Gene Profiling, Clustering, and Networking

Alfred O. Hero III

University of Michigan, Ann Arbor, MI

http://www.eecs.umich.edu/~hero

Mar. 2005

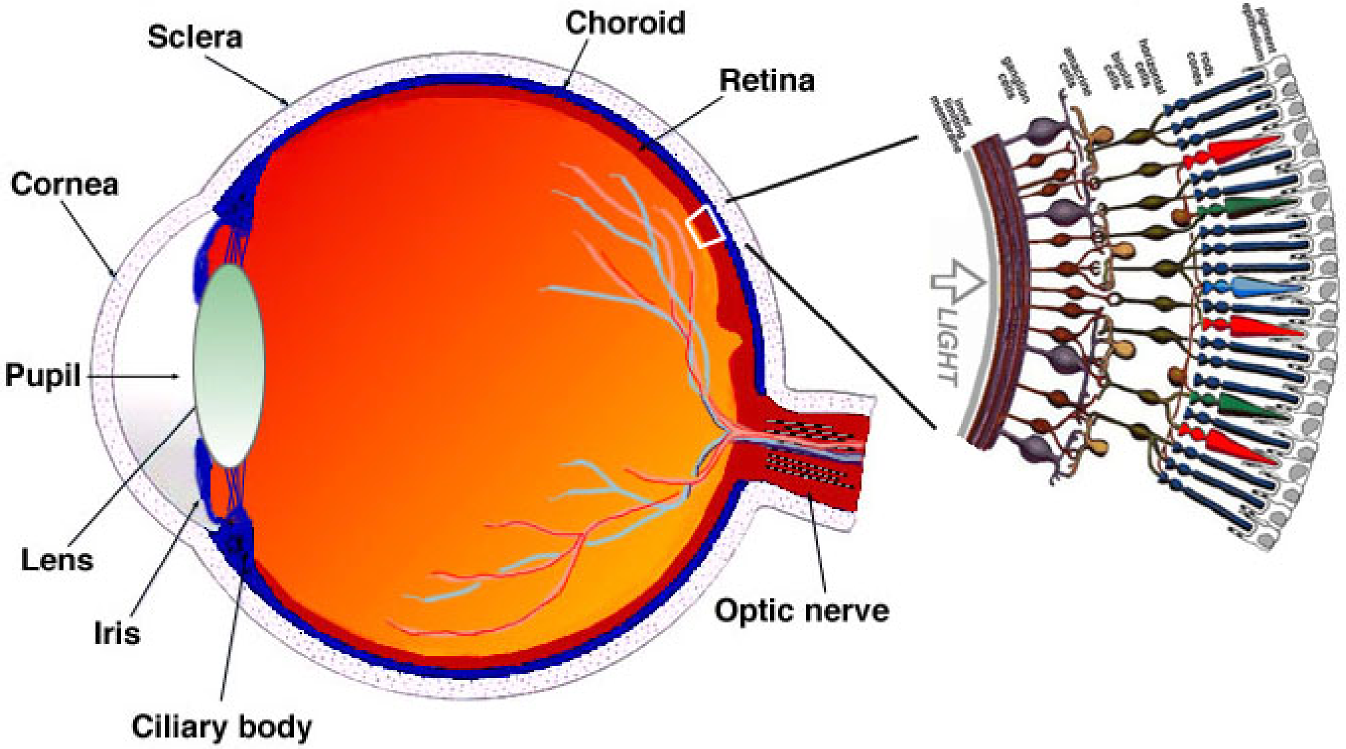
1. Genomics, transcriptomics and gene microarrays
2. Preprocessing of gene microarray data
3. Screening differentially expressed genes
4. Clustering gene co-regulation patterns
5. Conclusions

Acknowledgements

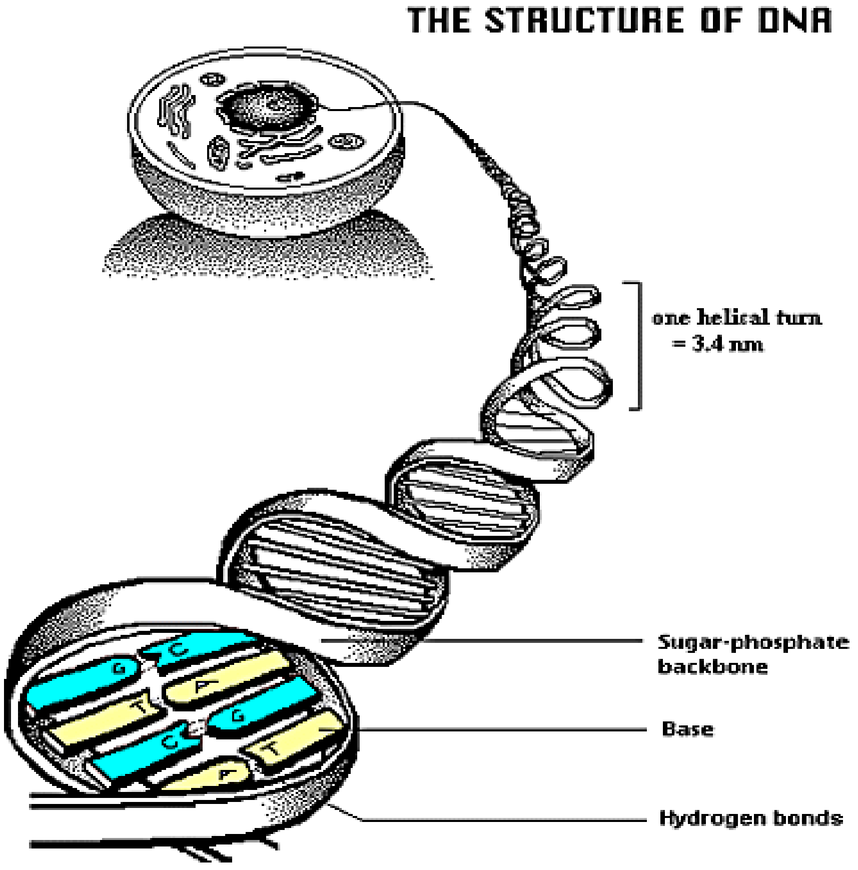
- Anand Swaroop, Ophthalmology, UM
- David States, Bioinformatics, UM
- Alan Mears, Univ Ottawa (CA)
- Gilles Fleury, ESE, France
- Debashis Ghosh, Biostatistics, UM
- Terry Speed, Statistics, UCB
- Jindan Yu, BME, UM
- Dongxiao Zhu, Bioinformatics, UM
- Yuezhou Jing, Dept. Statistics, UM

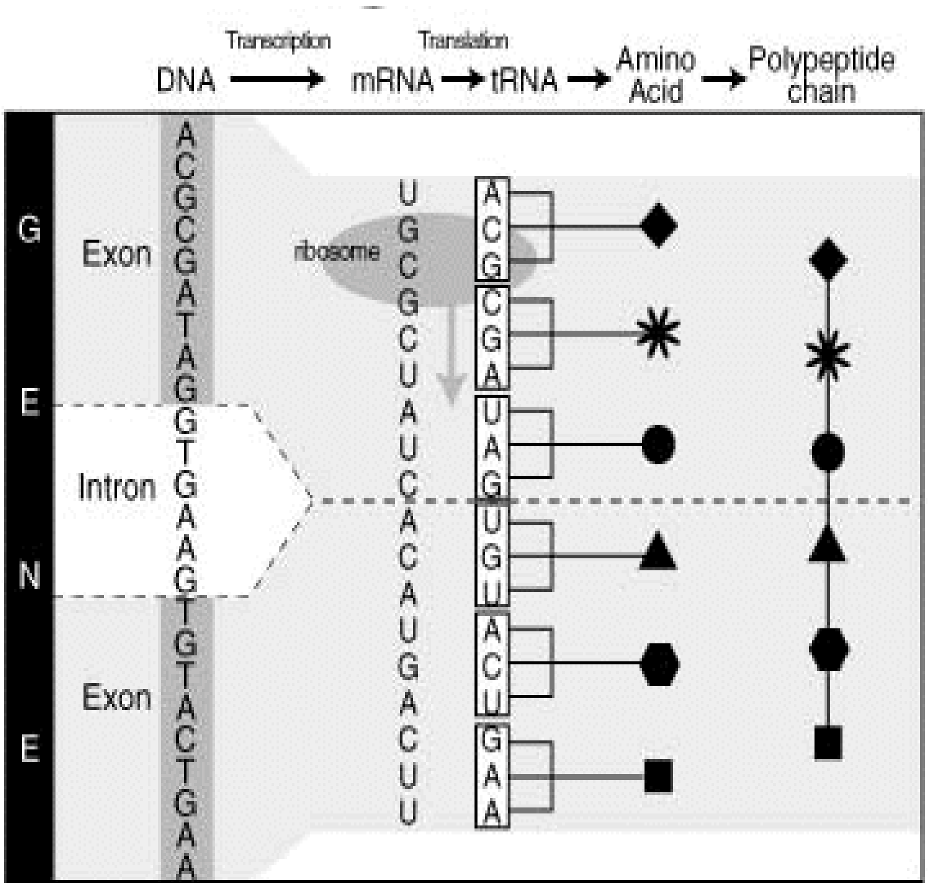
Some Biological Questions

- What is genetic basis for photoreceptor development, aging, and degeneration?
- What are patterns of gene expression in the retina over time?
- What genes mediate development of rods and cones?



1. Genomics, Transcriptomics and Gene Microarrays



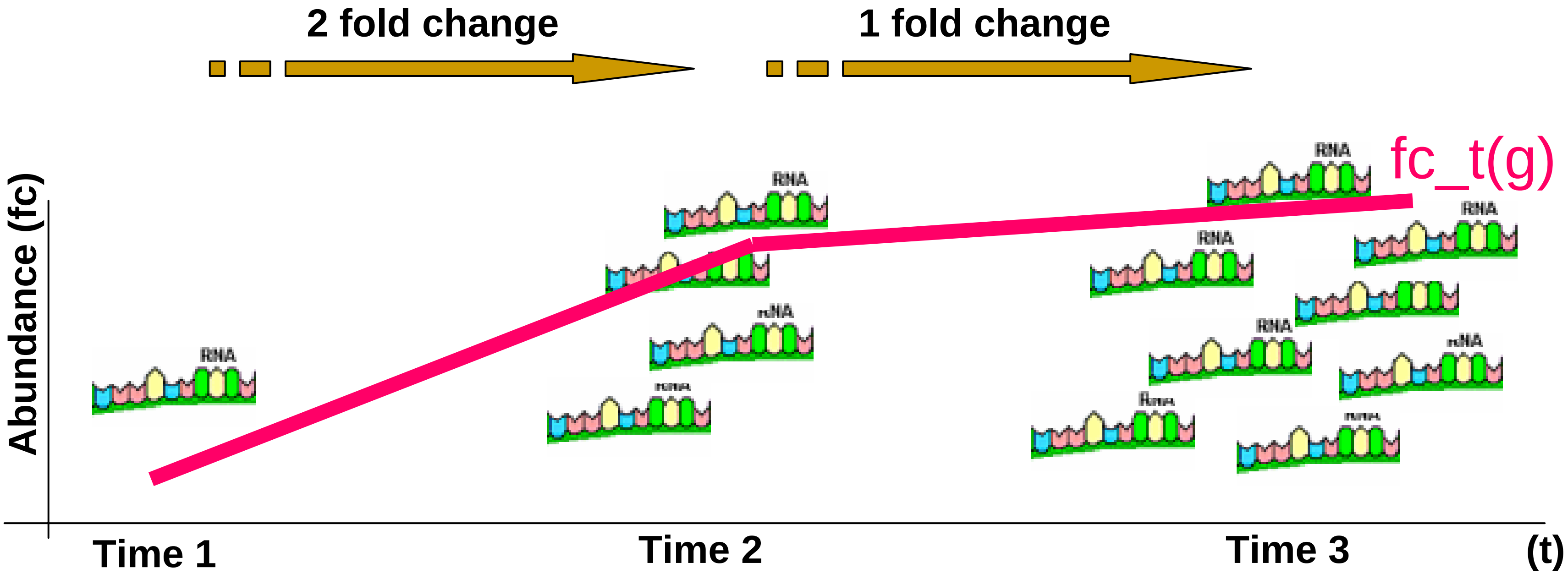


http://www-stat.stanford.edu/~susan/courses/s166/node2.html

http://www.genome.gov/

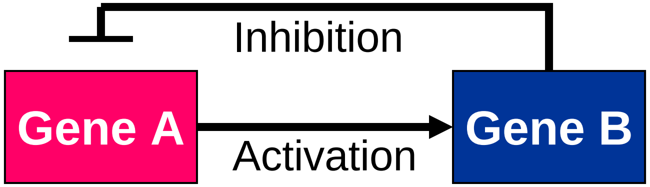
Transcriptomics: Gene expression profiling

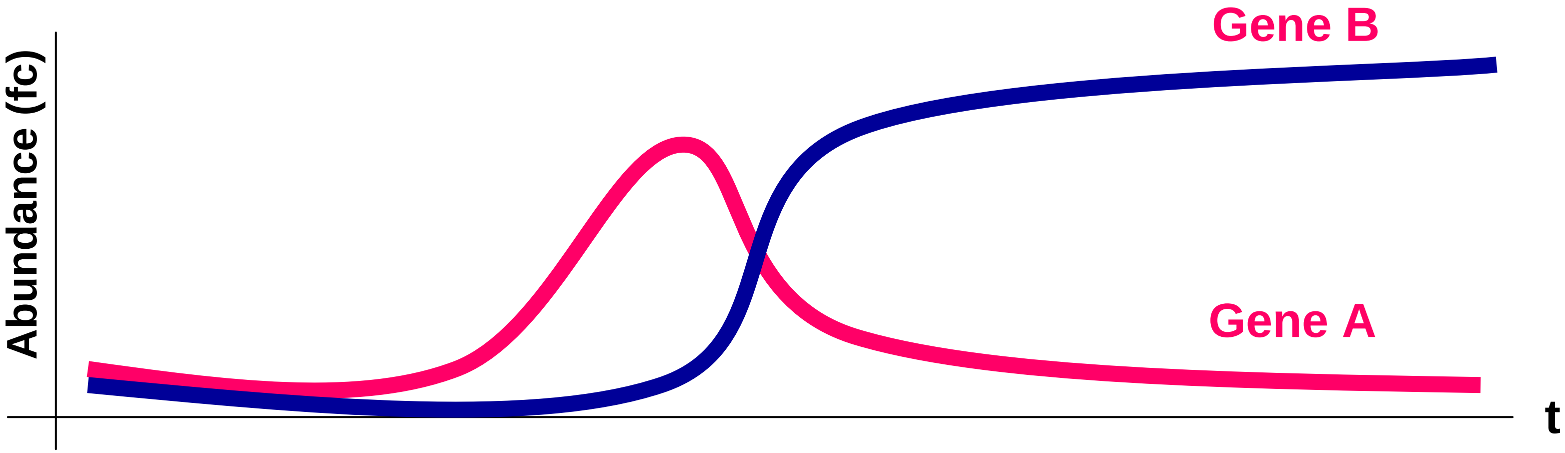
What is pattern of gene activation/inactivation over time, tissue, therapy, etc?



Discovery of Genetic Circuits

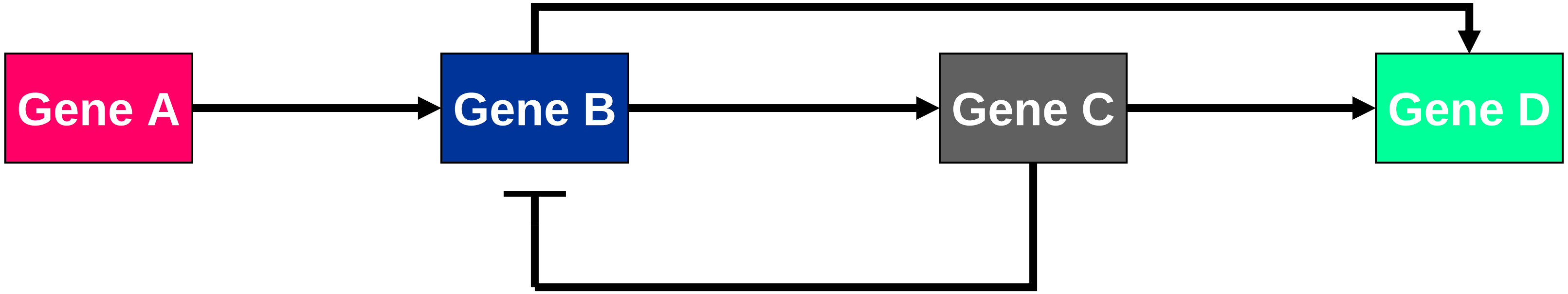
How do genes regulate (activate/inhibit) each other's expression levels over time?





