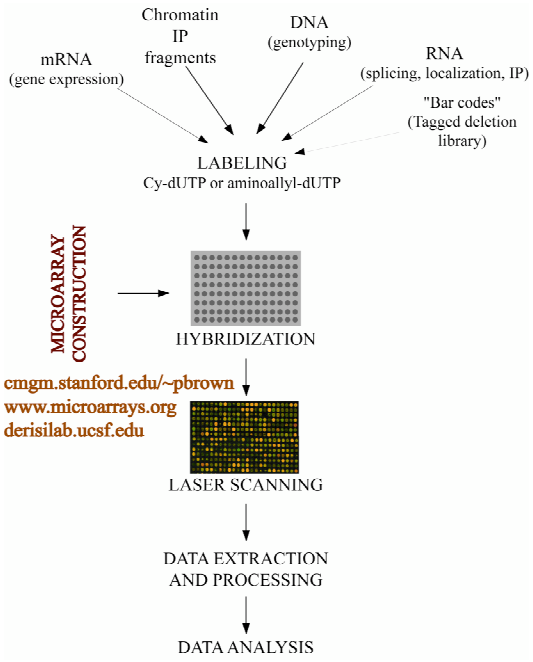


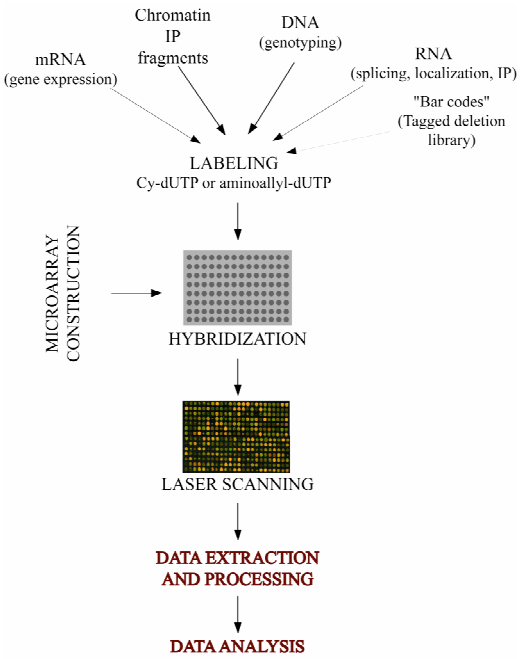


Spotted DNA Arrays

Spots can now be printed with center to center spacing of less than 100um, allowing for more than 150,000 spots to be printed on a standard glass slide.

A good robot can now print 50,000 spots on 200 slides in 24


Chromatin
IP
fragments
DNA
(genotyping)
mRNA
(gene expression)
RNA
(splicing, localization, IP)
"Bar codes"
(Tagged deletion
library)
LABELING
Cy-dUTP or aminoallyl-dUTP
MICROARRAY
CONSTRUCTION
HYBRIDIZATION
cmgm.stanford.edu/~pbrown
www.microarrays.org
derisilab.ucsf.edu
LASER SCANNING
DATA EXTRACTION
AND PROCESSING
DATA ANALYSIS


Chromatin
IP
fragments
DNA
(genotyping)
mRNA
(gene expression)
RNA
(splicing, localization, IP)
"Bar codes"
(Tagged deletion
library)
LABELING
Cy-dUTP or aminoallyl-dUTP
MICROARRAY
CONSTRUCTION
HYBRIDIZATION
LASER SCANNING
DATA EXTRACTION
AND PROCESSING
DATA ANALYSIS

Microarray hybridization design: which samples to compare?