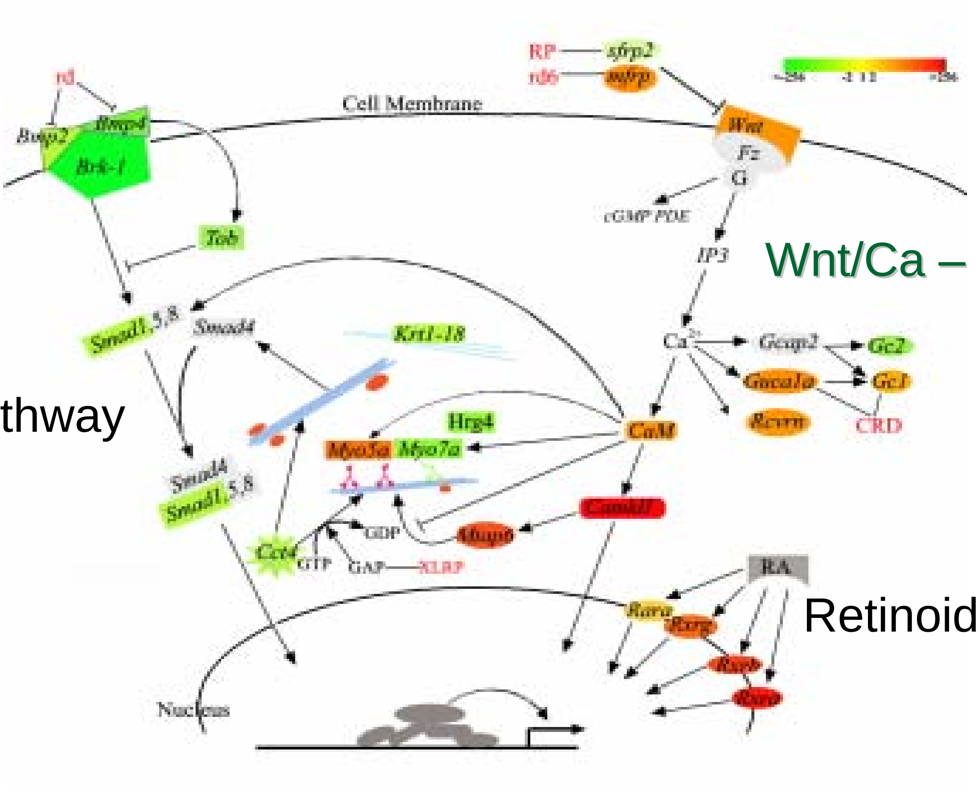
Discovery of Genetic Pathways

What sequence of gene interactions lead to a specific metabolic/structural (dys)function



Discovery of Gene Regulation Networks

What are the networks of gene pathways that co-regulate gene expression of an organism?



Wnt/Ca – calmodulin pathway

Bmp pathway

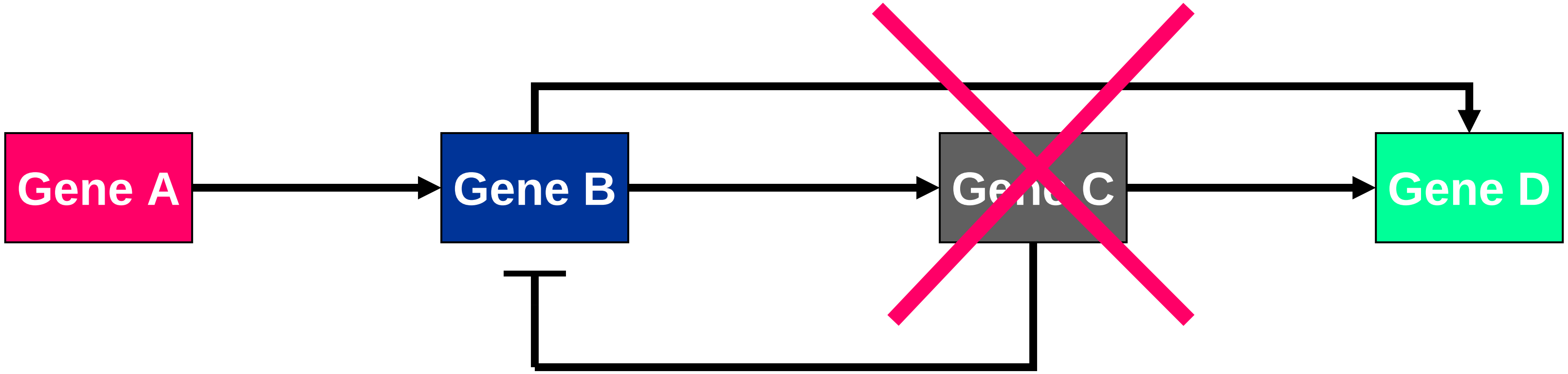
Retinoid acid pathway

Draft Pathways for Photoreceptor Function

Source: J. Yu, UM BioMedEng Thesis Proposal (2002)

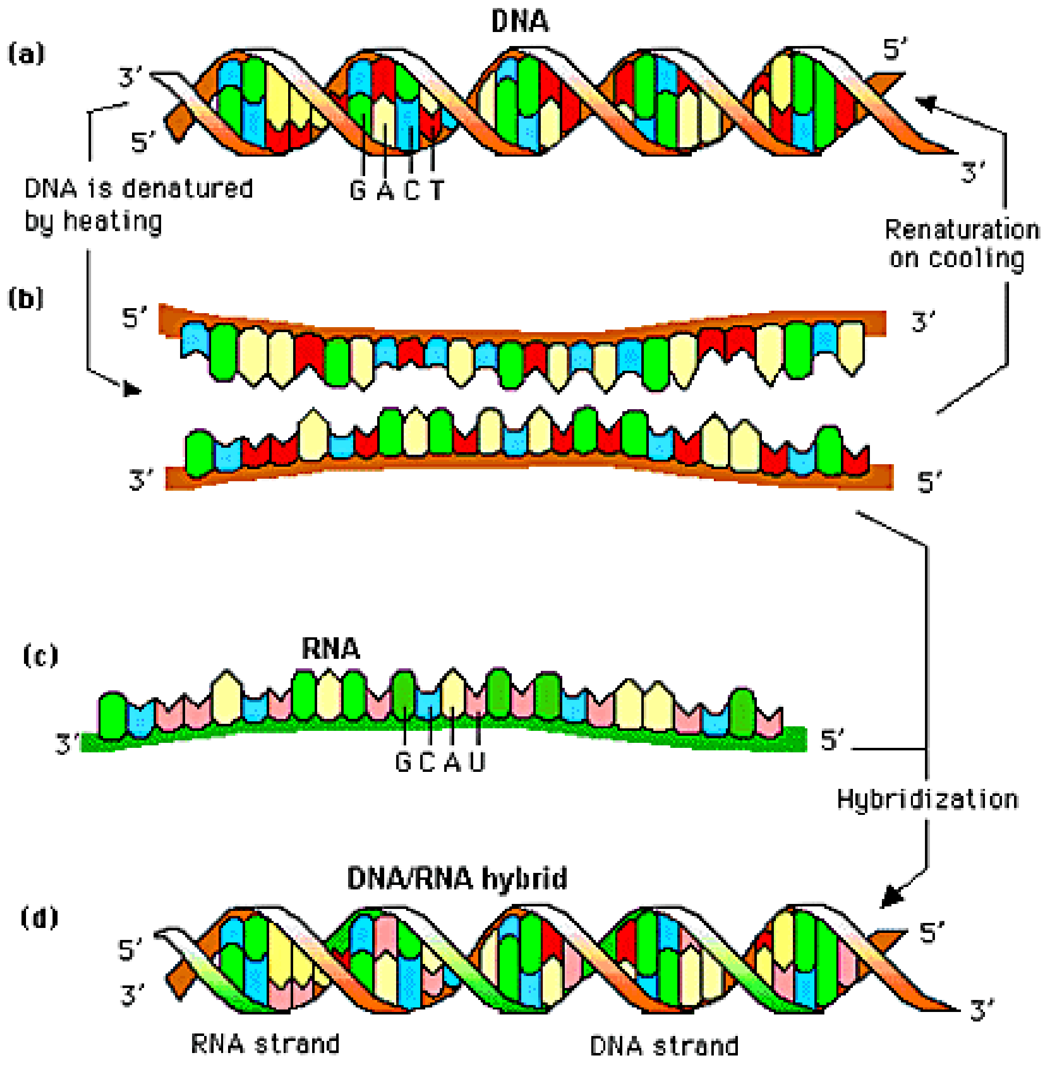
Experimental Design for Structure Discovery

- Treatment level experiments: aging, starvation, drugs
- Gene knockout experiments: create a mutant organism



- Issues:
  - For a network of G genes require 2^G knockouts per time point to explore full co-regulation network.
  - Experimental replication is necessary (“large p small n”)
  - There are other factors affecting gene expression: co-expression level,environment, protein-protein interactions…

Fundamental probing tool: hybridization

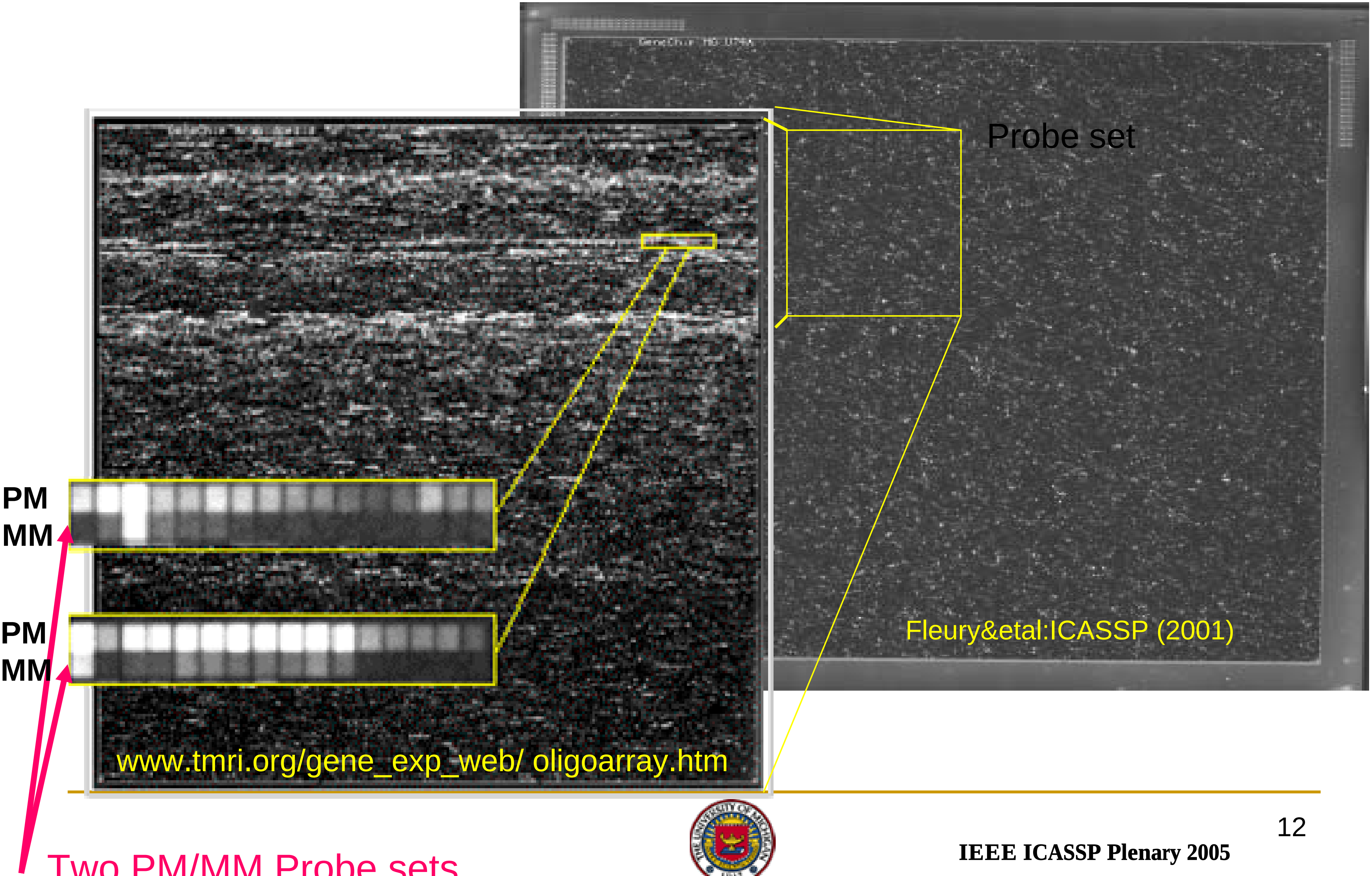


Nucleic Acid Hybridization

Gene Microarrays

- Two principal gene microarray technologies:
  - Oligonucleotide arrays: (Affymetrix GeneChips)
    - Matched and mismatched oligonucleotide probe sequences photoetched on a chip
    - Dye-labeled RNA from sample is hybridized to chip
    - Abundance of RNA bound to each probe is laser-scanned
  - cDNA spotted arrays: (Brown/Botstein)
    - Specific complementary DNA sequences arrayed on slide
    - Dye-labeled sample mRNA is hybridized to slide
    - Presence of bound mRNA-cDNA pairs is read out by laser scanner
- **10,000-50,000 genes can be probed simultaneously**

Oligonucleotide GeneChip (Affymetrix)



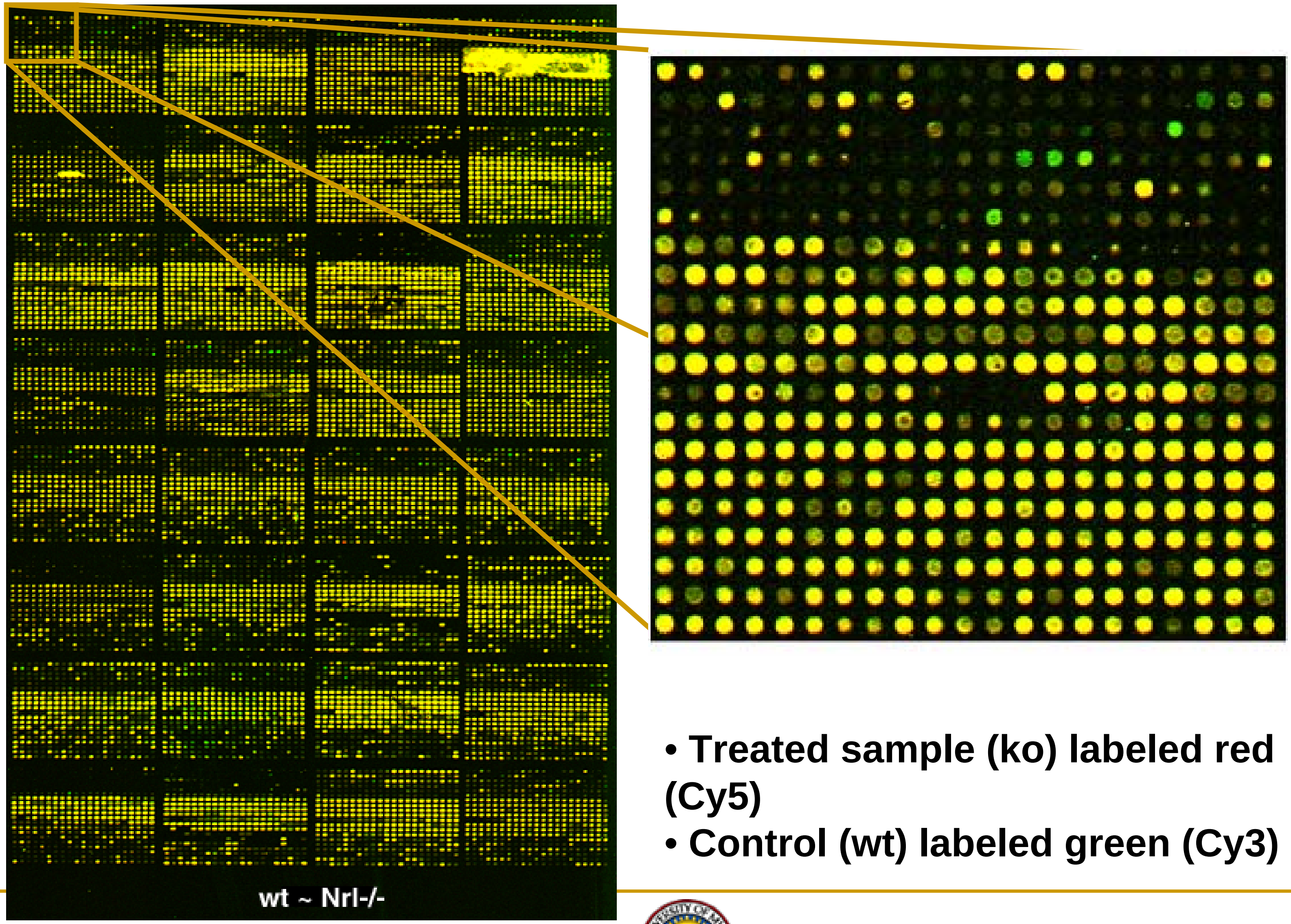
Probe set

Fleury&etal:ICASSP (2001)