Only requirement: good signal in reference channel

1. Directly compare two samples
   -- Tissue A vs. Tissue B
   -- Timepoint 60 min vs. Timepoint 0
2. Compare one sample to pool of samples
   -- Tissue A vs. pool of all tissues
   -- Timepoint 60 min vs. pool of timepoints
3. Compare each sample to genomic DNA
4. Loop design: successive comparisons
   -- S1 vs S2 S2 vs S3 S3 vs S4 S5 vs S6 S6 vs S1

Deconvoluting microarray data: ratios of ratios

Eg. Timecourse of mRNA samples vs. genomic DNA

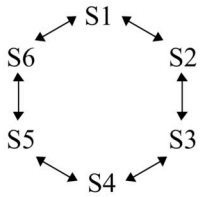
Array 1: $\frac{\text{mRNA (t = 0 min)}}{\text{genomic DNA}}$ Array 2: $\frac{\text{mRNA (t = 60 min)}}{\text{genomic DNA}}$

$$\frac{\text{Array 1}}{\text{Array 2}} = \frac{\frac{\text{mRNA (t = 60 min)}}{\text{genomic DNA}}}{\frac{\text{mRNA (t = 0 min)}}{\text{genomic DNA}}} = \frac{\text{mRNA (t = 60 min)}}{\text{mRNA (t = 0 min)}}$$

Deconvoluting microarray data: ratios of ratios

Eg. Loop design: successive comparisons

S1 vs S2 S2 vs S3 S3 vs S4 S5 vs S6 S6 vs S1



Drawbacks:

-- Need to do lots of deconvolutions to extract direct comparison information:

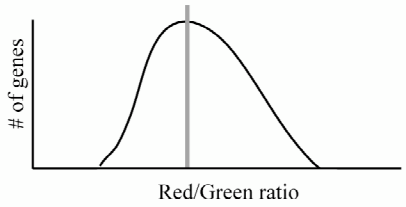
to compare S4 to S1:

$$\frac{\left[\dfrac{1}{\dfrac{S3}{S4}} \Big/ \dfrac{S2}{S3}\right]}{\dfrac{S1}{S2}} = \frac{S4}{S1}$$

Therefore missing data on one array = missing data on ALL of the arrays and error gets distributed during deconvolutions

NORMALIZATION METHODS

1. Assumption: average gene will not change in expression:
   adjust one channel intensity such that average gene expression change R/G = 1



2. Normalization by cell number (not RNA mass)

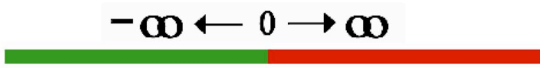
3. Dope labeling reactions with controls of known abundance with corresponding target spotted onto microarrays

4. Regional normalization across microarray
   (Terry Speed method) Yang et al. 2002 *NAR*

Data space: log vs. linear space
Want an equal scale for "red" and "green" spots

Red/Green

$\log_2$ (Red/Green)



Data selection:

- Ideally based on experimental reproducibility
  - -- t-test (with Bonferroni multiple-test correction)
  - -- SAM package (Tusher et al. 2001; R. Tibsharani website)
  - -- ANOVA
- Often based on arbitrary cutoffs
  - -- transcripts that change > arbitrary fold-cutoff
  - -- transcripts that change >cutoff in arbitrary # of experiments

Always do initial control experiments to define your own variability

How many replicates??

Answer: depends on the desired confidence
(and personal reproducibility)

Should do a minimum of 2-3 replicates for simple expression experiments

-- can get increased confidence in timecourse experiments
(nonrandom patterns of gene expression over time)

-- can get increased confidence in functionally-related genes
(eg. 135 ribosomal proteins acting in concert)

Based on the initial variance can do additional replicates if desired

Observation in many different organisms:

Genes that encode functionally related proteins are
often coexpressed at the level of transcript abundance

Therefore, a common goal in gene expression analysis is
to identify similarly expressed genes

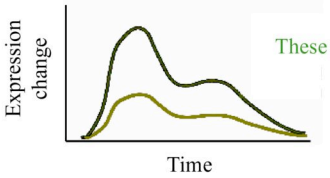
First step in doing this: choose a similarity metric

SIMILARITY METRIC:

gene expression pattern = n-dimentional vector

Euclidean distance:
shortest distance
between two points

Pearson correlation
(cosine-angle distance):
angle between two vectors
(ie the direction they point)



These patterns would be correlated using
Pearson correlation but not
Euclidean distance

** Using the WEIGHTED Pearson correlation is very useful
for large datasets (to underweight highly similar experiments)

COMPUTATIONAL METHODS OF ANALYSIS

Excellent review by J. Quakenbush 2001 *Nature Reviews-Genetics*

Nearest neighbors of a query gene

Clustering: grouping similarly expressed genes together

-- Hierarchical clustering

-- Self Organizing Maps (SOMs)

-- K-means clustering

More complicated algorithms:

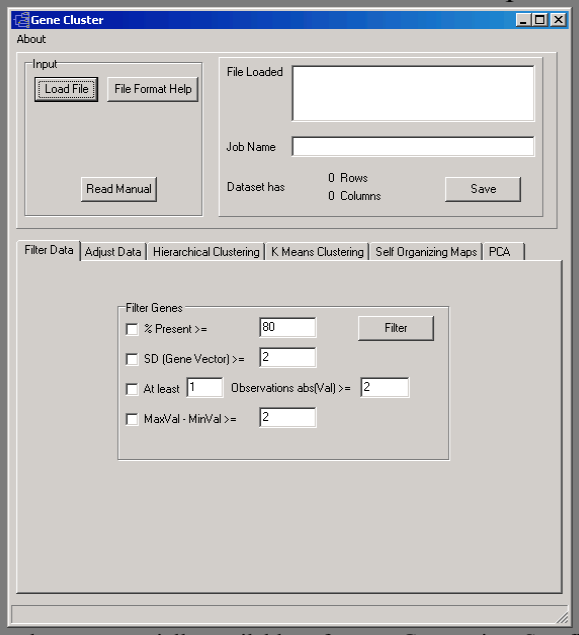
-- Support Vector Machines (SVMs)

-- Principal Component Analysis (PCA = SVD)