www.tmri.org/gene_exp_web/ oligoarray.htm

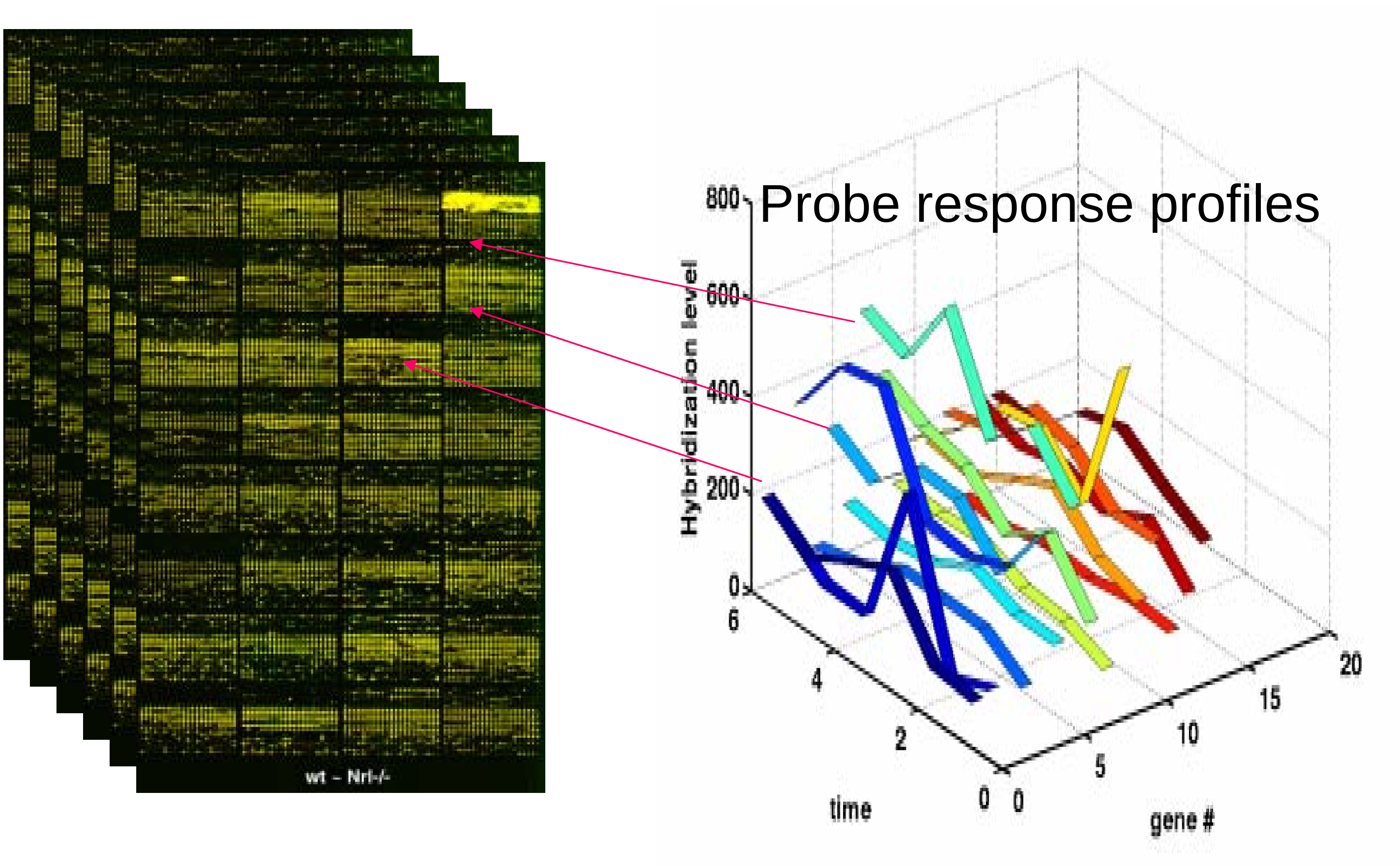
Two PM/MM Probe sets

cDNA spotted array



- Treated sample (ko) labeled red (Cy5)
- Control (wt) labeled green (Cy3)

Add Treatment Dimension: Expression Profiles


Probe response profiles

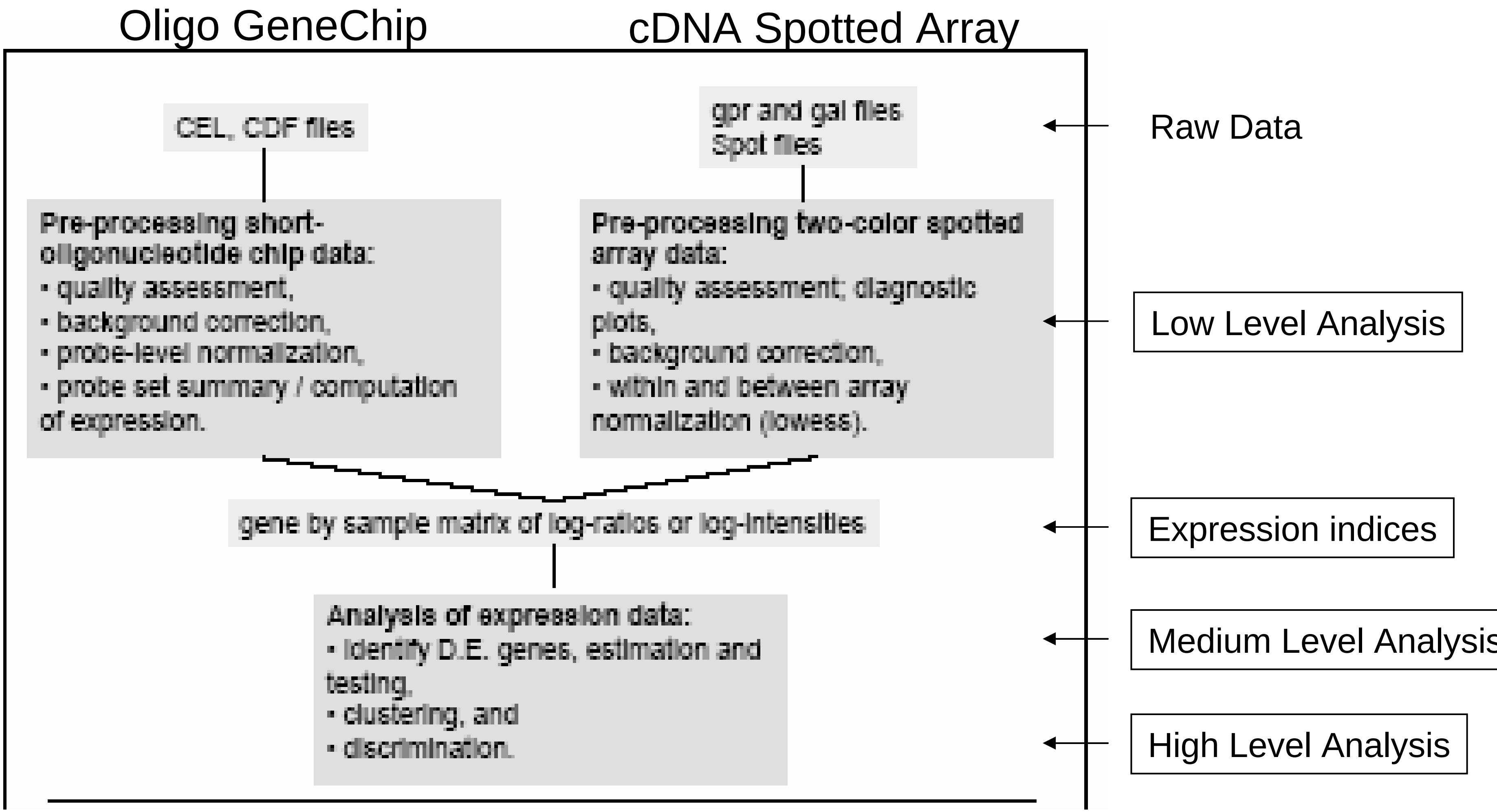
Sources of Experimental Variability

- **Population** – wide genetic diversity
- **Cell lines** - poor sample preparation
- **Slide Manufacture** – slide surface quality, dust deposition
- **Hybridization** – sample concentration, wash conditions
- **Cross hybridization** – similar but different genes bind to same probe
- **Image Formation** – scanner saturation, lens aberrations, gain settings
- **Imaging and Extraction** – misaligned spot grid, segmentation

Microarray data is intrinsically statistical and replication is necessary

Churchill (2002)

2. Preprocessing of Gene Microarray Data



Oligo GeneChip

cDNA Spotted Array

CEL, CDF files

gpr and gal files
Spot files

Pre-processing short-oligonucleotide chip data:
- quality assessment,
- background correction,
- probe-level normalization,
- probe set summary / computation of expression.

Pre-processing two-color spotted array data:
- quality assessment; diagnostic plots,
- background correction,
- within and between array normalization (lowess).

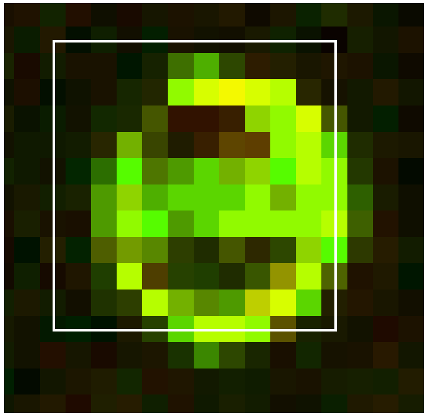
gene by sample matrix of log-ratios or log-intensities

Analysis of expression data:
- identify D.E. genes, estimation and testing,
- clustering, and
- discrimination.

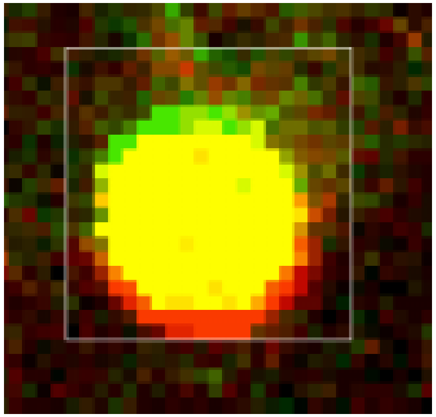
Source: Jean Yee Hwa Yang Statistical issues in design and analysis microarray experiment. (2003)

Image Processing: cDNA Spot Extraction

- **Addressing** – Locate “center of description” for each spot
- **Spot Segmentation** – Classification of pixels either as signal or background.
- **Spot Quantification** – Estimation of hybridization level/ratio of spot



Grid misalignment



Laser Misalignment

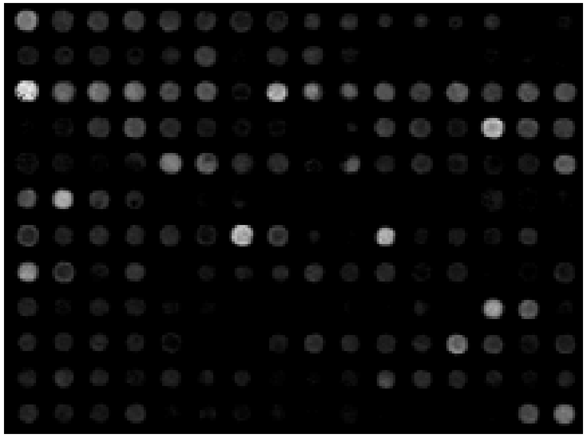
Source: C. Ball, Stanford Microarray Database

Refs: Spotfire, ScanAnalyze, GenePix, Quantarray, Spot

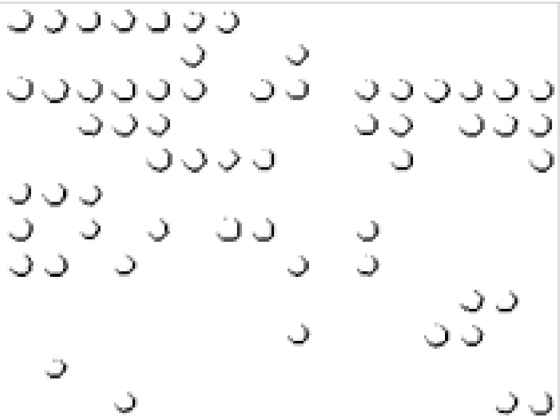
Spot Segmentation

- Threshold based
- Boundary based
  - Fixed circle
  - Adaptive circle (*used in QuantArray*)
  - Fixed Spot Mask (*used in ScanAlyze*)
- Region based
  - Seeded Region Growing (*used in Spot*)
- Active contours: level set algorithms
- Morphological operators: watershed segmentation

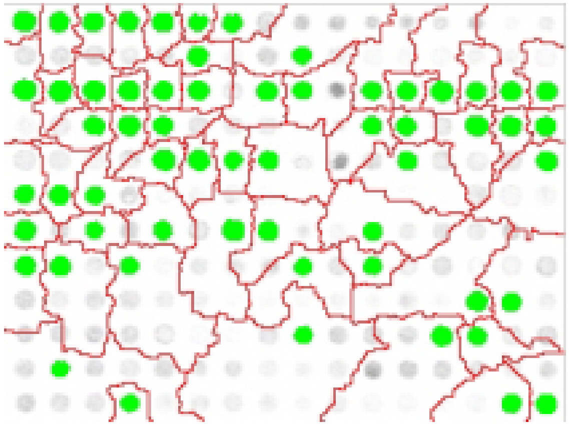
Segmentation via Morphological Operators



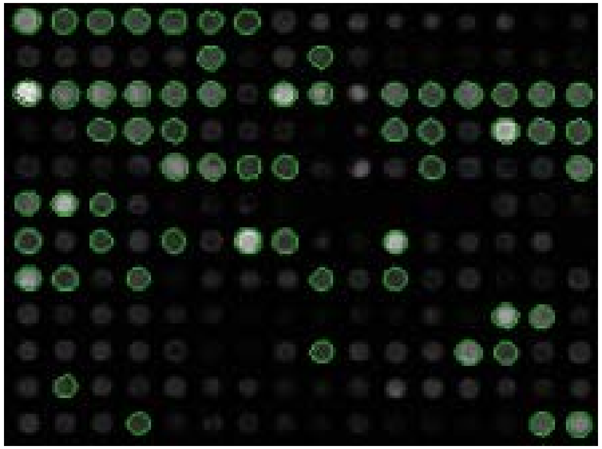
Original Image



Alternate-Sequential Filtered



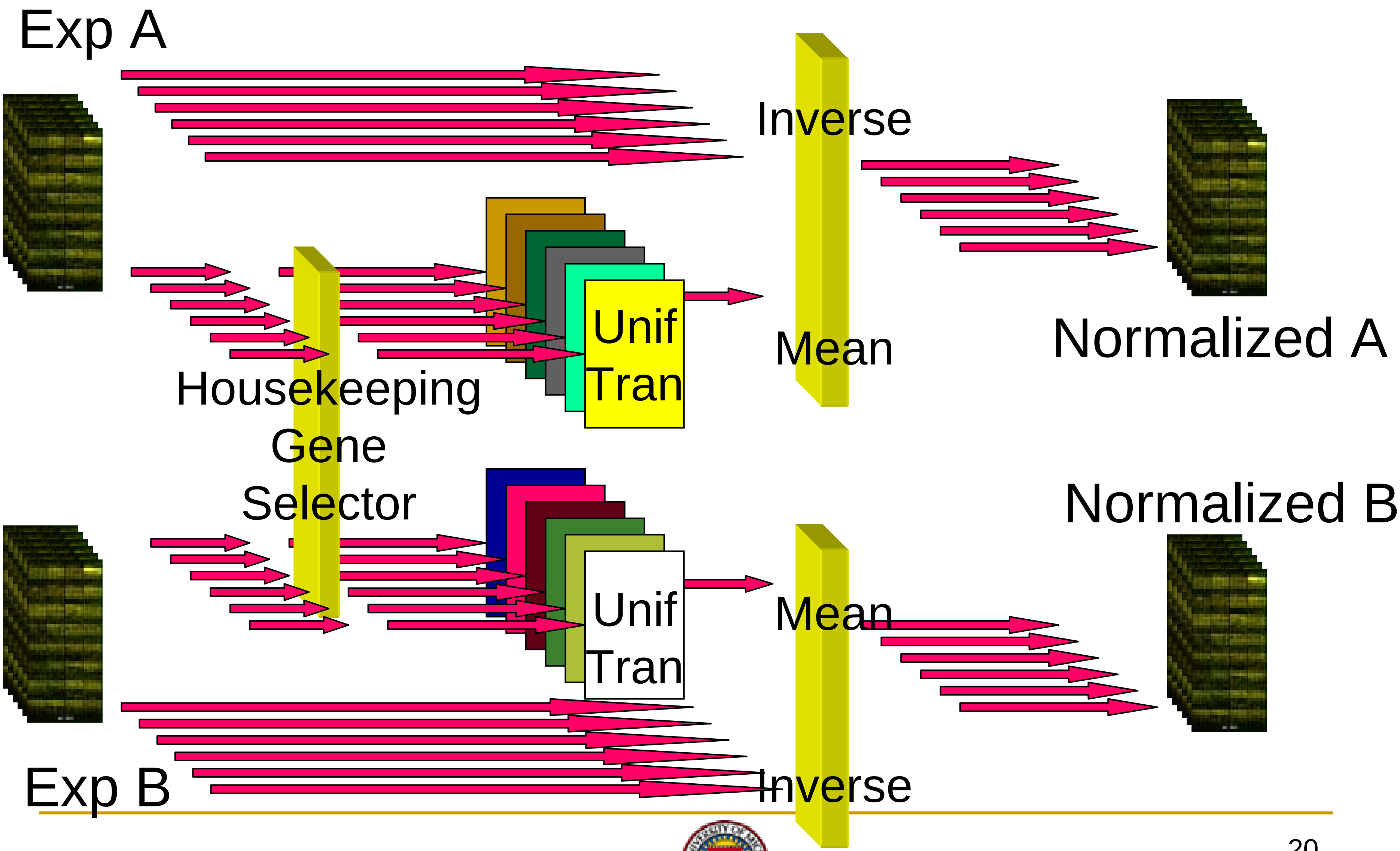
Watershed Transformed



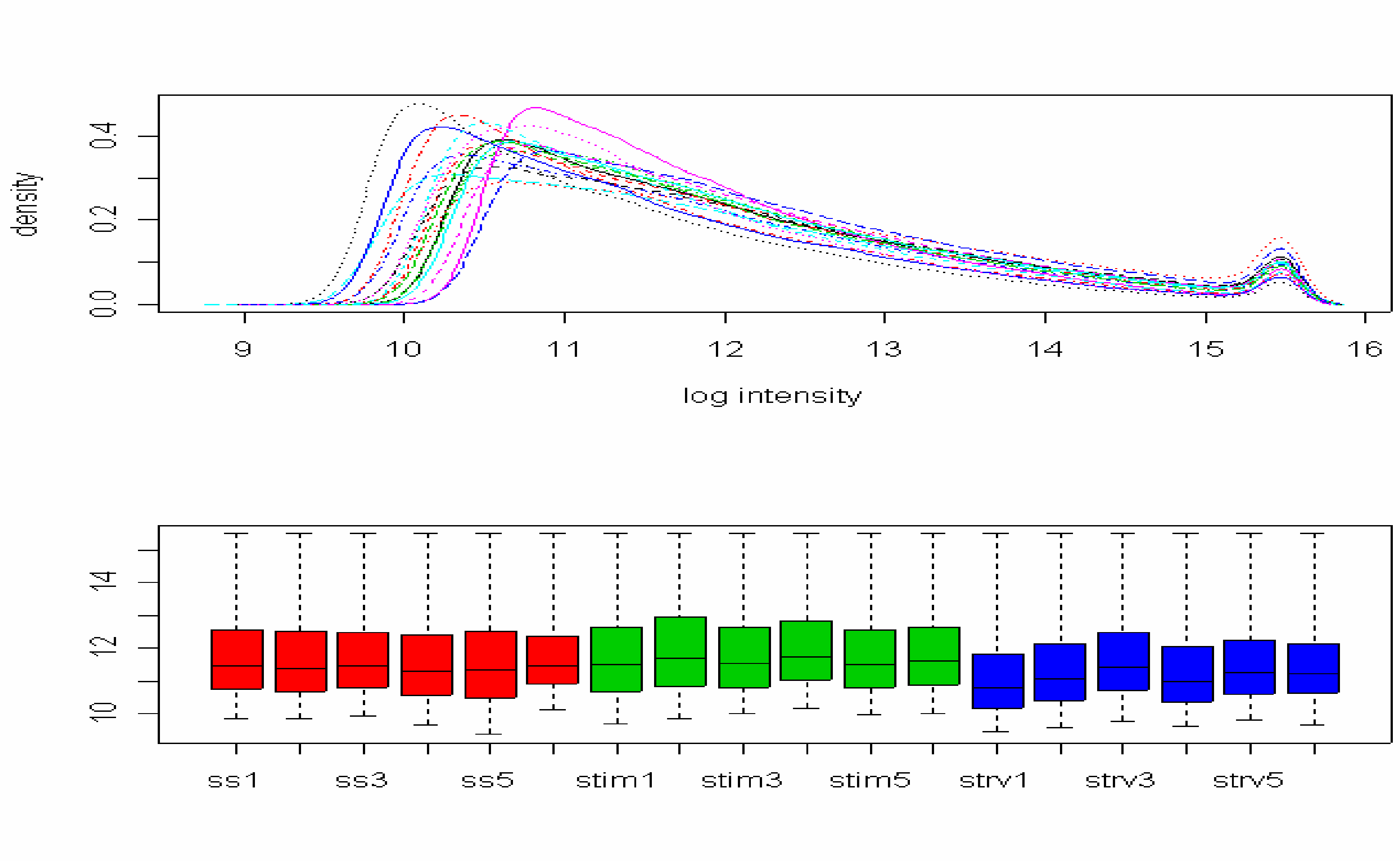
Final Segmented Image

Ref: Siddiqui&etal:Asilomar-02

A vs B Microarray Normalization



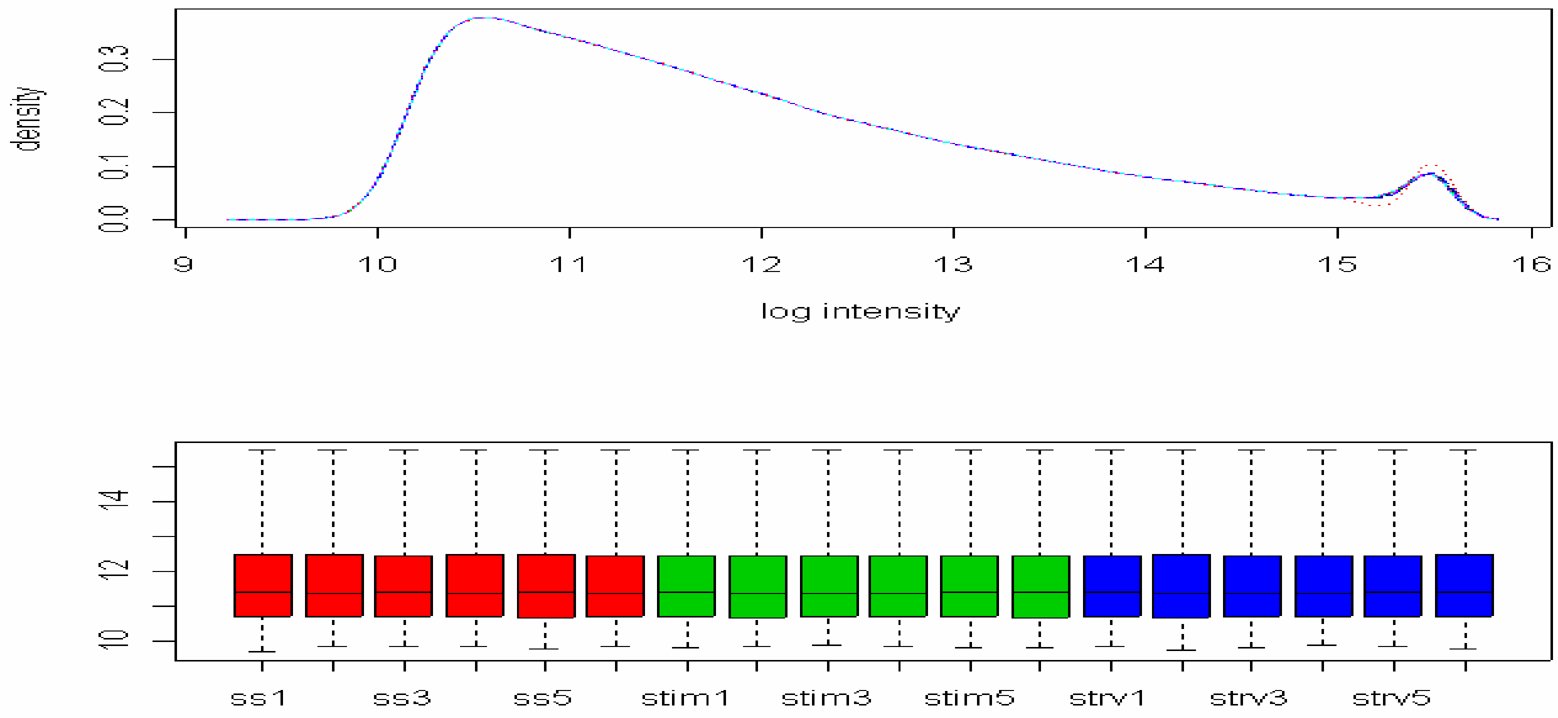
Pooled Microarray Normalization



Graphs are generated using R plot function hist() and boxplot()

Data: Lemon WJ et al. 2002

Post-Normalization Histogram



Graphs are generated using R plot function hist() and boxplot()

Data: Lemon WL et al. 2002

Extracting Expression Indices

- Each probe response level in microarray can be modeled via general mixed model

$$y_{gtr} = f_{gt}(\beta) + \rho_{gt}(\beta) Z_r + \sigma_{gt}(\beta)\epsilon_{gtr}$$

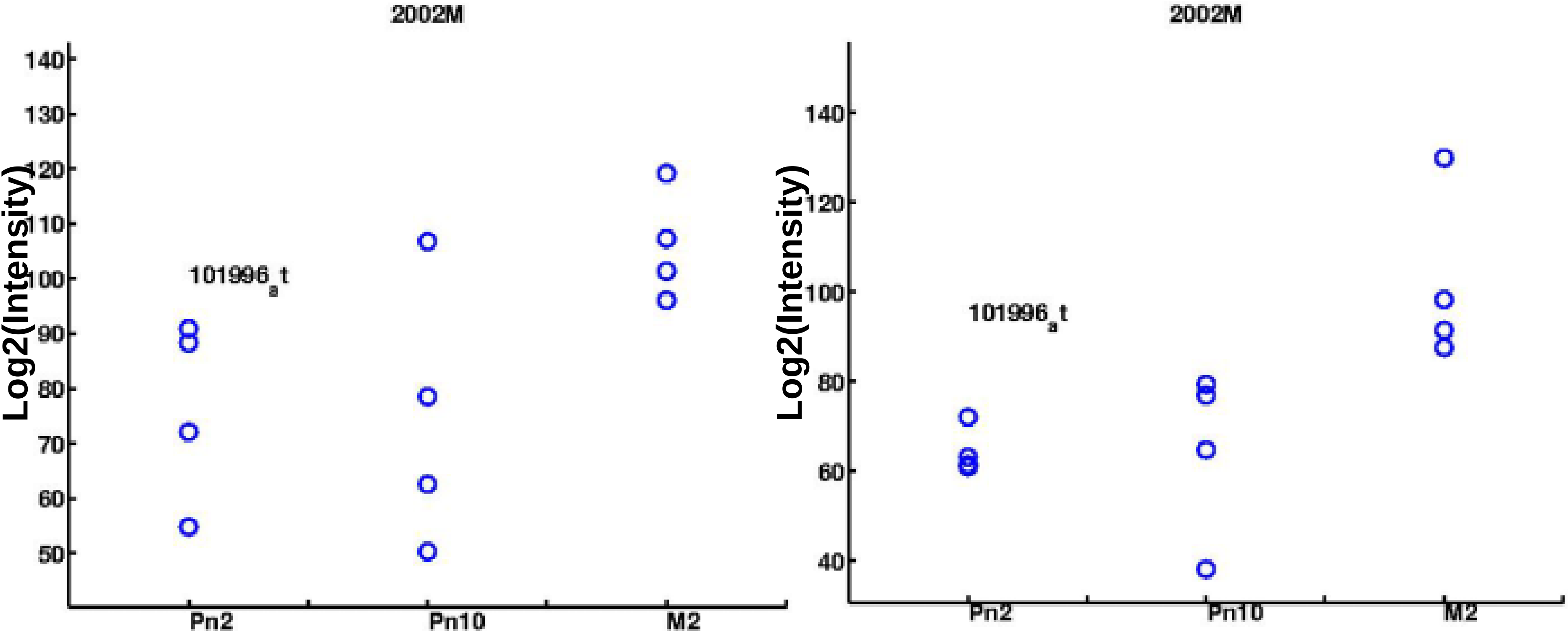
- g=gene probe index, t=timepoint, r=replicate
- $f_{gt}(\beta)$ is fixed effect
- $\sigma_{gt}(\beta) Z_r$ is random effect that may correlate t,g
- $\sigma_{gt}(\beta)\epsilon_{gtr}$ is noise component
- Special cases: MAS5, DChip, RMA. SMA, GEE
- Model similar to those used in array signal processing, statistical imaging, and other SP applications.

3. Screening Differentially Expressed Genes

12 knockout/wildtype mice in 3 groups of 4 subjects (24 GeneChips)

Knockout

Wildtype



$$\max_t\{\overline{K}_t(g) - \overline{W}_t(g)\} > \text{fcmin}$$

Biological vs Statistical Significance

- Biological significance refers to foldchange being sufficiently large to be biologically meaningful or testable, e.g. testable by RT-PCR

$$|\mathsf{fc}(g)| > \mathsf{fcmin}$$

- Statistical significance refers to foldchange being different from zero

$$\mathsf{fc}(g) \neq 0$$

Hero,Fleury,Mears,Swaroop:JASP2003

Single Comparison Test

- Let $\mathrm{fc}_t(g)$ = foldchange of gene ‘g’ at time point ‘t’.
- We wish to test the hypotheses:

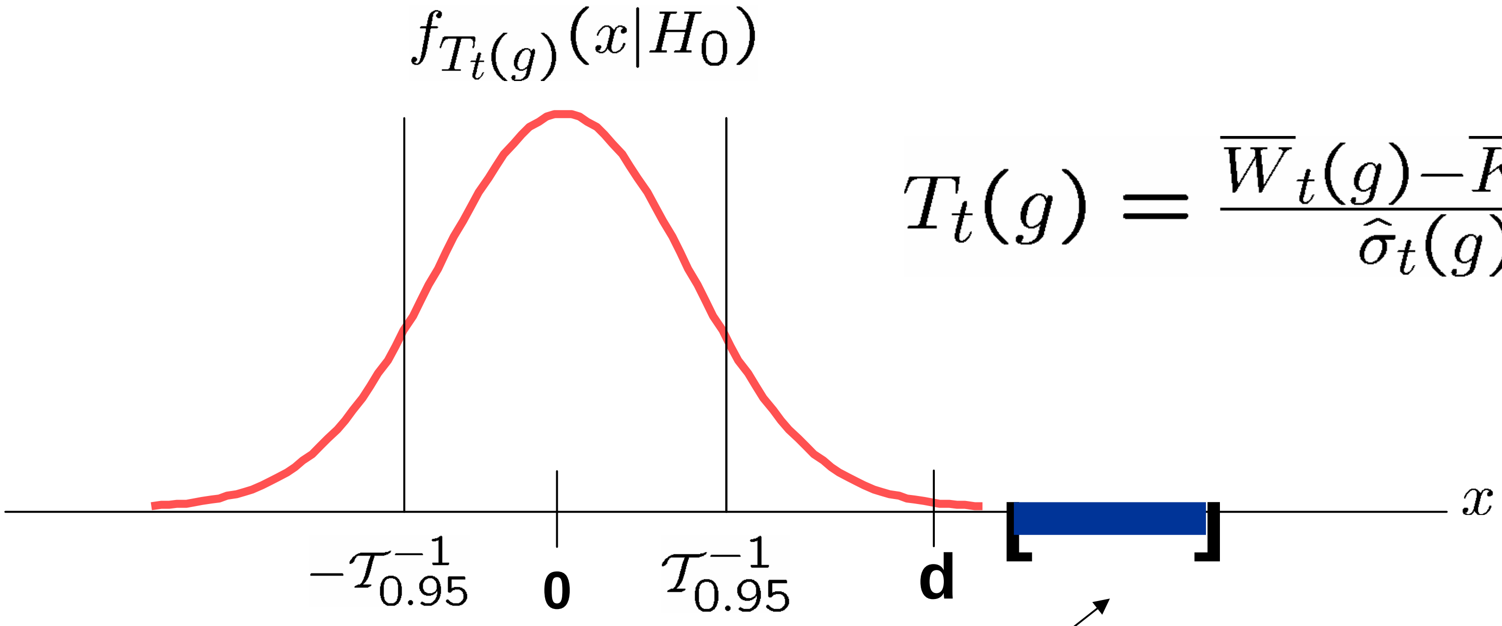
$$\begin{aligned} H_0(g,t) &: \quad |\mathsf{fc}_t(g)| \leq |d| \\ H_1(g,t) &: \quad |\mathsf{fc}_t(g)| > |d| \end{aligned}$$

- d = minimum acceptable difference (MAD)
- Method: confidence interval test

Hero,Fleury,Mears,Swaroop:JASP2003

Confidence Interval Test: Single Comparison

- Biologically&statistically **significant** differential response at 10% level of significance