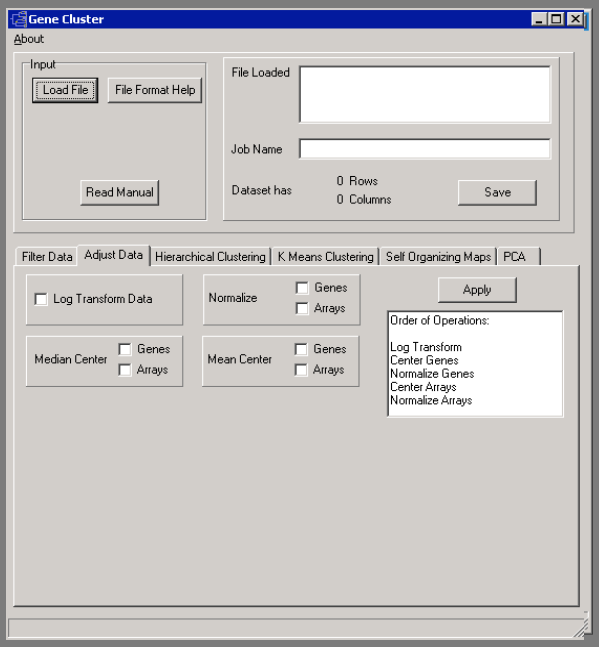
-- Bayesian Networks

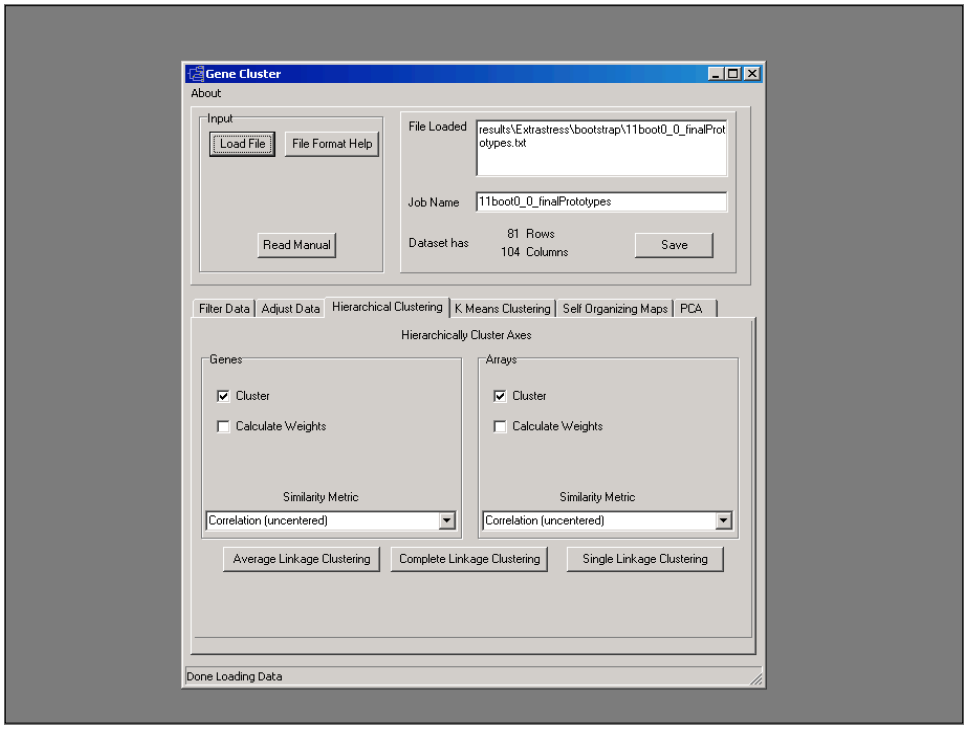
-- all kinds of heuristic algorithms

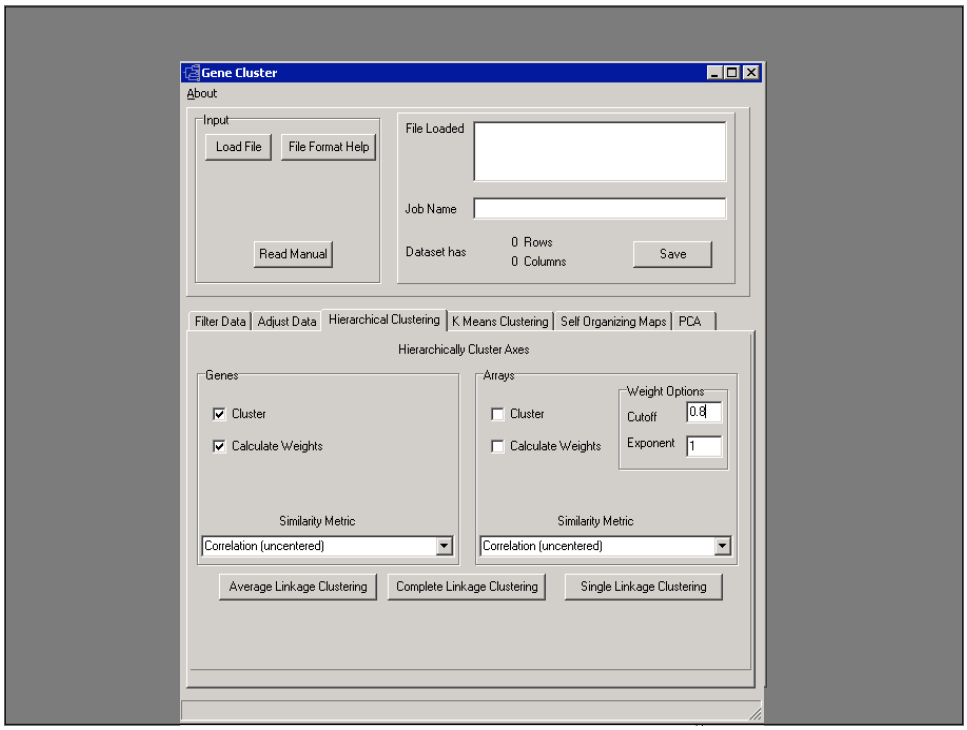
Mike Eisen's Cluster software: available at http://rana.lbl.gov