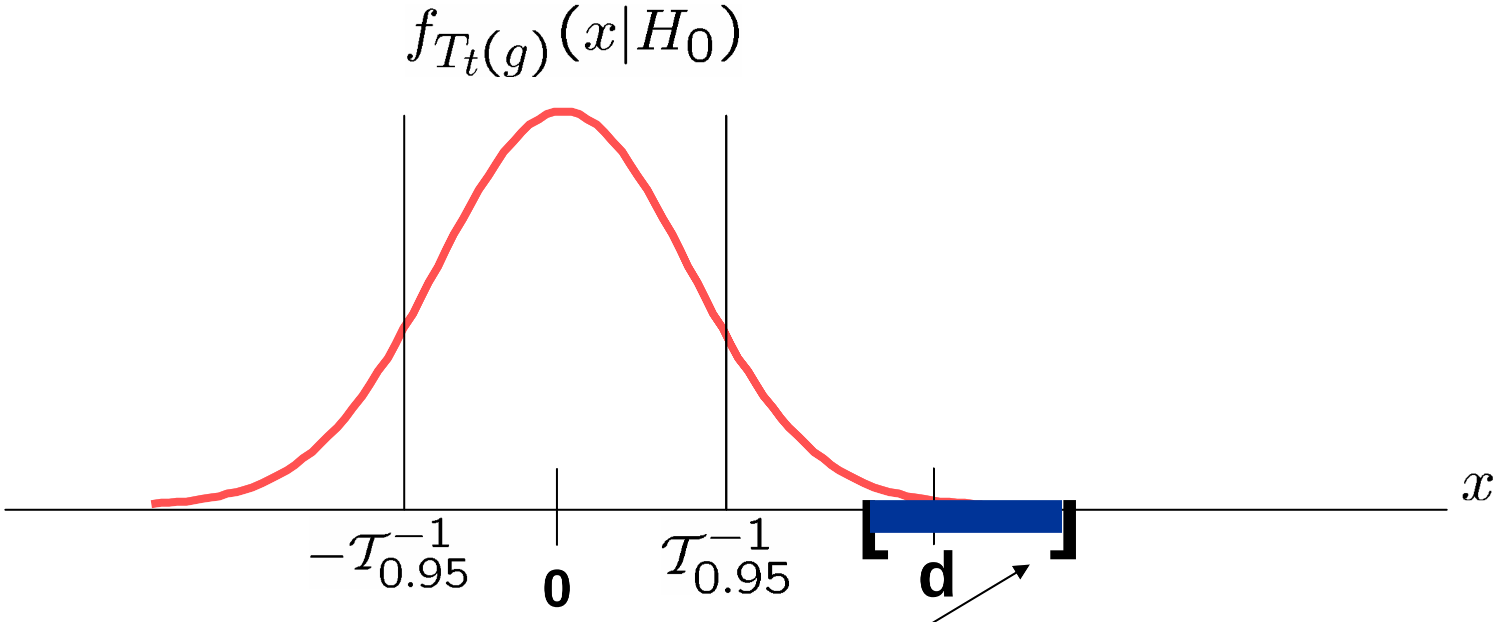


$$T_t(g) = \frac{\overline{W}_t(g) - \overline{K}_t(g)}{\hat{\sigma}_t(g)}$$

Conf. Interval on $\mathsf{fc}_t(g)$ of level 1-alpha

Confidence Interval Test: Single Comparison

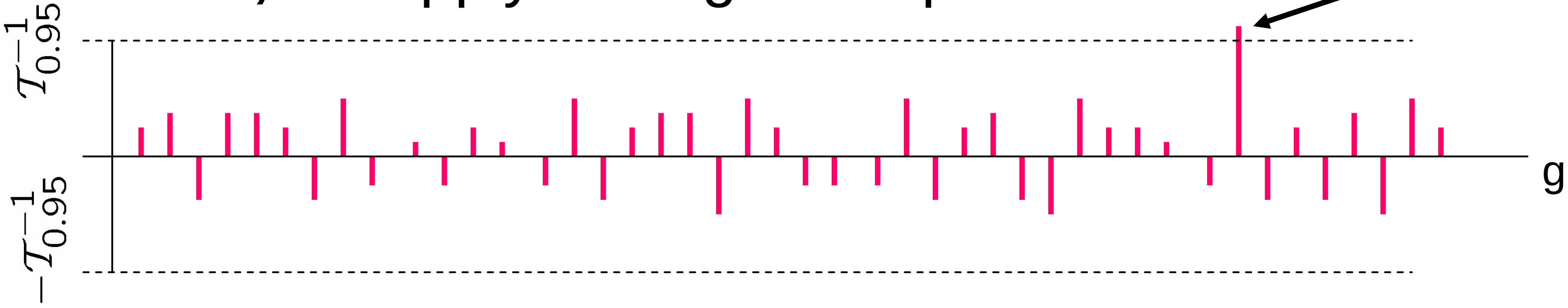
- Statistically significant but biologically **insignificant** fc



Conf. Interval on $\mathrm{fc}_t(g)$ of level 1-alpha

Multiple Comparisons: FWER, FDR

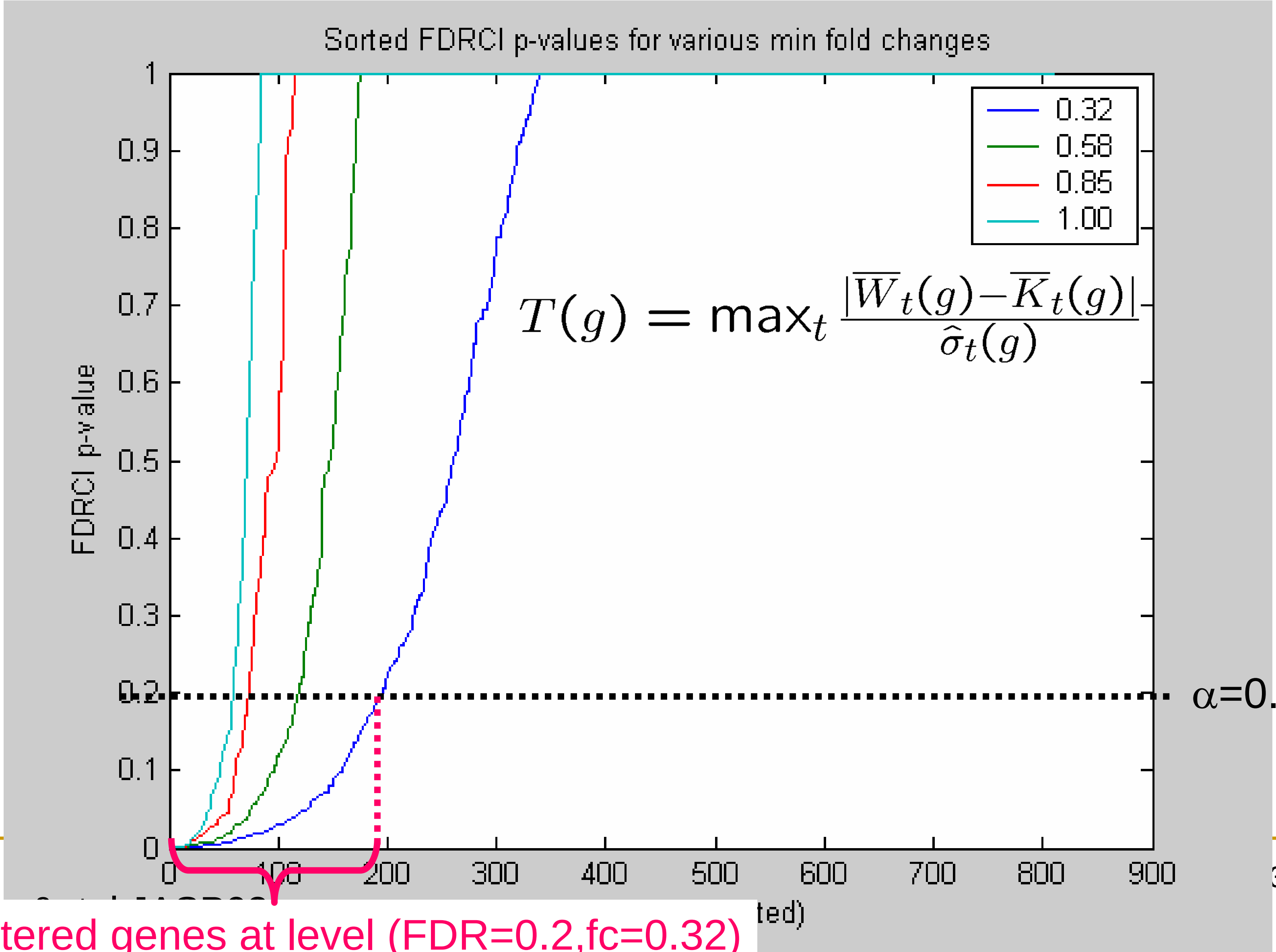
- **Pvalue, CI** apply to single comparison:



False positive

- **FWER, FDR** and **FDRCI** depend on **{T(g), g=1, … G}**.
  - FWER: familywise error rate
    - Avg number of experiments yielding at least one false positive
  - FDR: false discovery rate (Benjamini&Hochburg:1996)
    - Avg proportion of false positives in experiments
  - FDRCI: (1-α) CI on discovered fc (Benjamini&Yekutieli:2002)
    - Avg. proportion of CIs that cover true fc in a given experiment

Sorted FDRCI pvalues for ko/wt study



$$T(g) = \max_t \frac{|\overline{W}_t(g) - \overline{K}_t(g)|}{\hat{\sigma}_t(g)}$$

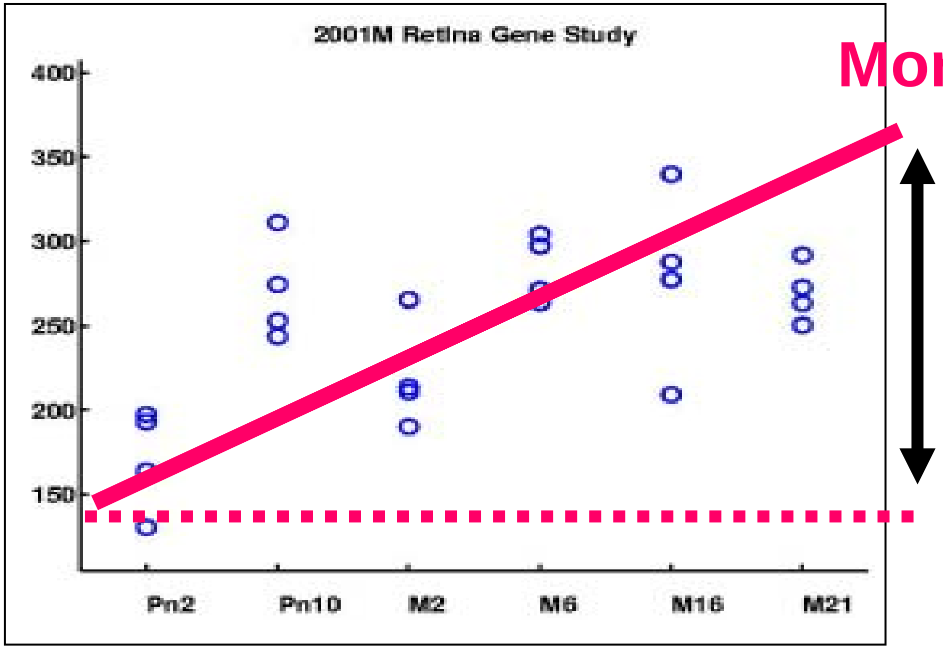
α=0.2

Filtered genes at level (FDR=0.2,fc=0.32)

Ref:

Screening Gene Expression Profiles

- Max foldchange is only one possible criterion of interest
- Objective: find the 250-300 genes having the most significant foldchanges wrt multiple criteria
- Example: Retinal aging study



Monotonic?

Large end-to-end foldchange?

Multi-objective Optimization Approach

- Rarely does a linear order exist with respect to more than one ranking criterion, as in

$$|\mathsf{fc}_1(g_1)| > |\mathsf{fc}_1(g_2)| > \ldots > |\mathsf{fc}_1(g_p)|$$

- However, a partial order is usually possible

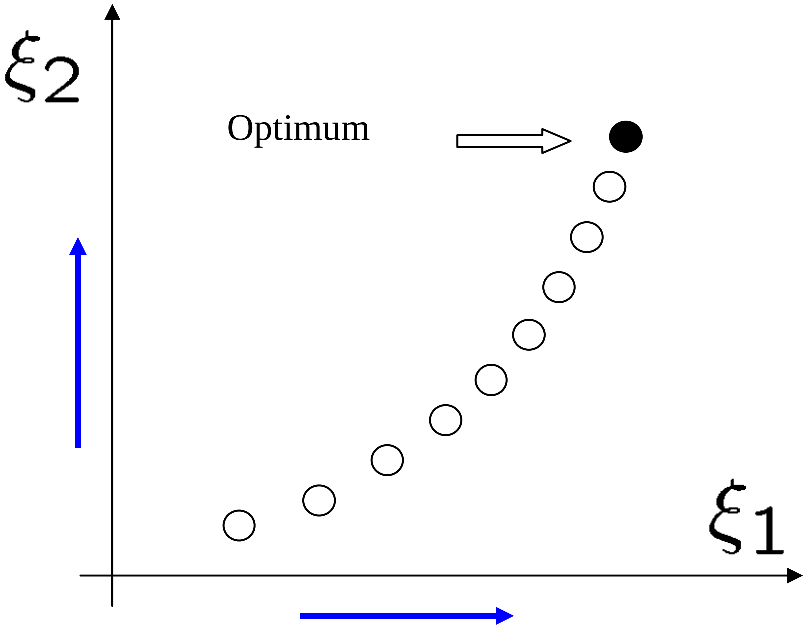
$$\{\mathsf{fc}_1(g), \ldots, \mathsf{fc}_6(g)\}_{g \in \mathcal{G}_1} > \ldots > \{\mathsf{fc}_1(g), \ldots, \mathsf{fc}_6(g)\}_{g \in \mathcal{G}_q}$$

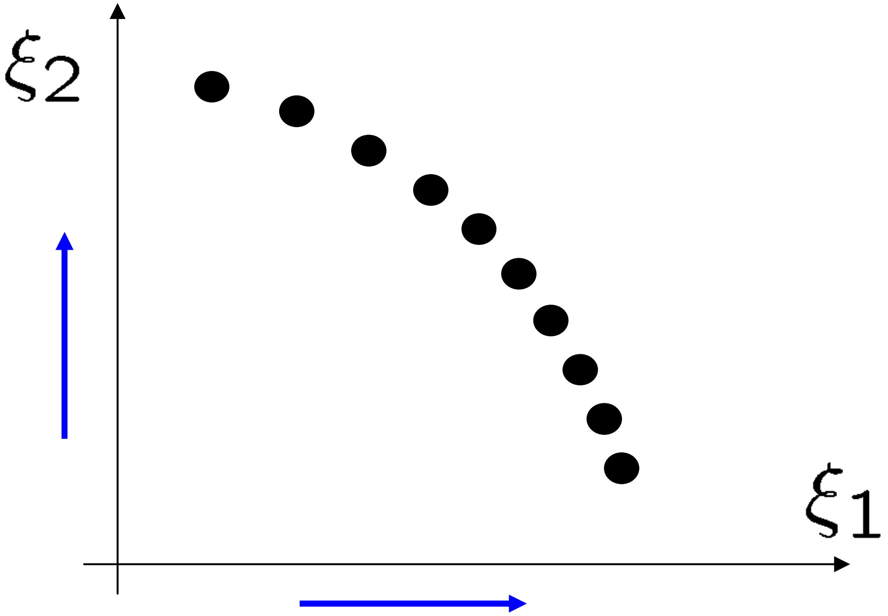
Illustration: two extreme cases

$\xi_1(g) = fc_6(g) - fc_1(g)$ - end-to-end criterion

$\xi_2(g) = \min_t\{fc_t(g) - fc_{t-1}(g)\}$ -increasing criterion

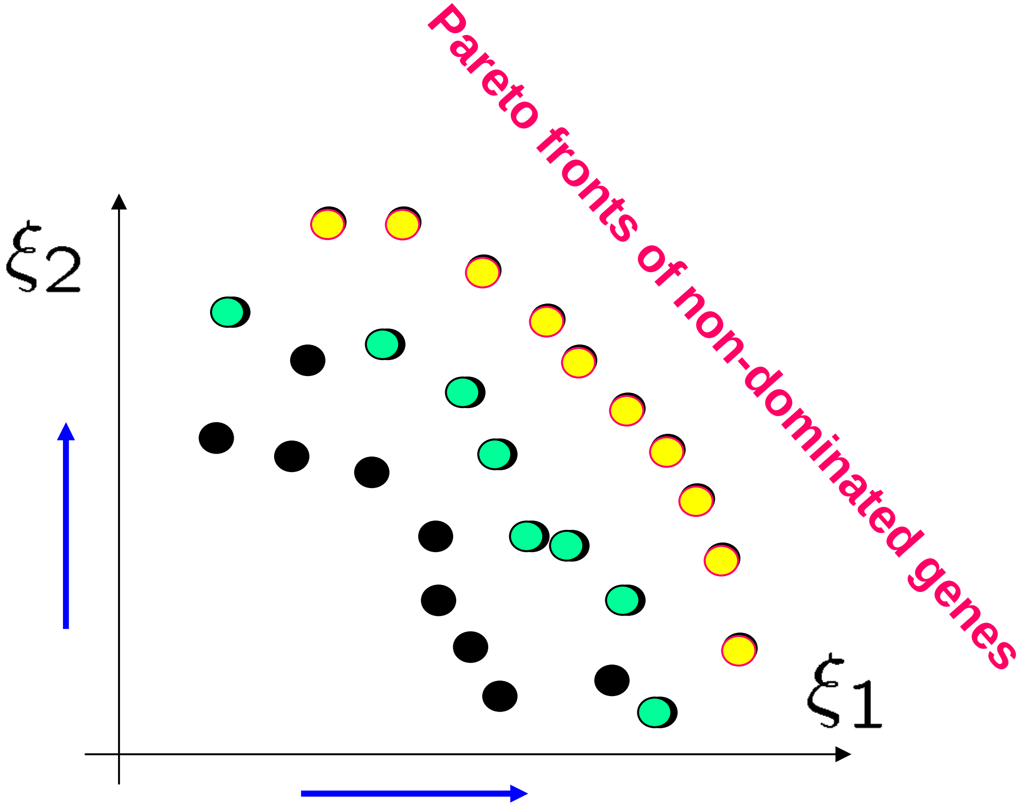
- A linear ordering exists
- No linear ordering exists