Also other commercially-available software: Genespring, Spotfire




Gene Cluster
About
Input
Load File
File Format Help
Read Manual
File Loaded
results\Extrastress\bootstrap\11boot0_0_finalPrototypes.txt
Job Name
11boot0_0_finalPrototypes
Dataset has
81 Rows
104 Columns
Save
Filter Data
Adjust Data
Hierarchical Clustering
K Means Clustering
Self Organizing Maps
PCA
Hierarchically Cluster Axes
Genes
Cluster
Calculate Weights
Similarity Metric
Correlation (uncentered)
Arrays
Cluster
Calculate Weights
Similarity Metric
Correlation (uncentered)
Average Linkage Clustering
Complete Linkage Clustering
Single Linkage Clustering
Done Loading Data


Gene Cluster
About
Input
Load File
File Format Help
Read Manual
File Loaded
Job Name
Dataset has
0 Rows
0 Columns
Save
Filter Data
Adjust Data
Hierarchical Clustering
K Means Clustering
Self Organizing Maps
PCA
Hierarchically Cluster Axes
Genes
Cluster
Calculate Weights
Similarity Metric
Correlation (uncentered)
Arrays
Cluster
Calculate Weights
Weight Options
Cutoff
0.8
Exponent
1
Similarity Metric
Correlation (uncentered)
Average Linkage Clustering
Complete Linkage Clustering
Single Linkage Clustering

Gene and Array weighting for clustering analysis as described in Cluster manual (rana.lbl.gov)

$$L_i = \sum_{\substack{\text{overall all rows } j \\ \text{where } d < \text{cutoff}}} \frac{(\text{cutoff} - \text{dist}(i,j))^n}{\text{cutoff}} \qquad W_i = \frac{1}{L_i}$$

If two arrays are identical, their corr = 1 and distance = 1-corr = 0

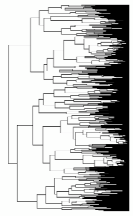
For exponent n = 1

$$L = \frac{(0.2 - 0)}{0.2} + \frac{(0.2 - 0)}{0.2} = 2 \qquad W = 0.5$$

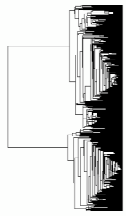
HIERARCHICAL CLUSTERING: groups similarly expressed genes together by building a tree from the "bottom up"

Different Methods of Tree Building (eg. clustering)