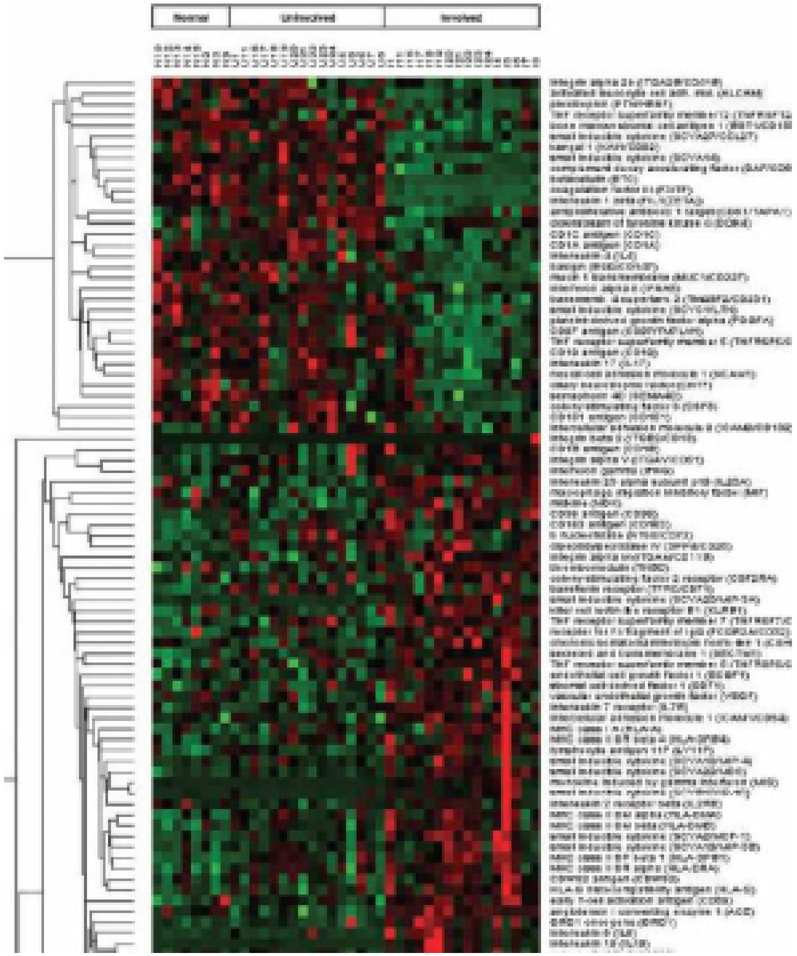


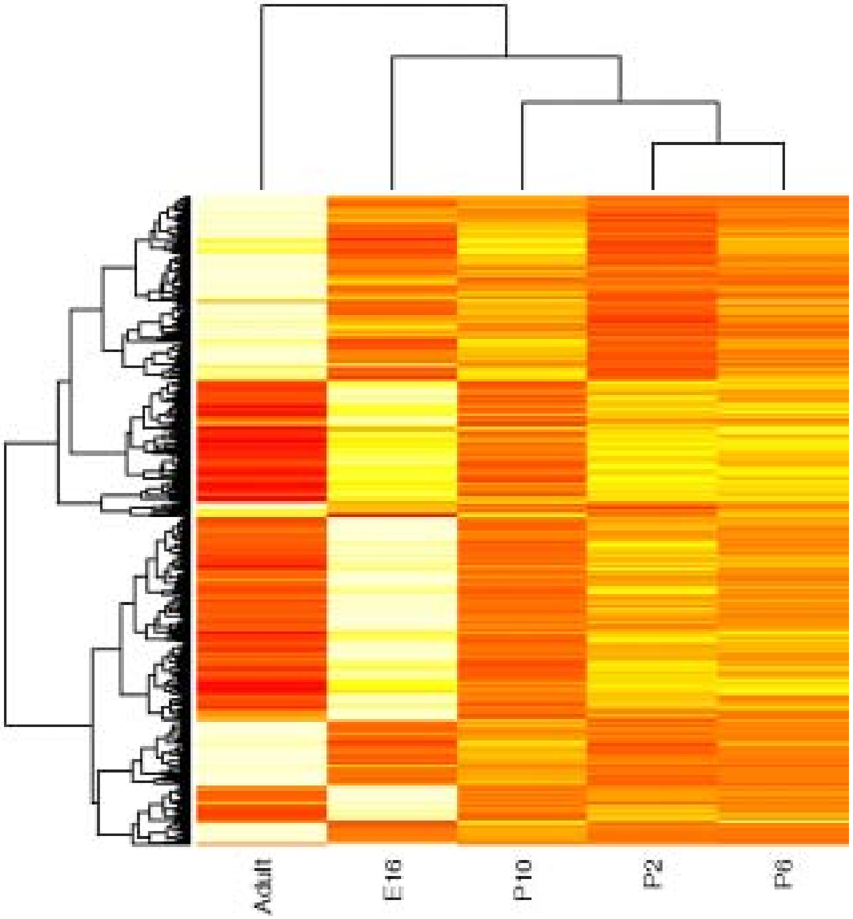
Pareto Front Analysis (PFA)

- Rank genes by peeling of successive Pareto Fronts



4. Clustering Gene Expression Patterns





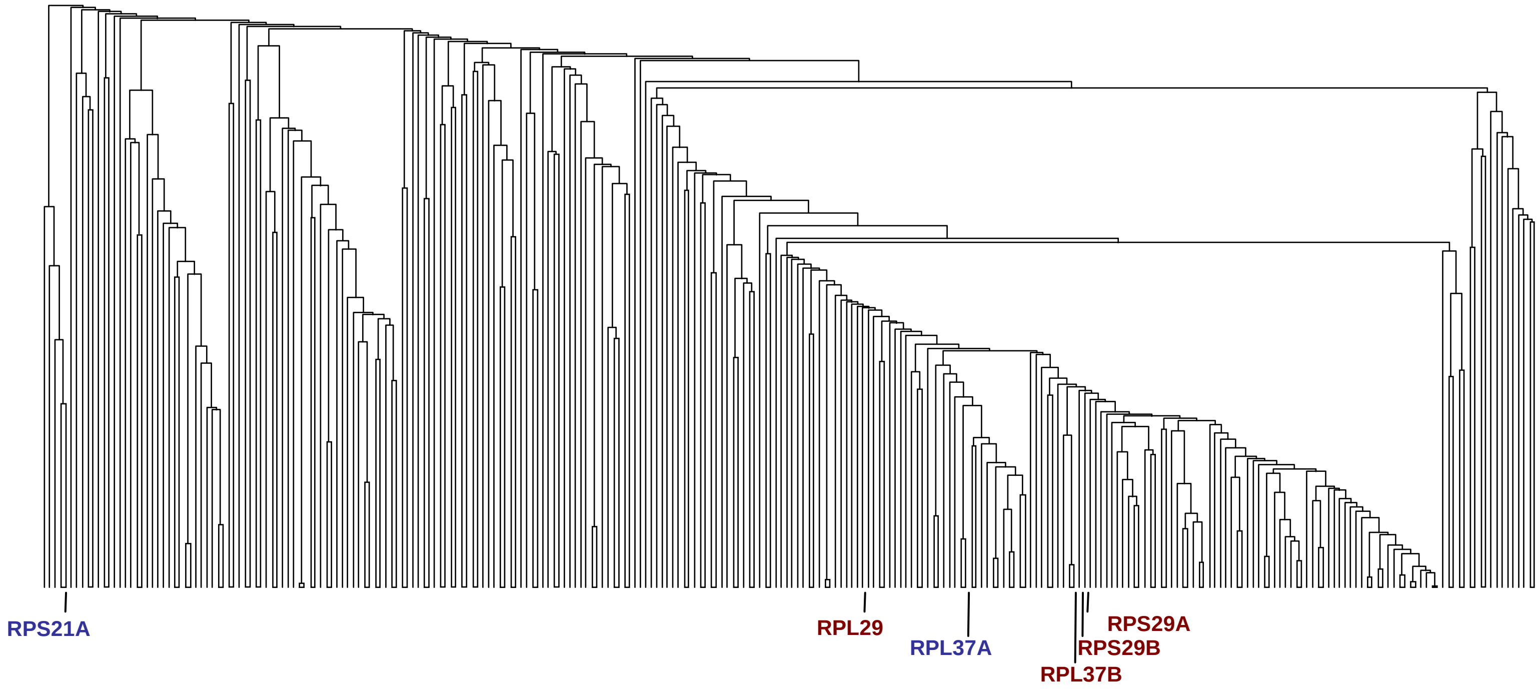
- Gene expression levels over multiple conditions are required for pathway studies
- Requires symmetric similarity metric, e.g pairwise profile correlation

Source: Wing Wong Lab, Stanford (left)

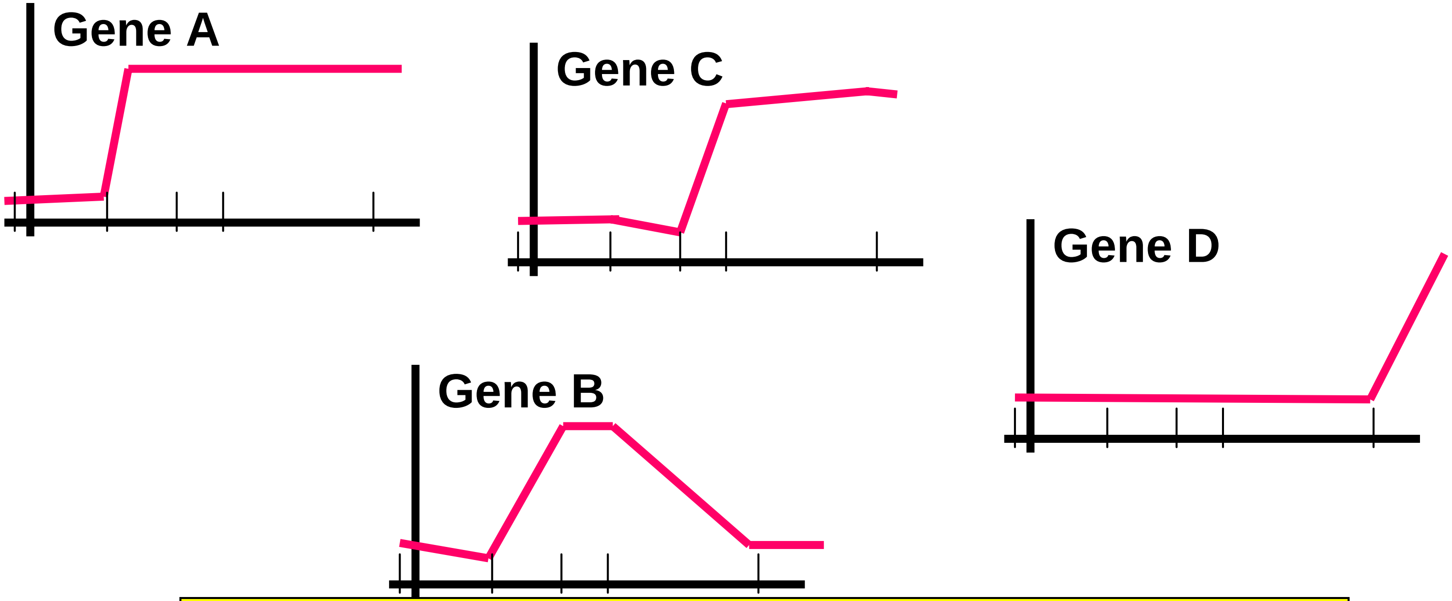
Swaroop Lab, Michigan (right)

Drawback of Traditional Clustering

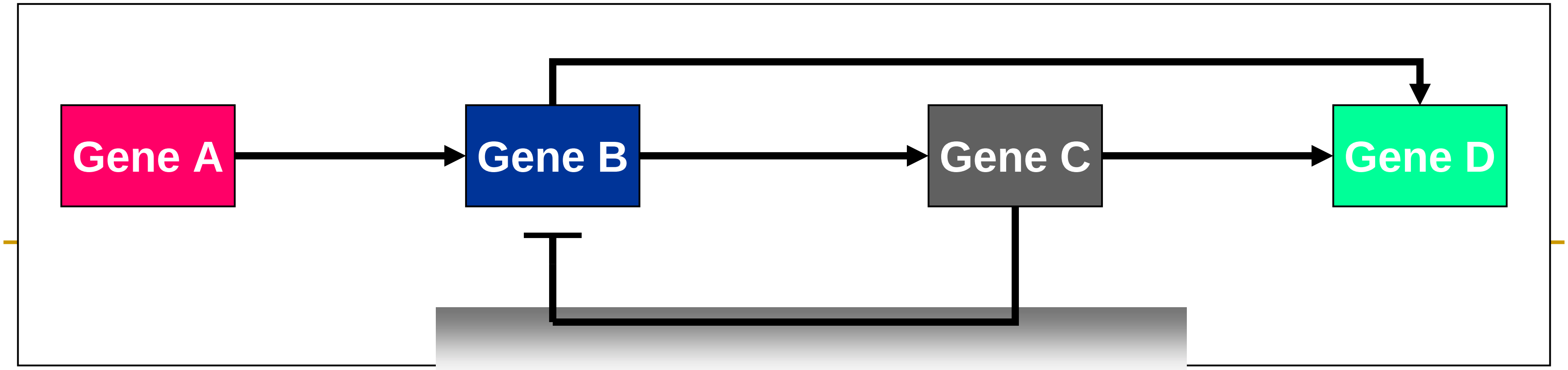
- Clustering using pairwise correlation fails to account for transitive co-expression (Zhou etal 2002)



Extraction of Co-Regulation Circuits



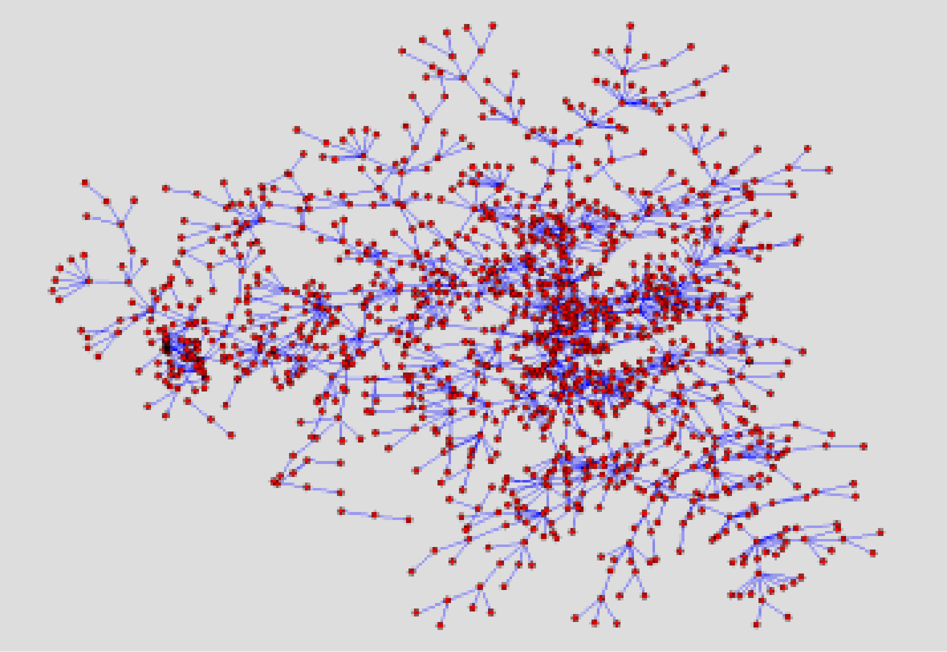
$$p(\mathcal{X}) = \prod_{gt} p(x_{gt} | \mathcal{X}_{gt}^{-})$$



Modeling co-Regulation Networks

- Relevance networks
  - Edge = strong correlation
- Dependency networks
  - Directed edge = strong partial correlation
- Dynamical dependency networks
  - Directed edge = strong partial correlation
- Bayesian networks
  - Profiles are quantized to small number of bits
  - One bit quantization = boolean networks

Network Constrained Clustering



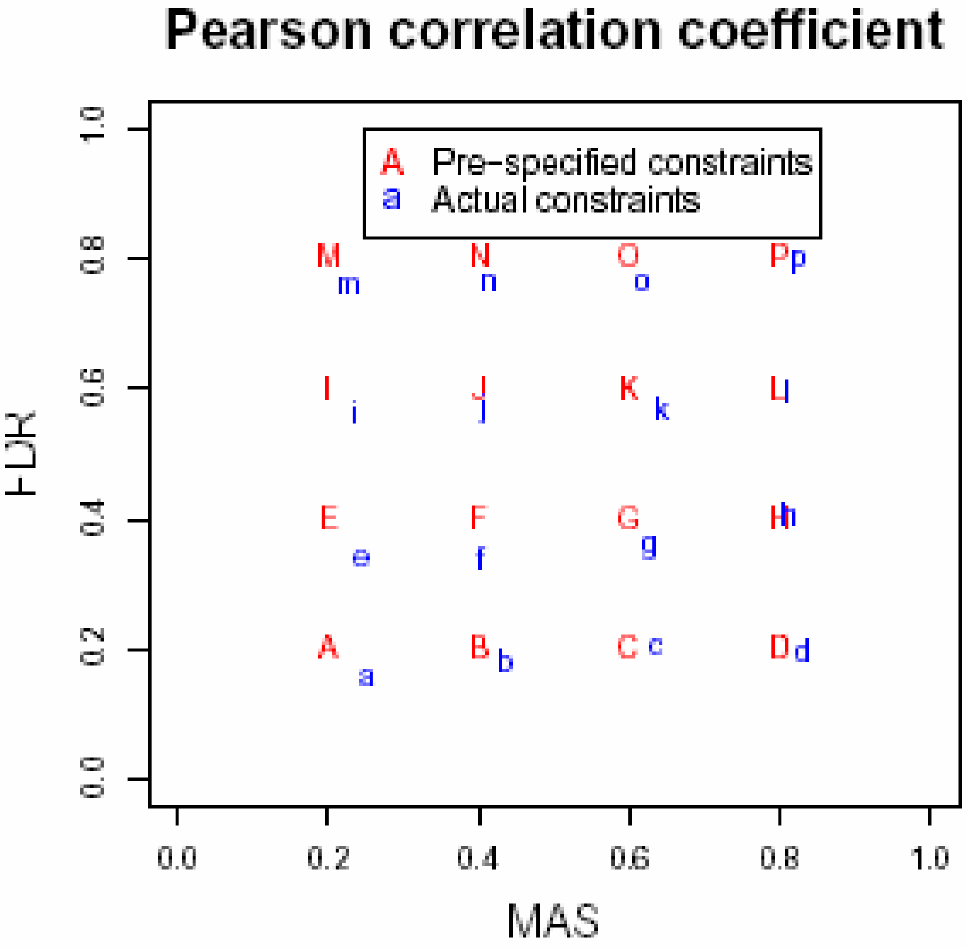
- If topology were known could use to improve clustering
- Otherwise suffer from combinatorial explosion:

$$p = 2^{\binom{G}{2}}$$

- Soln: FDRCI edge screening

FDRCI Edge Screen Procedure

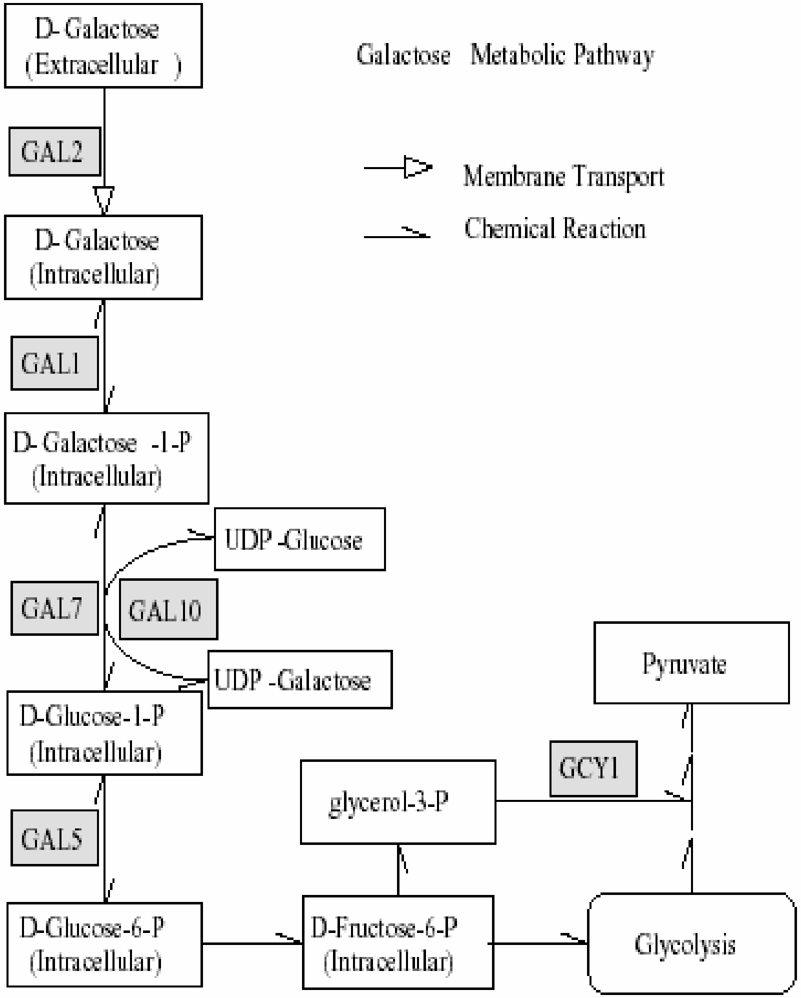
- Fix FDR level and MAS level on discovered edges
- Construct FDRCI's of desired FDR level on edge strengths
- Accept edge if FDRCI exceeds MAS



Dongxiao Zhu, A. Hero, S. Qi, JCB, 2004.

Yeast Galactose Metabolism Experiment

- 10 different yeast strains (9 gene knock-outs and 1 wild type) incubated in either GAL-inducing or non-inducing media (Ideker et al. 2001).
- 9 gene knock-outs are GAL1, GAL2, GAL3, GAL4, GAL5, GAL6, GAL7, GAL10, GAL80.
- 5935 gene 2-channel cDNA array. Reference channel is dilution "wild-type + galactose"

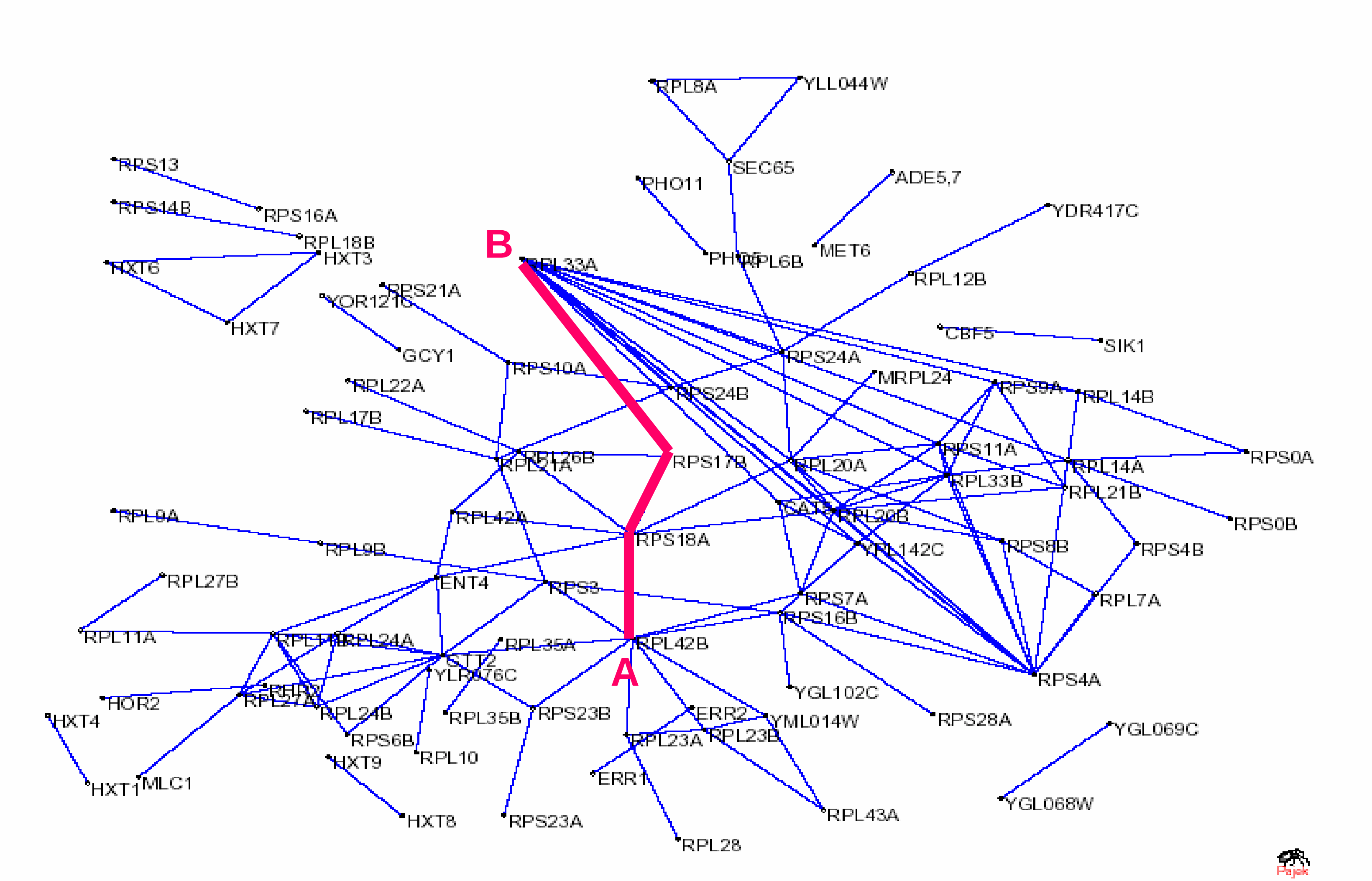


D- Galactose (Extracellular)
Galactose Metabolic Pathway
GAL2
Membrane Transport
Chemical Reaction
D- Galactose (Intracellular)
GAL1
D- Galactose -1-P (Intracellular)
UDP -Glucose
GAL7
GAL10
UDP -Galactose
Pyruvate
D-Glucose-1-P (Intracellular)
GCY1
glycerol-3-P
GAL5
D-Glucose-6-P (Intracellular)
D-Fructose-6-P (Intracellular)
Glycolysis

Dongxiao Zhu, A. Hero, S. Qi, JCB, 2004.

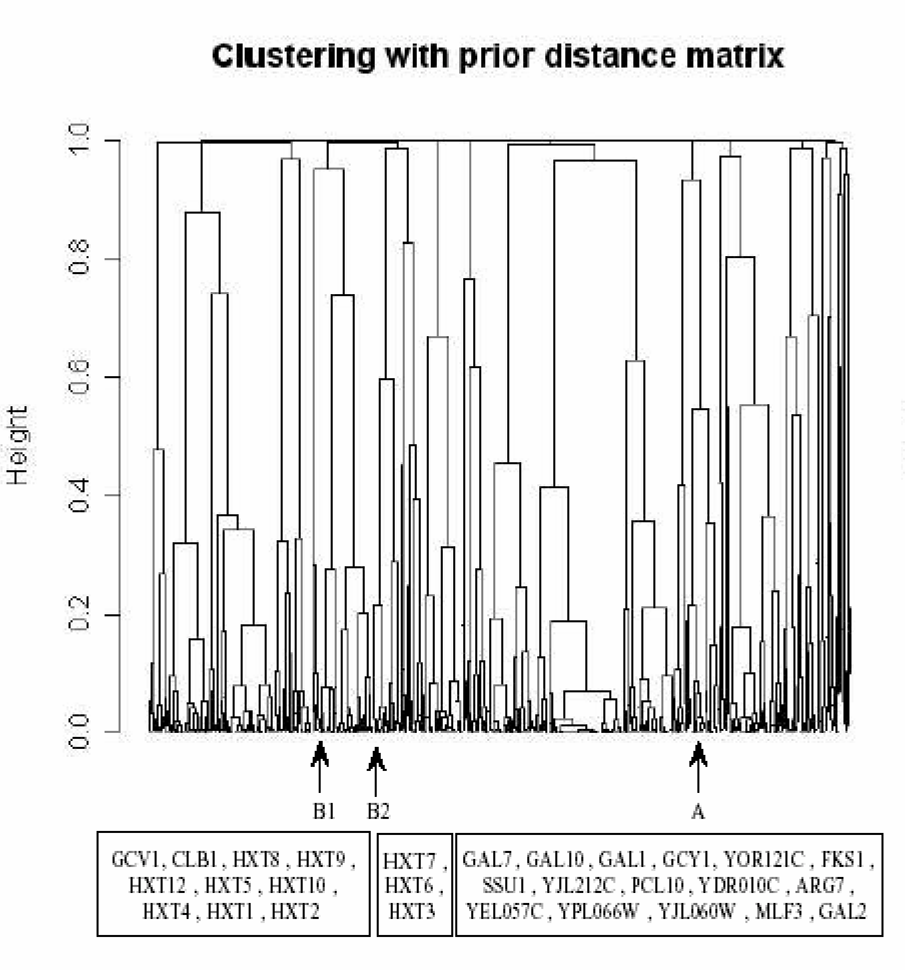
Relevance Network Visualization

(FDR <= 0.05, MAS = 0.9)



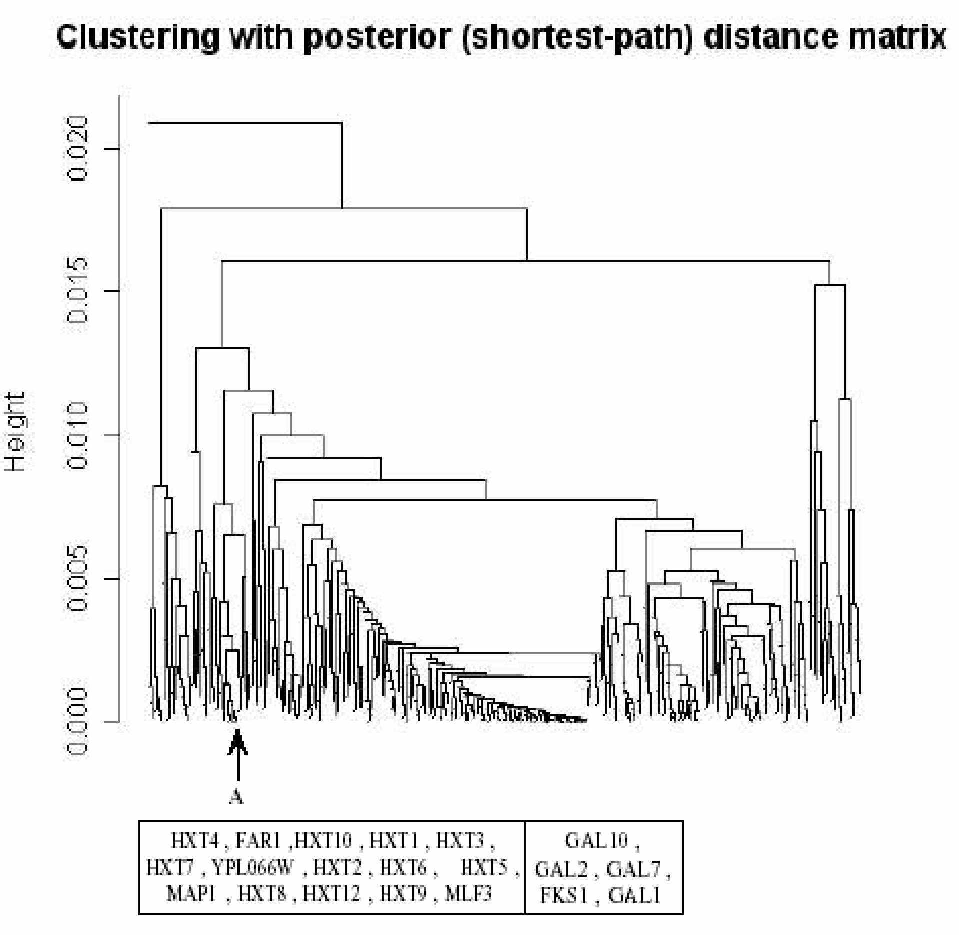
Dongxiao Zhu, A. Hero, S. Qi, JCB, 2004.