1. Single-linkage clustering

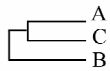


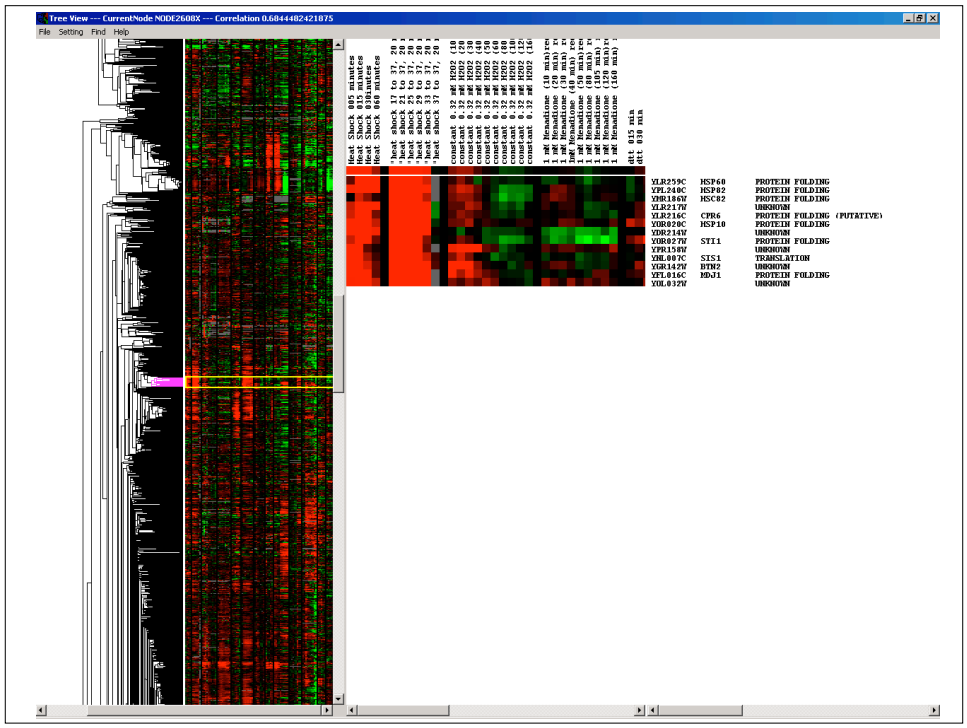
2. Complete-linkage clustering



3. Average-linkage clustering

D?





Many people are initially overwhelmed with the massive volume of data

Tips for initial data exploration

1. Make a first pass through the data, looking at thumbnail image, to an overview of the global gene expression programs

2. Next, go back through the data, cluster by cluster
   -- digest the gene expression pattern of each cluster
   -- glance at the genes in the cluster and look for relationships (functional, regulatory, etc)

3. Finally, go through the data in detail and focus on individual genes and small groups of genes

A useful function: the hypergeometric distribution

Can calculate the probability of observing at least q related objects in a cluster of l objects, based on the total number of related objects (M) in the genome of N genes.

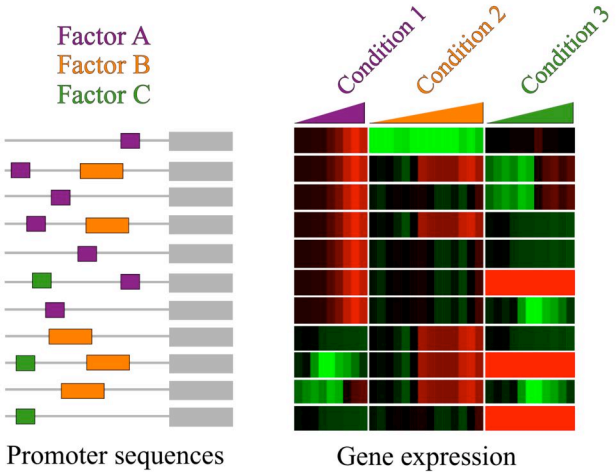
$$P = \sum_{i=q}^{l} \frac{\binom{M}{i}\binom{N-M}{l-i}}{\binom{N}{l}}$$

Example: Probability of observing 8 protein folding chaperones in a cluster of 15 genes, when there are 20 chaperones in a genome of 6,000 genes