Network Constrained Clustering



Clustering with prior distance matrix

GCV1, CLB1, HXT8, HXT9, HXT12, HXT5, HXT10, HXT4, HXT1, HXT2	HXT7, HXT6, HXT3	GAL7, GAL10, GAL1, GCY1, YOR121C, FKS1, SSU1, YJL212C, PCL10, YDR010C, ARG7, YEL057C, YPL066W, YJL060W, MLF3, GAL2



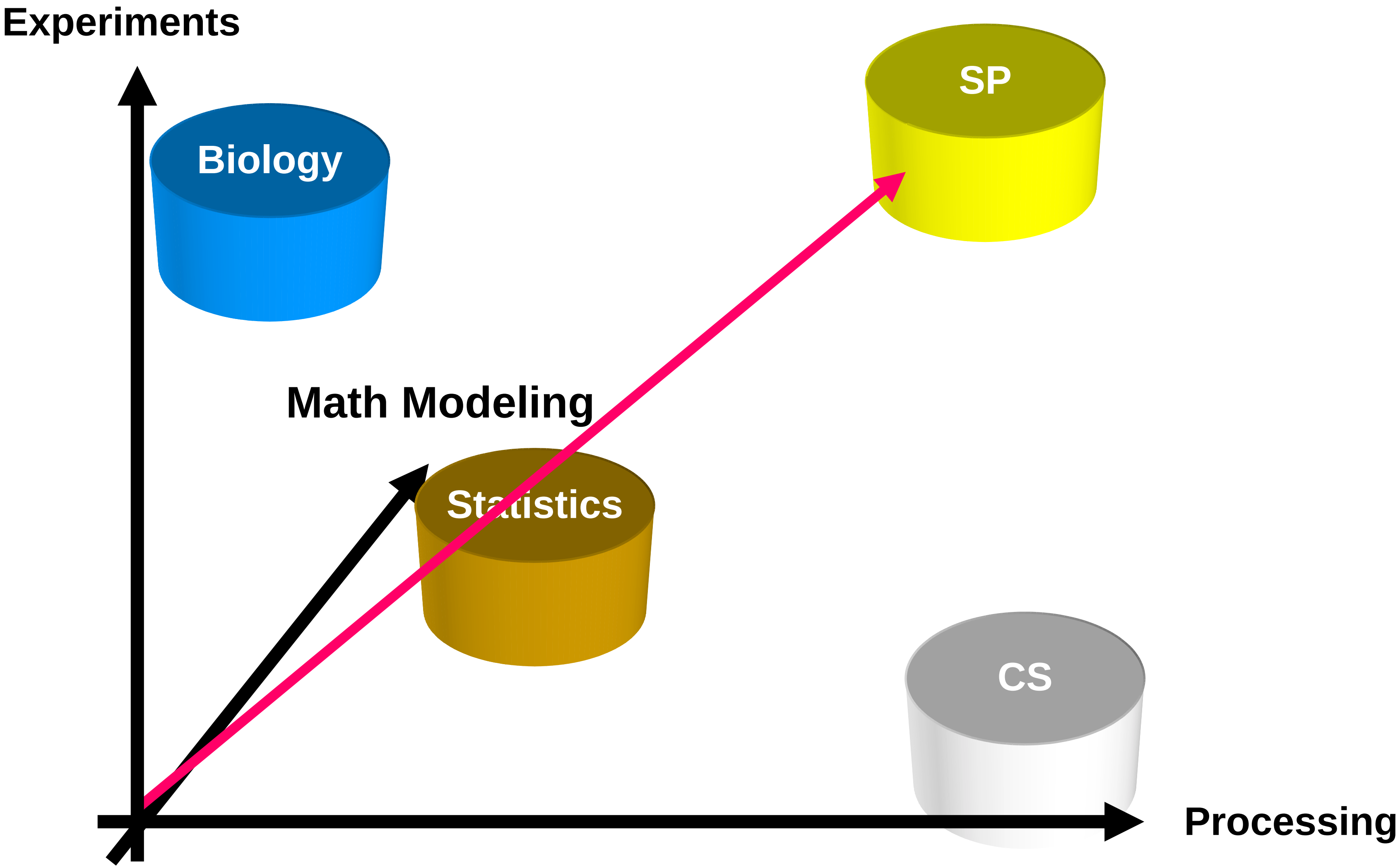
Clustering with posterior (shortest-path) distance matrix

HXT4, FAR1, HXT10, HXT1, HXT3, HXT7, YPL066W, HXT2, HXT6, HXT5, MAP1, HXT8, HXT12, HXT9, MLF3	GAL10, GAL2, GAL7, FKS1, GAL1

Horizons: Transcriptomics/Proteomics Technology

- Higher throughput cDNA/GeneChip microarrays
- Suspension microarrays
- Microscale “Lab on a chip”
- Protein-protein arrays
- Nuclear magnetic resonance spectroscopy
- In vivos molecular imaging: reporter genes

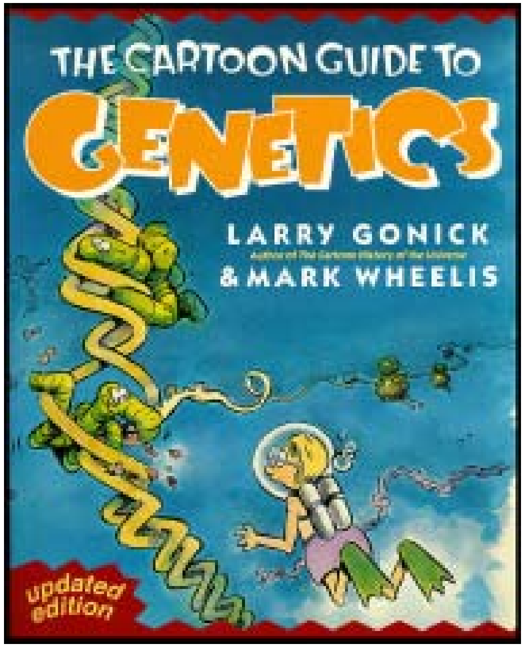
Where does SP fit?



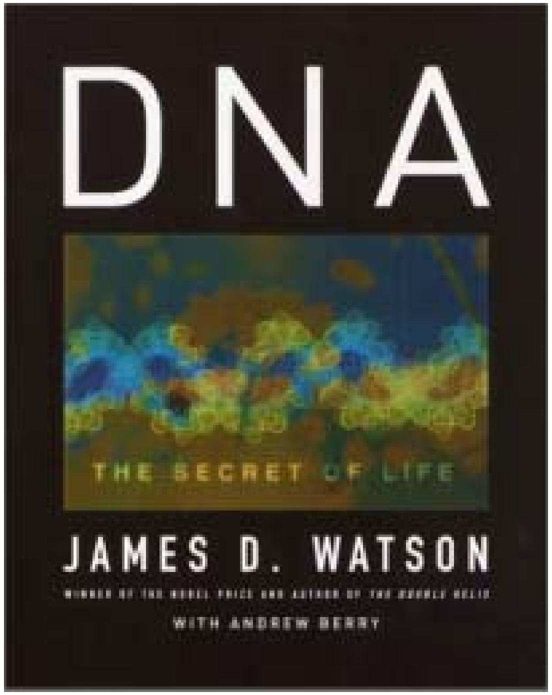
Signal Processing Opportunites

- There is room for new SP approaches
  - Non-modularized analysis: task-driven and top-down?
  - Active waveform design: sequential design of experiments?
  - Internet Tomography: gene network topology discovery?
  - MIMO: spatio-temporal wideband array processing?
  - Channel optimization: optimal gene layout on microarray?
- New technology is appearing that offers opportunities for SP'ers to develop models/algorithms
- There is still some low lying fruit!
- Collaboration with a biological scientist is essential in order to have impact

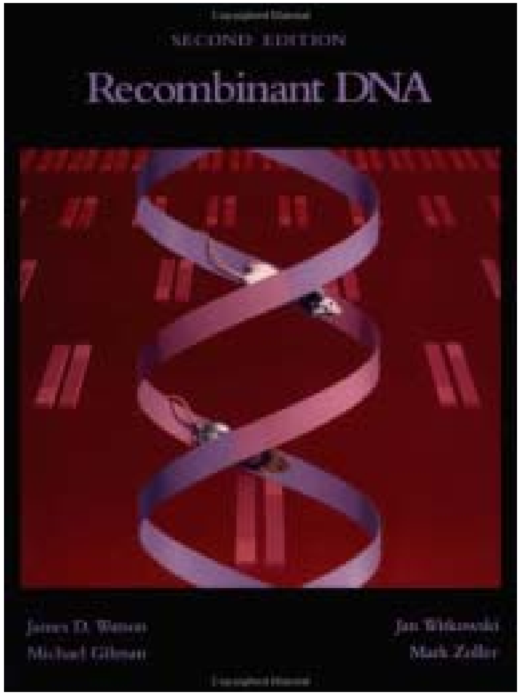
Where to learn more?



Genetics, the painless way, 1991

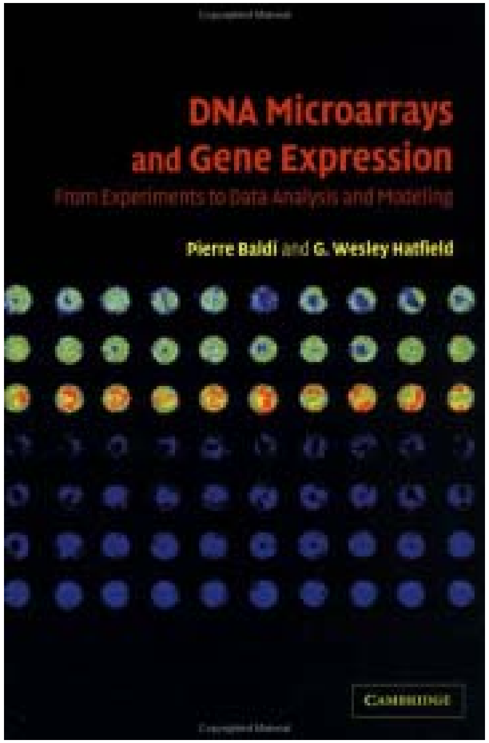


Historical overview by one of the pioneers, 2003

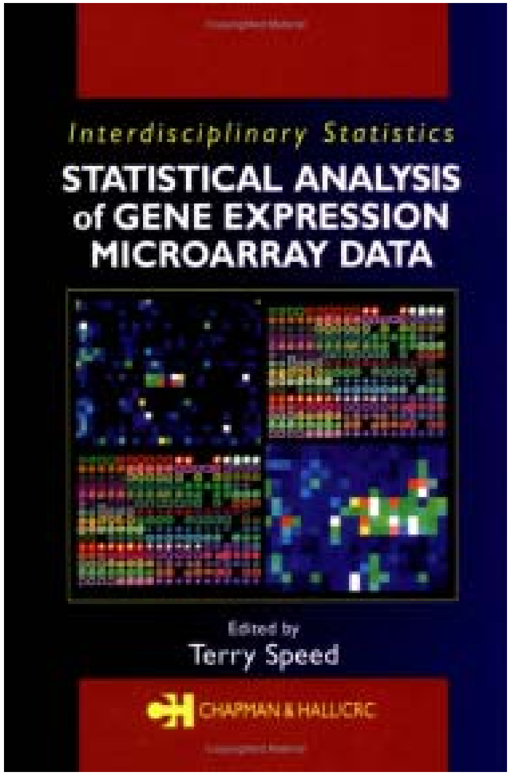


Basic undergraduate texbook, 1992

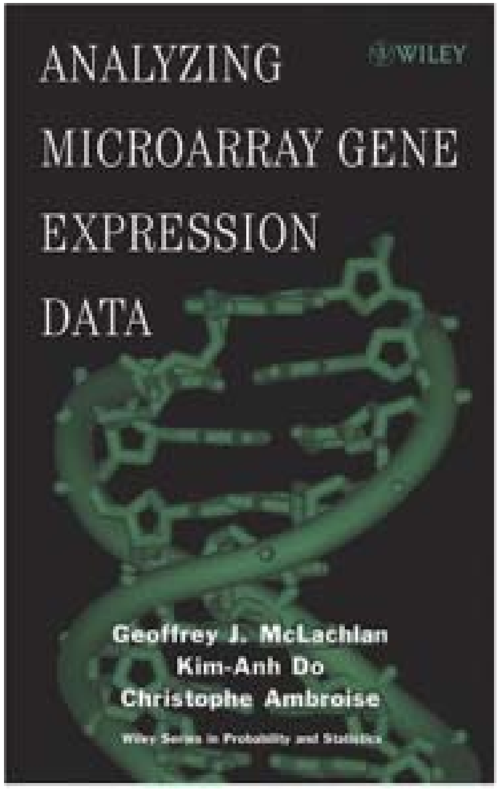
Where to learn more?


DNA Microarrays
and Gene Expression
From Experiments to Data Analysis and Modeling
Pierre Baldi and G. Wesley Hatfield
CAMBRIDGE

Overview of microarray technology and analysis, 2003

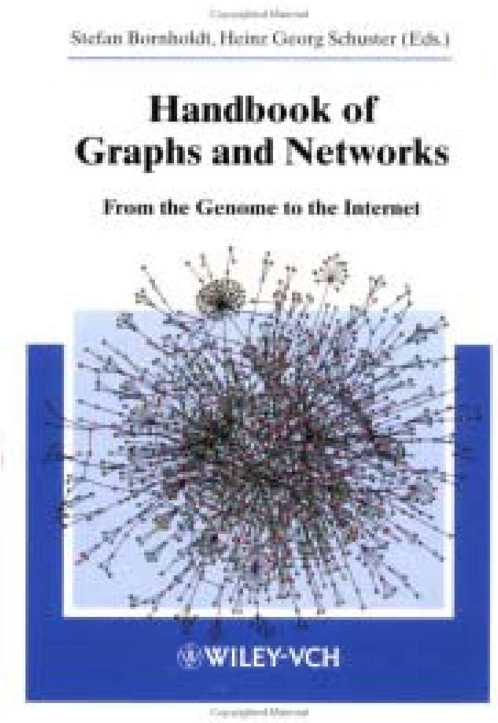


Edited volume on principal statistical techniques of microarray analysis, 2003



Textbook aimed at biostatisticians, 2004

Where to learn more?



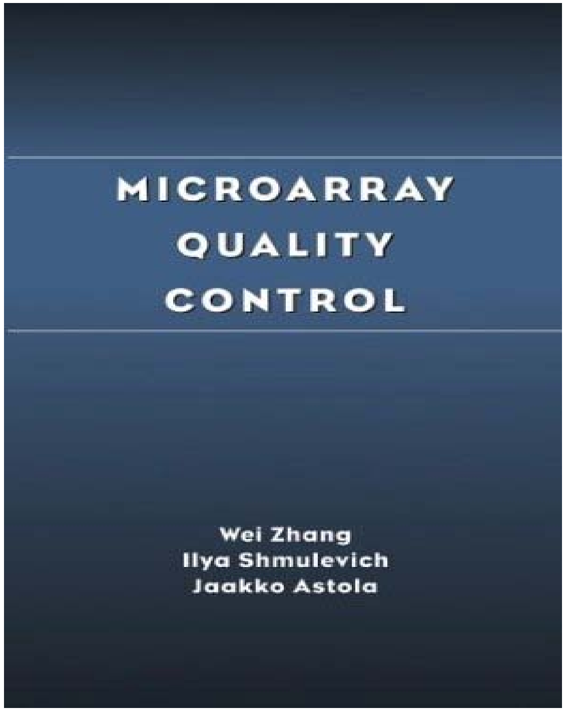
Stefan Bornholdt, Heinz Georg Schuster (Eds.)

Handbook of Graphs and Networks

From the Genome to the Internet

WILEY-VCH

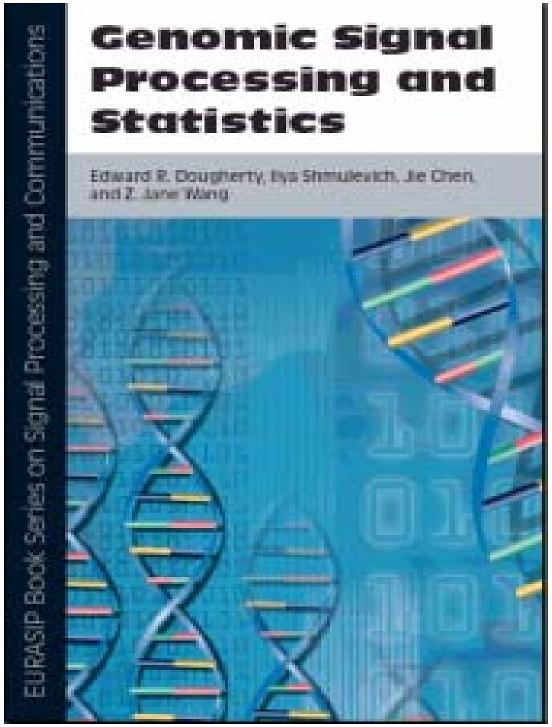
Edited monograph on
random graphs
In nature, 2005



MICROARRAY QUALITY CONTROL

Wei Zhang
Ilya Shmulevich
Jaakko Astola

To appear soon!



Genomic Signal Processing and Statistics

Edward R. Dougherty, Ilya Shmulevich, Jie Chen, and Z. Jane Wang

EURASIP Book Series on Signal Processing and Communications

To appear soon!

Future venues

IEEE Signal Processing Society

IEEE International Workshop on Genomic Signal Processing and Statistics, 2005

Sunday, 22 May 2005 – Tuesday, 24 May 2005
Newport, Rhode Island USA

Call for Papers

[illegible]

Areas of Interest (but not limited to):

[illegible]

Deadlines:

[illegible]

Contact Information: [illegible]

GENSIPS 2005 – Newport:

- Ray Liu, Jaako Astola
- Workshop dates: May 22 - 24, 2005
- Early registration ends March 30

Final acceptance notification: November 1, 2005 Final manuscript due: Dece

IEEE Transactions on Signal Processing Special Issue on Genomic Signal Processing

- Submission deadline: May 1, 2005
- Publication date: Sept. 2006

5. Conclusions

- Gene filtering: accounting for biological and statistical significance
- Gene ranking: can involve optimization over multiple criteria
- Gene co-regulation networks: discover co-dependent gene profiles that can aid in clustering
- Statistical signal and image processing approaches can have impact
- References to UM work and software presented here: http://www.eecs.umich.edu/~hero/bioinfo.html