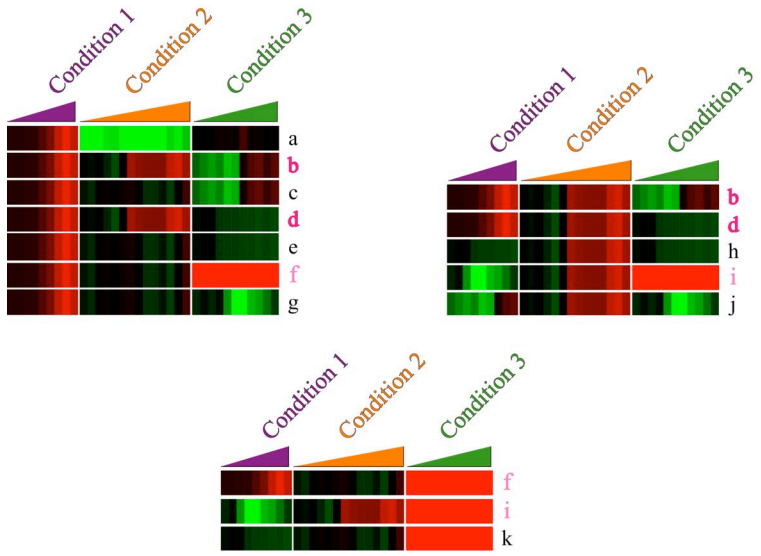
$$P = \sum_{i=q}^{l} \frac{\binom{20}{i}\binom{6000-20}{15-i}}{\binom{6000}{15}} = 2 \times 10^{-17}$$

8/15 chaperones in cluster = 53%
20/6000 chaperones in genome = 3%
= 18-fold enrichment in cluster

Different sets of genes are coregulated under different conditions



Allow genes to belong to more than one group



Modified Fuzzy k-means clustering

Based on the method developed by Bezdek *et al.* c. 1980

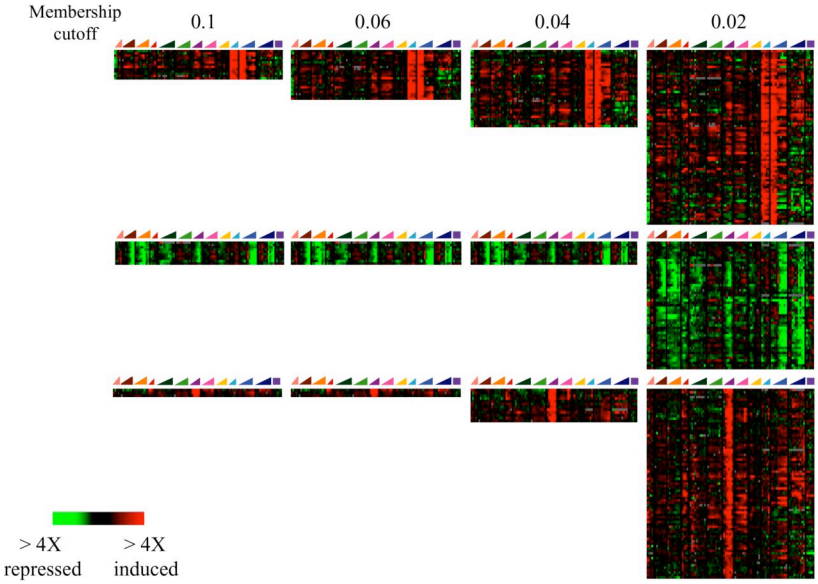
Heuristically modified to analyze yeast genomic expression data

Algorithm output:

- A list of cluster nodes (represented by average expression patterns)
  - A matrix of each gene's membership to each cluster node

Genes are assigned to each cluster to which they are similar above a user-defined membership cutoff

Genes are assigned to nodes using a sliding membership cutoff



FuzzyK is a C++ command line program that runs on Linux

The results can be viewed with the PERL program *FuzzyExplorer*

Both are available at http://rana.lbl.gov/FuzzyK

Final words about microarray data analysis

1. While analyzing data, always be aware of:
   -- exact features of the experiment
   -- normalization method

2. Be careful about gene annotations:
   -- often limited
   -- sometimes incorrect
   -- many genes have multiple functions that are hard to capture

3. Remember that cells may experience the experimental conditions beyond your interpretation …
   -- be aware of pleiotropic conditions
   -- be aware of secondary effects of conditions

4. The power of comparison: many responses may not be specific to your conditions